Recalage d'un système multi-vues pour la texturation automatique de modèles 3D simplifiés, en milieu urbain

Arnaud Le Troter[1] Alexandre Amalric[2] Jean-Patrice Noell[2] Bruno Serra[3]

[1] Laboratoire des Sciences de l'Information et des Systèmes, Equipe Image & Modèles (I&M)
ESIL - Campus de Luminy - case 925, 163 avenue de Luminy, 13288 Marseille Cedex 9
[2] PIXXIM S.A. 73E rue Perrin Solliers 13006 Marseille
[3] Thalès Alenia Space 100 Bd du Midi B.P.99 06322 Cannes La Bocca Cedex - France

arnaud.le-troter@univmed.fr, {a.amalric, jp.noell}@pixxim.fr, bruno.serra@thalesaleniaspace.com

Résumé

Nous présentons dans cet article une méthode globale de texturation de façades sur des modèles 3D polyhédriques de villes. Une première partie sera dédiée à la description des contraintes de prises de vues en milieu urbain, et présentera une méthodologie de prises de vues contiguës géo-référencées issues d'un dispositif multi-caméras préalablement calibrées. Ce dernier a été conçu de telle sorte qu'il autorise le mosaïcage d'images et facilite l'appariement d'images successives. La seconde partie propose une méthode itérative de recalage d'un système multi-vues ; L'objectif étant de recaler automatiquement des courtes séries de photos, en ne considérant que les couples de photos successives, et de façon indépendante. Pour chaque série, la première étape est le recalage manuel du dispositif. La seconde étape consiste à l'estimation de la géométrie épipolaire de chaque couple de photos. L'originalité de notre approche se situe dans l'estimation itérative des facteurs d'échelle successivement calculés dans le but de minimiser l'éloignement entre les points 3D issus de la photogrammétrie et le modèle 3D polyhédrique du bâti.

Mots clefs

Géo-référencement de photos, mosaïcage d'images, géométrie épipolaire, photogrammétrie, texturation, villes virtuelles 3D

1 Introduction

La réalisation de nouveaux logiciels puissants dans le domaine de la reconstruction 3D de villes virtuelles par de grands industriels comme Google et Microsoft, suscite depuis quelques années un intérêt croissant dans les domaines de la recherche et de l'industrie. Google Earth, Microsoft Live Maps, et plus récemment le Géoportail français, proposent des représentations 3D de villes virtuelles à grandes échelles, dont les particularités sont la génération automatique de grands volumes de données peu détaillés, permettant ainsi une visualisation temps-réel et ergonomique. Ces logiciels pêchent encore au niveau du réalisme des vues au sol, en particulier par la mauvaise qualité des textures de façades. Des solutions alternatives comme Google Street View ou encore Mappy permettent de visualiser des photos réelles de façades, mais dans un mode de navigation restreint. Le point fort et commun de ces approches est leur aspect non supervisé.

Le projet "Ville de Cannes en images virtuelles 3D" (modélisation 3D hyperréaliste de la ville de Cannes) réalisé entre août 2004 et juin 2006 permet d'illustrer l'intérêt de l'automatisation de ce travail. L'état final du prototype est une maquette 3D très détaillée (figure 1) sur l'ensemble du territoire de la commune de Cannes [1] qui comporte trois points innovants :

- une visualisation d'un volume "gigantesque" de données en temps réel, en utilisant des techniques avancées de "pagination" des données graphiques,
- une interface entre une application 3D "temps réel" et un Système d'Information Géographique (SIG),
- un rendu 3D réaliste permettant une forte immersion virtuelle.

Figure 1 – *Maquette de la ville de Cannes*

Cependant, la réalisation opérationnelle du projet, effectuée dans le contexte d'un faible niveau d'automatisation, a été caractérisée par un fort coût humain (25 homme x mois pour une commune de 21 Km^2). Le réalisme de la maquette 3D a été limité par plusieurs facteurs liés au volume des données à gérer (2300 ha de terrain, 217 Km de voirie, 22000 bâtiments) : par exemple, il n'aurait pas

été raisonnable, économiquement parlant, de travailler de manière détaillée sur tous les bâtiments que compte la ville de Cannes. De fait, le rendu des textures et la précision de la texturation sont d'un niveau acceptable, mais très perfectible : de nombreuses textures sont génériques.

