Super-résolution de texture pour la reconstruction 3D fine

Calum BURNS, Aurélien PLYER, Frédéric CHAMPAGNAT

ONERA - The French Aerospace Lab
Chemin de la Hunière, 91120 Palaiseau, France
calum.burns@onera.fr

Résumé – Nous proposons une méthode permettant de produire un atlas de texture haute résolution pour modèles 3D en exploitant pleinement l'information contenue dans une séquence vidéo dense, via des techniques de super-résolution. Nous analysons et surmontons l'imprécision intrinsèque des méthodes de reconstruction multi-vue. Par rapport à des méthodes concurrentes s'appuyant sur la correction de pose et l'affinage géométrique, notre méthode de recalage subpixellique corrige l'erreur de recalage dans le domaine image et est ainsi beaucoup plus efficace. Nous illustrons l'intérêt de notre méthode en augmentant la résolution en texture d'un modèle 3D produit par un logiciel de reconstruction à l'état de l'art.

Abstract – We describe a method for producing a high quality texture atlas for 3D models by fully exploiting the information contained in dense video sequences via super-resolution techniques. The intrinsic precision limitations of multi-view reconstruction techniques are analysed and overcome. Compared to similar methods that rely on camera pose correction and geometry refinement, our subpixel image registration technique directly corrects the registration error in the image domain is much more efficient. We illustrate our method by enhancing the texture resolution of a 3D model produced by a state of the art reconstruction software.

1 Introduction

La reconstruction 3D multi-vue atteint désormais un niveau de maturité industrielle : de nombreux logiciels commerciaux, comme ceux proposés par Agisoft et Pix4D, permettent à des utilisateurs non-experts de produire des modèles 3D large-échelle de qualité. Ces reconstructions s'appuient généralement sur des capteurs haut de gamme comme des LIDAR ou des appareils photos de type DSLR, montés sur un trépied et déplacés autour de la scène. Ces protocoles d'acquisition s'avèrent peu pratiques pour des scènes de grande taille à géométrie complexe. Dans un tel contexte, l'acquisition vidéo est une perspective intéressante, en particulier lorsque le capteur est monté sur un véhicule autonome tel qu'un drone [2]. Cependant, de nombreux logiciels de l'état de l'art ne sont pas adaptés à l'utilisation de séquences vidéos : la combinatoire engendrée par le grand nombre d'images implique un coût de calcul élevé pour les algorithmes utilisés et la reconstruction 3D ne peut être réalisée qu'avec un sous-ensemble de la séquence. Nous proposons ici d'utiliser les images restantes afin d'augmenter la qualité de la texture du modèle reconstruit. L'intérêt de la texture est de permettre de visualiser des détails fins du modèle numérisé qui ont été perdus dans le bruit géométrique de la reconstruction.

Notre contribution est d'exploiter les images d'une séquence vidéo afin d'augmenter la résolution de la texture du modèle 3D produit par un logiciel à l'état de l'art. Pour ce faire nous utilisons une méthode de super-résolution (SR) [3, 4, 7].

Afin de dissocier la résolution en texture de la résolution géométrique d'un modèle, de nombreux auteurs génèrent un espace de texture 2D qui paramétrise le modèle 3D en définissant une intensité en tout point de la surface : c'est l'atlas de texture. Une façon de produire cet atlas est d'utiliser les images d'entrée, liant ainsi la résolution en texture à la résolution native de ces images [5][12]. Notre objectif est d'améliorer la résolution de l'atlas de texture par des méthodes de SR vidéo [3, 4, 7]. L'idée sous-jacente à la SR est d'exploiter le mouvement subpixellique au sein des images d'une séquence vidéo afin d'augmenter la résolution spatiale et produire une image haute résolution : "l'image SR". Un tel résultat ne peut être obtenu que lorsque le déplacement inter-image est estimé avec une précision subpixellique. En étendant ces idées aux capteurs RGB-D, [10] propose une technique de SLAM estimant de façon simultanée une image d'intensité SR et une carte de profondeur SR à partir des données d'une Kinect.

