Recalage Modèle/Image pour des Prises de Vue Obliques

B. Debaque (*,**), M. Deseilligny (***) et G. Stamon (*)
*SIP-CRIP5, UFR-Mathématiques-Informatique, Université Paris V
45 rue des St-Pères 75006 Paris
**Thales-ISR Imagery and Geographic Information,
66-68 Av. P. Brossolette 92247 Malakoff Cedex, France
***Institut Géographique National-MATIS, 2 Av. Pasteur, 94165 St-Mandé France
debaque@math-info.univ-paris5.fr

April 10, 2001

Abstract

Nous proposons une méthode permettant d'estimer de façon optimale la position et l'orientation d'un modèle tridimensionnel sur une image. Ce problème survient lorsque l'on cherche à faire coïncider un modèle 3D avec des primitives extraites d'une image dans le plan image. Or, nous savons que cet alignement peut-être compromis pour plusieurs raisons. Nous nous intéressons donc dans cet article aux distorsions géométriques causées par les effets de perspectives. Ces phénomènes apparaissent surtout lorsque la distance entre les objets en avant et en arrière de la scène est importante (cas typique des prises de vue aérienne oblique).

Une solution basée sur un alignement 2D des contours des bâtiments est proposée. Un alignement initial à géométrie simplifiée permet de régler l'alignement sur un minimum global à la fois dans l'espace des correspondances et dans l'espace des transformations. Finalement, un modèle de capteur plus complexe est calculé afin d'augmenter la précision du recalage. L'erreur de mise en correspondance est une mesure globale prenant en compte les paramètres du capteur ainsi que l'incertitude du modèle 3D. Nous employons un estimateur robuste basé sur le critère de la norme L1 pour réduire l'effet des faux appariements.

L'algorithme a été testé avec succès et les performances de la méthode ont été évaluées sur des images simulées d'un modèle 3D de site urbain.

1 Introduction

Le cadre de notre étude est la reconnaissance de formes à base de modèles. Plus particulièrement, nous nous attachons au problème de reconnaissance d'un modèle de scène urbaine extrait d'images à faible résolution, dans une série d'images obliques à haute résolution. Cette problématique touche de près ou de loin de nombreuses applications encore difficiles à résoudre comme la texturation automatique de bâtiments, la mise à jour de bases de données cartographiques, la poursuite de cibles en phase terminale ou la simulation d'images d'interférences pour la téléphonie mobile.

Etant donné que l'algorithme s'inscrit dans un outil plus global de production de données cartographiques, l'aspect opérationnel est important et l'intervention humaine dans la chaîne de traitement n'est pas négligée. Cet article traite donc spécifiquement de la phase de reconnaissance d'un modèle 3D dans une image obtenue d'une caméra non métrique. Une solution rapide, précise et quantifiable est principalement attendue dans un contexte de production. En résumé, la problématique est la suivante: **étant donnée une image de scène et un modèle 3D CAD de cette scène, trouver une correspondance optimale de manière à aligner le modèle sur des objets reconnus de la scène.**

Durant la phase de reconnaissance, une projection est établie entre l'image et le modèle de scène, cette projection doit être suffisamment précise pour satisfaire la contrainte de consistance du point de vue [11]. Aussi, ce dernier problème se divise en deux sous étapes :

- trouver une transformation générale pour estimer de manière optimale le modèle de capteur
- réduire l'explosion combinatoire des correspondances modèle/image.

La reconnaissance de formes à base de modèles a été beaucoup étudiée dans le passé [12]. Néanmoins, dans le contexte particulier de la reconnaissance de scènes urbaines, quelques difficultés supplémentaires peuvent apparaître: les bâtiments peuvent avoir une structure assez complexe avec beaucoup de détails architecturaux, les images présentent de nombreux artéfacts non modélisés, où les objets qui y sont observés ont une géométrie et une radiométrie variables d'une image à l'autre, mais aussi des différences d'échelle

entre objets en fond et en avant de scène [6]. Aussi, la plupart des systèmes conçus pour l'extraction de bâtiments sont inappropriés à l'imagerie oblique [13]. Il est donc difficile d'inférer des primitives de haut niveau par la seule observation de l'image pour ce type de prise de vue. Dans notre problématique, le modèle 3D de la scène est connu (jusqu'à un certain degré et est dépendant des erreurs de construction), ce n'est pas un problème d'indexation (trouver une forme parmi un ensemble de formes), mais plutôt un problème de recalage.

