Simulation de point de vue pour la mise en correspondance et la localisation

Pierre Rolin, Marie-Odile Berger, Frédéric Sur

LORIA, UMR CNRS 7503
Université de Lorraine
INRIA Nancy Grand Est
pierre.rolin@loria.fr

RÉSUMÉ. *On considère le problème de la localisation d'une caméra à partir d'un modèle non structuré obtenu par un algorithme de type structure from motion. Dans ce modèle, un point est représenté par ses coordonnées et un ensemble de descripteurs photométriques issus des images dans lesquelles il est observé. La localisation repose sur l'appariement de points d'intérêt de la vue courante avec des points du modèle, sur la base des descripteurs. Cependant le manque d'invariance des descripteurs aux changements de point de vue rend difficile la mise en correspondance dès que la vue courante est éloignée des images ayant servi à construire le modèle. Les techniques de simulation de point de vue, comme ASIFT, ont récemment montré leur intérêt pour la mise en correspondance entre images. Cet article explore l'apport de ces techniques pour enrichir le modèle initial par des descripteurs simulés et évalue le bénéfice respectif de simulations affines et homographiques. Nous montrons en particulier que la simulation augmente la proportion de bons appariements et la précision du calcul de pose et permet de calculer une pose là où l'approche basée uniquement sur les descripteurs SIFT échoue.*

ABSTRACT. *We consider the problem of camera pose estimation from a scene model obtained beforehand by a structure from-motion (SfM) algorithm. The model is made of 3D points, each one of them being represented by its coordinates and a set of photometric descriptors such as SIFT, extracted from some of the input images of the SfM stage. Pose estimation is based on the matching of interest points from a test view with model points, using the descriptors. Descriptors having a limited invariance with respect to viewpoint changes, such an approach is likely to fail when the test view is far away from the images used to construct the model. Viewpoint simulation techniques, as ASIFT, have proved effective for wide-baseline image matching. This paper explores how these techniques can enrich a scene model by adding descriptors from simulated views, and evaluate the respective benefits of affine and homographic simulations. In particular we show that viewpoint simulation increases the proportion of correct correspondances, and permits pose estimation in situations where the approach based on the sole SIFT descriptors simply fails.*

MOTS-CLÉS : *calcul de pose, simulation de point de vue, appariement à un modèle 3D.*

KEYWORDS: *pose estimation, viewpoint simulation, 3D model matching.*

DOI:10.3166/TS.32.169-194

1. Introduction

L'estimation de la pose dans un environnement connu est un problème primordial, par exemple pour l'initialisation de pose (Collet *et al*., 2009), la relocalisation dans une approche de type SLAM en cas de perte du suivi (Williams *et al*., 2007), et de façon générale pour les applications de localisation (Schindler *et al*., 2007) ou de réalité augmentée (Gordon, Lowe, 2006). Dans cet article nous nous intéressons à l'estimation de la pose à partir de correspondances entre des points d'intérêt extraits d'une vue test et des points d'un modèle 3D non structuré, comme (Williams *et al*., 2007 ; Gordon, Lowe, 2006 ; Rothganger *et al*., 2006). Le modèle de la scène utilisé est un nuage de points obtenu à partir d'un ensemble d'images en utilisant un algorithme de type *structure from motion* (SfM) (Wu *et al*., 2011 ; Wu, 2011). Un tel algorithme commence par apparier des points d'intérêt dans les images en utilisant des descripteurs photométriques. Les chaînes de descripteurs issus de plusieurs images sont ensuite utilisées pour simultanément estimer la pose des caméras et les positions des points 3D par triangulation et ajustement de faisceaux. Les points du modèle sont associés à l'ensemble des descripteurs ayant servi à les reconstruire. Plusieurs façons de représenter les points du modèle sont envisagées dans la littérature, notamment l'utilisation de patchs invariants (Rothganger *et al*., 2006) ou de mots visuels construits à partir de descripteurs photométriques (Bhat *et al*., 2011 ; Irschara *et al*., 2009). Dans cet article nous utilisons des ensembles de descripteurs SIFT (Lowe, 2004), comme (Gordon, Lowe, 2006). Chaque point 3D du modèle est donc associé à la classe des descripteurs SIFT présents dans la chaîne de correspondances utilisée pour le reconstruire.

Estimer la pose d'une caméra à partir de la vue test en utilisant le modèle consiste à résoudre le problème *Perspective-n-Points* (DeMenthon, Davis, 1995 ; Lepetit, Fua, 2005 ; Lepetit *et al*., 2009) pour un ensemble de correspondances entre la vue test et le modèle. Cette approche est limitée par l'invariance des descripteurs photométriques, qui est limitée à des changements d'orientation de 30° (Moreels, Perona, 2007). Si la vue test est trop éloignée des vues réelles, la mise en correspondance des descripteurs SIFT ne produit plus un ensemble de correspondances suffisant pour résoudre le problème PnP.

1.1. État de l'art et contribution

Dans cet article, nous proposons d'enrichir la description des points 3D en générant par simulation des descripteurs additionnels qui correspondent à des points de vue éloignés de ceux des images ayant servi à la reconstruction initiale du modèle. Nous montrons au travers de plusieurs expériences qu'enrichir ainsi le modèle augmente le degré d'invariance de la description des points 3D. Cela facilite l'appariement, et donc le calcul de pose, lorsque la scène présente un fort changement d'aspect dans la nouvelle vue dont on cherche la pose. La simulation de points de vue a déjà montré son utilité dans le cadre de la mise en correspondance entre deux images présentant un changement de point de vue important, cf. ASIFT (Morel, Yu, 2009) ou dans une moindre mesure FERNS (Ozuysal *et al*., 2010). Dans ASIFT ou FERNS, la simula-

tion est faite en utilisant des transformations affines. Dans notre cas, en supposant la scène localement plane, toutes les vues d'une région autour d'un point 3D sont liées par des homographies avec le modèle sténopé ou des transformations affines avec le modèle orthographique. Nos descripteurs simulés seront donc générés à partir de vues synthétisées par un certain nombre de transformations de l'un de ces deux types, de manière à simuler un déplacement de la caméra dans des positions non représentées dans les images initiales. Ceci est illustré sur les figures 1 et 2. Une approche similaire est envisagée dans (Kushnir, Shimshoni, 2012) et (Wu *et al.*, 2008), mais les vues simulées sont uniquement fronto-parallèles.

Dans (Hsiao *et al.*, 2010), la simulation est utilisée pour améliorer la reconnaissance d'objets. Un modèle des objets à reconnaître est construit par un algorithme de type *SfM*, en utilisant une grande variété de directions de vue (25 points de vus régulièrement répartis sur un cercle autour de l'objet). Les simulations effectuées pour compléter ces points de vue sont faites au sens d'ASIFT : c'est-à-dire de manière globale sur une image. Contrairement à ce qui est proposé dans notre méthode, la géométrie de l'objet n'est pas prise en compte pour faire les simulations. L'ensemble des descripteurs associés à chaque point du modèle est réduit à un ensemble d'élément représentatifs par une approche de type *mean-shift* (Comaniciu, Meer, 2002).

Dans (Irschara *et al.*, 2009) la simulation est utilisée dans le cadre de la localisation d'une caméra dans un grand environnement. Le modèle de la scène est, comme dans notre approche, issu d'un algorithme de type *SfM*, mais les environnements considérés sont beaucoup plus vastes et le nombre de vues de construction utilisées plus important. La classe des descripteurs associés à chaque point du modèle est réduite à quelques éléments représentatifs. La mise en œvre de la simulation est peu décrite, il semble que les vues virtuelles sont générées globalement comme dans ASIFT. La mise en correspondance image/modèle utilise des correspondances entre images qui sont ensuite validées géométriquement, contrairement à notre approche qui met directement en correspondance les descripteurs de l'image avec ceux du modèle. Les changements de point de vue considérés sont relativement faibles, contrairement à ceux présents dans nos expériences.

La méthode que nous proposons permet de localiser une caméra par rapport à un modèle obtenu par *SfM*, même lorsque la pose cherchée est très éloignée des vues ayant été utilisées pour construire le modèle. Cette invariance aux changements de points de vue est obtenue en complétant la description de chaque point du modèle par des descripteurs issus de patchs simulés prenant en compte la géométrie locale du point. Nos expériences montrent que l'ajout de 25 points de vue virtuels permet de calculer des poses là où des méthodes basées uniquement sur SIFT (Gordon, Lowe, 2006) échouent. De façon plus générale, la simulation augmente le taux de correspondances correctes entre l'image et le modèle, ce qui permet de converger plus rapidement vers la pose cherchée.