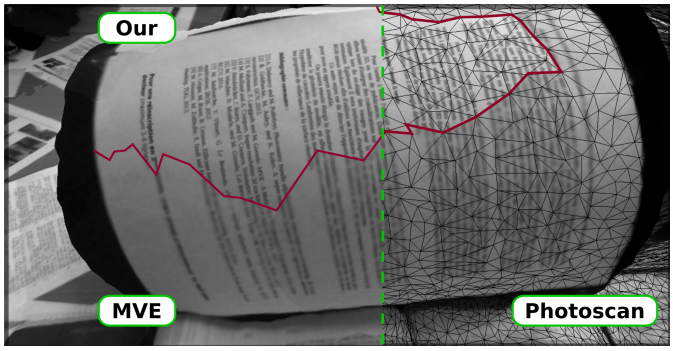

FIGURE 1 – Résultat de notre méthode de TSR sur une portion de texture (en rouge), sur un maillage généré par Photoscan et texturé par [12] de la librairie MVE ; réalisé sur la corbeille.

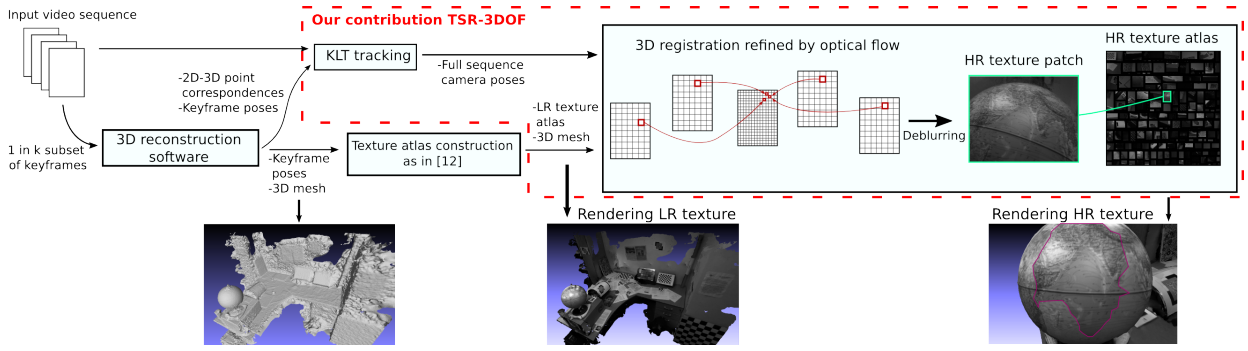

FIGURE 2 – Vue d'ensemble de la chaîne algorithmique avec illustration du globe monde reconstruit.

L'estimation de texture haute résolution pour des modèles 3D est un problème complexe, car la résolution géométrique du modèle est en général plus faible que la résolution du pixel [6]. Ainsi la précision géométrique des modèles n'est pas suffisante pour permettre un recalage subpixellique des images.

Peu de travaux sont dédiés à la SR de texture de modèles 3D. Goldluecke *et. al.* [6] s'attaquent au problème en optimisant une fonction de coût globale dépendant de la texture SR, des paramètres caméra et d'une carte de déplacement surfacique. Les résultats sont impressionnants au prix d'un coût calculatoire élevé : 2h d'optimisation pour traiter une base de 50 images sur un modèle 3D d'environ 10cm de haut.

Maier *et. al.*[9] applique un traitement SR sur un sous-ensemble d'images clefs de sa séquence vidéo afin de produire un atlas de texture SR. Les auteurs estiment une valeur d'intensité pour chaque pixel de l'image clef SR en effectuant une médiane pondérée des valeurs des pixels correspondants dans chaque image voisine. Ce recalage est basé uniquement sur la géométrie estimée de la scène et des poses de la caméra et est donc sujet aux imprécisions des méthodes de reconstruction multi-vue, que l'on tente donc d'éliminer par filtrage médian.

