Simulation de point de vue pour la localisation d'une caméra à partir d'un modèle non structuré

Pierre Rolin, Marie-Odile Berger, Frédéric Sur

Simulation de point de vue pour la localisation d'une caméra à partir d'un modèle non structuré

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. En particulier la simulation augmente la proportion de bons appariements et la précision du calcul de pose.*

Mots Clef

Calcul de pose, simulation de point de vue, mise en correspondance.

Abstract

We consider the problem of camera localization from a non structured model obtained by a structure from motion algorithm. In this model, a point is represented by its coordinates and a set of photometric descriptors extracted from the images in which it is observed. Localization is based on matching interest points from the current view with model points, using the descriptors. However, the limited invariance of descriptors with respect to viewpoint changes makes this matching step difficult when the current view is far away from the images used to construct the model. Viewpoint simulation techniques, like ASIFT, have proved effective for matching images. This article explores how these techniques can improve the model by adding simulated descriptors and compares the contribution of affine and homographic models. Simulation increases the proportion of correct matches and the accuracy of the estimated pose.

Keywords

Pose estimation, viewpoint simulation, point matching.

1 Introduction

L'estimation de la position et de l'orientation (la *pose*) d'une caméra dans un environnement connu est un problème à la base de nombreuses applications en géolocalisation [23] ou réalité augmentée [6]. Dans cet article, nous nous intéressons à l'estimation de la pose à partir de correspondances entre points d'intérêt d'une vue courante et points d'un modèle tridimensionnel de la scène observée, construit préalablement comme dans [6] ou [22]. Le modèle est supposé ici non structuré : il s'agit d'un nuage de points obtenu à partir d'un ensemble d'images à l'aide d'un logiciel implantant un algorithme de type *structure from motion* [26, 25]. Un tel algorithme consiste à commencer par apparier des points d'intérêt dans les images (ces points appariés sont alors la projection du même point 3D dans les images), puis à estimer simultanément la pose des caméras et la position relative des points 3D. L'appariement des points d'intérêt est basé sur la similarité de descripteurs de la photométrie dans une région autour du point considéré. Différents descripteurs sont possibles : par exemple des patches invariants [22] ou des mots visuels [2, 11]. On utilise ici des descripteurs SIFT [15] comme dans [6]. Chaque point 3D du modèle est donc associé à l'ensemble des descripteurs ayant servi à apparier les points d'intérêt en lesquels il se projette dans les images initiales.

L'estimation de la pose d'une nouvelle vue consiste à résoudre le problème *Perspective-n-Points* (PnP) [3, 13, 14] à partir d'appariements entre des points d'intérêt de cette vue et les points du modèle 3D, selon la similarité entre descripteurs (comme dans [6]). La limite de cette approche est l'invariance réduite des descripteurs photométriques, même de ceux réputés affine-invariants [16]. Par exemple, SIFT est théoriquement invariant aux similitudes. Lorsque la nouvelle vue présente un trop fort changement d'aspect par rapport aux images ayant servi à construire le modèle, les descripteurs SIFT des points d'intérêt de cette vue ne peuvent donc plus être appariés de manière fiable avec les descripteurs associés aux points 3D du modèle et trop peu

de correspondances fiables peuvent être établies pour résoudre le problème PnP.

Contribution. Dans cet article, nous proposons d'enrichir la description des points 3D en générant des descripteurs synthétiques additionnels, correspondant à des points de vue éloignés de ceux des images ayant servi à la reconstruction initiale du modèle. L'objectif est d'augmenter le degré d'invariance de la description des points 3D, de manière à faciliter 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 [17] ou dans une moindre mesure FERNS [20]. Dans ces approches, la scène est implicitement supposée plane et la caméra affine de manière à ce que les deux images soient liées par une transformation affine. 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 ou des transformations affines avec l'hypothèse de la caméra affine. Nos descripteurs synthétiques 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 peu présentes dans les images initiales. Ceci est illustré sur les figures 1 et 2. Notre approche diffère donc fondamentalement de celle d'ASIFT dans le sens où nos simulations sont guidées par la géométrie de la scène, qui n'est pas prise en compte dans ASIFT. Notons qu'une approche similaire est employée dans [12] ou [27], mais les auteurs génèrent uniquement des vues fronto-parallèles après rectification. Les auteurs de [10] utilisent également une simulation de point de vue pour améliorer la reconnaissance d'objets, et ceux de [11] pour la localisation d'une caméra dans un grand environnement. Dans les deux cas les descripteurs issus des vues simulées sont intégrés dans un vocabulaire visuel.