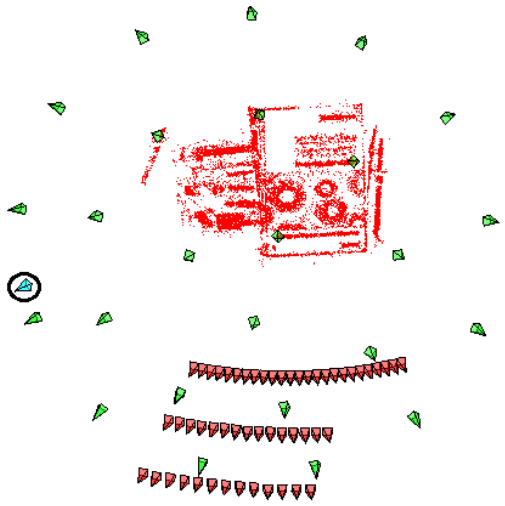

Figure 1. Le modèle 3D de la scène (points rouge), les caméras ayant servi à le construire (en rouge pâle), une caméra éloignée dont on chercherait la pose (en cyan, entouré), et les caméras virtuelles (en vert), réparties ici sur une demi-sphère. Les caméras virtuelles permettent de générer de nouveaux descripteurs pour chaque point du modèle

(a) (b)

Figure 2. La vue test (a) et la vue réelle la plus proche de celle-ci (b). Notons le fort changement de point vue. La vue (a) pourra difficilement être appariée au modèle à partir de descripteurs issus de vues comme (b)

1.2. Vocabulaire et conventions

Dans l'ensemble de l'article, nous appelons *vue réelle* une vue ayant servi à la reconstruction de la scène, *vue virtuelle* (ou *patch virtuel*) une vue (ou un patch) obtenu par simulation affine ou homographique à partir d'une vue réelle et *vue test* une vue de la scène dont on veut calculer la pose. La *classe* des descripteurs associés à un point 3D désigne l'ensemble des descripteurs utilisés pour reconstruire ce point ainsi que ceux ajoutés après l'étape de simulation.

Dans toutes les figures, les points de vue utilisés pour la reconstruction de la scène sont en rouge, les points de vue virtuels en vert, les points de vue test en cyan et les points de vue calculés en bleu.

1.3. Plan de l'article

Dans la section 2 nous détaillons la simulation par transformation affine ou homographique. La section 3 explique comment le modèle non structuré est enrichi à l'aide des descripteurs simulés, et comment nous procédons à l'appariement image/modèle permettant de déterminer la pose. La section 4 présente une étude expérimentale et une comparaison des modèles affines et homographiques.

2. Simulation de points de vue dans un monde localement plan

Nous supposons disposer d'un modèle d'une scène, constitué d'un nuage de points, et que chacun de ces points est associé à un ensemble de descripteurs SIFT provenant des vues réelles dans lesquelles il a été repéré. Nous supposons également que la scène est localement plane autour des points 3D, et que l'on a associé à chaque point le vecteur normal du plan sur lequel il se trouve. Étant donnée une vue réelle d'une zone plane autour d'un point 3D, comment synthétiser une vue de cette zone à partir d'une nouvelle position de caméra, afin d'en extraire un nouveau descripteur SIFT ?

Si on modélise les caméras comme des sténopés, deux vues d'un même plan sont liées par une homographie. Dans le modèle de caméras affines (lorsque la profondeur de la scène est faible devant la focale), les deux vues sont liées par une transformation affine. Les auteurs de (Morel, Yu, 2009 ; Ozuysal *et al.*, 2010) montrent que cette simplification est souvent suffisante. En effet, comme une transformation affine est une approximation au premier ordre d'une homographie, des transformations affines ou homographiques d'une petite zone de l'image sont visuellement proches. Néanmoins les descripteurs SIFT sont souvent extraits sur des disques de plusieurs dizaines de pixels de rayon, pour lesquels l'approximation affine n'est plus valide dès que l'angle entre les vues est assez grand (plus grand que 30°).

2.1. Cas des homographies

Soient deux caméras représentées par leurs matrices de projection $P_1 = K_1[R_1|T_1]$ et $P_2 = K_2[R_2|T_2]$ (où K_i est la matrice des paramètres intrinsèques pour un capteur à pixels carrés, et R_i, T_i déterminent la pose dans un repère commun, $i \in \{1, 2\}$). Considérons un plan de l'espace d'équation $n^T X + d = 0$ (où n est un vecteur normal au plan, d un paramètre réel, et X des coordonnées d'un point de l'espace). La transformation induite par le plan entre les deux caméras est alors l'homographie H donnée par l'équation homogène (Hartley, Zisserman, 2004) :

$$H = K_2(R - Tn^T/d)K_1^{-1} \tag{1}$$

où $R = R_2R_1^T$ et $T = -R_2(C_2 - C_1)$ (où le centre optique C_i vérifie $C_i = -R_i^T T_i$, $i \in \{1, 2\}$.)

Remarquons que dans le cas où les deux caméras partagent le même axe optique et que celui-ci porte le vecteur n, cette homographie se réduit à une similitude.

Si P_1 est la matrice de projection d'une caméra réelle, P_2 celle d'une caméra virtuelle, et I_1 et I_2 les images du plan dans ces deux caméras, alors $HI_1 = I_2$, soit :

$$K_2R_2(R_1^T + (C_2 - C_1)n^T/d))K_1^{-1}I_1 = I_2. \tag{2}$$

Rappelons que la matrice R_2 s'écrit $R_2 = R_Z(\kappa)R_Y(\phi)R_X(\omega)$ où (X, Y, Z) est un repère orthonormé tel que Z est l'axe optique de la caméra et (κ, ϕ, ω) sont les angles d'Euler associés. Les descripteurs SIFT étant supposés invariants par similitude (plane), on voit que toute rotation autour de l'axe optique ou tout changement de focale de la caméra 2 fournira les mêmes descripteurs. Donc la pose de la caméra virtuelle n'a besoin d'être fixée qu'à une rotation selon l'axe optique près, et la focale est arbitraire. Comme il l'a été souligné dans (Morel, Yu, 2011), ce raisonnement sur des images idéales continues reste valable pour des images discrètes sous réserve de respect de la condition de Shannon-Nyquist. Néanmoins la position de la caméra est ici importante (T_2 intervient dans (2)).

La donnée du plan, d'une pose de caméra réelle, et de la pose de la caméra virtuelle (à une rotation selon l'axe optique près) permet de simuler avec l'équation (2) une vue de laquelle nous allons extraire un descripteur SIFT.

2.2. Cas des transformations affines

Dans le cas de deux caméras affines, notons $(\lambda_i, \psi_i, t_i, \phi_i)$ les éléments caractéristique de la caméra $i \in \{1, 2\}$ dans un repère associé à un plan repéré par son vecteur normal n (figure 3). Les angles ϕ_i et θ_i sont respectivement la longitude et la latitude de l'axe optique de la caméra. Le paramètre $t_i = 1/cos(\theta_i)$ est le tilt de la caméra. Le paramètre ψ_i correspond à la rotation de la caméra autour de son axe optique et λ_i au zoom. La transformation induite par le plan entre une vue fronto-parallèle de ce plan et la vue i est donnée par la transformation affine suivante (Morel, Yu, 2009 ; Ozuysal *et al.*, 2010) :

$$A_i = \lambda_i \begin{pmatrix} \cos(\psi_i) & -\sin(\psi_i) \\ \sin(\psi_i) & \cos(\psi_i) \end{pmatrix} \begin{pmatrix} t_i & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cos(\phi_i) & -\sin(\phi_i) \\ \sin(\phi_i) & \cos(\phi_i) \end{pmatrix}. \quad (3)$$

Par composition, la transformation affine induite par le plan entre les deux caméras est :

$$A = A_2 A_1^{-1}. \quad (4)$$

Avec les mêmes notations que dans le cas des homographies, $AI_1 = I_2$ soit $A_1^{-1} I_1 = A_2^{-1} I_2$. L'invariance aux similitudes des descripteurs SIFT nous permet d'écrire que toutes les valeurs de $\psi_1, \psi_2, \lambda_1, \lambda_2$ fournissent les mêmes descripteurs SIFT, que l'on choisit donc arbitrairement à $\psi_1 = \psi_2 = 0, \lambda_1 = \lambda_2 = 1$.

Ainsi la donnée des positions relatives (t_i, ϕ_i) des caméras réelles et virtuelles par rapport à la normale à une partie plane de la scène permet de simuler une vue avec l'équation (4) de laquelle on extraira un descripteur SIFT.

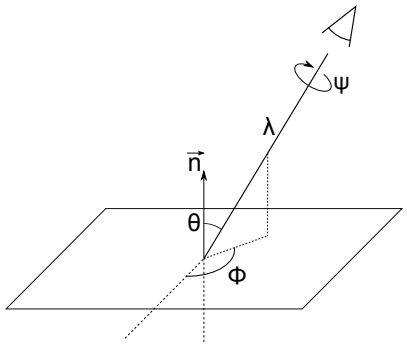

Figure 3. Position d'une caméra affine par rapport à la normale d'un morceau de plan, avec les notations de l'équation (3) où $t = 1/\cos(\theta)$

2.3. Résumé

Pour chaque point du modèle 3D associé à une direction normale, et pour chaque position de caméra virtuelle, on génère une vue (selon une transformation homographique ou affine selon la méthode choisie), puis on extrait un descripteur SIFT dans cette vue que l'on associe au point 3D. Un exemple de simulation est visible sur la figure 4.