Dans ce papier nous présentons une méthode originale de super-résolution de texture (TSR). Par rapport aux approches précédentes, nous surpassons la précision limitée intrinsèque aux algorithmes de reconstruction 3D en corrigeant les erreurs de recalage directement dans le domaine image plutôt que dans le domaine 3D comme proposé dans [6]. Notre correction des erreurs de recalage se fait de façon efficace grâce à un algorithme de flot optique rapide [11], et on se distingue ainsi des champs de déplacement coûteux décrits dans [6] ou des cartes de profondeur affinées proposées en [9].

Nous illustrons notre méthode de TSR sur une séquence vidéo monoculaire en affinant la texture d'un modèle produit par un logiciel de reconstruction commercial.En section 2 nous présentons les principales étapes de notre chaîne algorithmique. Puis nous discutons des résultats de cette chaîne TSR en section 3.

2 Super-résolution de texture

Une vue d'ensemble de notre chaîne de production de modèle 3D avec texture SR est esquissée en Fig.2. Nous détaillons les points importants dans cette section.

Notre méthode se construit autour d'un logiciel de reconstruction multi-vue (RMV) tel Photoscan ou MVE [12]. La RMV produit un maillage 3D de la scène à partir d'un sous ensemble des images d'entrée : les *images clefs*. Le sous-échantillonnage temporel de la vidéo d'origine est nécessaire du fait que les logiciels RMV ne passent pas à l'échelle sur le grand nombre d'images d'une séquence vidéo dense. La RMV fournit également les poses des images clefs qui seront utilisées dans la construction de l'*atlas de texture* : il s'agit de déterminer, pour chaque facette triangulaire du maillage 3D, l'image clef la plus apte à lui apporter son information de texture. Nous utilisons la formulation par Champ de Markov aléatoire proposé dans [12] qui cherche à choisir la vue ayant la meilleure qualité image tout en préservant au mieux une labellisation identique des facettes voisines afin de minimiser les coutures visibles.

En utilisant l'atlas de texture généré, la résolution en texture est dissociée de la résolution géométrique du modèle. Nous pouvons donc augmenter la résolution en texture sans modifier le modèle 3D. Notre objectif est d'appliquer des méthodes de SR sur les images clefs en utilisant l'ensemble des images de la séquence vidéo initiale. Pour ceci il faut d'abord effectuer un recalage avec une précision subpixellique de chaque image avec l'image clef le plus proche. Ensuite les informations des images voisines recalées sont fusionnées pour augmenter la résolution de l'image clef.

Le recalage subpixellique aurait pu être fait avec un simple algorithme de flot optique, mais nous obtenons un recalage plus robuste et une meilleure résolution de texture en utilisant l'information de profondeur donnée par le maillage 3D. Il est donc nécessaire en premier lieu d'estimer la pose 3D des images voisines.

La RMV fournit des correspondances entre les points d'intérêt 2D des images clefs et les points 3D du maillage. Nous propageons ces correspondances aux images voisines avec un trackeur KLT [1]. Nous appliquons ensuite sur ces nouvelles correspondances un algorithme PnP itératif basé sur une optimisation Lenvenberg-Marquard afin d'estimer la pose 3D de la caméra voisine. Nous assurons la robustesse de ce processus par un filtrage par matrice fondamentale des points suivis ainsi qu'un filtrage RANSAC des correspondances 2D/3D estimées.

A partir de la pose caméra, chaque pixel des images voisines est projeté sur le maillage 3D puis rétroprojeté dans l'image

clef la plus proche. Nous obtenons ainsi les coordonnées 2D correspondantes dans le référentiel de l'image clef. Cependant, ce recalage initial ne peut être effectué avec la précision subpixellique nécessaire puisque le maillage 3D n'est qu'une approximation affine par morceaux de la surface réelle. Cette approximation peut engendrer une erreur de recalage allant jusqu'à 1 ou 2 pixels, ceci est insuffisant pour effectuer de la SR.