L'objet de cet article est de présenter une méthode de texturation de façades peu supervisée permettant d'augmenter le rendement dans la production d'une ville virtuelle.

2 Problématique

Il y a deux façons d'aborder le problème de la texturation d'un modèle 3D polyédrique. La première approche consiste à sélectionner, segmenter et "redresser" des régions d'intérêt dans l'image, puis à les associer aux faces correspondantes du modèle 3D. La seconde approche consiste à recaler précisément la caméra (ou l'image résultante) avec le modèle 3D. Cela nécessite une étape fastidieuse de calibrage de caméra mais qui, une fois effectuée, permet de texturer l'ensemble des faces visibles dans le champ de vision de la caméra. Chacune de ces méthodes a ses avantages et inconvénients. La première correspond plus à une démarche pragmatique, et par conséquent est très supervisée. La seconde répond plus à des attentes d'automatisation. Elle est donc peu supervisée, et peut s'étendre au problème du recalage multi-caméras, ceci souvent au détriment de la robustesse et de la précision. En effet, les méthodes de recalage de caméras nécessitent des phases d'optimisation complexes, et, de plus, les modèles 3D (notamment ceux issus de la photogrammétrie aérienne) ne reproduisent pas toujours précisément la réalité.

La texturation spécifique de façades soulève dès le départ le problème de l'acquisition et de la gestion de données à grande échelle. La phase initiale de notre étude a donc consisté à élaborer un dispositif de prises de vues géo-référencées (par GPS), permettant de structurer l'ensemble des données concernées (section 3).

Nous avons mis en œuvre une méthodologie de prises de vues contiguës, à la fois compatible avec des procédés photogrammétriques nécessitant de forts recouvrements inter-photos, et optimisées de manière à ce que la couverture d'une ville entière se fasse dans des délais raisonnables (parcours de quelques jours). L'objectif était de trouver un bon compromis entre, d'une part la quantité de données acquises, et, d'autre part la qualité des photos et de leurs recouvrements, tout en tenant compte des contraintes induites par l'environnement urbain (circulation, accessibilité).

Les méthodes classiques de photogrammétrie consistent à reconstruire un modèle 3D (en général un nuage de points non structuré) uniquement à partir d'images (PhotoTourism[2], Debevec[3], Pollefey[4]). Notre problématique diffère de ce type d'approche, car nous disposons déjà d'un modèle 3D polyédrique du bâti, construit manuellement à partir de vues stéréoscopiques aériennes. Nous avons donc utilisé les outils photogrammétriques pour résoudre le problème du recalage de caméras avec un modèle 3D. On peut cependant noter que des travaux récents dans ce même contexte ont déjà été effectués avec une efficacité certaine, notamment le recalage de données SIG (modèle 3D, GPS) avec des sources vidéos [5][6].

Connaissant les paramètres d'un dispositif de prise de vue (focale, position, orientation, distorsion), chaque photo peut être modélisée sous la forme d'une caméra virtuelle. Le recalage précis de ces caméras virtuelles dans le modèle 3D une fois effectué, la texturation est ramenée à un problème plus simple : la projection des faces du modèle 3D dans les "plans image" des caméras virtuelles et la rétroprojection de ces régions sur le modèle 3D.

Cependant, il ne faut pas sous-estimer l'impact de la précision du recalage sur la qualité de la projection de la texture. En annexe, nous illustrons la répercussion de l'erreur d'un géo-référencement GPS d'une photo (d'au moins 5 mètres [5]), cumulée à l'erreur de l'orientation du dispositif (de l'ordre de 5°) dans le processus de texturation mono-caméra. Nous avons mis en évidence les limites d'un processus d'extraction d'une texture de façade dans une seule image, et de correction des position et orientation brutes d'une unique caméra. La rectification de ces données par des procédés de traitement et d'analyse d'image s'avère non fiable au delà d'une précision de l'ordre du mètre pour le positionnement et du degrés pour l'orientation.

3 Méthodologie de prises de vues et contraintes en milieu urbain

Nous souhaitons texturer des bâtiments situés à la fois en bordure de route mais aussi dans de petites rues piétonnes inaccessibles en voiture. Deux types de dispositifs de prises de vues (figure 2) ont été utilisés et ont permis de construire des jeux de données conséquents sur la ville de Cannes. Tous deux sont munis d'un système de géo-référencement GPS et d'estimation de l'orientation, permettant ainsi la reconstruction de la trajectoire empruntée.

