Estimation de la pose d'une caméra dans un environnement connu à partir d'un recalage 2D-3D

Danda Pani Paudel, Cédric Demonceaux, Adlane Habed, Pascal Vasseur

Estimation de la pose d'une caméra dans un environnement connu à partir d'un recalage 2D-3D

Danda Pani Paudel[1] Cédric Demonceaux[1] Adlane Habed[2] Pascal Vasseur[3]
[1] Le2i UMR 6306 CNRS, Université de Bourgogne
[2] ICube UMR 7357 CNRS, Université de Strasbourg
[3] LITIS EA 4108, Université de Rouen
Danda-Pani.Paudel@u-bourgogne.fr

Résumé

Nous proposons une méthode directe de recalage robuste 2D-3D permettant de localiser une caméra dans un environnement 3D connu. Il s'agit d'un problème rendu particulièrement difficile par l'absence de correspondances entre les points 3D du nuage et les points 2D. A cette difficulté, s'ajoute la différence d'échelle entre le nuage 3D connu et le nuage 3D reconstruit à partir d'images qui, de plus, peut contenir des points aberrants et des occultations. Notre méthode consiste en l'optimisation d'une fonctionnelle de manière itérative en deux étapes : estimation de la pose de la caméra et mise en correspondance 2D-3D. Ainsi, nous obtenons une méthode d'estimation conjointe de la localisation et de la reconstruction en minimisant les erreurs de reprojection dans les images tout en préservant la structure 3D de la scène. Les problèmes d'occultations et d'échelle sont surmontés grâce à un histogramme alors que les points aberrants sont gérés par un M-estimateur robuste.

Mots Clef

Estimation de pose, reconstruction 3D, SfM.

Abstract

We propose a robust and direct 2D-to-3D registration method for localizing 2D cameras in a known 3D environment. In this case, localizing the cameras remains a challenging problem that is particularly undermined by the unknown 2D-3D correspondences, outliers, scale ambiguities and occlusions. Once the cameras are localized, the Structure-from-Motion reconstruction obtained from image correspondences is refined by means of a constrained nonlinear optimization that benefits from the knowledge of the scene. We also propose a common optimization framework for both localization and refinement steps in which projection errors in one view are minimized while preserving the existing relationships between images. The problems of occlusion are handled by employing a scale histogram while the effect of data inaccuracies is minimized using an M-estimator-based technique.

Keywords

Pose estimation, 3D reconstruction, SfM.

1 Introduction

Les méthodes de reconstruction de scène à partir du mouvement d'une caméra (Structure-from-Motion : SfM) évoluant dans un environnement inconnu estiment généralement la pose de la caméra et la reconstruction 3D en minimisant l'erreur de reprojection dans l'image sur toutes les prises de vue. Pour une meilleure localisation de la caméra, il est hautement souhaitable de bénéficier de la connaissance de la scène lorsque celle-ci est disponible. En effet, pour localiser la caméra dans la scène à un instant t donné, il est préférable d'utiliser une information 3D fiable donnée à l'instant $t-1$ plutôt que d'effectuer un ajustement de faisceaux à chaque nouvelle image. Pour ce faire, nous devons recaler un nuage de points 3D connu avec des données 2D.

Le problème de recalage 2D-3D est traité dans la littérature suivant deux approches : les méthodes directes et celles indirectes. Les méthodes directes reposent sur la mise en correspondance de caractéristiques telles que des points, des lignes, des plans, des points de fuite et l'utilisation de boites englobantes entre les images et la scène 3D. Les méthodes d'appariement basées sur les points proposées dans [10, 5] nécessitent une caractérisation de chaque point 3D de la scène par un descripteur SIFT. Ces descripteurs sont alors comparés aux descripteurs présents dans l'image. La mise en correspondance peut être compromise par l'absence de ces descripteurs dans les points de la scène fournis ainsi que par la variabilité des conditions d'éclairage pendant les acquisitions 2D et 3D. Les méthodes qui nécessitent des caractéristiques de plus haut niveau, telles que les lignes [1], des plans [11] et des boîtes englobantes [6], ne sont généralement applicables que pour des scènes de type Manhattan. Les méthodes basées sur la détection du ciel [9], ainsi que les méthodes reposant sur un modèle 3D prédéfini [2] ont, elles aussi, une portée limitée. Les méthodes indirectes sont réalisées soit par un recalage 3D-3D soit en recherchant des paramètres de recalage appropriés. Elles consistent en un algorithme ICP (Iterative Closest Point) rigide ou non-rigide entre la reconstruction