3. Mise en œuvre

Un modèle non structuré est construit et les points associés à un ensemble de descripteurs SIFT et au vecteur normal au plan sous-jacent (section 3.1), puis des descripteurs associés à des vues simulées sont ajoutés (section 3.2). La pose d'une nouvelle vue peut ensuite être estimée à partir de ce modèle enrichi (section 3.3).

Figure 4. Un exemple de simulation. Les vues simulées de la couverture du livre (affine à gauche et homographique à droite) sont obtenues à partir de la vue réelle. La simulation par homographie ressemble d'avantage à la vue test

3.1. Construction du modèle

Le logiciel VisualSFM (Wu, 2011) est utilisé pour générer un ensemble de points $\mathcal{P}$ de la scène tridimensionnelle, chaque point étant associé à la classe des descripteurs SIFT extraits des images dans lesquelles il est vu. Le logiciel permet également de générer une reconstruction dense de la scène basée sur (Furukawa, Ponce, 2010). Nous utilisons ce modèle dense pour générer en chaque point de $\mathcal{P}$ une estimation de la normale en considérant le plus petit vecteur propre d'une analyse en composantes principales des coordonnées de ses k-plus proches voisins (Hoppe *et al.*, 1992). La normale est orientée vers les caméras dans lesquelles le point considéré est repéré. Nous n'utilisons plus la reconstruction dense dans la suite.

3.2. Ajout de descripteurs simulés

3.2.1. Position des caméras virtuelles.

La position des caméras virtuelles est choisie de manière à compléter les points de vue des caméras ayant permis de construire le modèle. Comme on l'a vu dans la sec-

tion 2, le cas affine ne nécessite que de positionner les caméras sur une demie-sphère orientée par la normale considérée, alors que le cas homographique nécessiterait de préciser leur distance par rapport à la scène.

Dans cette étude préliminaire nous placerons les caméras virtuelles dans les mêmes positions dans les deux cas : il s'agit de vingt-cinq positions régulièrement réparties sur une demi-sphère s'appuyant sur un plan moyen de la scène, de rayon égal à la distance de la plus proche caméra à la scène, comme dans la figure 1 ; les caméras sont dirigées vers le barycentre de la scène. Nous simulons donc un grand nombre de directions d'observation de la scène, mais pas de variations de la distance de la caméra à la scène. Néanmoins, les expérience présentées dans la section 4.2.2 montrent que ces simulations son suffisantes pour calculer des poses relativement éloignées des vues de reconstruction et des vues virtuelles.

Cet échantillonnage est arbitraire pour le moment, mais devra à terme être défini en fonction de la géométrie de la scène et des points de vue utilisés pour construire le modèle.

3.2.2. Choix de la vue utilisée pour la simulation et extraction d'un descripteur SIFT.

Étant donné un point du modèle 3D (associé à des descripteurs venant de plusieurs vues réelles) et un point de vue à simuler, il faut également choisir à partir de quelle vue réelle réaliser la simulation. Parmi les vues dans lequel le point 3D est visible, la vue à partir de laquelle la simulation est réalisée est la plus proche angulairement du point de vue qu'on veut simuler, ce qui est un choix classique pour limiter l'influence des spécularités.

La simulation produit une imagette de taille 100×100 pixels centrée sur un point du modèle, qui correspond à l'apparence de ce point observé à partir d'une caméra virtuelle. L'algorithme SIFT permet alors d'extraire des couples de points d'intérêt et descripteurs dans cette imagette. On ajoute alors à la liste des descripteurs de ce point 3D le descripteur extrait de l'imagette dont le point d'intérêt est le plus proche de la position théorique de la projection du point 3D, si cette distance est inférieure à 10 pixels. Ce seuil correspond à une distance de reprojection typique des points du modèle obtenu par SfM.

3.3. Estimation de la pose

3.3.1. Correspondances image/modèle

On commence par extraire les descripteurs SIFT de la nouvelle vue. La méthode de mise en correspondance utilisée est celle proposée dans (Gordon, Lowe, 2006). Pour apparier un point d'intérêt p_1 de la nouvelle vue à un point 3D, on considère les distances d_1 et d_2 du descripteur SIFT de p_1 aux deux plus proches classes de descripteurs. Si d_1/d_2 est inférieur à un seuil λ on retient la correspondance. La recherche des plus proches voisins est accélérée comme dans (Gordon, Lowe, 2006) par une recherche approchée (Mount, Arya, 2010).

3.3.2. Perspective-n-Points

Le calcul de pose se fait par une estimation robuste de type RANSAC (Fischler, Bolles, 1981) basée sur l'algorithme PnP proposé dans (Hesch, Roumeliotis, 2011). Bien entendu, plus la proportion de correspondances correctes dans l'étape précédente est grande, plus le nombre d'itérations requises dans RANSAC peut être diminué.

4. Étude expérimentale

Les expériences suivantes montrent qu'en présence de fortes variations de direction de vue ou de profondeur la simulation de point de vue améliore considérablement l'estimation de la pose. La pose peut être calculée dans des situations où une approche basée uniquement sur SIFT, telle que celle de (Gordon, Lowe, 2006), échoue. Plus généralement, pour un nombre fixé d'itérations de RANSAC, la pose est calculée avec plus de précision en utilisant la simulation. À la fin de cette section nous discutons les problèmes de temps de calcul et les améliorations envisageables.

4.1. Protocole expérimental

La méthode proposée est évaluée sur quatre séquences d'images : la séquence numéro 2 de la base *Robot Data Set* avec la première illumination proposée (la reconstruction de la scène est présentée dans la figure 1 et les positions des caméras utilisées dans la figure 5) et trois séquences personnelles, illustrées dans la figure 6. Ces séquences sont composées d'images de taille 1600×1200 pixels et les scènes associées sont globalement planes par morceaux et centrées objet.

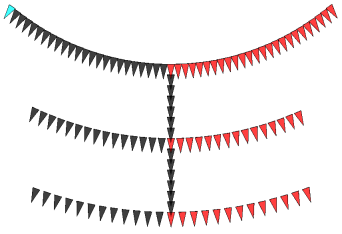

Figure 5. Les positions des 119 caméras de la base Robot Data Set. En rouge les caméras servant à la reconstruction par SfM, en cyan la caméra de test

Toutes les expériences utilisent le même protocole. Un modèle 3D de la scène est construit avec VisualSfM (section 3.1). La pose d'une vue test est calculée (section 3.3) dans trois scénarios : **S** où le modèle est la reconstruction obtenue par SfM sans simulation, **A** où le modèle de **S** est enrichi par des descripteurs issus de simulations affines (section 2.2), et **H** où le modèle de **S** est enrichi par des descripteurs issus de simulations homographiques (section 2.1).

Pour comparer les trois scénarios, 100 poses sont calculées pour la même vue test dans chaque cas en utilisant le même nombre d'itérations de RANSAC. La variabilité de ces 100 poses est évaluée visuellement. Lorsque ces poses sont superposées, nous calculons également l'écart type (reporté dans les figures). L'échelle étant un paramètre libre de toute reconstruction SfM, les écart types sont exprimés en pourcentage de la distance à la scène. De plus, pour chaque expérience, des contours des objets de la scène sont reprojetés dans la vue test en utilisant les poses calculées.

Comme les taux d'*inliers* dans les correspondances image/modèle sont très variables d'une séquence à une autre (e.g., de 4 % à 23 % pour le scénario **S**), nous utilisons un nombre d'itérations de RANSAC différent pour chaque séquence. Cependant, pour rendre possible la comparaison de la variabilité, le même nombre d'itérations est utilisé pour les trois scénarios.

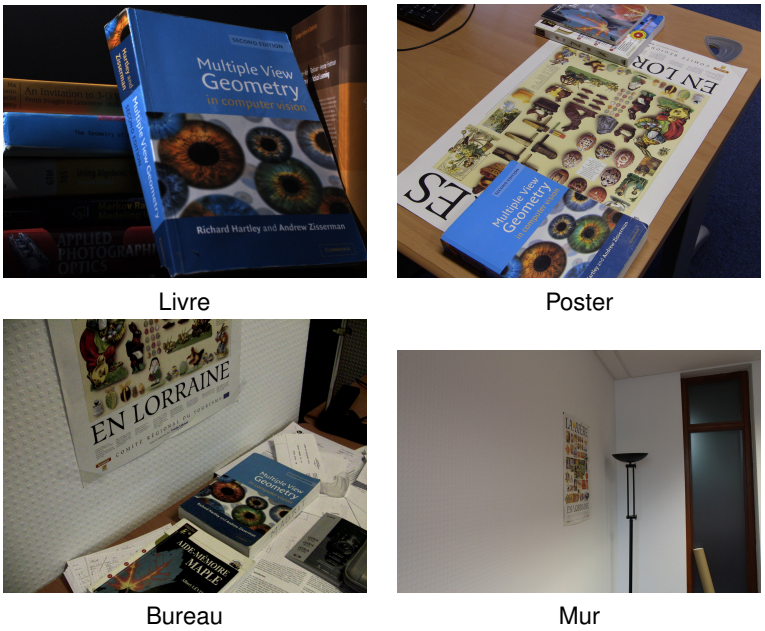

Figure 6. Images représentatives des quatre séquences. Livre *vient de (Aanæs* et al.*, 2012). Les autres séquences sont personnelles*

4.2. Amélioration du calcul de pose dans les modèles enrichis

4.2.1. Robustesse du calcul de pose aux changements de direction de vue

Nous montrons ici que la simulation de point de vue améliore significativement la précision des poses calculées lorsque la vue test est éloignée des vues réelles et a donc un aspect très différent.