De plus, nous souhaitons exploiter une série d'images pas nécessairement contigus et présentant des distorsions optiques importantes, aussi l'emploi des techniques stéréoscopiques pour ce cas précis n'a pas été retenu. Certaines primitives peuvent ne plus apparaître d'une image à l'autre et leur forme peut varier radicalement d'un point de vue à un autre. Un exemple typique serait de mettre en correspondance une image aérienne verticale et terrestre. Par exemple, lorsque nous prenons l'avion il nous est difficile de reconnaître certaines structures qui nous apparaissaient évidentes en voiture. En conséquence, une technique de recalage basée sur des correspondances 3D vers 2D est adoptée.

2 Recherche de correspondances 2D

Les primitives sont appariées selon une recherche récursive entre l'espace des transformations et l'espace des correspondances. Le but étant d'obtenir des alignements entre des primitives extraites de l'image et du modèle de scène. Une première transformation est établie grâce à un minimum de paires homologues, les primitives du modèle de scène encore non traitées sont reprojetées sur l'image [16]. Néanmoins, la combinatoire entre alignements possibles peut devenir rapidement élevée. Nous devons donc établir des techniques efficaces pour le calcul et la vérification des transformations [9].

Carceroni et Brown optent pour une solution générique précise et suggèrent des méthodes numériques pour la recherche de la transformation optimale [2]. Que l'on emploie une méthode à base de dérivées ou une méthode par complexité croissante du modèle de projection, une mesure est minimisée selon des valeurs observées et prédites. Néanmoins, nous élargissons le problème strict de recherche du point de vue à un problème de recherche d'une transformation plus générale pour ainsi prendre en compte les caméras non métriques (d'orientation interne inconnue).

Concernant le choix du type de primitives à apparier, celui-ci dépendra du degré souhaité de fiabilité et de la nature du problème. Dans un travail récent, nous avons montré que la recherche de primitives de type coins n'est pas nécessairement évidente dans des scènes complexes, étant donné que la localisation précise de ce genre de primitive est loin d'être garantie [3]. Dans cet article, nous étendons le recalage à base de points à un recalage à base de droites. L'utilisation de primitives locales tels que les segments extraits des contours est employée de par leur stabilité dans les images encombrées et complexes. Néanmoins, ce choix se fait au détriment d'une perte du contenu sémantique par rapport à des structures de type parallélogramme ou autres au contenu sémantique plus riche. La perte d'exhaustivité des formes des primitives que l'on est à même d'extraire augmente le risque d'apparition des faux appariements.

3 Gestion des faux appariements

Les faux appariements peuvent provenir des occlusions ou des données non suggestives (artéfacts, bruits, etc.). Durant le processus d'estimation des paramètres de projection, les observations erronées ("blunders") peuvent être reliées aux faux appariements. En rappelant la théorie classique de l'estimation, une attitude communément admise lors de l'élimination des observations erronées consiste à utiliser des observations redondantes, l'exactitude devenant secondaire [14].

Une donnée erronée peut être comprise comme une erreur imprévisible de grande magnitude. Les données erronées sont relieés aux données aberrantes ("outliers") lorsqu'elles sont considérées comme des observations bien séparées de l'amas central des observations.

4 Méthodologie

Dans le but d'obtenir une solution unique, la figure 1. illustre l'algorithme complet de la méthode proposée. L'algorithme comprend deux phases distinctes : une phase d'extraction et une phase de recalage. Ces deux étapes sont complémentaires l'une de l'autre puisque l'étape d'extraction produit en bout de ligne un ensemble d'hypothèses de correspondances et la phase de recalage réduit l'espace de recherche des hypothèses de correspondances à une solution acceptable.

4.1 Alignement initial

La première étape consiste à établir une transformation faible dans le but de réduire la combinatoire d'un algorithme automatique futur. Nous assumons que le capteur est de type sténopé. Pour avoir une estimation grossière du point de vue nous employons une affinité et l'étendons à une transformation projective complète. Le modèle de caméra final est un modèle projectif générique qui retrouve les cinq paramètres internes et les six paramètres de localisation.