Figure 2 – *Dispositifs de prises de vues utilisés, pédestre mono-caméra (à gauche), véhicule multi-caméras (à droite)*

3.1 Prises de vues géo-référencées

Le GPS est le système de géo-référencement qui nous a paru le plus adapté à notre problématique, car plus souple dans son utilisation que des systèmes comprenant des odomètres ou des centrales inertielles. Il fournit par défaut, le positionnement en latitude et longitude, mais malheureusement de façon souvent imprécises en milieu urbain (Notamment à cause de l'effet "canyon" : occultation d'un satellite par le relief - un bâtiment par exemple - , écho du signal contre une surface, etc...). Afin d'augmenter la précision des données de géo-référencement, le récepteur GPS a été connecté à un récepteur différentiel DGPS. Un gyroscope sert de compas (pour l'estimation de l'orientation).

3.2 Contraintes d'une prise de vues en milieu urbain

Réaliser une bonne prise de vue en milieu urbain revient à trouver un compromis entre la largeur du champ de vision, l'éloignement de la caméra à la façade, le parallélisme du plan image et de la façade.
Sur la figure 3, nous présentons trois prises de vues (réalisées avec le dispositif pédestre (figure 2) d'une même façade) et la représentation des caméras virtuelles correspondantes, recalées dans le modèle 3D. La première photo est trop éloignée de la façade, par conséquent la résolution de la texture est trop faible. La seconde est la meilleure prise de vue au regard du meilleur compromis précédemment énoncé. La troisième s'avère inutilisable, une partie de la façade n'étant pas dans le champ de vision de la caméra. Dans la réalité, la possibilité de s'éloigner d'une cible est très souvent limitée ou contrainte par des obstacles présents dans la scène (bâtiments, voitures, etc...). Cette figure illustre la difficulté d'effectuer une bonne prise de vue en milieu urbain avec un simple objectif.

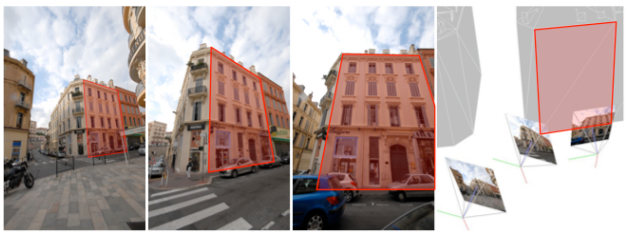

Figure 3 – *Trois prises de vues d'une façade (marquée en rouge) et leurs positions relatives dans le modèle 3D*

Pour remédier à ce problème, nous avons utilisé un dispositif composé de plusieurs caméras (jusqu'à six boîtiers) disposées autour d'un axe offrant un champ de vision de 360° (figure 2), et permettant la création de vues panoramiques de qualité.

Dans notre expérimentation, nous nous sommes limités à l'utilisation de triplets de caméras (figure 4).

Figure 4 – *Exemple d'une prise de vue d'un triplet de caméras (trois images et une mosaïque du triplet)*

On nomme mosaïque d'images, une projection des plans image du triplet sur un plan (sur la figure 4, le champ de vision a été limité à 150°). Une mosaïque est modélisable comme une caméra virtuelle, au même titre que chacune des caméras du triplet la définissant. Les boîtiers ont été placés autour d'un axe de rotation de sorte que leurs axes optiques soient concourants sur cet axe. Le dispositif a été préalablement calibré pour automatiser la génération de mosaïques.

3.3 Méthodologie de séries de prises de vues

La méthodologie d'une série de prises de vues est schématisée sur la figure 5. Il convient d'estimer la transition entre deux prises de vues de triplets consécutifs. Pour cela, le recouvrement des champs de vision d'au moins un couple de caméras issues des deux triplets est nécessaire. Sur la figure 5, nous avons hachuré la zone de recouvrement d'une caméra du $triplet_i$ (cam3) et d'une caméra du $triplet_j$ (cam1) et représenté la transition $(RTs_d)_{ij}$ qui exprime la position relative du $triplet_j$ par rapport à la position du $triplet_i$ (à un facteur d'échelle s_d près).