SfM déduite du mouvement de la caméra et le nuage de points 3D de la scène connue. Toutefois, un tel recalage n'est pas simple en raison du facteur d'échelle inconnu de la reconstruction. Cette ambiguïté peut, par exemple, être résolue par une extension de l'algorithme des 4 points [3]. Une autre approche consiste à utiliser l'information mutuelle [14] ou à segmenter l'image [12]. Ces méthodes estiment la position de la caméra image par image de façon indépendante sans tenir compte des contraintes géométriques multivues (telles que la contrainte épipolaire). Enfin, en ce qui concerne la pose de la caméra par ajustement itératif, la méthode proposée dans [11] fournit de très bons résultats dans un environnement connu. Cependant, cette méthode repose uniquement sur les plans de la scène et suppose que le recalage initial 2D-3D a déjà été effectué.

Dans cet article, nous proposons une méthode directe de recalage 2D-3D à partir de prises de vues calibrées et d'une scène connue. Pour ce faire, nous modélisons le problème par une optimisation non linéaire sous contraintes tenant compte des connaissances *a priori*. Cette approche raffine la position des caméras à partir d'un recalage frustre et d'une localisation approximative de la première caméra et ne requiert aucune hypothèse sur la structure 3D de l'environnement. Nous considérons que nous disposons d'une mise en correspondance 2D-2D entre les points images pour chaque prise de vue mais avec une correspondance 3D inconnue. A notre connaissance, il n'existe pas de méthode utilisant conjointement les informations 2D des images et l'information 3D de la scène sans mise en correspondance préalable entre les points 2D et les points 3D. Notons que les méthodes telles que les ajustements de faisceaux sur scène connue [13] et PnP [4] nécessitent une correspondance 2D-3D *a priori*. En pratique, les correspondances 2D-2D inter-images peuvent être obtenues en utilisant des descripteurs tels que SIFT. A partir d'une pose approximative de la première caméra, le recalage est réalisé en minimisant une erreur de projection dans les vues qui doit préserver la structure 3D de la scène et les correspondances entre les paires d'images. Les correspondances 2D-3D sont déterminées ici de telle sorte que chaque paire de points appariés dans les images donne un point 3D dont la distance à la scène est minimale. Cette mesure est dérivée de la géométrie épipolaire et donc indépendante de l'échelle relative. Une telle mesure permet d'éviter le problème de l'échelle qui se pose lors de la reconstruction. La véritable échelle relative peut ensuite être récupérée par la construction d'un histogramme où les correspondances 2D-3D votent pour leurs échelles relatives. Nos expériences montrent que la précision de cette méthode est significativement meilleure qu'une approche par ajustement de faisceaux couramment utilisée.

Ce papier est organisé comme suit : la Section 2 introduit les notations et outils nécessaires. Dans la Section 3, nous modélisons le problème d'optimisation 2D-3D et proposons une résolution itérative. Des expérimentations synthétiques et réelles sont présentées dans la Section 4. Enfin, la Section 5 conclut notre travail.

2 Notations et position du problème

Soit $X^i, i = 1 \dots n$ les points 3D connus de la scène définis dans le repère monde O_w. Notre but est de localiser un ensemble de p caméras calibrées de repère local respectif $O_1, O_2 \dots O_p$ tenant compte de la connaissance des points 3D du repère monde. Soit R et t la rotation et la translation, de la première caméra relativement au repère monde O_w. Si x_1^j et $x_2^j, j = 1 \dots m$ sont les points en correspondance entre 2 vues, la position relative de la deuxième caméra par rapport à la première caméra (R', t') peut être obtenue à l'aide de la matrice essentielle entre ces deux vues [8]. Notons $\widetilde{X}^j, j = 1 \dots m$ les points 3D reconstruits à partir de ces deux vues dans O_1. Chaque matrice de rotation R est représentée par un vecteur de quaternions 4×1 q. Les points 3D et 2D sont représentés par des vecteurs 3×1. La fonction de correspondance entre les points 2D et 3D est notée ϕ. Par exemple, nous notons $\phi(j)$ la fonction qui projette chaque point 2D $x_1^j \leftrightarrow x_2^j$ vers le point 3D correspondant X^i. La distance entre deux matrices de rotation R_1 et R_2 est mesurée par la norme spectrale de leur différence $|||R_1 - R_2|||$.