Afin de fournir le recalage pixel à pixel nécessaire à la SR, Goldluecke *et. al.* [6] corrigent l'erreur dans le domaine 3D. Nous observons qu'il est plus efficace de corriger cette erreur directement dans le domaine image en utilisant un algorithme de flot optique. Nous estimons donc d'abord un recalage approximatif en utilisant le modèle 3D et les poses caméra, qui sert ensuite pour initialiser l'estimation du flot optique entre l'image voisine et l'image clef. Afin de rester efficace sur un grand nombre d'images, nous choisissons l'algorithme eFolki [11] pour l'estimation du flot optique. Cet algorithme dense et multi-échelle de type Lucas-Kanade favorise l'efficacité calculatoire tout en maintenant une précision satisfaisante. Afin d'augmenter d'avantage la robustesse de notre estimation de flot optique, nous implémentons la validation "aller/retour" proposée dans [11] : nous estimons également le flot optique inverse (de l'image clef à l'image voisine). La somme des flots aller et retour doit être nulle, nous fixons donc un seuil qui permet d'éliminer la contribution de pixels dont la somme des flots aller/retour est trop élevée.

Une fois estimé un recalage subpixellique pour chaque image voisine, nous construisons l'image clef SR par inversion d'un modèle de formation image (MFI). Cet MFI définit la relation entre chaque image observée I_k^{BR}, dite aussi image basse résolution (BR), et l'image SR I^{HR}. Une MFI générale est proposée dans [3] :

$$I_k^{BR} = DBW_k I^{HR} \quad (1)$$

Où B est le noyau Gaussien modélisant la fonction d'étalement du point (PSF). W_k est l'opérateur de warp modélisant le mouvement pixellique entre l'image k et l'image de référence. D est l'opérateur de sous-échantillonnage défini par le facteur de SR noté L. Notre objectif est d'implémenter un algorithme de SR efficace en coût de calcul afin de traiter de longues séquences vidéo. Pour ceci, nous utilisons une approximation du MFI décrit ci-dessus, proposé dans [4] :

$$I_k^{BR} = DW_k BI^{HR} \quad (2)$$

En théorie, l'hypothèse que $BW_k = W_k B$ n'est valide que dans le cas de mouvements de translation pure, mais Hardie *et. al.* [7] ont montré qu'en pratique ce modèle peut être étendu à des mouvements plus généraux. L'avantage de cette approximation est que l'opérateur de sous-échantillonnage est directement appliqué à l'opérateur W_k. De ce fait, l'inversion de ce modèle ne nécessite qu'un recalage inter-image BR. Nous pouvons donc utiliser directement le mouvement tel qu'il a été estimé par notre méthode. Ceci évite le stockage des flots haute-résolution (HR) et des interpolations coûteuses de flots BR. Par ailleurs, cette formulation permet de séparer l'inversion du MFI en deux sous-tâches : une étape dite de "Shift&Add" qui produit une image HR floue suivi d'une étape de déconvolution [4].

L'étape Shift&Add (S&A) est une stratégie de vote. Chaque pixel natif est recalé sur la grille HR. On assigne à chaque pixel HR la valeur moyenne des pixels BR qui ont voté à ses coordonnées. On note que le mouvement des pixels BR est tronqué au pixel HR le plus proche, ce qui engendre une erreur de recalage supplémentaire qui peut être compensée en choisissant un facteur de SR élevé. L'image HR intermédiaire obtenue est notée $\widehat{I}_{HR}$.

L'étape finale est le défloutage. Pour ce faire on minimise la fonctionnelle qui suit par méthode du gradient conjugué :

$$E_{deconvQ}(I_{HR}) = \|\widehat{I}_{HR} - BI_{HR}\|_W^2 + \lambda\|\nabla I_{HR}\|^2 \quad (3)$$

Où W est une matrice de poids diagonale, chaque coefficient diagonal correspondant au nombre de pixels BR qui ont voté pour le pixel HR en question. De cette manière les pixels HR ayant reçu le plus de votes se voient attribués un poids plus fort lors de la déconvolution.