Plan de l'article. Dans la section 2 nous détaillons la simulation par transformation affine ou homographie. La section 3 explique comment le modèle non-structuré est enrichi à l'aide des descripteurs synthétiques, et comment nous procédons à l'appariement image/modèle permettant de déterminer la pose. La section 4 présente une première 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 3D, et que chacun de ces points est associé à un ensemble de descripteurs SIFT provenant des vues initiales dans lesquelles il a été repéré. Le modèle sera suffisamment petit pour que cette représentation soit réaliste (comme dans [6]) sans avoir besoin de construire une représentation compacte des descripteurs [11]. Nous

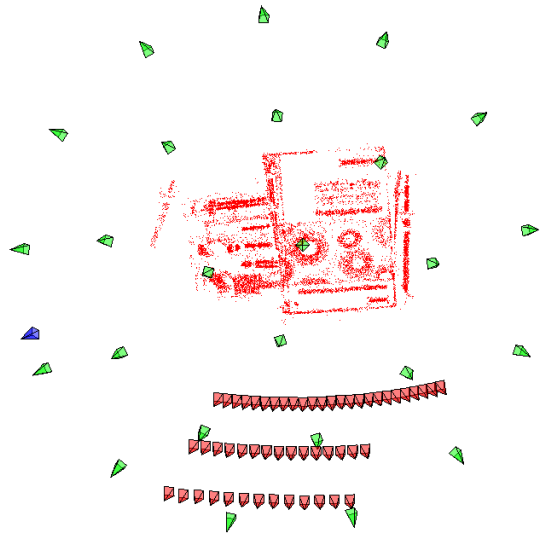

FIGURE 1 – Le modèle 3D de la scène (points rouge), les caméras ayant servi à le construire (en rouge pâle), une caméra éloignée dont on chercherait la pose (en bleu), et les caméras virtuelles (en vert), réparties ici sur une demi-sphère. Les caméras virtuelles permettant de générer de nouveaux descripteurs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

(a) (b)

FIGURE 2 – La vue dont on cherche la pose (a) et la vue la plus proche de celle-ci parmi celles utilisées pour construire le modèle (b). Notons le fort changement de point vue.

supposons également que la scène est localement plane autour des points 3D, et que l'on peut associer un vecteur normal à chaque point. 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 [17, 20] 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 quasiment indiscernables. 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°).

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 [7] :

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

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 et 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 :

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

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, aux problèmes de résolutions d'image près bien sûr. 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.

Cas des transformations affines. Dans le cas de deux caméras affines, notons, avec les notations de la figure 3, $(\lambda_i, \psi_i, t_i, \phi_i)$ les éléments caractéristique de la caméra $i \in \{1, 2\}$ dans un repère associé à un plan repéré par son vecteur normal n. La transformation induite par le plan entre une vue fronto-parallèle de ce plan et la vue i est donnée par la transformation affine [17, 20] :

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

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

$$A = A_2A_1^{-1}. \quad (4)$$

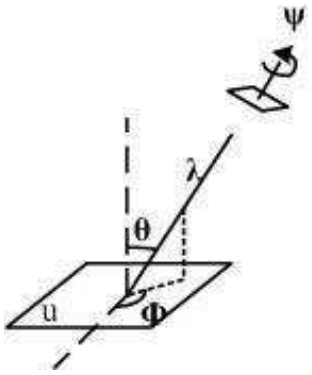

FIGURE 3 – Position d'une caméra affine par rapport à la normale d'un morceau de plan, avec les notations de l'équation (3) où $t = 1/\cos(\theta)$. Image extraite de [17].

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 $\psi_1 = \psi_2 = 0, \lambda_1 = \lambda_2 = 1$.
Donc 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.

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).