3 Recalage 2D-3D

3.1 Formulation du problème

Les relations entre les points 2D et 3D sont décrites dans Fig. 1. Le produit scalaire entre la normale au plan $[t']_\times R' x_2$ et le vecteur $RX + t$ définissant le même plan aboutit à la relation :

$$f(R, t, R', t') = (RX + t)^T [t']_\times R' x_2 = 0. \quad (1)$$

Comme le vecteur x_1 doit être colinéaire avec le vecteur $RX + t$,

$$RX + t = \alpha x_1. \quad (2)$$

Cette colinéarité se traduit par

$$g(R, t) = ||[x_1]_\times (RX + t)||^2 = 0. \quad (3)$$

De plus, la contrainte épipolaire entre les deux vues est exprimée par

$$h(R', t') = x_1^T [t']_\times R' x_2 = 0. \quad (4)$$

alors que (3) est exprimée dans le repère de la première caméra, (1) exprime une contrainte de la seconde caméra relativement au repère monde. De même, (4) exprime la position de la seconde caméra par rapport à la première. Les équations (1), (3) et (4) sont évidemment redondantes. Cependant, en présence de bruits et de correspondances inconnues, toutes ces contraintes permettent d'obtenir un système plus robuste. Ainsi, les trois équations sont incorporées dans notre schéma d'optimisation sous contraintes. Considérons désormais le problème de

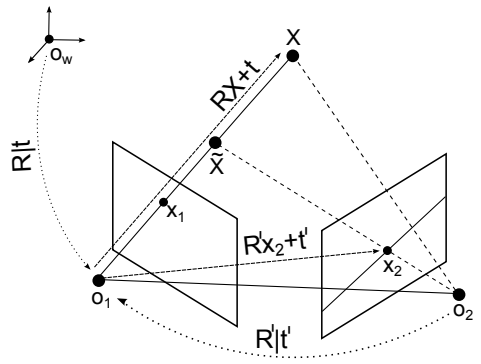

FIGURE 1 – Triangulation.

localisation des deux caméras à partir de correspondances 2D-2D ($x_1^j \leftrightarrow x_2^j$) connues et 2D-3D ($x_1^j \leftrightarrow x_2^j \leftrightarrow X^{\phi(j)}$) inconnues. Le problème de recalage 2D-3D revient à déterminer la fonction ϕ optimale. Par conséquent, R, t, R', t' et ϕ sont solutions de :

$$\begin{aligned} \min_{q,t,q',t',\phi} \quad & \sum_{j=1}^{m} \{(RX^{\phi(j)}+t)^T [t']_\times R' x_2^j\}^2 \\ \text{telles que} \quad & |||[x_1^j]_\times (RX^{\phi(j)}+t)||^2 = 0, \\ & \{(x_1^j)^T [t']_\times R' x_2^j\}^2 = 0,\ j = 1 \dots m \\ & ||q||^2 = 1,\ ||q'||^2 = 1,\ ||t'||^2 = 1. \end{aligned} \tag{5}$$

Le problème (5) considère que chaque point de l'image admet un correspondant 3D dans la scène. En pratique, 2 problèmes se posent : (a) plusieurs points 3D se situent sur la même ligne de vue de la caméra et satisfont donc la contrainte de la géométrie épipolaire et aboutissent à des ambiguïtés pour la correspondance 2D-3D, (b) des points 3D reconstruits par les données 2D, non connus *a priori*, aboutissent à des mauvaises correspondances 2D-3D. Ces deux problèmes sont surmontés en associant un poids pour chaque correspondance en fonction de l'histogramme d'échelle. A partir des correspondances 3D-3D, $\widetilde{X}^j \leftrightarrow X^{\phi(j)}, j = 1 \dots m$, l'échelle relative de la reconstruction est

$$s(j) = \frac{||\widetilde{X}^j||}{||RX^{\phi(j)}+t||}, \quad j = 1 \dots m. \tag{6}$$