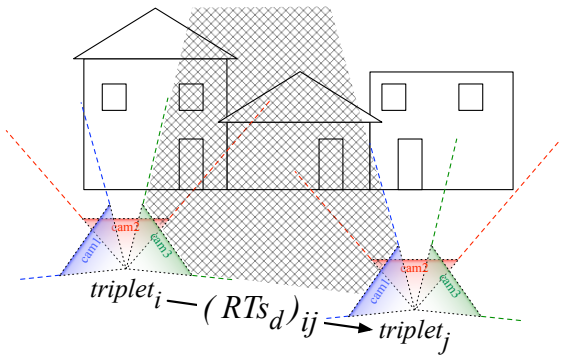

Figure 5 – *Représentation de deux triplets successifs de caméras et leur zone de recouvrement*

4 Recalage d'un système multi-vues

La première étape consiste en un recalage initial de l'ensemble des caméras à l'aide des données de positionnement GPS brutes. Cette étape fournit une estimation grossière de la position et de l'orientation de la caméra dans le repère géo-référencé du modèle 3D, pour chaque prise de vue.

Cette estimation est nécessaire car elle permet d'apparier des faces du modèle 3D avec les façades correspondantes de l'image. Sur la figure 7, nous illustrons la vue subjective d'une caméra virtuelle initialisée par les données brutes (image centrale) et le résultat d'un recalage manuel de caméra (image de droite). En moyenne, l'écart entre la position initiale et celle recalée est de l'ordre de 5 mètres circulaire et de 5 degrés.

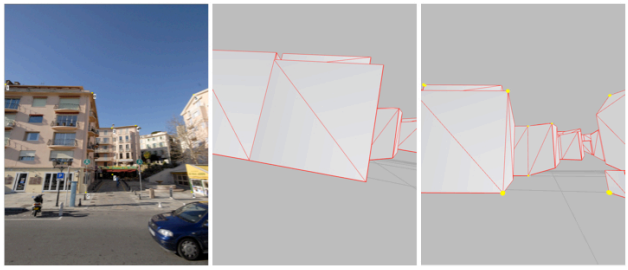

Figure 6 – *Recalage manuel d'une caméra à partir d'une vue issue du positionnement GPS*

La seconde étape correspond au recalage automatique de triplets successifs de caméras par la résolution d'un système d'optimisation. Les grandes lignes de notre méthode sont les suivantes :

- Un premier $triplet_i$ est précisément calibré puis recalé sur le modèle 3D (section 4.1)
- On cherche des points de correspondance fiables (pt_i) et (pt_j) dans les images issues respectivement des triplets successifs i et j par la méthode de SIFT (section 4.2).
- Pour estimer la position du $triplet_j$ (relative à celle du $triplet_i$), on calcule la transformation $(RTs_d)_{ij}$ définie à un facteur d'échelle près $(s_d)_{ij}$ (section 4.3).
- En faisant l'hypothèse qu'une majorité des points des nuages (pt_i) et (pt_j) sont des points situés sur la façade, on estime ensuite le facteur d'échelle $(s_d)_{ij}$ en introduisant le modèle 3D (section 4.4)
- Le $triplet_j$ est alors considéré comme recalé par rapport au modèle 3D et on réitère ces opérations pour estimer successivement les positions des caméras suivantes.

4.1 Calibrage et recalage d'une caméra

Le but principal du calibrage est la détermination d'informations tri-dimensionnelles à partir des images, qui, elles, sont bi-dimensionnelles. Les images dépendent de la structure de la scène, mais aussi des caractéristiques des caméras qui les ont acquises. Quant à la géométrie d'une caméra, nous distinguons sa géométrie externe (sa position et son orientation : paramètres extrinsèques) de sa géométrie interne (l'ensemble de ses paramètres internes ou intrinsèques), décrivant ses propriétés optiques et autres, tels la distance focale ou la taille des pixels. On dit qu'une caméra est calibrée lorsque sa géométrie interne est connue, et recalée lorsqu'on estime sa géométrie externe. Le modèle de camera que nous avons utilisé pour le calibrage et recalage est le modèle de Zhang [7], qui permet aussi de corriger la distorsion radiale.