3.1 Construction du modèle

Le logiciel VisualSFM [25] est utilisé pour générer un ensemble de points $\mathcal{P}$ de la scène tridimensionnelle, chaque point étant associé à l'ensemble des descripteurs SIFT extrait des images dans lesquelles il est vu. Nous parlerons de *classe* de descripteurs associée à un point 3D. Le logiciel permet également de générer une reconstruction dense de la scène basée sur [5]. 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

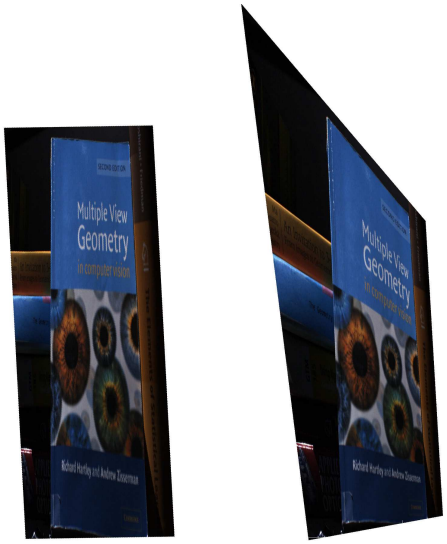

FIGURE 4 – Un exemple de simulation. Les vues simulées de la couverture du livre correspondent à celles que l'on obtiendrait de la position de la caméra dont on cherche la pose (en bleu dans la figure 1), à partir de l'image réelle la plus éloignée. À gauche : simulation par transformation affine ; à droite : simulation par homographie. À une légère rotation près, la simulation par homographie ressemble davantage à la vraie vue (dans la figure 2 a).

proches voisins [9]. 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 synthétiques

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 section 2, le cas affine ne nécessite que de positionner les caméras sur une demi-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 des plus proches caméras à la scène, comme dans la figure 1 ; les caméras sont dirigées vers le barycentre de la scène. Si cela nous permet de simuler (à la discrétisation de l'ensemble des directions de vue près) toutes les vues dans le modèle affine, cela ne permet pas de simuler une caméra s'approchant ou s'éloignant de la scène dans le modèle homographique. Néanmoins, d'après la remarque suivant l'équation 1, cela permet tout de même de simuler les déplacements le long des axes optiques des caméras, pour les plans de la scène orthogonaux à ces axes.

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.

Choix de la vue utilisée pour la simulation et extraction d'un descripteur SIFT. Étant donnés 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.
La simulation produit une imagette 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 dans nos expériences.

3.3 Estimation de la pose

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 [6]. 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 (en pratique $\lambda = 0,6$). La recherche des plus proches voisins est accélérée comme dans [6] par une recherche approchée [18].

Perspective-n-Points. Le calcul de pose se fait par une estimation robuste de type RANSAC [4] basée sur l'algorithme PnP proposé dans [8]. 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

Nous présentons dans cet article des résultats préliminaires issus de la séquence 2 de la base d'image *Robot Data Set* [1] qui contient des images de très bonne résolution (1600×1200 pixels) caractéristique des caméras modernes, ainsi qu'une calibration fine des caméras. La figure 2 montre deux images issues de la base. La procédure de la section 3.1 permet de construire le modèle de la scène à partir d'une partie des images fournies (positions de caméras montrées en rouge dans la figure 1), et on utilise une des images non utilisées pour déterminer sa pose, que l'on peut donc confronter à une « vérité terrain » (caméra bleue dans la figure 1).
Trois scénarios de calcul de pose sont confrontés : **A/** on applique l'algorithme de la section 3.3 sans simulation

(c'est l'algorithme de [6]) ; **B/** on enrichit préalablement les classes de descripteurs par simulation affine à partir des vingt-cinq positions de caméras virtuelles comme décrit dans la section 3.2 ; **C/** idem, avec simulation homographique. Le modèle initial se compose de 15269 points et 225207 descripteurs au total. Les simulations affines ajoutent 178455 descripteurs et les simulations homographiques en ajoutent 161763. Les performances des trois scénarios sont comparées sur 100 estimations indépendantes de la pose de l'image test qui présente un fort changement de point de vue par rapport aux vues permettant de construire le modèle.

Après l'étape de mise en correspondance image/modèle, on peut profiter du fait que la base de donnée nous fournit aussi une estimation précise de la pose que l'on cherche pour vérifier quelles sont les correspondances réellement correctes. En effet, étant donnée une correspondance image/modèle, on peut projeter le point 3D dans la vue, en utilisant la pose de la vérité terrain. En considérant la distance entre cette projection et le point 2D on décide que l'association est correcte si elle est inférieure à 5% de la diagonale de l'image. Le taux moyen de correspondances correctes est de 23% dans le scénario **A**, 30% dans le scénario **B** et 37% dans le scénario **C**. On voit donc que la proportion de correspondances correctes augmente significativement grâce à la simulation, et davantage dans le cas de la simulation homographique qu'affine.