En théorie, tous les points ont la même échelle $s(j) = s(i)\ \forall \{i,j\} \epsilon 1 \dots m$. Pour obtenir cette échelle, nous contruisons l'histogramme des échelles $H(u)$, $u = 1 \dots b$, l'échelle retenue correspond au mode de cet histogramme. Si u_{max} est le mode de l'histogramme, les poids sont alors distribués comme suit :

$$w(j) = \begin{cases} 1 & \text{si} \quad s(j) \epsilon H(u_{max}) \\ 0 & \text{sinon.} \end{cases} \tag{7}$$

En outre, l'effet des erreurs dans les données 3D est réduit par l'introduction d'un estimateur robuste. Par conséquent, le problème d'optimisation (5), incluant l'estimation robuste et pondérée par l' histogramme, est ré-écrit ainsi :

$$\begin{aligned} \min_{q,t,q',t',\phi} \quad & \sum_{j=1}^{m} w(j) \rho((RX^{\phi(j)}+t)^T [t']_\times R' x_2^j) \\ \text{tel que} \quad & w(j)\rho(|||[x_1^j]_\times (RX^{\phi(j)}+t)||) = 0, \\ & \rho((x_1^j)^T [t']_\times R' x_2^j) = 0,\ j = 1 \dots m \\ & |||q||^2 = 1, ||q'||^2 = 1, |t'||^2 = 1. \end{aligned} \tag{8}$$

où $\rho(x)$ est la fonction de Tukey :

$$\rho(y) = \begin{cases} \frac{y^6}{6} - \frac{\xi^2 y^4}{2} + \frac{\xi^4 y^2}{2} & \forall\, |y| < \xi \\ \frac{\xi^6}{6} & \text{sinon} \end{cases} \tag{9}$$

dont la fonction d'influence est $\psi(y) = y\left(\xi^2 - y^2\right)^2$ $\forall |y| < \xi$ et 0 sinon.
Notons que la fonction de coût et la première contrainte ne considèrent que la partie connue de la scène. Cependant, la deuxième contrainte tient compte de la partie 3D inconnue de la scène. Les paramètres de recalage optimaux sont obtenus en résolvant de manière itérative ce problème d'optimisation. A chaque itération le problème se décompose en deux parties : (a) le recalage 2D-3D et (b) la localisation des caméras.

3.2 Recalage 2D-3D

Une paire de caméra est localisée dans la scène de manière itérative à partir des paramètres de recalage R, t et ϕ. Ceci est réalisé en résolvant le problème global :

$$\begin{aligned} \min_{R,t,\phi} \quad & \sum_{j=1}^{m} w(j) \{(RX^{\phi(j)}+t)^T [t']_\times R' x_2^j\}^2 \\ \text{tel que} \quad & w(j)|||[x_1^j]_\times (RX^{\phi(j)}+t)||^2 = 0, j = 1 \dots m. \end{aligned} \tag{10}$$

Pour une rotation R et une translation t données, ϕ est solution de :

$$\phi(j) = \underset{i\,=\,1 \dots n}{\text{argmin}}\ d(R,t,X^i,x_1^j,x_2^j),\ j = 1 \dots m, \tag{11}$$

avec $d(R,t,X,x_1,x_2)$ qui représente l'erreur de reprojection :

$$\begin{aligned} d(R,t,X,x_1,x_2) = & ||[x_1]_\times (RX+t)||^2 \\ & + ((RX+t)^T [t']_\times R' x_2)^2. \end{aligned} \tag{12}$$

Par la suite, la position de la caméra 1 est solution de :

$$\begin{aligned} & \{R^*, t^*\} = \\ & \underset{R,t}{\text{argmin}} \quad \sum_{j=1}^{m} w(j) \{(RX^{\phi(j)}+t)^T [t']_\times R' x_2^j\}^2 \\ & \text{tel que} \quad w(j)|||[x_1^j]_\times (RX^{\phi(j)}+t)||^2 = 0, j = 1 \dots m. \end{aligned} \tag{13}$$

Comme ce problème est linéaire en R et t, il est résolu par décomposition en valeur singulière.