4.2 Descripteur SIFT pour l'appariement d'images

La recherche de points de correspondance dans les couples d'images est une des étapes les plus sensibles de notre méthode car elle conditionne fortement la qualité du recalage. En particulier, les points de correspondance sont le support du calcul de la géométrie épipolaire pour l'estimation des transitions de caméra, et aussi pour l'intégration du modèle 3D. En effet, les images de façades sont souvent composées de motifs répétés (fenêtres par exemple), qui peuvent engendrer de fausses correspondances.

Les points de correspondance sont recherchés par la méthode SIFT[8]. Chaque image est l'objet d'une analyse en termes de points d'intérêt, chacun d'eux caractérisé par un descripteur multi-dimensionnel invariant aux facteurs d'échelles, aux rotations et robuste aux changements de point de vue et d'illuminations. La mise en correspondance des points d'intérêt nécessite une mesure performante de similarité entre descripteurs ; Pour cela nous avons utilisé une méthode d'appariement vectoriel sur les 128 attributs caractéristiques. L'appariement se fait en utilisant la méthode proposée par Lowe ([8]) qui recherche pour chaque point d'intérêt d'une image, le point sur l'autre image le plus proche, au sens de la distance vectorielle. Pour améliorer ce processus, nous avons introduit la contrainte épipolaire, en ne conservant que les couples de points positionnés sur leurs droites épipolaires associées.

4.3 Estimation de la transformation géométrique entre deux caméras

La transformation géométrique entre deux caméras relève de la géométrie épipolaire. Elle s'écrit sous la forme d'une matrice 3x3 de rang 2. On appelle cette matrice la matrice fondamentale [9] lorsque les paramètres intrinsèques de la caméra sont inconnus et la matrice essentielle lorsqu'ils sont préalablement estimés. Cette matrice définit à un facteur d'échelle près la transformation entre deux caméras. Pour l'estimation de la matrice fondamentale, nous avons limité nos expérimentations aux algorithmes classiques LMEDS (Least Median of Squares) et RANSAC (Random Sample Consensus) appliqués à la méthode des huit points [10] parmi les centaines d'appariements proposés par SIFT.

4.4 Détermination automatique du facteur d'échelle

Les étapes suivantes sont mises en oeuvre pour estimer la valeur du facteur d'échelle :

- Application du détecteur SIFT sur les trois images résultantes des trois caméras des $triplet_i$ et $triplet_j$
- Recalage du $triplet_i$ par la mise en correspondance manuelle d'au moins sept points des images du $triplet_i$ avec des points 3D du modèle.
- Estimation de la géométrie épipolaire entre les $triplet_i$ et $triplet_j$ (quatre couples de caméras sont prédéfinis pour le calcul : $(cam2_i, cam1_j)$, $(cam2_i, cam2_j)$, $(cam3_i, cam1_j)$ et $(cam3_i, cam2_j)$, voir figure 5).
- Validation de la mise en correspondance des points (pt_i) et (pt_j) par l'application de la contrainte épipolaire.
- Estimation de la transformation $(RTs_d)_{ij}$ par la méthode RANSAC avec l'algorithme des huit points.
- Calcul itératif de la position du $triplet_j$ relative à celle du $triplet_i$ pour des facteurs d'échelle $(s_d)_{ij}$ avec d variant entre 0 et D (D étant la distance maximum parcourue entre deux prises de vues, environ 10 mètres) et estimation des points $(pt_{ij})_d$ intersection des faisceaux passant par les points (pt_i) et (pt_j) et les centres optiques respectifs des triplets i et j.
- Calcul des projections (pt_{3D}) de (pt_i) sur le modèle 3D
- Algorithme des moindres carrés avec (pt_{3D}) et $(pt_{ij})_d$