Après avoir appliqué RANSAC pour l'estimation de la pose, on observe que les ensembles de consensus maximaux regroupent en moyenne 25 correspondances dans le scénario **A**, 25 dans le scénario **B** et 43 dans le scénario **C**. Les descripteurs obtenus par simulation sont en moyenne utilisés dans 28% des correspondances de l'ensemble de consensus dans le scénario **B**. Cette proportion passe à 49% dans le scénario **C**. Cela montre que les correspondances apportées par les descripteurs simulés sont très présentes dans les ensembles de consensus, et entrent en jeu dans l'estimation de la pose par résolution de PnP. Notons qu'on ne peut pas comparer directement le nombre des correspondances dans les ensembles de consensus car la qualité de la pose dépend davantage de la répartition spatiale de ces correspondances que de leur nombre.

On observe que les poses calculées avec les modèles enrichis (scénarios **B** et **C**) sont plus fiables qu'avec le modèle initial (scénario **A**), en particulier avec un faible nombre d'itérations pour RANSAC (cf. figures 5 et 6). Ceci s'explique par la plus grande proportion d'appariements image/modèle corrects dans les modèle simulés. Avec 500 itérations de RANSAC l'écart-type de la position de la pose calculée est 3 fois plus faible dans le cas homographique (scénario **C**) que dans le cas affine (scénario **B**). Avec 1000 itérations ce facteur passe à 2. Lorsque le nombre d'itérations de RANSAC continue d'augmenter tous les modèles finissent par produire des poses acceptables, puisqu'il y a dans ce cas précis suffisamment de correspondances correctes avec le modèle initial.

Dans cette expérience, un phénomène remarquable se produit. En effet dans chacun des scénarios les poses calculées se répartissent en trois catégories : des poses proches de celle attendue, des poses totalement fausses et un groupe de poses erronées qui s'avèrent être dues à un motif répété (figure 7). Le motif répété est présent sur deux plans différents de la scène et son apparence est similaire entre les deux vues, ce qui n'est pas le cas du reste de la scène. Ce motif devient donc la seule source de correspondances image/modèle. Ceci conduit à des poses erronées avec des ensembles de consensus relativement grands, de l'ordre de grandeur de ceux des poses correctes. Il s'agit d'un phénomène mis en évidence par exemple dans [21]. Cependant, comme les « bons » ensembles de consensus sont sensiblement plus importants avec le modèle enrichi par simulation homographique, RANSAC les sélectionne et permet un calcul de pose correct. En d'autres termes, la simulation permet d'apparier davantage de motifs non-répétés qui permettent de désambiguïser les correspondances entre motifs répétés. Dans le contexte de l'appariement de paires d'images, cette propriété est discutée pour ASIFT dans [19, 24] et pour un algorithme basé sur une rectification plane dans [12]. Dans le contexte du calcul de pose, des expériences complémentaires sont bien sûr nécessaires pour conforter ces premières constatations.

5 Conclusion et perspectives

Nous avons discuté dans cet article de l'apport potentiel de la simulation de point de vue pour le calcul de pose à partir d'un modèle non-structuré de scène dans lequel les points 3D sont associés à des descripteurs SIFT. Deux types de simulations du changement de point de vue d'un plan, affine et homographique, sont envisagées. Nos premiers résultats expérimentaux suggèrent une amélioration du taux de bons appariements image/modèle par simulation, et ensuite une estimation plus fiable de la pose, davantage robuste à la présence de motifs répétés dans la scène.

Concernant le temps de calcul, si la génération des vues synthétiques fait que la construction du modèle enrichi est assez longue (quoique des simplifications sont envisageables), l'appariement image/modèle par l'algorithme de plus proches voisins approchés, et donc le calcul de pose, ne sont pas significativement allongés.