3.3 Estimation robuste de la pose des caméras

Le recalage approximatif précédent permet d'initialiser une optimisation plus robuste non linéaire tenant compte des problèmes soulignés en Section 3.1. Cette étape permet de raffiner la pose de la première caméra ainsi que la pose de la seconde caméra en tenant compte des connaissances 3D sur la scène. On résoud ainsi le problème robuste par moindres carrées pondérés itératifs suivant

$$\begin{aligned}\{q^*, t^*, q'^*, t'^*\} = \\ \underset{q,t,q',t'}{\text{argmin}} \quad & \sum_{j=1}^{m} w(j)\rho((R_q X^{\phi(j)} + t)^T [t']_\times R_{q'} x_2^j) \\ \text{tel que} \quad & w(j)\rho(|||[x_1^j]_\times (R_q X^{\phi(j)} + t)||) = 0, \\ & \rho((x_1^j)^T [t']_\times R_{q'} x_2^j) = 0,\ j = 1 \dots m \\ & ||q||^2 = 1, ||q'||^2 = 1, ||t'||^2 = 1. \end{aligned} \quad (14)$$

3.4 Algorithme

A partir d'une initialisation frustre des paramètres de la première caméra $\{R_0, t_0, R'_0, t'_0\}$, et la matrice essentielle entre les deux caméras, l'algorithme estime itérativement $\{R_k, t_k, R'_k, t'_k, \phi_k\}$ solutions de (8). Chaque itération consiste :

1. **Alignement des caméras** : chaque caméra est alignée avec la scène 3D à partir de la position initiale $\{R_{k,0}, t_{k,0}\} = \{R_k, t_k\}$. Chaque itération ($l = 0 \dots r$) procède de la façon suivante :
   (a) calcul des correspondances 2D-3D (11) ;
   (b) construction de l'histogramme d'échelle et calcul des poids $w(j), j = 1 \dots n$;
   (c) mise à jour de la position de la caméra 1 (13).
2. **Raffinement simultané** : à partir de $\{R_{k,r}, t_{k,r}, R'_k, t'_k, \phi_{k,r}\}$, la pose des deux caméras est obtenue par résolution de (14).

3.5 Normalisation et pose

Pour des raisons de stabilité numérique, les points de la scène 3D sont normalisés de telle sorte que la distance entre le centre de gravité du nuage de points et la première caméra soit approximativement égal à 1. Si l'estimation initiale de la première pose de la caméra est $\{R_0, t_0\}$, la normalisation consiste à poser $x_n^i = (R_0 X^i + t_0)/||t_0||$, , $i = 1 \dots n$. Après cette transformation , R_0 et t_0 sont simplifiés en $I_{3\times 3}$ et $0_{3\times 1}$ respectivement. Si les paramètres de recalage optimaux obtenus à partir de l'optimisation sont R^*, t^*, R'^* et t'^* ; R' et t' sont mis à jour en R'^* et t'^*, et R, t sont mis à jour en $R^* R_0$ et $R^* t_0 + ||t_0|| t^*$. De plus, on normalise également les données lors de l'estimation robuste c'est-à-dire y dans l'équation (9) est mis à l'échelle avec deux fois sa valeur médiane et $\xi = 1$. L'algorithme est interrompu lorsque l'incrément de la pose entre deux itérations consécutives $k-1$ et k est non significatif.

3.6 Généralisation au cas multi-vues

Dans le cas multi-vues ($p > 2$ caméras), la problème est alors modélisé par

$$\begin{aligned}\underset{\substack{q_l, q'_l, \phi, \\ t_l, t'_l}}{\min} \quad & \sum_{l=1}^{p}\sum_{j=1}^{m} w_l(j)\rho((R_l X^{\phi(j)} + t_l)^T [t'_l]_\times R'_l x_{l+1}^j) \\ \text{tel que} \quad & w_l(j)\rho(|||[x_l^j]_\times (R_l X^{\phi(j)} + t_l)||) = 0, \\ & \rho((x_l^j)^T [t'_l]_\times R'_l x_{l+1}^j) = 0,\ j = 1 \dots m \\ & ||q_l||^2 = 1, ||q'_l||^2 = 1, \\ & ||t'_l||^2 = 1,\ l = 1 \dots p-1. \end{aligned} \quad (15)$$

La résolution de ce problème s'apparente à un ajustement de faisceaux [13] pour lequel nous ne modifions pas les points 3D. Dès qu'une paire de caméra est localisée, la pose de la deuxième caméra permet d'initialiser l'algorithme pour la paire suivante. Notons que comme le raffinement de la position des caméras est contraint aux données 3D connues, les erreurs introduites entre les paires suivantes ne sont pas propagées dans l'estimation des positions suivantes. De plus, seule une estimation grossière de la première caméra suffit dans le cas multi-vues.

