Vers la modélisation 3D flexible avec une caméra catadioptrique

Toward Flexible 3D Modeling using a Catadioptric Camera

Maxime Lhuillier
LASMEA-UMR 6602, UBP/CNRS
24 avenue des Landais, 63177 Aubière
Maxime.Lhuillier@univ-bpclermont.fr, maxime.lhuillier.free.fr

Résumé

La modélisation 3D automatique à partir d'une séquence d'images catadioptriques est un sujet relativement peu traité jusqu'à présent, bien que ce problème soit classique pour des images perspectives. Toutes les approches (catadioptriques) précédentes sont limitées à la reconstruction dense à partir de quelques points de vues, et la majorité d'entre elles nécessitent la calibration de la caméra. Cet article propose une méthode capable de traiter des séquences de plusieurs centaines d'images sans connaissance précise de la calibration. Dans ce contexte, un même point 3D de la scène peut être visible et reconstruit par un grand nombre d'images à des précisions très variables. La plus grande partie de cet article traite donc de la sélection des points reconstruits, un problème largement ignoré dans les travaux précédents. Les résumés des étapes d'estimation de la géométrie et de stéréo dense sont aussi donnés. Les expériences incluent la reconstruction de modèles 3D de scènes d'intérieur et d'extérieur, ainsi qu'une visite virtuelle dans une ville (vidéo disponible sur le web).

Mots Clef

Modélisation basée image, caméra catadioptrique, incertitude, système de reconstruction 3D automatique.

Abstract

Fully automatic 3D modeling from a catadioptric image sequence has rarely been addressed until now, although this is a long-standing problem for perspective images. All previous catadioptric approaches have been limited to dense reconstruction for a few view points, and the majority of them require calibration of the camera. This paper presents a method which deals with hundreds of images, and does not require precise calibration knowledge. In this context, the same 3D point of the scene may be visible and reconstructed in a large number of images at very different accuracies. So the main part of this paper concerns the selection of reconstructed points, a problem largely ignored in previous works. Summaries of the structure from motion and dense stereo steps are also given. Experiments include the 3D model reconstruction of indoor and outdoor scenes, and a walkthrough in a city (video available in the web).

Keywords

Image-Based Modeling, Catadioptric Camera, Uncertainty, Automatic 3D Reconstruction System.

1 Introduction

Le calcul de modèles 3D photo-réalistes à partir d'une séquence d'images est un sujet de recherche à long terme en vision et en synthèse d'images par ordinateur. Le besoin minimal pour une visite interactive est le rendu de la scène pour toutes les directions de vues horizontales lorsque l'observateur se déplace sur le sol. Ceci suggère un grand champ de vue pour les images données, pour lesquelles plusieurs types de caméras sont possibles [3] : les caméras catadioptriques, fish-eye, ou des systèmes rigides de plusieurs caméras pointant dans plusieurs directions. Puisque l'on souhaite capturer toute scène (intérieure et extérieure) accessible à un piéton, le matériel d'acquisition doit être tenu à la main ou sur la tête, peu encombrant, et (si possible) bon marché. On choisit ici une caméra catadioptrique car elle satisfait à toutes ces contraintes.

Le principal inconvénient de ce choix est la faible résolution comparée à une caméra standard (perspective) pour un champ de vue donné. On souhaite limiter ce problème grâce à des photos prises individuellement à l'aide d'une caméra équiangulaire. La qualité d'une photo (résolution, bruit) est meilleure que celle d'une image vidéo. Une caméra équiangulaire est choisie parmi d'autres car elle est conçue pour répartir au mieux la résolution du champ de vue dans toute l'image. Ces deux choix ont principalement deux conséquences. D'abord, une séquence d'images prises individuellement nécessite un peu d'effort et de patience : l'utilisateur doit alterner un pas en avant dans la scène et la prise d'une photo (non floue) en appuyant sur un bouton. Ensuite, une caméra équiangulaire n'est pas une caméra centrale. Un modèle non-central complique toutes les méthodes impliquées, car les rayons lumineux (droites) s'échappant de la caméra ne s'intersectent pas en un point unique comme pour une caméra perspective. Cet article montre cependant que les résultats obtenus avec ce matériel sont très intéressants, et que l'approximation par un modèle de caméra central suffit dans beaucoup de cas.