3 Résultats

Pour nos expérimentations nous avons scanné un bureau 2m x 1m sur lequel étaient déposés de multiples objets texturés. La caméra utilisée a une optique de focale 5.5mm travaillant à $f/2.8$ montée sur un détecteur Bayer e2V 1/18". Nous utilisons des images monochromatiques en sous-échantillonnant les images d'entrée de 1600×1200 à 800×600 afin de ne garder que la moitié du canal vert de la matrice Bayer. Notre dataset comporte environ 3200 images, 1 image sur 20 est utilisée comme image clef pour la reconstruction géométrique. Nous utilisons le logiciel "Photoscan" de Agisoft pour produire un modèle 3D de la scène ainsi que les poses des images clef.
Les résultats de SR présentés par la suite sont obtenus par la fusion de 30 images voisines. Le facteur de SR L est fixé à 6. Pour le défloutage nous utilisons un noyau Gaussien avec $\sigma = 0.7$ pixels BR et fixons le paramètre de régularisation $\lambda = 0.1$. Ces paramètres ont été choisis empiriquement.

En figure 3 nous comparons la méthode de SR décrite dans[9] à notre méthode de SR avec 3 variantes de recalage d'images : recalage uniquement par le modèle 3D et les poses des caméras ("TSR-3D"), recalage par flot optique seul ("TSR-OF") et recalage 3D raffiné par flot optique ("TSR-3DOF"). Les image LR interpolées à la taille de l'image SR sont présentées sur la gauche. Nous notons tout d'abord que la méthode TSR-3DOF surpasse [9] et TSR-3D qui sont des méthodes reposant uniquement sur l'estimation de la géométrie 3D de la scène. Ceci démontre l'intérêt de notre affinage par flot optique. Sur la corbeille courbe, TSR-3DOF et TSR-OF ont des performances similaires, mais TSR-3DOF surpasse TSR-OF sur le globe. Ceci s'explique par le fait que sur cette zone complexe, on arrive à recaler plus de pixels dans l'image clef avec une initialisation du mouvement par la 3D, la reconstruction de la texture en est alors meilleure. Pour une portion d'atlas correspondant à une