4 Experimentations

Nous avons testé notre méthode à partir de données synthétiques et réelles. Nos résultats avec les données de synthèse sont comparés à ceux d'une ICP avec la méthode des 5 points [8]. C'est-à-dire que dans un premier temps, nous calculons la structure 3D à partir des données 2D (SfM) puis nous recalons le nuage de points calculé sur le nuage connu (ICP). Les résultats obtenus sur deux benchmarks réels et une scène générée à l'aide d'une Kinect sont comparés aux approches SfM.

4.1 Simulations

Nous avons généré un ensemble de 800 points 3D aléatoires dispersés sur la surface de quatre faces d'un cube de dimension $[-10\ ,\ 10]^3$. Les caméras ont été placées à environ 20 ± 2 de l'origine du cube avec des rotations générées aléatoirement tout en regardant à peu près vers le centre de gravité de la scène. Tous les points de la scène ont été projetés sur des images de taille 256×256 avec un *skew* de zéro et une focale de 100 au centre de l'image. Les données 2D ont été obtenues en bruitant les données par un bruit blanc. Sur les 800 points, 400 sont aléatoirement utilisés pour localiser la seconde caméra par rapport à la première à l'aide d'une technique SfM classique. Au cours de ce processus, la moitié des points est rejetée pour réduire au minimum l'effet des valeurs aberrantes menant ainsi à la reconstruction de 200 points. Les mêmes données ont été utilisées pour notre méthode pour effectuer le recalage et le raffinement des positions. Nous avons exécuté 100 tests pour chaque écart type de bruit (de 0 à 2,0).
L'initialisation grossière de la méthode est réalisée en bruitant la position de la première caméra par un bruit de

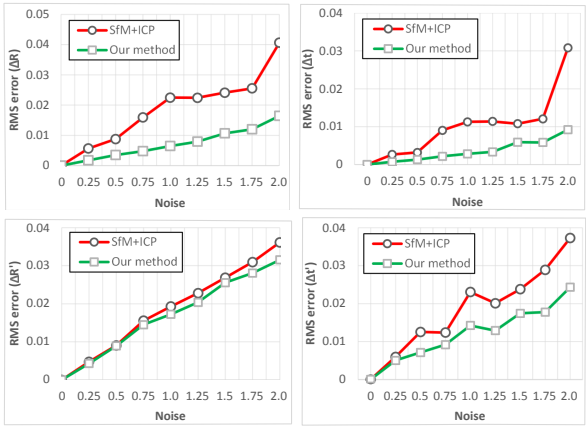

FIGURE 2 – Comparaison de notre approche avec une SfM+ICP en fonction du bruit dans les données.

$[0.05\,0.075]^c$ sur les trois angles de la rotation. Nous avons introduit ces erreurs sur R pour observer la convergence de notre méthode itérative. De même, des erreurs de $\pm 5\%$ sont introduites en translation. Notons que ces erreurs sont très importantes car la scène est relativement loin des caméras. L'histogramme a été construit avec 10 classes après avoir éliminé les échelles inférieures à 0,1 et supérieure à deux fois la médiane. Dans un premier temps, nous avons retenu les meilleures estimations possibles de R, R' et t en utilisant une SfM et une ICP. Comme l'ICP ne peut être réalisée sans connaissance de l'échelle relative, l'échelle est calculée en supposant que la reconstruction 3D déduite du SfM est equi-répartie sur l'ensemble de la scène 3D fournie. Notons que notre méthode ne nécessite pas ces informations supplémentaires pour déduire l'échelle. Afin de comparer avec notre approche, nous comparons les erreurs résiduelles des estimations de ΔR, Δt, $\Delta R'$, and $\Delta t'$ par rapport à la vérité terrain. La figure 2 montre ces erreurs en fonction du niveau de bruit dans les données. Nous pouvons observer que notre méthode améliore significativement les estimations par rapport à une approche SfM+ICP.