L'approche présentée ici est complètement automatique étant donnée la séquence d'images acquises avec la caméra se déplaçant dans la scène : (1) estimation de la géométrie complète à l'aide d'une méthode de "structure from motion", et (2) reconstruction d'un modèle 3D par une méthode de stéréo dense multi-vues et par fusion de modèles locaux. Par "structure from motion", on entend l'estimation automatique, robuste et optimale de la structure de la scène (un nuage de points 3D épars) et du mouvement de la caméra (tous les paramètres extrinsèques et quelques paramètres intrinsèques) à partir des images. Le modèle 3D obtenu est une liste de triangles texturés en 3D qui approximent la partie visible de la scène, et on le visualise avec des outils standards. Bien que cette approche soit maintenant "classique" pour une caméra perspective tenue à la main [15, 17, 10], de tels modèles 3D n'avaient pas été obtenus jusqu'ici avec une caméra catadioptrique.

1.1 Travaux précédents

Il y a eu beaucoup de travaux sur la reconstruction de modèles 3D à partir d'une séquence d'images. Les contextes sont nombreux : capteurs actifs ou passifs, points de vues aériens ou terrestres, mouvements de caméras généraux ou particuliers (ex : tables tournantes), objets généraux ou particuliers (ex : visages), méthodes manuelles ou automatiques. L'objectif ici n'est pas de passer en revue tout les cas, mais de se focaliser uniquement sur les plus proches travaux.

Les auteurs [7, 1, 5] reprojettent leurs images originales sur des cylindres virtuels (images panoramiques) pour appliquer une méthode de stéréo dense avec la contrainte épipolaire. La première tentative [7] fusionne des images acquises avec une caméra perspective qui pivote autour d'un axe vertical, et les suivantes [1, 5] reprojettent des images catadioptriques (centrales). Un suivi de points et l'algorithme des 8 points [6] sont utilisés par [7, 1] pour estimer la géométrie des paires d'images cylindriques successives. Ensuite, les facteurs d'échelle relatifs à deux paires d'images sont obtenus en mesurant les distances entre les positions de la caméra ou par odométrie [7, 1]. La géométrie complète de la séquence est alors connue. Enfin, ces deux approches utilisent une méthode de stéréo multi-baseline inspirée de [16]. Le troisième travail [5] utilise une méthode active pour calculer la géométrie de la séquence et une méthode de stéréo dense par coupure de graphe (ainsi que de nombreux post-traitements). Une méthode manuelle simple est aussi proposée par [4], mais le point important ici est l'introduction d'une méthode de sélection de paire d'images pour reconstruire un point avec la meilleure "fiabilité" (ou "reliability"). Ceci est un problème clef (négligé par les autres auteurs) pour la modélisation 3D à partir d'images catadioptriques, car un même point 3D peut être visible et reconstruit dans un grand nombre d'images à des résolutions et baselines très diverses.

Comme cela a déjà été dit, il faut avant tout estimer la géométrie [6]. Les principes sont bien connus maintenant. Cependant, les systèmes de reconstruction automatiques, robustes et optimaux (c'est à dire incluant un ajustement de faisceaux global) ne sont pas si nombreux si la seule donnée est une longue séquence d'images acquise par une caméra catadioptrique générale et une faible connaissance de la calibration. Ceci contraste avec le cas des caméras perspectives. Jusqu'à présent, les travaux les plus avancés sur ce sujet étaient ceux de [19] et [14]. Un ajustement de faisceaux (AF) est appliqué une seule fois à la fin de la procédure d'estimation de la géométrie étant donnée une connaissance précise de la calibration [19], bien qu'il soit reconnu qu'une approche hiérarchique pour l'AF [6] améliore considérablement la précision et la robustesse simultanément. L'approche non-calibrée [14] s'effectue en deux temps : (1) estimer la géométrie complète de la séquence pour un modèle approché de caméra centrale et (2) améliorer cette géométrie en remplaçant le modèle central par un modèle non central intégrant la connaissance du miroir.