De manière formelle, une transformation projective de l'espace projectif P^4 à P^3 peut s'écrire [7]

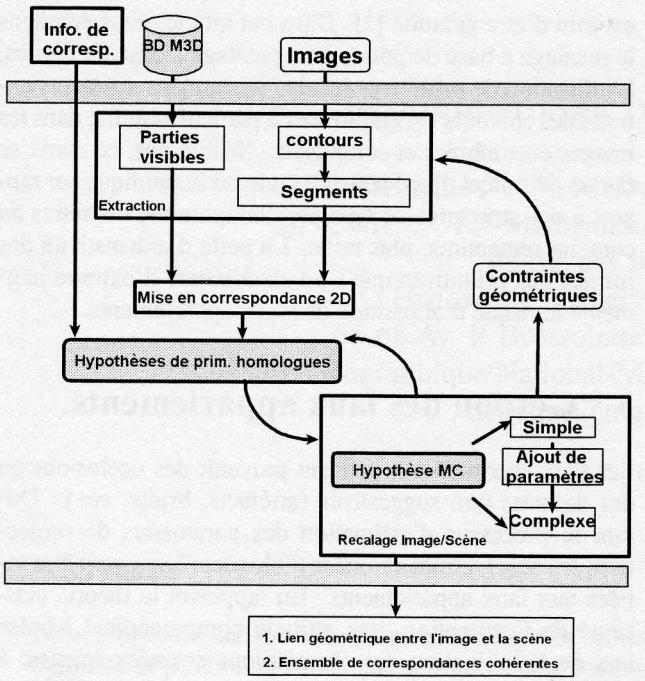

Figure 1: Structure de l'algorithme

$$\begin{aligned} \Pi_p \quad &: \quad P^4 \rightarrow P^3 \qquad (1) \\ \mathbf{M}_i \quad &\Rightarrow \quad s_i \begin{bmatrix} \mathbf{m}_i \\ 1 \end{bmatrix} = \mathbf{W}_j^T \begin{bmatrix} \mathbf{M}_i \\ 1 \end{bmatrix}, \quad i = 1, \ldots, n \end{aligned}$$

Où $\mathbf{W}_j$ est une matrice non-singulière de taille $(3+1) \times (2+1)$ définie de manière unique, et $\mathbf{m}_i = \begin{bmatrix} u_i & v_i \end{bmatrix}^T$ est la i^e coordonnée image observée, $\mathbf{M}_i = \begin{bmatrix} x_i & y_i & z_i \end{bmatrix}^T$ est le point objet correspondant et s_i est un facteur échelle. Une transformation projective de P^4 à P^3 à $(3+1) \times (2+1) - 1$ degrés de liberté. Ainsi, le calcul d'une transformation projective est le calcul des onzes papramètres de $\mathbf{W}_j$. La solution n'est unique puisque six points donnent (6×2) équations, aussi il y a une légère redondance. Une affinité n'a que $(3+1) \times (2+1) - (1+4)$ degrés de liberté. Le calcul d'une affinité est le calcul des huits coefficients de Λ_j. La solution, encore une fois, n'est pas unique puisque quatre points apportent (4×2) équations. Λ_j n'est que la représentation de $\mathbf{W}_j$ moins sa dernière ligne.

La matrice de covariance *a posteriori* des paramètres du capteur est obtenue des équations suivantes. L'équation 1 peut être réecrite de manière plus compacte (comme la i^e équations d'observation)

$$\mathbf{m}_i = \Pi_{proj} \cdot \mathbf{M}_i \qquad (2)$$

où Π_{proj} est la matrice de projection (que l'on emploie Λ_j ou $\mathbf{W}_j$), en présence de bruits, la i^e erreur de mesure s'écrit alors

$$e_i = \mathbf{m}_i - \Pi_{proj} \cdot \mathbf{M}_i \qquad (3)$$

la contrainte géométrique sur les paramètres du capteur s'inscrit dans un formalisme d'optimisation locale de l'équation 4 dans l'espace des transformations