Nous présentons d'abord les résultats sur la séquence Livre (figure 1) pour laquelle la pose réelle de la vue test est connue. Il est donc possible de déterminer si une correspondance 2D/3D est correcte ou non, en reprojetant le point 3D en utilisant la pose de la vérité terrain. Si la distance de reprojection est inférieure à 20 pixels la correspondance est considérée correcte (ce seuil correspond à $\mu + 3\sigma$ avec μ et σ respectivement la moyenne et l'écart type de l'erreur de reprojection de l'étape SfM ; les images sont de taille 1600×1200 pixels). Dans cette expérience la proportion de correspondances correctes est de 23 % dans le scénario **S**, 30 % dans le scénario **A** et 37 % dans le scénario **H**.

La figure 7 montre la répartition des correspondances 2D/3D parmi les vues réelles et simulées dans le scénario **H**. Le point de vue qui contribue le plus au calcul de pose est virtuel et proche de la caméra test. Globalement, les points de vue simulés produisent 85 % de l'ensemble de consensus de RANSAC. Ces graphes illustrent la pertinence de l'approche proposée et l'augmentation du taux d'inliers obtenue grâce aux simulations.

Les résultats du calcul de pose sont illustrés dans les figures 8 ($N = 500$) et 9 ($N = 1\,000$). Les poses estimées sont visuellement plus précises dans les scénarios **A** et **H** que dans le scénario **S**. Avec 500 itérations dans RANSAC, le calcul de la pose échoue dans **S**, alors que les résultats sont corrects dans **H**. En augmentant le nombre d'itérations à 1 000, la variabilité de la pose n'est que légèrement réduite dans **S** alors que dans **H** toutes les poses calculées sont superposées.

Un phénomène remarquable se produit dans **A** (et dans une moindre mesure dans **S**). Dans cette expérience les poses calculées se répartissent en trois catégories : la plupart des poses sont proches du point de vue attendu, quelques unes sont totalement fausses et un groupe de poses erronées se trouve face à la couverture du livre. Cet ensemble d'erreurs est provoqué par un motif répété de la scène, à savoir l'œil de la couverture qui apparaît également sur la tranche du livre. La reprojection des bords de la couverture dans la figure 9 illustre bien le phénomène. Dans ce cas, les simulations homographiques produisent plus de correspondances en dehors de ce motif répété, ce qui permet d'obtenir des poses correctes dans **H**. L'influence des motifs répétés est discutée par exemple dans (Noury *et al.*, 2010 ; Sur *et al.*, 2013 ; Roberts *et al.*, 2011).

Ces expériences ont été reproduites sur les séquences Poster et Bureau avec des résultats similaires, voir figures 10 et 11. Dans tous les cas présentés, la simulation améliore la précision de l'estimation de la pose, ce qui est illustré par la meilleure superposition des positions de caméra estimées ou des quadrilatères correspondant à la projection de contours 3D de la scène par les caméras estimées.

4.2.2. Robustesse du calcul de pose aux variations de distance par rapport à la scène

Comme expliqué dans la section 2, la simulation utilisant le modèle de caméra affine est indépendante de la distance du point de vue simulé à la scène. Bien que la simulation par homographie dépende, elle, de cette distance, touts les points de vue simulés sont choisis à la même distance de la scène. L'objectif de cette expérience est

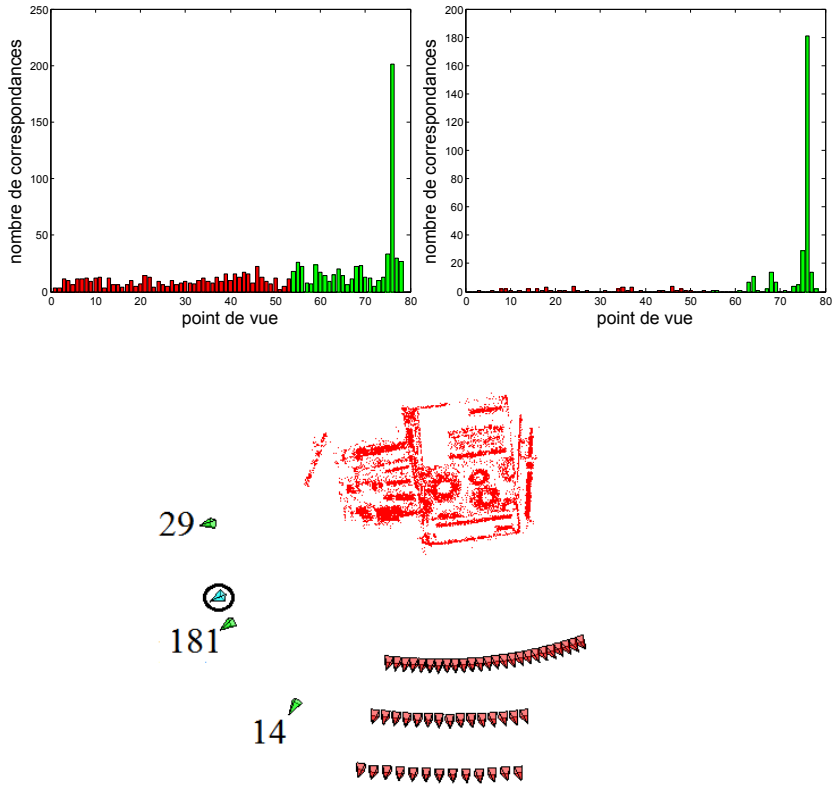

Figure 7. Séquence Livre *: nombre de correspondances associées à chaque point de vue (réel en rouge, virtuel en vert), pour l'ensemble des correspondances image/modèle (en haut à gauche) et dans l'ensemble de consensus de RANSAC (en haut à droite). Les points de vue contribuant le plus restent les mêmes, et sont proches de la pose cherchée. Les trois points de vue contribuant le plus et le nombre de correspondances associées sont montrés en bas*

de mettre en évidence l'influence de ce choix sur le calcul de pose lorsque le point de vue test est beaucoup plus éloigné de la scène que les points de vues utilisés pour la reconstruction.

Un modèle de la scène est construit à partir de 6 caméras orientées vers un poster (en rouge dans la figure 12). Cette scène a été choisie pour mettre en évidence l'apport de la simulation : nous avons besoin d'une caméra test non alignée avec l'axe optique des autres caméras et qui n'observe pas le poster en vue frontale, de telle sorte que la transformation résultante soit une homographie non réduite à une similarité.

Les vues de test sont donc prises avec un changement de direction de vue relativement faible mais de fortes variations de profondeur, voir figures 12 et 13. Le nombre d'itérations de RANSAC est $N = 300$ pour toutes ces expériences. Nous ne détaillons