Parmi les perspectives ouvertes par ces premiers travaux, nous envisageons d'abord de modifier l'appariement image/modèle qui est rendu particulièrement difficile par le grand nombre de descripteurs ajoutés au modèle. Il est actuellement uniquement basé sur un critère photométrique (distance entre descripteurs, l'un étant éventuellement synthétique). On pourrait cependant également utiliser des informations géométriques (contrainte de visibilité, de spatialité) pour améliorer la mise en correspondance. En particulier, chaque descripteur présent dans le modèle est associé à un point de vue particulier. Si les résultats sont corrects, les descripteurs retenus doivent correspondre à des directions de vue voisines, puisque les plus semblables

d'entre eux devraient correspondre à des vues proches de celle dont on calcule la pose.
D'autre part, l'échantillonnage des directions de vues à simuler doit être adapté aux images initiales servant à construire le modèle de la scène.

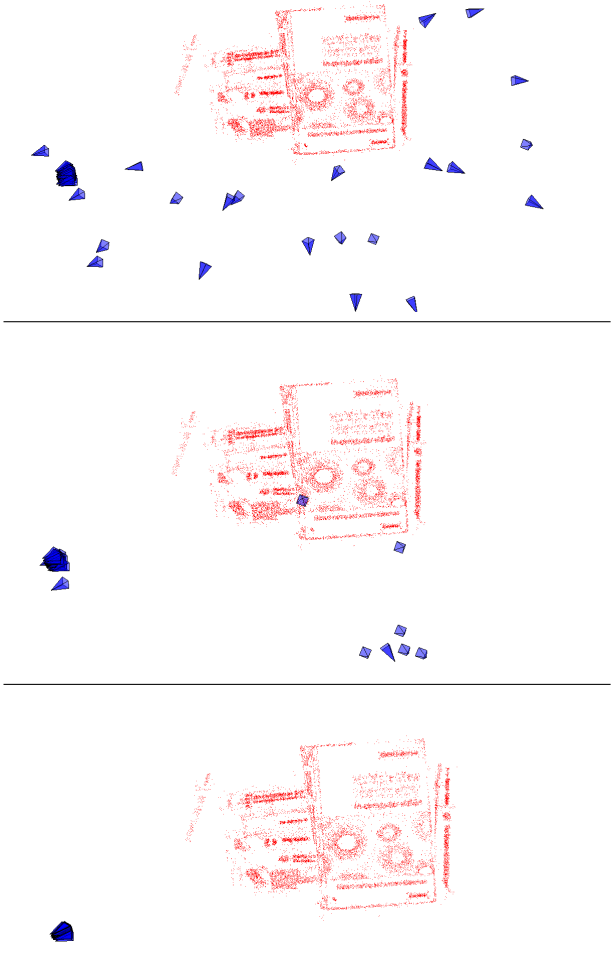

FIGURE 5 – 100 poses calculées avec 500 itérations de RANSAC sur le modèle initial (scénario **A**, en haut : un grand nombre de poses sont erronées) ; enrichi par simulation affine (scénario **B**, au milieu : beaucoup de poses sont correctement déterminées, près de la pose attendue à gauche de la scène, et un certain nombre autour d'une position erronée face au livre) ; enrichi par simulation homographique (scénario **C**, en bas : toutes les poses sont déterminées à proximité de la pose attendue).