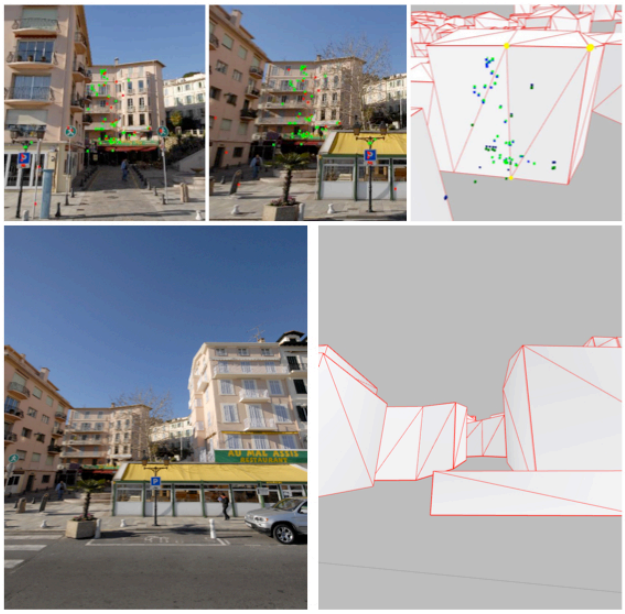

Figure 7 – *Etapes du processus de recalage automatique d'une caméra d'un couple stéréoscopique*

La figure 7 illustre certaines étapes de l'estimation automatique d'une pose de caméra. Les points verts sont les points de correspondance validés par la contrainte épipolaire, à l'inverse des points rouges. Les points bleus sont les intersections avec le modèle 3D.

La figure 8 montre l'application de la méthode sur une série de six prises de vues successives. Le triplet de caméra le plus à gauche a été recalé manuellement et les cinq autres triplets ont été successivement recalés de manière automatique.

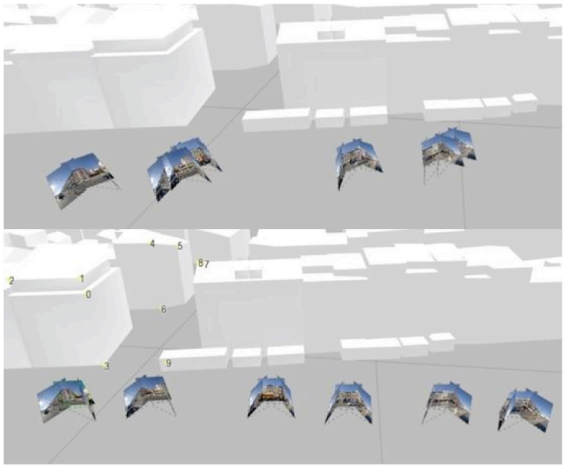

Figure 8 – *Positions de six triplets avant et après correction partie supérieure : localisation des triplets de manière grossière par géo-référencement. Partie inférieure : répartition finale après correction*

5 Conclusions et Perspectives

La méthode de texturation de modèle 3D présentée dans cet article est fondée sur un système multi-vues panoramiques. Une base de données d'images conséquente nous a permis de valider et de qualifier les limites de notre méthode de texturation de façades. L'étape automatique de recalage de caméras se limite en moyenne à des séries d'une dizaine de triplets. Pour pallier la dérive du système et pour garantir une meilleure production, il est possible d'interagir de manière ergonomique en introduisant des nouveaux couples de points d'ancrages.

Contrairement à des approches photogrammétriques classiques où l'objectif est la reconstruction d'un modèle 3D (nuage de points non structuré et très dense) à partir d'un ensemble d'images, notre approche est originale dans le sens où nous introduisons un modèle 3D simplifié et structuré (arrêtes de faîtages, gouttières, etc...) à chaque nouvelle estimation de positionnement de caméras ; L'idée étant de minimiser les erreurs intrinsèques aux diverses données utilisées qui sont souvent imprécises (modèle 3D polyhédrique issu de photogrammétrie aérienne, nuage de points 3D issu de photogrammétrie au sol, données GPS). Ce module est en cours d'intégration dans la chaîne de production de la société PIXXIM et permet d'une part l'amélioration du réalisme d'une maquette, et d'autre part un meilleur positionnement commercial de la réalisation de villes virtuelles 3D.

Annexe

Dans cette annexe, nous mesurerons la répercusion des erreurs éventuelles du géo-référencement d'une photo (donc de la position et de l'orientation d'une caméra) dans l'image pour estimer jusqu'à quel point notre processus d'analyse d'images peut pallier ces imprécisions.