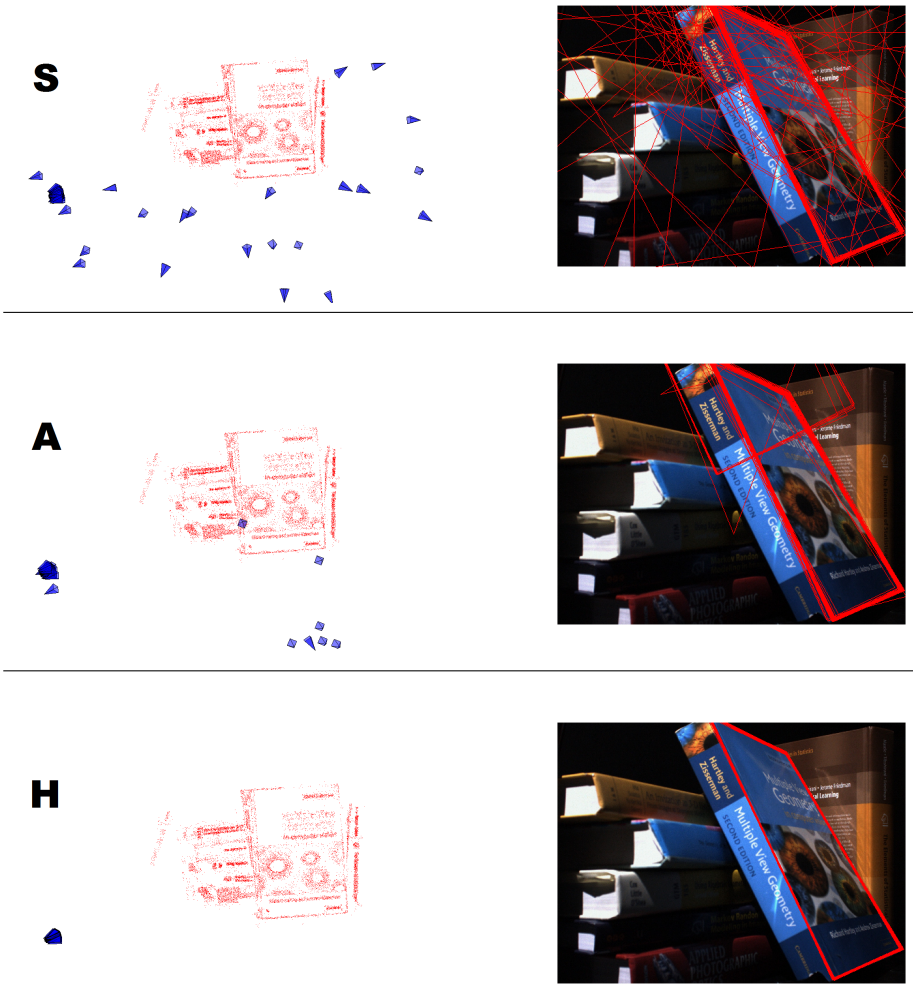

Figure 8. Séquence Livre *: 100 poses calculées avec N = 500 itérations de RANSAC, et la reprojection des bords de la couverture en utilisant ces 100 poses. Dans le scénario **H** l'écart type de la position de la caméra est 0,31 % de la distance à la scène*

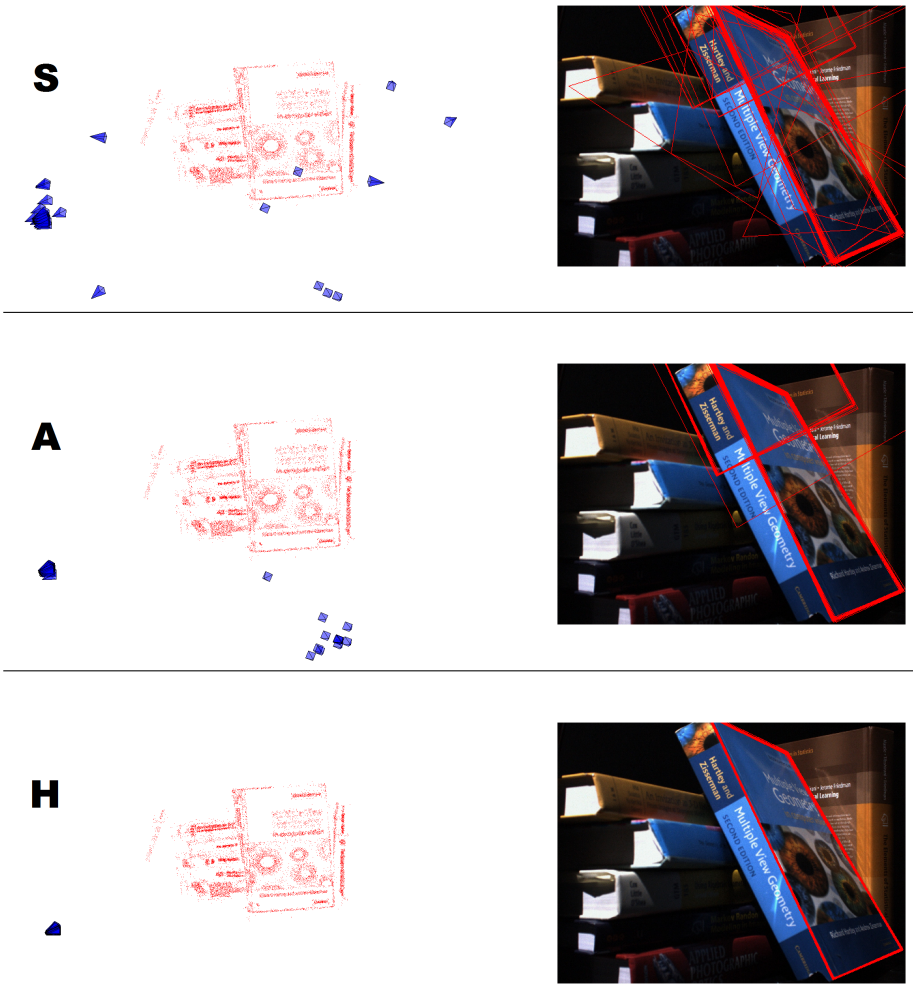

Figure 9. Séquence Livre *: 100 poses calculées avec N = 1 000 itérations de RANSAC, et la reprojection des bords de la couverture en utilisant ces 100 poses. Dans le scénario* ***H*** *l'écart type de la position de la caméra est 0,29 % de la distance à la scène*

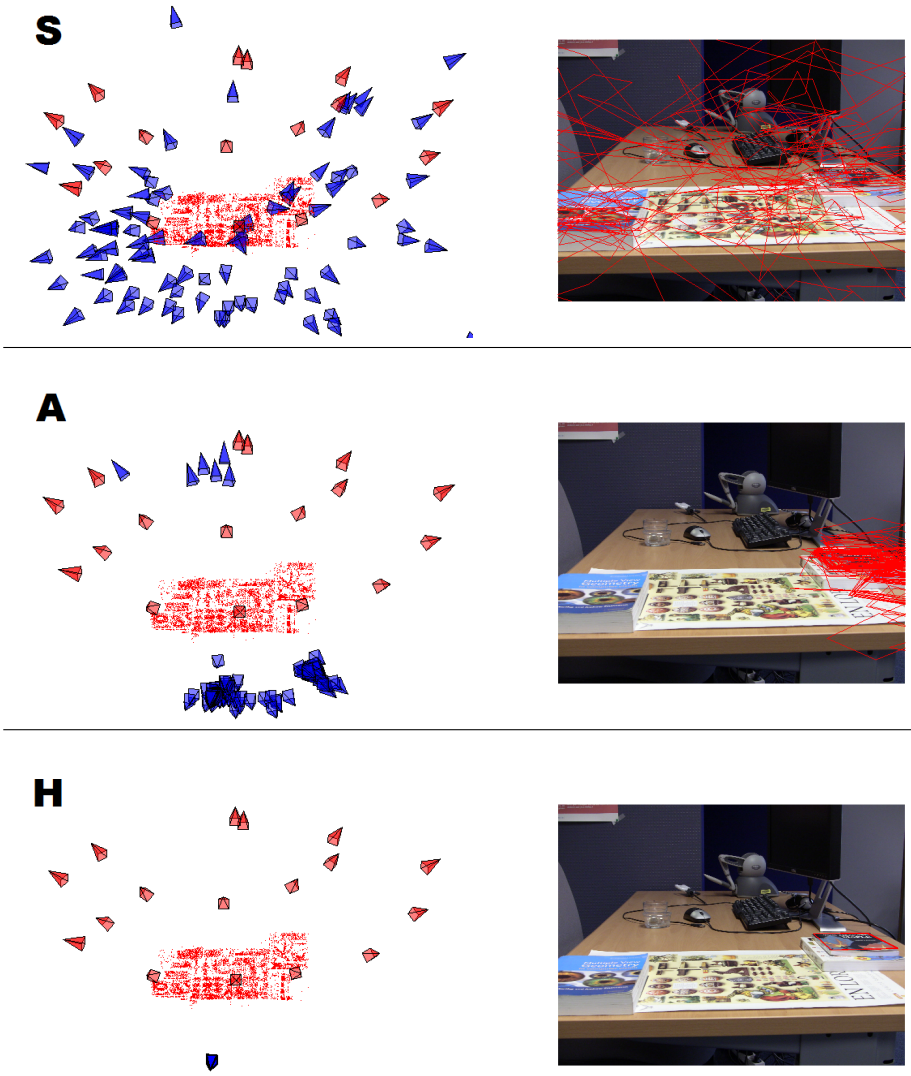

Figure 10. Séquence Poster *: 100 poses calculées avec N = 1 000 itérations de RANSAC, et la reprojection des bords du livre en utilisant ces 100 poses. Dans le scénario* ***H*** *l'écart type de la position de la caméra est 0,07 % de la distance à la scène*