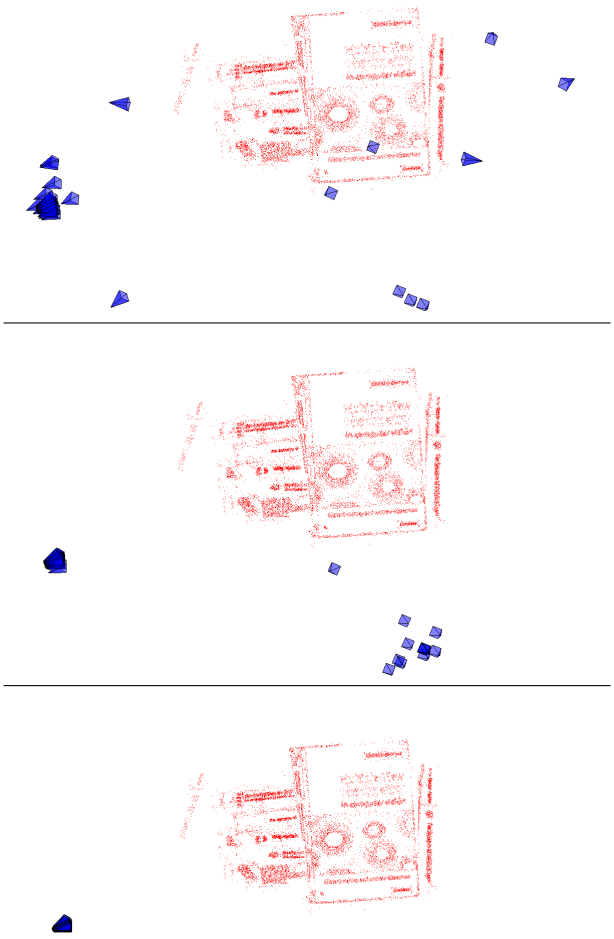

FIGURE 6 – 100 poses calculées avec 1000 itérations de RANSAC sur le modèle initial (scénario **A**, en haut) ; enrichi par simulation affine (scénario **B**, au milieu) ; enrichi par simulation homographique (scénario **C**, en bas).

FIGURE 7 – Une trentaine d'appariements SIFT est due à la présence d'un motif répété entre la vue dont on cherche la pose (à gauche) et une des vues du modèle (à droite). Du point de vue du calcul de pose, la position de la caméra peut alors être inférée comme si la tranche était confondue avec la couverture, et la caméra se retrouve dans le petit groupe face à la couverture dans les figures 5 et 6.

Références

[1] H. Aanæs, A.L. Dahl, and K. Steenstrup Pedersen. Interesting interest points. *International Journal of Computer Vision*, 97(1) :18–35, 2012.

[2] S. Bhat, M.-O. Berger, and F. Sur. Visual words for 3D reconstruction and pose computation. In *Joint 3DIM/3DPVT Conference (3DIMPVT)*, pages 326–333, Hangzhou, China, may 2011.

[3] D.F. DeMenthon and L.S. Davis. Model-based object pose in 25 lines of code. *International Journal of Computer Vision*, 15(1-2) :123–141, 1995.

[4] M. Fischler and R. Bolles. Random sample consensus : A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6) :381–395, 1981.

[5] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(8) :1362–1376, 2010.

[6] I. Gordon and D.G. Lowe. What and where : 3D object recognition with accurate pose. In J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors, *Toward Category-Level Object Recognition*, volume 4170 of *Lecture Notes in Computer Science*, pages 67–82. Springer, 2006.

[7] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.

[8] J.A. Hesch and S.I. Roumeliotis. A direct least-squares (DLS) method for PnP. In *International Conference on Computer Vision (ICCV)*, pages 383–390, Barcelona, Spain, 2011.

[9] H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, and W. Stuetzle. Surface reconstruction from unorganized points. In *Computer Graphics (SIGGRAPH '92 Proceedings)*, volume 26, pages 71–78, 1992.

[10] E. Hsiao, A. Collet, and M. Hebert. Making specific features less discriminative to improve point-based 3D object recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

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

[12] M. Kushnir and I. Shimshoni. Epipolar geometry estimation for urban scenes with repetitive structures. In *Asian Conference on Computer Vision*, pages 163–176, 2012.

[13] V. Lepetit and P. Fua. Monocular Model-Based 3D Tracking of Rigid Objects : A Survey. *Foundations and Trends in Computer Graphics and Vision*, 1(1) :1–89, 2005.

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

[15] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2) :91–110, 2004.

[16] P. Moreels and P. Perona. Evaluation of features detectors and descriptors based on 3D objects. *International Journal of Computer Vision*, 73(3) :263–284, 2007.

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

[18] D.M. Mount and S. Arya. ANN : A library for approximate nearest neighbor searching. `http://www.cs.umd.edu/~mount/ANN/`, 2010.

[19] N. Noury, F. Sur, and M.-O. Berger. How to overcome perceptual aliasing in ASIFT ? In *International Symposium on Visual Computing (ISVC), part I*, volume LNCS 6453, pages 231–242, Las Vegas, Nevada, USA, 2010.

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

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

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

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

[24] F. Sur, N. Noury, and M.-O. Berger. Image point correspondences and repeated patterns. Research Report 7693, INRIA, 2011.

[25] C. Wu. VisualSFM : A visual structure from motion system. `http://homes.cs.washington.edu/~ccwu/vsfm/`, 2011.

[26] C. Wu, S. Agarwal, B. Curless, and S.M. Seitz. Multicore bundle adjustment. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3057–3064, 2011.

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