$$\min \left[\sum \rho(\frac{e_i}{\sigma_i}) \right] \qquad (4)$$

où ρ est un critère, par exemple le critère des Moindres Carrés, et σ_i l'écart-type des observations. Pour des fonctions **LP** (Linéaire pour les Paramètres), les propriétés stochastiques des paramètres du capteur est obtenue de la matrice de covariance

$$\mathbf{\Sigma}_p = \mathbf{R}^T \mathbf{\Sigma} \mathbf{R} \qquad (5)$$

où $\mathbf{\Sigma}$ est la matrice de covariance des observations, et $\mathbf{R}$ est la matrice des coefficients obtenus après manipulations algébriques et éventuellement linéarisation de l'équation 3, voir [5].

4.2 Choix de la répartition des points initiaux

L'hypothèse affine de la transformation permet d'aligner globallement les primitives du modèle et des images. Nos essais sur des données simulées ont montrés que dans le voisinage des quatre points initiaux, l'approximation affine etait encore valide. Cette présomption est émise pour des modèles denses (comportant de nombreuses primitives rapprochées). Néanmoins, la validité locale de la transformation affine pour des projections coniques comportant des distorsions optiques (de faible amplitude) non modélisées est fonction de la géométrie et de la localisation initiale des points. Concernant la répartition exacte des points dans l'espace et sur le plan image, nous savons que:

- les quatres points ne peuvent être coplanaires, auquel cas le rang de Λ_j est inférieur strictement à 8 (cas dégénéré) [4]

- la validité de la transformation se répartit selon la localisation des points du recalage et tend à devenir uniforme au fur et à mesure que les points occupent plus d'espace dans l'image (simulations effectuées sur une centaine de points avec des configurations variables des points de recalage). Néanmoins, lorsque la scène occupe tout l'image, les déformations liées à l'approximation affine rendent impossible une quelconque reconnaissance.

On choisira alors des points initiaux de recalage localisés sur l'image de manière à rendre l'affinité valide dans cette région. Pour guider l'utilisateur dans ce choix, il est possible de fournir une estimation de la qualité du recalage par le calcul *a posteriori* des variances des paramètres. On choisira alors des configurations pour lesquelles les magnitudes sont minimales.

4.3 Obtention d'une transformation plus complexe

Pour obtenir une transformation projective complète de l'eq. 1., au minimum deux observations sont insérées dans la solution. Ces observations sont insérées en cherchant l'espace des correspondances tout en minimisant les valeurs possibles de l'espace des transformations. Cette sélection permet de tirer la meilleur estimée (en termes de proximité à la fois dans l'espace des contraintes des paramètres capteur et dans l'espace des correspondances) parmi un groupe d'observations.

4.3.1 Création de l'espace des correspondances

L'espace des correspondances est établi à partir d'un **GAP** (Graphe d'Adjacence des Primitives), reliant segments du modèle de scène et segments image (voir Fig. 2). A chaque segment du modèle est assigné une région d'incertitude. Cette région d'incertitude est calculée par propagation des erreurs à travers l'eq. 2, on obtient alors pour chaque point un écart-type, on calcule ensuite l'enveloppe convexe pour chaque paire de points. Chaque segment image extrait touchant à cette région sera pris en compte.

Le **GAP** est un graphe attribué qui consiste en un ensemble de noeuds N caractérisant l'ensemble des segments et un ensemble d'arcs E caractérisant l'ensemble des paires non-ordonnées (s_i, s_j) représentant les relations de voisinage entre les deux segments ayant pour vecteur d'attributs relationnel $\overrightarrow{x}$. De manière formelle, $GAP = \{N, E\}$, est un graphe orienté où chaque noeud représente une primitive modèle pouvant avoir une ou plusiseurs adjacences avec des primitives image. Si une telle adjacence existe elle est représentée par un arc pointant du noeud modèle au noeud image.