1.2 Contributions

La partie 2 présente un modèle de caméra centrale et résume la méthode d'estimation de la géométrie. Les différences avec [14] sont : séquences longues grâce à un AF hiérarchique, une fonction de distorsion radiale générale avec une connaissance approchée des paramètres intrinsèques (au lieu de la connaissance du type de modèle avec des paramètres intrinsèques inconnus). La partie 3 présente une méthode pour obtenir un modèle 3D "local" de la scène étant données trois images de la séquence en utilisant une méthode de stéréo dense. Les images cylindriques [7, 1, 5] ne sont pas utilisées ici dans l'étape de stéréo dense pour faciliter la reconstruction 3D du sol. La partie 4 décrit une généralisation de la sélection de paire d'images pour reconstruire un point : on remplace le critère heuristique sur deux vues [4] (qui utilise un angle entre deux rayons) par un critère général multi-vues (qui utilise des ellipsoïdes d'incertitudes). Les morceaux les plus précis de tout les modèles locaux sont sélectionnés par ce critère, et sont fusionnés ensuite dans le modèle 3D global de la scène. Ce sujet a été négligé par les auteurs précédents, bien que le même point 3D puisse être estimé à des précisions très variables (dans le contexte catadioptrique) selon les images sélectionnées pour la reconstruction. On note que la sélection est une étape de pré-traitement pour des méthodes de fusion comme [2, 18], ce n'est pas une méthode concurrente. Des expérimentations sur des scènes intérieures et extérieures sont proposées dans la partie 5 (pas de scènes extérieures dans les travaux précédents [7, 1, 4, 5]). Cet article est la traduction en français de [12].

2 Estimation de la géométrie

Dans cette partie, la méthode automatique de calcul de la géométrie est brièvement résumée (plus de détails dans [11]). Les hypothèses suivantes sont nécessaires : (1) un miroir de révolution dont les sections circulaires inférieures et supérieures sont visibles (2) une caméra pers-

pective avec des pixels carrés (3) un même axe de symétrie pour le miroir et la caméra perspective (4) une calibration constante et (5) une connaissance approchée des deux angles définissant le champ de vue.

2.1 Modèle de caméra central

Le modèle de caméra est défini par son orientation R (une rotation), son centre $t \in \Re^3$ ($\Re$ est l'ensemble des réels), tout les deux exprimés dans le repère de la scène, et une projection centrale $p : \Re^3 \backslash \{0\} \rightarrow \Re^2$. En utilisant la même notation X pour un point 3D (fini) et ses coordonnées dans le repère de la scène, la direction du rayon lumineux allant de t vers X dans le repère de la caméra est donné par $d = R^\top(X - t)$. La projection dans l'image de X est $p(d)$. Le modèle est symétrique par rapport à l'axe Oz du repère de la caméra : l'image catadioptrique est entre deux cercles concentriques de rayons r_{up} et r_{down}, et il y a une fonction positive décroissante r telle que

$$p(x, y, z) = r(\alpha(x, y, z)) \begin{pmatrix} \frac{x}{\sqrt{x^2+y^2}} & \frac{y}{\sqrt{x^2+y^2}} \end{pmatrix}^\top$$

avec $\alpha(x, y, z)$ l'angle entre l'axe Oz et le rayon lumineux de direction $d = \begin{pmatrix} x & y & z \end{pmatrix}$. Soient α_{up} et α_{down} les deux angles qui définissent le champ de vue. On a $\alpha_{up} \leq \alpha \leq \alpha_{down}$ et $r_{down} \leq r(\alpha) \leq r_{up}$. La caméra est exactement équiangulaire si r est une fonction affine.

2.2 Initialisation de la calibration

Grâce aux hypothèses, les sections circulaires basse et haute du miroir se projettent dans les images en cercles concentriques. D'abord ces cercles sont détectés et estimés dans chaque image catadioptrique avec les méthodes RANSAC et Levenberg-Marquardt appliquées sur des contours régulièrement polygonisés (dans ce papier, la stabilité de détection est améliorée en supposant que les cercles sont immobiles dans la séquence complète). Ensuite, la fonction $r(\alpha)$ initiale est définie par la fonction affine telle que $r(\alpha_{up}) = r_{up}$ et $r(\alpha_{down}) = r_{down}$. Les rayons de cercles r_{down}, r_{up} sont obtenus dans l'étape précédente, et les angles $\alpha_{down}, \alpha_{up}$ définissant le champ de vue sont donnés par le constructeur du miroir. Ces angles ne sont pas exactement connus puisqu'ils dépendent de la position relative entre le miroir et la caméra perspective.

2.3 Initialisation de la géométrie

Des points de Harris sont détectés et appariés dans chaque paires d'images successives de la séquence à l'aide du score de corrélation ZNCC, sans contrainte épipolaire. Les directions des rayons correspondants sont aussi obtenues grâce à l'initialisation de la calibration. Ensuite, les matrices essentielles pour ces paires d'images sont estimées avec RANSAC (en utilisant l'algorithme des 7 points [6]) et raffinées par Levenberg-Marquardt. Des points 3D sont aussi reconstruits pour chaque paire. Plusieurs de ces points 3D sont détectés dans trois images, et sont utilisés pour initialiser les échelles 3D relatives entre deux paires d'images successives. Enfin, la géométrie de la séquence complète est obtenue à l'aide de plusieurs ajustements de faisceaux (AF) appliqués hiérarchiquement pour fusionner toutes les géométries partielles [6]. Ces AF raffinent simultanément les paramètres des points 3D et des poses de la caméra en minimisant la somme des carrés des erreurs de reprojections.