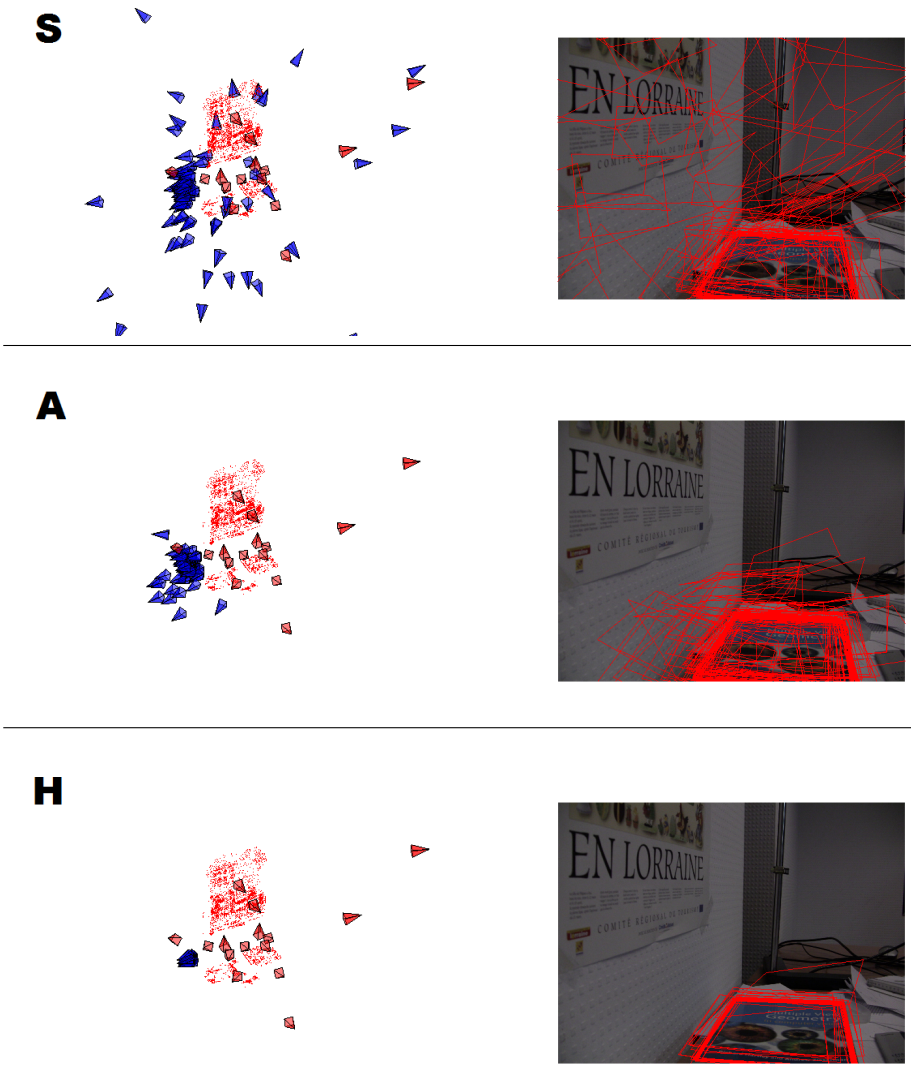

Figure 11. Séquence Bureau *: 100 poses calculées avec N = 5 000 itérations de RANSAC, et la reprojection des bords du livre de droite en utilisant ces 100 poses. Dans le scénario* ***H*** *l'écart type de la position de la caméra est 3,04 % de la distance à la scène*

que les scénarios **S** et **H**, le scénario **A** produisant les même résultats que **S**. En effet le modèle de transformation affine ne prend pas en compte les transformations liées à un changement de profondeur.

La figure 14 montre les résultats dans le scénario **S**. On constate qu'une bonne estimation de la pose n'est possible que dans le cas où la vue test est proche des vues réelles, ce qui est le cas des vues 1 et 2. Par contre pour la vue 3 la précision est largement moindre. La figure 15 montre les résultats dans le scénario **H**. On constate que la pose est estimée avec précision dans l'ensemble des cas, les poses étant visuellement superposées.

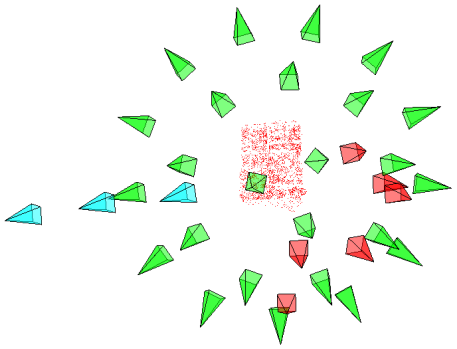

Figure 12. Séquence Mur *: position des caméras de reconstruction (rouge), des caméras virtuelles (vert) et des point de vue test (cyan) 1, 2 et 3*

4.3. Ambiguïté due aux vues symétriques

Comme remarqué dans (Yu, Morel, 2011), avec un modèle de caméra affine un plan a la même apparence observé avec deux points de vue symétriques par rapport à la normale au plan (cf. figure 16), à une rotation d'image de 180° près. C'est ce qui justifie de ne simuler que par l'intermédiaire de caméras virtuelles situées sur un demi hémisphère dans l'algorithme ASIFT.

Dans notre cas, la scène est composée de plusieurs plans et il n'y a donc pas de raison a priori pour se limiter à un demi hémisphère pour placer les points de vue virtuels. Cependant, dans certaines scènes dominées par un plan (Poster et Mur) on peut clairement observer l'influence de cette symétrie (figures 18 et 17).

Dans la séquence Poster, on observe que les points de vue contribuant le plus à la mise en correspondance image/modèle sont un point de vue virtuel proche de la pose test et un autre point de vue virtuel symétrique du premier (figure 17).

Dans la séquence Mur, les points d'intérêts extraits de la vue test sont concentrés dans une faible portion de l'image. Ces points d'intérêts ont la même apparence

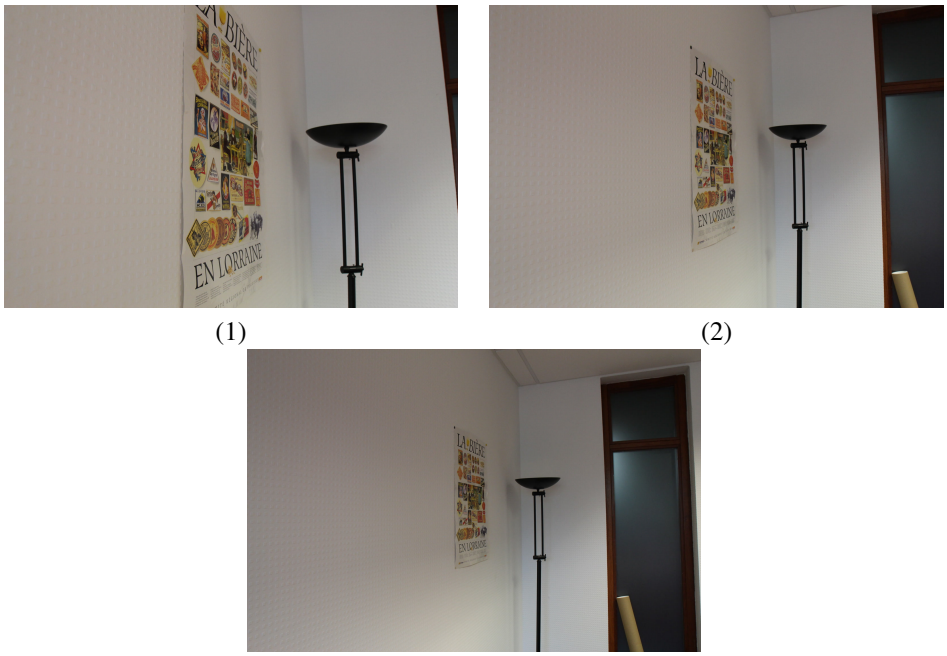

Figure 13. Séquence Mur *: les trois vues de test utilisées pour évaluer la robustesse du calcul de pose par rapport à la distance à la scène*

avec la pose test correcte et la pose symétrique, et leur répartition ne permet plus de différencier les deux (figure 18).

4.4. Temps de calculs et perspectives d'optimisation

Le tableau 1 donne pour chaque séquence la taille des modèles utilisés et les temps de calcul des étapes de mise en correspondance et de calcul de pose. Le code est exécuté sous Matlab sur un processeur Intel Core i7 sans optimisation. Les temps de calcul sont raisonnables pour un prototype ; cependant l'utilisation d'heuristiques pour améliorer ces temps est en cours d'étude. Nous en détaillons certaines ici.

Une première idée pour diminuer le temps de calcul serait d'échantillonner naïvement le modèle, mais cette approche n'est pas envisageable. En effet nous avons observé que l'approche basée sur les plus proches voisins produit un grand nombre de correspondances fausses à cause de points qui sont détectés dans l'image mais qui n'existent plus dans le modèle.