Soit Ξ l'ensemble des interprétations entrant en compétition entre primitives modèle et image, étant donné un ensemble de caractéristiques $\overrightarrow{x} = [L, O, C]$ globalement discriminantes, un ensemble de correspondances m ayant différentes qualités d'appariements sera écrit $M = \{\overrightarrow{x_1}, \ldots, \overrightarrow{x_m}\}$. Chaque objet peut donc être ordonné dans une liste selon le degré de ressemblance. Néanmoins, l'ambiguïté se propage pendant le processus de mesure et cela pour plusieurs raisons : d'une part elle vient de l'incertitude liée à la définition de l'objet observé et des attributs, d'autre part du manque d'exactitude de l'extraction

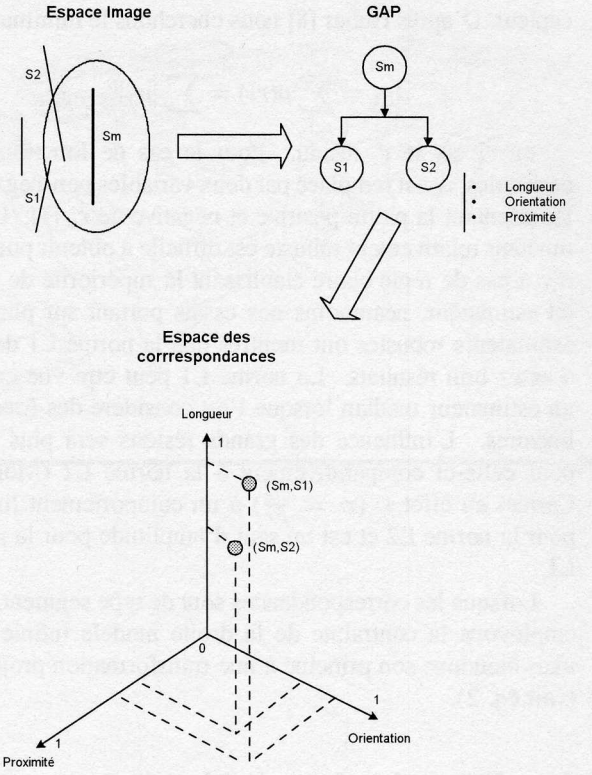

Figure 2: Création de l'espace des correspondances

des primitives image. Aussi une qualité d'appariement est établie de manière à gérer les erreurs de correspondance.

Un arc à trois attributs décrivant une adjacence:

- Difference de longueur: $L = \|s_i\| - \|s_j\|$, tel que $0 \leqslant L \leqslant 1$ (est normalisé selon le plus grand L),
- Différence d'orientation: $O = \widehat{(s_i, s_j)}$ tel que $0 \leqslant O \leqslant 1$ (est normalisé avec $1 \equiv \pi/2$),
- Proximité: $C = \|(milieu(s_i), milieu(s_j))\|$, tel que $0 \leqslant C \leqslant 1$ (est normalisé selon le plus grand C).

Trois étapes sont nécessaires pour construire le graphe: 1) projection du modèle 3D sur l'image, 2) sélection des primitives images et 3) construction des noeuds et calcul des attributs(voir Fig. 2).

Finalement, les paires de segments ayant leur attribut à l'intérieur d'un certain seuil seront sélectionnées. En pratique, on calculera la moyenne des trois atttributs $E(\overrightarrow{x_m})$ comme étant la mesure de ressemblance unique entre primitives adjacentes.

4.3.2 Robustification dans l'espace des transformations

La présence des faux appariements nous amène à employer un estimateur robuste pour l'estimation des paramètres du

capteur. D'aprés Huber [8] nous cherchons le minimum de

$$J_{L1} = \sum \rho(r_i) = \sum \|r_i\|_1 \quad (6)$$

où r_i est le i^e résidu. Pour le cas de fonctions non dérivables, r_i est remplacé par deux variables non-négatives, représentant la partie positive et négative de r_i [1]. Un estimateur relativement robuste est difficile à obtenir puisqu'il n'y a pas de règle claire établissant la supériorité de tel ou tel estimateur, néanmoins nos essais portant sur plusieurs estimateurs robustes ont montrés que la norme L1 donnait d'assez bon résultats. La norme L1 peut être vue comme un estimateur médian lorsque l'on considère des fonctions linéaires. L'influence des grands résidus sera plus petite pour celle-ci comparativement à la norme L2 (Moindres Carrés) en effet ψ ($\psi = \frac{\partial \rho}{\partial r}$) à un comportement linéaire pour la norme L2 et est un saut d'amplitude pour la norme L1.