4.2 Données réelles

Dans la première expérience réelle, nous avons construit la scène 3D à partir d'images acquises avec un capteur 3D (Kinect). Cette scène a ensuite été sous-échantillonnée à environ 50 000 points comme le montre la figure 3 (à gauche). Une fois la 3D de la scène acquise, un ballon de football de taille standard a été placé dans cette même scène et deux images de taille 1080×1920 ont été capturées par une caméra mobile. Ces images et leurs 1198 correspondances sont présentées dans la figure 3. Nous avons sélectionné manuellement 14 points (TI) à partir des coins de l'icosaèdre (Fig. 4 (à droite)) afin d'évaluer la qualité de la reconstruction.

Pour une analyse quantitative, les paramètres géométriques suivants des points reconstruits TI sont calculés : (i) LS : erreur RMS de la longueur des côtés. (ii) AH : erreur RMS des angles intérieurs d'hexagones. (iii) AP : erreur

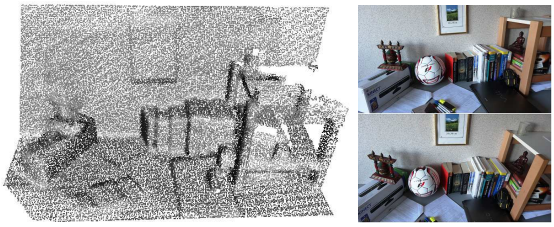

FIGURE 3 – Gauche : 3D acquis par une Kinect ; Droite : paire d'images.

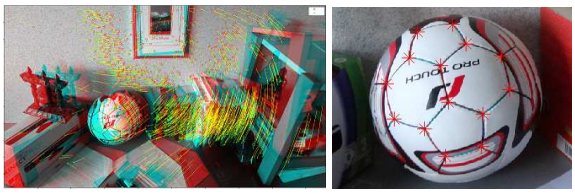

FIGURE 4 – Gauche : Correspondances ; Droite : points caractéristiques utilisés comme vérité terrain.

RMS des angles internes de pentagones. (iv) A- HP : erreur RMS d'angles dièdres entre hexagones et pentagones. (v) A- HH : angle dièdre entre deux hexagones (vi) CS : Circonférence de la sphère. Le tableau 1 compare ces paramètres avec la norme de la FIFA par rapport à une SfM classique. Ceci est un exemple de fusion 2D-3D, dans lequel la totalité de la reconstitution des données à partir de deux points de vue de la caméra n'est pas présente dans la scène d'origine. Cet exemple démontre également que notre approche est robuste aux problèmes d'occultation (objet placé dans la scène après l'acquisition 3D). En outre, même si les données 3D *a priori* ne sont pas très précises (acquisition Kinect), ce résultat montre l'intérêt d'utiliser cette information pour localiser les caméras dans la scène.

Nous avons également testé notre méthode avec les benchmarks Fontaine-P11 et Herz-Jesu-K7 (Fig. 5 disponible à `http://cvlabwww.epfl.ch/~Strecha`). Ces données sont constituées, respectivement, de 11 et 7 images de taille 3072×2048. Dans un premier temps, les reconstructions 3D pour chaque paire d'images consécutives sont obtenues par une méthode SfM classique. Tous ces résultats sont ensuite raffinés en utilisant notre méthode. Les résultats avant et après raffinement sont comparés à la vérité terrain dans le tableau 2. Les erreurs 3D présentées ici sont l'erreur moyenne RMS 3D de toutes les paires. Dans le cas multi-vues, le recalage est assujetti à

	LS (cm)	AP	AH	A-HP	A-HH	CS (cm)
SfM	0.20	4.26	2.00	6.19	140.19	76.25
Méthode proposée	0.11	2.94	0.86	3.34	139.20	73.10

TABLE 1 – Erreur de reconstruction 3D du ballon.

FIGURE 5 – Gauche : Fountaine-P11 ; Droite : Herz-Jesu-K7.

	Méthode	Fountaine	Herz-Jesu
$\Delta R'$(RMS)	SfM	0.0044	0.0072
	Notre approche	8.49e-4	0.0013
$\Delta t'$(RMS)	SfM	0.0404	0.0757
	Notre approche	0.0031	0.0052
3D error	SfM	0.0011	0.0025
	Notre approche	5.95e-4	0.0018

TABLE 2 – Comparaison entre SfM et notre approche.

l'accumulation d'erreurs et la dérive du facteur d'échelle. De plus, nous avons raffiné ces résultats en utilisant notre méthode ainsi qu' un ajustement de faisceaux (BA) [7]. Les résultats obtenus en utilisant notre approche se sont révélés significativement meilleurs que ceux de BA. Nous avons également affiné nos résultats en appliquant un BA après convergence de notre méthode. Les résultats obtenus à partir de BA, notre méthode et notre méthode puis BA sont présentés dans le tableau 3 . On observe que le BA effectué sur nos résultats s'écarte de la vérité terrain. Ceci est dû au fait que le BA ne prend en compte que les informations image et n'intègre ni les connaissances 3D, ni le bruit présent dans l'image et ni les erreurs de calibrage.