La mise en correspondance image/modèle prend du temps à cause du nombre important de descripteurs présents dans le modèle. Dans ce qui précède, le modèle est supposé suffisamment petit pour qu'il soit réaliste d'utiliser une représentation ex-

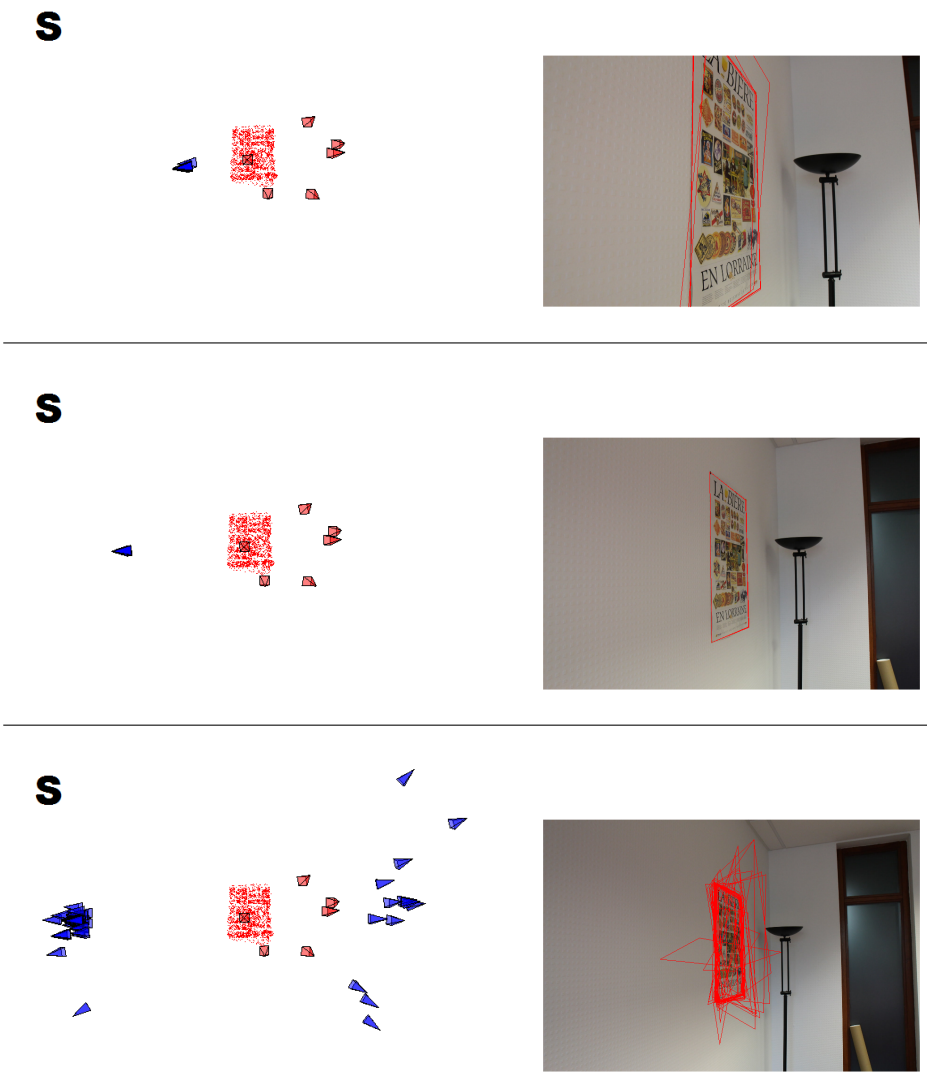

Figure 14. Séquence Mur *: 100 calculs de pose avec N = 300 itération de RANSAC pour les trois vues de test (voir 13) dans le scénario* **S**. *De gauche à droite : les vues test 1 à 3. L'écart type de la position de la caméra est 2,14 % de la distance à la scène pour la vue 1 et 0,12 % pour la vue 2*

Figure 15. Séquence Mur *: 100 calculs de pose avec N = 300 itération de RANSAC pour les trois vues de test (voir 13) dans le scénario **H**. De gauche à droite : les vues test 1 à 3. L'écart type de la position de la caméra est 0,07 % de la distance à la scène pour la vue 1, 0,02 % pour la vue 2 et 0,28 % pour la vue 3*

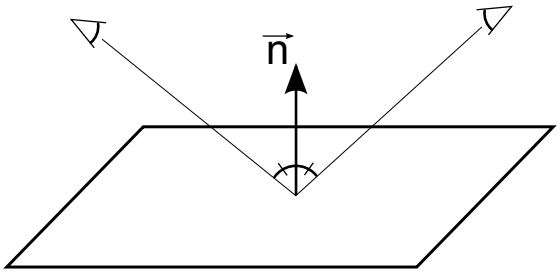

Figure 16. Deux points de vue symétriques selon la normale $\vec{n}$ pour lesquels, avec le modèle affine, le plan a la même apparence à une rotation selon l'axe optique de 180° près

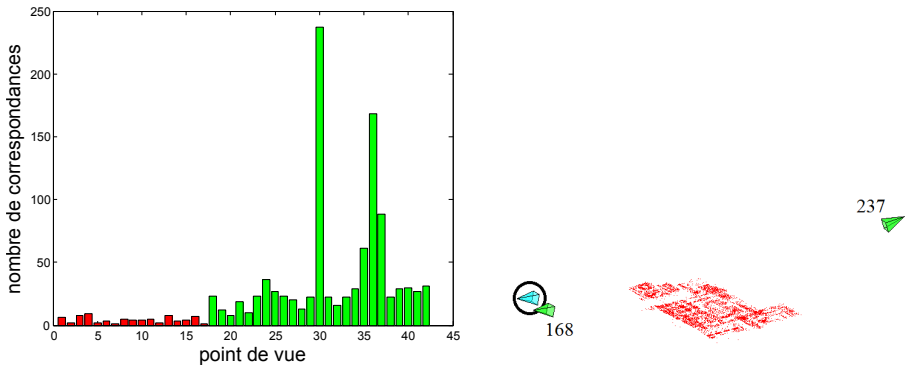

Figure 17. Séquence Poster *: nombre de correspondances associées à chaque point de vue (réel en rouge, virtuel en vert), pour l'ensemble des correspondances image/modèle (à droite). Les deux points de vue correspondant aux pics dans l'histogramme sont un point de vue proche de la pose test et le point de vue symétrique (à droite ;les contributions respectives des deux points de vue sont indiquées)*

haustive, c'est-à-dire conserver pour chaque point du modèle l'ensemble des descripteurs utilisés pour le construire, comme dans (Gordon, Lowe, 2006). Mais on peut envisager une représentation compacte, dans laquelle les classes de descripteurs associées aux points du modèle sont réduites à quelques éléments représentatifs, comme dans (Irschara *et al.*, 2009). Les classes que nous considérons possèdent peu d'éléments et sont peu bruitées, dans le sens où elles contiennent peu de descripteurs incohérents du fait du procédé de simulation. On y trouve des descripteurs isolés qui sont particulièrement discriminants. Ces descripteurs risquent d'être perdus si on quantifie

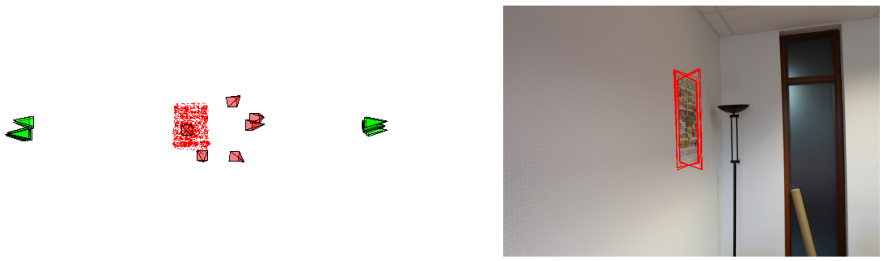

Figure 18. Séquence Mur *: 100 calculs de pose avec N = 200 itérations de RANSAC pour la vue 3. Deux groupes de poses sont calculés, qui correspondent respectivement à la pose test (gauche) et à la position symétrique (droite)*

*Tableau 1. Nombre de vues réelles (1), nombre de points 3D dans le modèle SfM (2), nombre de descripteurs dans les scénarios **S/A/H** (3), temps de calcul en secondes pour la mise en correspondance image/modèle (4), nombre de correspondances (5)*

	Livre	Poster
(1) nb de vues réelles	53	17
(2) nb de points 3D	15 269	7 552
(3) nb de descripteurs	225 207 / 403 662 / 386 970	47 643 / 161 596 / 224 923
(4) temps de calcul (s)	76,7 / 82,4 / 81,4	70,2 / 99,5 / 120,8
(5) nb de correspondances	1 272 / 809 / 1 097	1 144 / 1 293 / 1 092
	Bureau	**Mur**
(1) nb de vues réelles	17	6
(2) nb de points 3D	3 525	2 527
(3) nb de descripteurs	15 109 / 33 396 / 45 393	10 765 / 59 325 / 61 690
(4) temps de calcul (s)	11,0 / 16,9 / 22,3	3,0 / 10,2 / 10,2
(5) nb de correspondances	892 / 779 / 657	322 / 338 / 266

la classe, ce qui est particulièrement problématique si on utilise une approche de type plus proches voisins, comme souligné dans (Boiman *et al.*, 2008). Nos expériences suggèrent néanmoins que réduire la taille des classes est envisageable. Nous avons utilisé comme élément représentatif le medoïd de la classe, qui est l'élément de la classe qui minimise la distance moyenne aux autres éléments. Dans le cas de la séquence Livre par exemple, on observe qu'utiliser ce représentant pour faire la mise en correspondance ne réduit que légèrement le taux d'inliers : de 23 % à 22 % dans le scénario **S**, de 30 % à 28 % dans le scénario **A** et de 37 % à 35 % dans le scénario **H**. Bien que ces résultats soient encourageants, choisir un unique représentant est paradoxal avec notre méthode consistant à enrichir les classes de descripteurs. En revanche, l'idée de réduire le nombre d'éléments présents dans les classes est à retenir. L'avantage de cette méthode serait de rendre la mise en correspondance considérablement plus rapide.