Lorsque les correspondances sont de type segment, nous employons la contrainte de la droite modèle infinie [10], mais étendons son principe à une transformation projective (voir éq. 2).

4.4 Extraction des primitives image

Nous avons employé un algorithme traditionnel pour extraire des primitives de type segment. D'abord des sauts d'amplitude sont détectés grâce à un filtre récursif de Deriche. Ensuite, le squelette est calculé, celui-ci est convertit en contours puis en segments et arcs elliptiques. Les petits segments sont éliminés ou fusionnés. La taille du filtre de détection de contours peut aussi être établi en fonction de la position relative scène/capteur [15].

5 Expérimentations, résultats et évaluation

L'algorithme a été testé sur des images synthétiques d'un modèle de campus utilisé notamment dans le projet VITRA [17]. L'algorithme complet comprend les phases suivantes:

1. entrée du modèle 3D et de l'image
2. sélection manuelle d'un trièdre centré sur un coin de bâtiment (voir les points saisis en avant de la scène de la fig. 3.a)
3. sélection automatique focalisée des segments consistants à l'intérieur du plus grand cercle contenu dans le trièdre, puis création de nouveaux coins à partir des appariements de segments (extraction des segments et calcul du **GAP**, seuil à 0.2), voir fig. 3.b)
4. estimation robuste (norme L1) du modèle projectif, si aucun coin n'est trouvé l'utilisateur les entre manuellement,
5. alignement global entre le modèle et l'image
6. extraction des primitives image et calcul du **GAP**, sélection automatique des segments (seuil à 0.1)
7. recherche dans toute l'image d'appariements de segments valides par insertion/vérification (chaque segment est inséré si l'estimation des paramètres géométriques converge)
8. si pas de modification significante sur les paramètres projectif arrêt, sinon retour à l'étape 5.

Les tests ont portés sur une prise de vue simulée. Les quatre points servant à initialiser l'affinité sont choisis par l'utilisateur. La figure 4 présente l'alignement finale ainsi que les primitives trouvées comme appariement valides. A chaque projection du modèle sur l'image un calcul des parties cachées est mise en oeuvre. Ce calcul est une variante de l'algorithme du Z-Buffer, à la différence que l'information géométrique des primitives du modèle projeté est gardée en mémoire en plus de la distance de l'objet au centre de prise de vue (et ceci pour chaque pixel). Cette opération est passablement plus longue qu'un Z-Buffer ordinaire puisque des intersections entre segments peuvent être calculées.

La figure 5 présente différentes courbes sur l'évolution de la convergence de l'alignement. La moyenne des résidus vrais (résidus des points de vérification) est exprimée en pixels, on remarque qu'à la dernière étape (estimation projective 3) les résidus sont en dessous du pixel (0.8 pixels). On présente aussi les valeurs moyennes (sur une échelle entre 0 et 1, voir fig. 2) de la qualité de l'alignement, à la fois pour les homologues reconnus et les homologues inconnus (segments sélectionnés mais non-retenus lors de l'étape 7). On donne aussi le nombre de primitives sélectionnées à chaque étape. Le nombre d'homologues inconnus comptabilise le nombre d'homologues possibles, c-à-d. le nombre d'homologues formés des segments images appartenant à l'ellipse d'erreur du segment modèle. Une mesure plus exacte aurait été de comptabiliser le nombre des segments modèles visibles mais non appariés.

Bien que l'algorithme ait été testé sur des données simulées, certaines questions restent encore difficiles. Nottament, il serait interessant d'observer le comportement de l'algorithme en présence de distorsions géométriques non-modélisées, d'erreurs géométriques sur le modèle de scène et sur une mauvaise initialisation de la transformation affine.