2.4 Raffinements de la géométrie

Une fois que la géométrie complète est obtenue pour la fonction approchée $r(\alpha)$ définie précédemment, $r(\alpha)$ est redéfinie par un polynôme de degrés 3 dont les 4 coefficients sont à estimer. Un AF supplémentaire est appliqué pour estimer les $4 + 6c + 3p$ paramètres de la séquence (c est le nombre de poses de la caméra, p est le nombre de points 3D), tout en augmentant le nombre de points 3D et 2D consistants avec la géométrie.

3 Modèles 3D locaux

La géométrie de la séquence d'images catadioptriques a été estimée avec la méthode décrite dans la partie 2. On note que la contrainte épipolaire et la reprojection d'une des images sur une surface virtuelle (cylindre, cube ...) sont faciles à appliquer car un modèle de caméra central est utilisé (ceci n'est pas le cas avec un modèle non central).

Un modèle 3D local est reconstruit à partir de trois images : une image de référence ref, une image secondaire sec_1 avant et une image secondaire sec_2 après l'image de référence dans la séquence. Les trois images ne sont pas nécessairement consécutives ce qui permet de reconstruire des parties de la scène plus ou moins distantes. Évidement, une grande baseline augmente la précision pour les parties distantes, mais une petite baseline est aussi utile pour les parties les plus proches (notamment le sol) pour augmenter le champ de vue commun et simplifier la mise en correspondance.

Les auteurs précédents [7, 1, 5] reprojettent leurs images sur un cylindre virtuel pour appliquer des méthodes de stéréo dense avec la contrainte épipolaire. Les courbes épipolaires sont alors des sinusoïdes [13]. Le principal désavantage des images cylindriques est la grande distorsion d'images sur les parties du haut et du bas, qui correspondent respectivement au ciel et au sol. Ceci augmente la difficulté de la mise en correspondance pour ces composants de la scène. Pour cette raison, on préfère ici reprojeter une image catadioptrique sur les 6 faces d'un cube virtuel et appliquer une méthode de stéréo dense classique pour deux images sur deux faces parallèles de deux cubes. Les courbes épipolaires deviennent alors des droites parallèles, excepté pour les faces qui contiennent les épipoles : les droites intersectent l'épipole au centre d'une face. Par facilité, la méthode quasi-dense [9] est choisie.

La méthode pour un triplé d'images catadioptriques (ref, sec_1, sec_2) est donc la suivante. D'abord, on applique une méthode stéréo dense pour deux vues sur un cube de ref et cube de sec_1, mais aussi sur un cube de ref et cube de sec_2. Ensuite, les résultats de stéréo sont combinés dans

l'image catadioptrique ref. Pour chaque pixel de ref, les points correspondants de sec_1 et sec_2 sont obtenus grâce aux fonctions de reprojections entre cubes et images catadioptriques et aux fonctions induites par la mise en correspondance entre deux cubes. Un point 3D est obtenu pour le pixel courant de ref par intersection de trois rayons lumineux à l'aide de la méthode de Levenberg-Marquardt. Si l'erreur de reprojection est plus grande qu'un seuil (ou bien si l'un des trois rayons est manquant), on peut raisonnablement douter de la qualité de la mise en correspondance et aucun point 3D n'est retenu. Enfin, plusieurs trous (des pixels de ref sans points 3D estimés) sont remplis avec des points 3D par interpolation.