Notre méthode d'extraction automatique de texture de façades pour un système mono-caméra géo-référencé par GPS se décompose en cinq étapes :

- Correction automatique de la distorsion dans l'image.
- Caractérisation de faisceaux de droites horizontaux et verticaux pour la détection des points de fuites dans l'image par une transformée de Hough.
- Redressement de la perspective
- Recalage de la façade du modèle 3D projetée dans l'image sur les alignements et points remarquables les plus proches situés dans un anneau d'intérêt.
- Extraction de la texture prête à être plaquée sur le modèle 3D

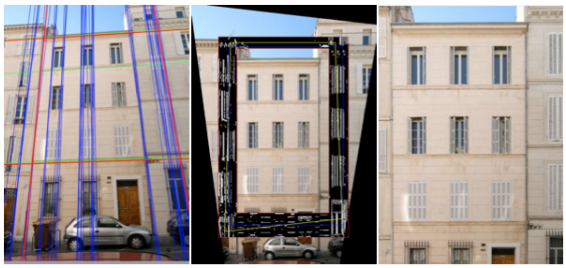

Figure 9 – *étapes du processus d'extraction automatique de texture*

La figure 9 illustre les limitations de cette approche. L'étape de redressement de la perspective est robuste mais l'erreur de projection de la façade dans l'image cumulée à la présence d'un masque (ici une voiture) ne permet pas de segmenter correctement la texture (sur l'image de droite, aucun bord de la texture ne correspond aux bords de la façade).

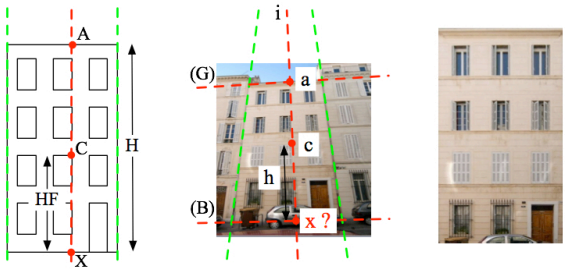

Figure 10 – *Recherche dans le modèle et dans l'image la conservation d'un bi-rapport*

Remarque : une transformation projective ne conserve pas les rapports mais uniquement les bi-rapports :

$$\frac{\|A-C\|}{\|X-C\|} = \frac{H}{HF} - 1 = \frac{\|a-c\|/\|a-i\|}{\|x-c\|/\|x-i\|}$$

Pour retrouver par exemple le bas de la façade dans l'image (la droite (B)), de nombreuses informations sont à connaître ou à extraire du modèle et de l'image de la façade (figure 10) : la hauteur H de la façade, la hauteur HF d'un point C de la façade et son correspondant c dans l'image, les points i et a dans l'image, (respectivement un point de fuite et l'intersection de la ligne de gouttière (droite (G)) et de la droite passant par les points c et i), un point x appartenant à la droite (B).

Références

[1] Site web de la maquette 3d de la ville de cannes, 2006. `http://3d.cannes.fr/`.

[2] Noah Snavely, Steven M. Seitz, et Richard Szeliski. Photo tourism : Exploring photo collections in 3d. Dans *SIGGRAPH Conference Proceedings*, pages 835–846, New York, NY, USA, 2006. ACM Press.

[3] Paul Ernest Debevec. *Modeling and rendering architecture from photographs*. Thèse de doctorat, 1996. Chair-Jitendra Malik.

[4] Marc Pollefeys et Luc Van Gool. From images to 3d models. *Commun. ACM*, 45(7) :50–55, 2002.

[5] T. Colleu, G. Sourimant, et L. Morin.. Une méthode d'initialisation automatique pour le recalage de données sig et vidéo. Dans *COmpression et REpresentation des Signaux Visuels*, Montpellier, France, novembre 2007. (CORESA 2007).

[6] M. Pollefeys, D. Nistér, J. M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S. J. Kim, P. Merrell, C. Salmi, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewénius, R. Yang, G. Welch, et H. Towles. Detailed real-time urban 3d reconstruction from video. *Int. J. Comput. Vision*, 78(2-3) :143–167, 2008.

[7] Zhengyou Zhang et Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 :1330–1334, 2000.

[8] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2) :91–110, 2004.

[9] Q. t. Luong. The fundamental matrix : Theory, algorithms, and stability analysis. *International Journal of Computer Vision*, 17 :43–75, 1996.

[10] R. I. Hartley. In defence of the 8-point algorithm. Dans *ICCV '95 : Proceedings of the Fifth International Conference on Computer Vision*, page 1064, Washington, DC, USA, 1995. IEEE Computer Society.