6 Conclusion

Dans cette étude le recalage d'une image avec un modèle 3D de scène urbaine a été proposé, en établissant un mod-

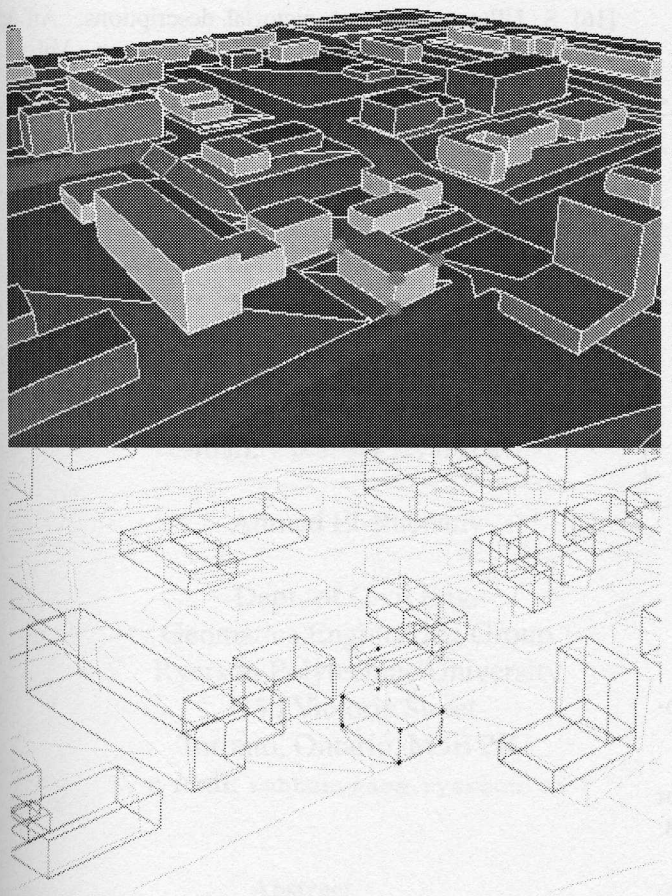

Figure 3: a) Extraction des segments sur l'image et sélection d'un trièdre b) Recalage affine initial

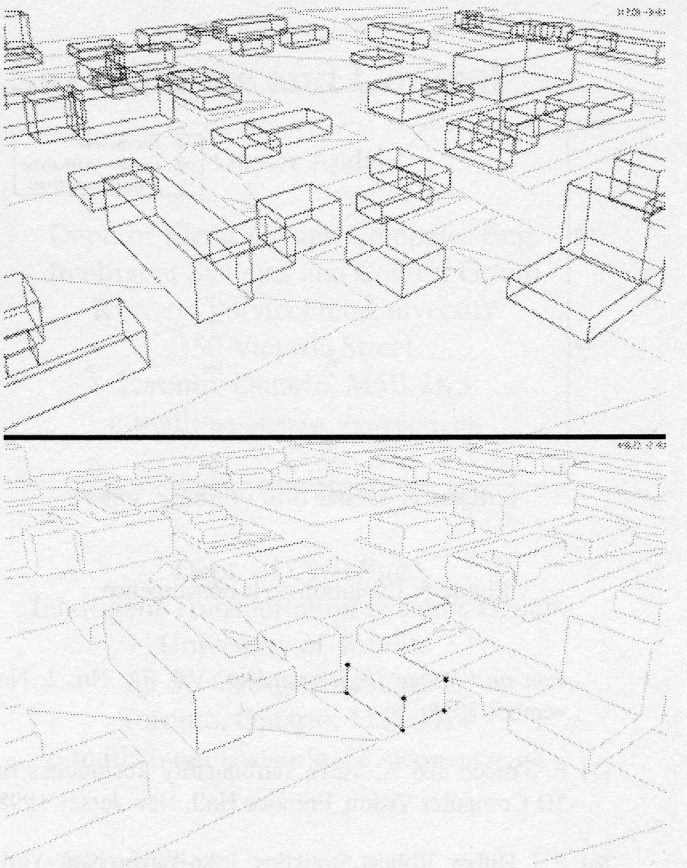

Figure 4: a) Alignement final, b) Primitives retenues pour l'alignement final

èle générique à partir d'une géométrie simplifiée. La méthode offre l'avantage d'estimer le point de vue à partir de peu de points initiaux, et ainsi aide à réduire la combinatoire lors d'un processus global de reconnaissance. Des appariements hypothétiques sont d'abord fournis pour démarrer l'algorithme, les faux appariements sont réduis par l'utilisation d'une mesure de ressemblance dans l'espace des correspondances et par l'emploi d'un estimateur robuste dans l'espace contraint des transformations. Néanmoins, d'autres tests devraient fournir une meilleure connaissance de la stabilité de la méthode vis-à-vis de la présence d'erreurs et sur le choix des points initiaux. L'utilisation des faux négatifs (échec à la reconnaissance de certaines primitives) comme moyen de vérification est aussi envisageable.