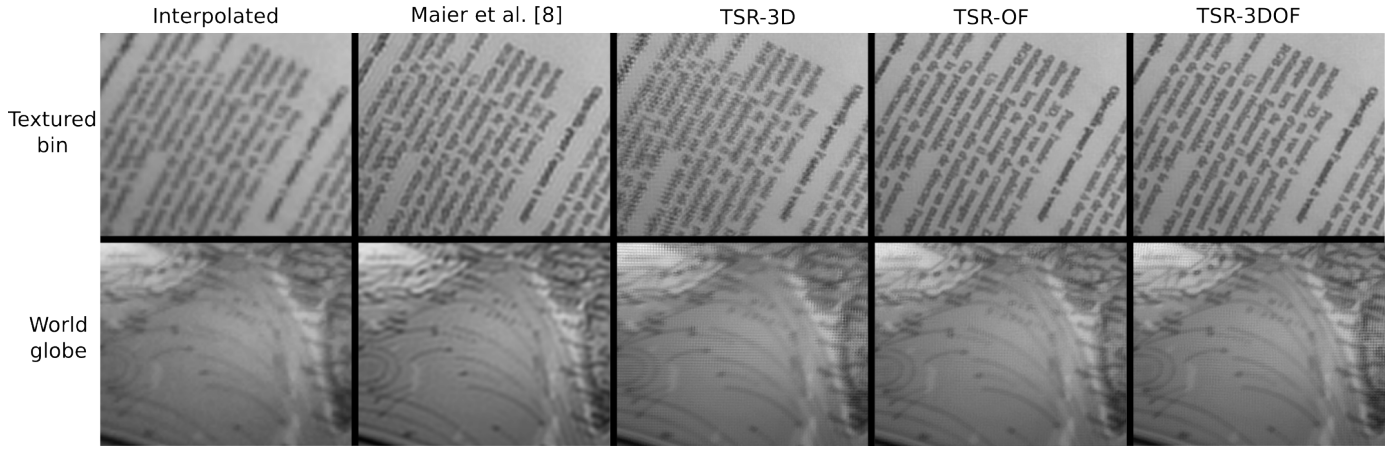

FIGURE 3 – Images clefs SR sur une portion de texture d'une corbeille courbe (haut) et un globe du monde (bas). De gauche à droite : image LR interpolée, fusion SR selon Maier et. al. [9], SR avec recalage par 3D seule, SR avec recalage par flot optique seul, SR avec 3D affiné par flot optique.

image clef, notre méthode de recalage tourne en 3 minute pour 30 images, et la déconvolution prend environ 3'30" pour une portion 1200×800 d'image clef HR.

4 Conclusion

Nous proposons dans ce papier une méthode originale de TSR qui améliore la résolution en texture de modèles 3D produits par des logiciels de reconstruction actuels à l'état de l'art. Pour ce faire, nous exploitons pleinement l'information contenue dans une séquence vidéo monoculaire dense. Les méthodes précédentes de TSR s'appuient sur un raffinage de la géométrie 3D de la scène reconstruite. Cette ré-optimisation devient rapidement coûteuse lorsqu'on cherche à atteindre la précision requise par les algorithmes de SR. Nous montrons que la précision requise pour la SR peut peut être atteinte en corrigeant l'erreur initiale par recalage géométrique de façon efficace directement dans le domaine image par un algorithme de flot optique. Dans la suite de nos travaux, nous souhaitons quantifier le gain en résolution apporté par notre méthode en généralisant des mesures telles celle définie dans [8] au contexte de la reconstruction 3D. Nous souhaitons également nous émanciper des limitations induites par les surfaces spéculaires en intégrant des modèles d'illumination qui nous permettront potentiellement d'estimer des textures d'albédo et de normales à haute résolution.

Références

[1] J.-Y. Bouguet. Pyramidal implementation of the lucas kanade feature tracker. *Intel Corporation, Microprocessor Research Labs*, 2000.

[2] Shreyansh Daftry, Christof Hoppe, and Horst Bischof. Building with drones : Accurate 3D facade reconstruction using MAVS. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 3487–3494. IEEE, 2015.

[3] M. Elad and A. Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE Transactions on Image Processing*, 6(12) :1646–1658, 1997.

[4] M. Elad and Y. Hel-Or. A fast super-resolution reconstruction algorithm for pure translational motion and common space-invariant blur. *IEEE Transactions on Image Processing*, 10(8) :1187–1193, Aug 2001.

[5] R. Gal, Y. Wexler, E. Ofek, H. Hoppe, and D. Cohen-Or. Seamless Montage for Texturing Models. *Computer Graphics Forum*, 2010.

[6] B. Goldluecke, M. Aubry, K. Kolev, and D. Cremers. A super-resolution framework for high-accuracy multiview reconstruction. 2014.

[7] R. C. Hardie and K. J. Barnard. Fast super-resolution using an adaptive Wiener filter with robustness to local motion. *Opt. Express*, 20(19) :21053–21073, Sep 2012.

[8] S. Landeau. Evaluation of super-resolution imager with binary fractal test target. volume 9249, pages 924909–924909–16, 2014.

[9] R. Maier, J. Stueckler, and D. Cremers. Super-resolution keyframe fusion for 3D modeling with high-quality textures. In *International Conference on 3D Vision (3DV)*, 2015.

[10] M. Meilland and A. Comport. Super-resolution 3D tracking and mapping. *ICRA*, 2013.

[11] A. Plyer, G. Le Besnerais, and F. Champagnat. Massively parallel Lucas Kanade optical flow for real-time video applications. *Journal of Real-Time Image Processing*, 2014.

[12] M. Waechter, N. Moehrle, and M. Goesele. Let there be color ! Large-scale texturing of 3D reconstructions. In *ECCV 2014*, pages 836–850. Springer International Publishing, 2014.