Pour finir, nous proposons une méthode pour accélérer RANSAC. La figure 7 montre un exemple de distribution des correspondances image/modèle parmi les points de vue réels et virtuels. Il apparaît que seuls quelques points de vue contribuent de façon significative lors du calcul de la pose, et que le taux d'inliers parmi les correspondances qui leurs sont associées est élevé. Une méthode que nous voulons investiguer pour accélérer RANSAC serait de biaiser le tirage des hypothèses de RANSAC en faveur de ces points de vue.

5. Conclusion

Cet article étudie l'utilisation de la simulation de point de vue pour enrichir un modèle non structuré dans le cadre du calcul de pose. Il présente à la fois un modèle théorique et une mise en œuvre expérimentale. Bien que cette étude se limite à quelques séquences, elle nous permet de tirer plusieurs enseignements. Premièrement, la simulation de point de vue permet de calculer une pose dans des situations où l'algorithme de (Gordon, Lowe, 2006) échoue, soit à cause d'une forte variation de direction de vue, soit à cause d'un fort changement de profondeur par rapport à la scène. Deuxièmement, dans un cas plus général, la simulation de point de vue permet d'estimer la pose avec une grande précision en utilisant un nombre réduit d'itérations de RANSAC, le taux de correspondances images/modèle correctes étant plus élevé. Enfin plusieurs perspectives d'optimisation ont été étudiées.

Le modèle homographique produit des résultats significativement meilleurs que le modèle affine : le taux d'*inliers* parmi les correspondances image/modèle est plus élevé et les ensembles de consensus obtenus plus importants, alors que les temps de calcul sont semblables entre les deux modèles.

Des travaux futurs sont nécessaires pour améliorer la mise en correspondance image/modèle. L'utilisation d'une représentation compacte pourrait permettre d'atteindre des temps de calcul sensiblement plus faibles. Nous avons atteint des taux d'*inliers* élevés dans l'étape de mise en correspondance, mais pour améliorer la précision de la pose il faudrait également étudier leur répartition dans l'image. Un critère heuristique dans le cadre de la reconstruction à deux vues, basé sur l'échelle des descripteurs SIFT, est proposé dans (Liu *et al.*, 2014). Il serait intéressant de l'étendre à notre problématique.

Bibliographie

Aanæs H., Dahl A., Pedersen K. (2012). Interesting interest points. *International Journal of Computer Vision*, vol. 97, n° 1, p. 18–35.

Bhat S., Berger M.-O., Sur F. (2011). Visual words for 3D reconstruction and pose computation. In *Proc. 3DimPVT*, p. 326-333.

Boiman O., Shechtman E., Irani M. (2008). In defense of Nearest-Neighbor based image classification. In *Proc. Conference on Computer Vision and Pattern Recognition.*

Collet A., Berenson D., Srinivasa S., Ferguson D. (2009). Object recognition and full pose registration from a single image for robotic manipulation. In *Proc. International Conference on Robotics and Automation*, p. 48-55.

Comaniciu D., Meer P. (2002). Mean shift : a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, p. 603 -619.

DeMenthon D., Davis L. (1995). Model-based object pose in 25 lines of code. *International Journal of Computer Vision*, vol. 15, n° 1-2, p. 123–141.

Fischler M., Bolles R. (1981). Random Sample Consensus : A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, vol. 24, n° 6, p. 381–395.

Furukawa Y., Ponce J. (2010). Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, n° 8, p. 1362–1376.

Gordon I., Lowe D. (2006). What and where : 3D object recognition with accurate pose. In J. Ponce, M. Hebert, C. Schmid, A. Zisserman (Eds.), *Toward category-level object recognition*, vol. 4170, p. 67-82. Springer.

Hartley R. I., Zisserman A. (2004). *Multiple view geometry in computer vision* (Second éd.). Cambridge University Press.

Hesch J., Roumeliotis S. (2011). A direct least-squares (DLS) method for PnP. In *Proc. International Conference on Computer Vision*, p. 383–390. Barcelona, Spain.

Hoppe H., DeRose T., Duchamp T., J.McDonald, Stuetzle W. (1992). Surface reconstruction from unorganized points. In *Computer graphics (SIGGRAPH '92 proc.)*, vol. 26, p. 71–78.

Hsiao E., Collet A., Hebert M. (2010). Making specific features less discriminative to improve point-based 3D object recognition. In *Proc. Conference on Computer Vision and Pattern Recognition*, p. 2653–2660.

Irschara A., Zach C., Frahm J.-M., Bischof H. (2009). From structure-from-motion point clouds to fast location recognition. In *Proc. Conference on Computer Vision and Pattern Recognition*, p. 2599-2606.

Kushnir M., Shimshoni I. (2012). Epipolar geometry estimation for urban scenes with repetitive structures. In *Proc. Asian Conference on Computer Vision*, p. 163-176.

Lepetit V., Fua P. (2005). Monocular model-based 3D tracking of rigid objects : A survey. *Foundations and Trends in Computer Graphics and Vision*, vol. 1, n° 1, p. 1–89.

Lepetit V., Moreno-Noguer F., Fua P. (2009). EPnP : An Accurate $O(n)$ Solution to the PnP Problem. *International Journal of Computer Vision*, vol. 81, n° 2, p. 155–166.

Liu Z., Monasse P., Marlet R. (2014). Match selection and refinement for highly accurate two-view structure from motion. In *Proc. European Conference on Computer Vision*, p. 818–833.

Lowe D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, vol. 60, n° 2, p. 91–110.

Moreels P., Perona P. (2007). Evaluation of features detectors and descriptors based on 3D objects. *International Journal of Computer Vision*, vol. 73, n° 3, p. 263–284.

Morel J.-M., Yu G. (2009). ASIFT : A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, vol. 2, n° 2, p. 438–469.

Morel J.-M., Yu G. (2011). Is SIFT scale invariant ? *AIMS Inverse Problems and Imaging*, vol. 5, n° 1, p. 115–136.

Mount D., Arya S. (2010). *ANN : A library for approximate nearest neighbor searching.* http://www.cs.umd.edu/~mount/ANN/.

Noury N., Sur F., Berger M.-O. (2010). How to overcome perceptual aliasing in ASIFT ? In *Proc. International Symposium on Visual Computing, part. 1*, p. 231–242.

Ozuysal M., Calonder M., Lepetit V., Fua P. (2010). Fast keypoint recognition using random ferns. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 32, n° 3, p. 448–461.

Roberts R., Sinha S., Szeliski R., Steedly D. (2011). Structure from motion for scenes with large duplicate structures. In *Proc. Conference on Computer Vision and Pattern Recognition*, p. 3137–3144.

Rothganger F., Lazebnik S., Schmid C., Ponce J. (2006). 3D object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *International Journal of Computer Vision*, vol. 66, n° 3, p. 231–259.

Schindler G., Brown M., Szeliski R. (2007). City-scale location recognition. In *Proc. Conference on Computer Vision and Pattern Recognition.*

Sur F., Noury N., Berger M.-O. (2013). An a contrario model for matching interest points under geometric and photometric constraints. *SIAM Journal on Imaging Sciences*, vol. 6, n° 4, p. 1956–1978.

Williams B., Klein G., Reid I. (2007). Real-time SLAM relocalisation. In *Proc. International Conference on Computer Vision.*

Wu C. (2011). *VisualSFM : A visual structure from motion system.* http://homes.cs.washington .edu/~ccwu/vsfm/.

Wu C., Agarwal S., Curless B., Seitz S. (2011). Multicore bundle adjustment. In *Proc. Conference on Computer Vision and Pattern Recognition*, p. 3057-3064.

Wu C., Clipp B., Li X., Frahm J.-M., Pollefeys M. (2008). 3D model matching with viewpoint-invariant patches (VIP). *Proc. Conference on Computer Vision and Pattern Recognition*.

Yu G., Morel J.-M. (2011). ASIFT : An algorithm for fully affine invariant comparison. *Image Processing On Line*, vol. 2011.

Article soumis le 8/12/2014
Accepté le 2/06/2015