References

[1] Barrodale, I., and Roberts, F. D. K., Solution of an Over-Determined System of Equations in the L1 Norm. 69, University of Victoria, BC. Canada, Department of Mathematics, Hannover 1972.

[2] R. L. Canceroni and C. M. Brown, Numerical Methods for Model-Based Pose Recovery, University of Manchester, Computer Science Department, Manchester TR659, 1997.

[3] B. Debaque, S. Gobert, G. Ruckebusch and G. Stamon, Strong-from-Weak Model Sensor Estimation Using Voronoi Diagrams. ICIAP '99 10th International Conference on Image Analysis and Processing, september 1999 Venice.

[4] O. Faugeras, Three-Dimensional Computer Vision, The MIT Press, Cambridge MA.

[5] R. Fletcher, Practical Methods of Optimization, Wiley & Sons, 1990.

[6] A. Gruen and R. Nevatia, Automatic Special Issue on Building Extraction from Aerial Images, *Computer Vi-*

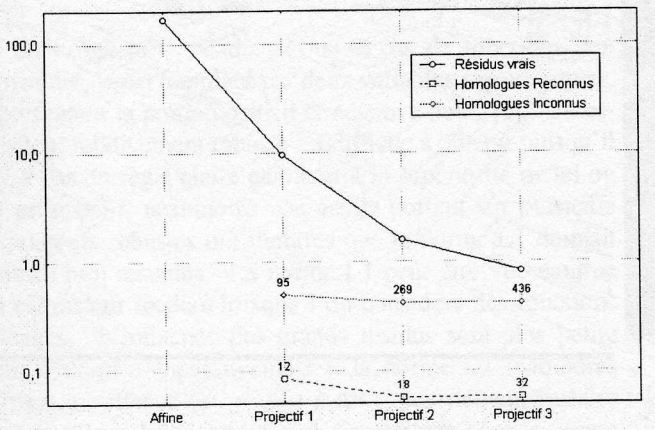

Figure 5: Evolution de la convergence

sion and Image Understanding, Vol. 72, No. 2, November 1998.

[7] E. Trucco and A. Verri, Introductory Techniques for 3D Computer Vision, Prentice Hall, New Jersey 1998.

[8] P. J. Huber, Robust Statistics, John Wiley, New-York, 1981.

[9] D. P. Huttenlocher, Three-Dimensionnal Recognition of Solid Objects from Two-Dimensionnal Image, Cambridge: MIT, 1988.

[10] R. Kumar and A.R. Hanson, Robust methods for estimating pose and a sensitivity analysis, TR David Sarnoff Research Center, Princeton, NJ 1994.

[11] D. G. Lowe, The viewpoint consistency constraint, *Int. J. Computer Vision*, vol. 1, pp. 57-72, 1987.

[12] A. R. Pope, Model-Based Object Recognition A Survey of Recent Research, 1994.

[13] J. A. Shufelt, Performance Evaluation and Analysis of Monocular Building Extraction From Aerial Imagery, *PAMI*, Vol. 21, No. 4, april 1999, pp. 311-326.

[14] P. Stefanovic, Blunders and Least Squares, *ITC Journal*, vol. 1, pp. 122-157, 1978.

[15] R.Thomas, M.V. Serfaty, R. Horak, G. Stamon. Vers un paramétrage local automatique d'opérateurs de vision précoce en imagerie aérienne oblique. Vision Interface'2000.

[16] S. Ullman, Aligning pictorial descriptions: An approach to object recognition, *Cognition*, Vol. 32, 1989,pp. 193-254.

[17] http://www.dfki.uni-sb.de/~butz/oogltools/models
http://www.dfki.uni-sb.de/vitra/