5 Conclusion

Dans ce travail, nous avons proposé une méthode permettant de localiser précisément deux ou plusieurs caméras dans un environnement connu. Nous avons démontré qu'il était possible de recaler précisément les images 2D avec la scène 3D en utilisant uniquement les points caractéristiques 2D. L'utilisation d'une scène 3D connue pour estimer la pose d'une caméra est essentielle pour parvenir à une telle précision. Pour ce faire, une méthode de recalage 2D-3D a été développée. Nous avons ainsi montré que lorsque la caméra se déplace dans un environnement connu, nous pouvions utiliser cette information afin d'améliorer l'estimation de pose de la caméra.

	Méthode	Fountaine	Herz-Jesu
ΔR(RMS)	BA	0.0436	0.0123
	Notre approche	0.0020	0.0067
	Notre approche + BA	0.0251	0.0080
Δt(RMS)	BA	0.0311	0.0402
	Notre approche	0.0019	0.0224
	Notre approche + BA	0.0172	0.0241
3D error	BA	0.0020	0.0069
	Notre approche	0.0015	0.0068
	Notre approche + BA	0.0020	0.0069

TABLE 3 – Comparaison entre notre approche avec un ajustement de faisceaux (BA) et un ajustement de faisceaux réalisé après convergence de notre méthode.

Références

[1] S. Christy and R. Horaud. Iterative pose computation from line correspondences. In *Comput. Vis. Image Underst.*, pages 137–144, January 1999.

[2] M. J. Clarkson, D. Rueckert, D. L.G. Hill, and D. J. Hawkes. Using photo-consistency to register 2D optical images of the human face to a 3D surface model. In *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1266–1280, November 2001.

[3] M. Corsini, M. Dellepiane, F. Ganovelli, R. Gherardi, A. Fusiello, and R. Scopigno. Fully automatic registration of image sets on approximate geometry. In *Int. J. Comput. Vision*, pages 91–111, March 2013.

[4] J.A. Hesch and S.I. Roumeliotis. A direct least-squares (DLS) method for PnP. In *ICCV*, 2011.

[5] J. Knopp, J. Sivic, and T. Pajdla. Avoiding confusing features in place recognition. In *ECCV*, 2010.

[6] L. Liu and L. Stamos. Automatic 3D to 2D registration for the photorealistic rendering of urban scenes. In *CVPR*, 2005.

[7] M.I. A. Lourakis and A.A. Argyros. SBA : A software package for generic sparse bundle adjustment. In *ACM Trans. Math. Software*, pages 1–30, 2009.

[8] D. Nistér. An efficient solution to the five-point relative pose problem. In *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 756–777, June 2004.

[9] Srikumar R., Sofien B., Peter S., and Matthew B. Geolocalization using skylines from omni-images. In *ICCV Workshops*, 2009.

[10] T. Sattler, B. Leibe, and L. Kobbelt. Fast image-based localization using direct 2D-to-3D matching. In *ICCV*, 2011.

[11] M. Tamaazousti, V. Gay-Bellile, S. N. Collette, S. Bourgeois, and M. Dhome. Nonlinear refinement of structure from motion reconstruction by taking advantage of a partial knowledge of the environment. In *CVPR*, 2011.

[12] A. Taneja, L. Ballan, and M. Pollefeys. Registration of spherical panoramic images with cadastral 3D models. In *3DIMPVT*, 2012.

[13] B. Triggs, P. Mclauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment - a modern synthesis. In *Vision Algorithms : Theory and Practice, LNCS*, pages 298–375. Springer Verlag, 2000.

[14] P. Viola and W. Wells, III. Alignment by maximization of mutual information. In *Int. J. Comput. Vision*, pages 137–154, September 1997.