4 Modèle 3D global

Soit L la liste de tout les modèles 3D locaux qui ont été calculés avec les méthodes décrites dans les parties 2 et 3. Certaines parties de ces modèles doivent être sélectionnées et fusionnées dans un modèle 3D global. Premièrement, la partie 4.1 définit l'incertitude virtuelle $U_l(P)$ comme une fonction pour tout point 3D P et tout modèle local l de L. L'incertitude virtuelle est l'incertitude standard lorsque P est reconstruit par l, et l'incertitude virtuelle étend l'incertitude standard pour les autres modèles locaux. Deuxièmement, la sélection sur les modèles 3D locaux est présentée dans la partie 4.2. Étant donné un point P reconstruit par un modèle local l_0, on veut savoir si l_0 est un des modèles locaux de L avec la plus petite incertitude virtuelle pour P. Si cela n'est pas le cas, P ne doit pas être retenu dans le modèle global car un meilleur modèle local est possible pour reconstruire P. Même si l_0 minimise l'incertitude virtuelle pour P, la qualité de P peut être trop mauvaise pour le modèle global (par exemple si P et tout les centres des caméras de l_0 sont alignés). Pour cette raison, des conditions de fiabilité sur P sont décrites dans la partie 4.3. Enfin, la partie 4.4 décrit comment utiliser le critère de sélection efficacement et obtenir concrètement un modèle global à partir de tout les modèles locaux de L.

4.1 Définition de l'incertitude virtuelle

L'incertitude virtuelle $U_l(P)$ pour un point 3D P et un modèle local l est définie de la façon suivante. Soit J_l le Jacobien de la fonction $\Re^3 \rightarrow \Re^{2k}$ qui projette P sur toutes les images $i_1 \cdots i_k$ à partir desquelles l est reconstruit. On note que $C_l(P) = \sigma^2(J_l^\top(P)J_l(P))^{-1}$ est la matrice de covariance associée au problème de reconstruction de P à partir des images $i_1 \cdots i_k$, en supposant des bruits gaussiens identiques et indépendants d'écart type σ pixels pour les erreurs de reprojections et des incertitudes nulles pour les caméras. Une estimation de σ est donnée par les erreurs de reprojections des intersections de rayons (partie 3). On définit l'incertitude virtuelle $U_l(P)$ comme la longueur du plus grand des demi-axes de l'ellipsoïde d'incertitude définit par la covariance $C_l(P)$ et une probabilité p.

Cette définition de $U_l(P)$ suppose que P est "visible" dans toutes les images de l. Si cela n'est pas le cas, on considère que l ne peut reconstruire P précisément et on choisit $U_l(P) = +\infty$. La définition de la visibilité dépend de la connaissance que l'on a de la surface globale à reconstruire. Évidement, un point visible P doit être dans les champs de vue de toutes les images de l. De plus, la normale de la surface en P définit un demi espace qui doit contenir le centre de chaque caméra de l. La surface globale à reconstruire est, elle aussi, une cause d'occultation possible pour P dans chaque point de vue de l, mais on ignore ce problème dans cet article.

L'incertitude virtuelle étend la "fiabilité" (ou "reliability") proposée par [4] : la fiabilité est seulement définie dans le cas $k = 2$ par $\frac{\pi}{2} - \arccos(|d_1.d_2|)$ avec d_1 et d_2 les directions des rayons qui passent par P et les centres des caméras i_1 et i_2, respectivement. Bien que intuitive, (les fiabilités les meilleures/les pires sont pour deux rayons orthogonaux/parallèles), la fiabilité ne dépend pas de la distance entre les centres des caméras. L'incertitude si.

4.2 Sélection sur les modèles 3D locaux

Le critère de sélection est défini à partir de l'incertitude virtuelle. Supposons que l'on ait un point P reconstruit par un modèle local l_0. On souhaite savoir si l_0 est un des meilleurs modèles locaux de L pour reconstruire P.

On peut calculer $U_l(P), \forall l \in L$ et trier ces valeurs dans l'ordre croissant. Si $U_{l_0}(P)$ est classé parmi les n premiers, l_0 est un des meilleurs modèles locaux. Un seuil $n > 1$ est utile pour augmenter la densité des points retenus dans le modèle global, et aussi pour tolérer certaines erreurs de mises en correspondance (les faux négatifs : les points qui auraient du être appariés et qui ne l'ont pas été). Cependant, les n meilleurs incertitudes virtuelles peuvent avoir des ordres de grandeurs très variables. On évite ce problème en estimant l'incertitude relative $U^r_{l_0}(P) = \frac{U_{l_0}(P)}{min_{l \in L} U_l(P)}$ et en retenant l_0 comme l'un des meilleurs modèles pour P si $U^r_{l_0}(P) \leq 1 + \epsilon$. Comme précédemment, un seuil $\epsilon > 0$ est utile pour augmenter la densité des points 3D du modèle global.

On note que ces critères de sélection sont indépendants du choix de la probabilité p et du bruit σ utilisés pour définir les incertitudes virtuelles : changer p ou σ revient à multiplier tout les $U_l(P)$ par un même coefficient.

4.3 Conditions de fiabilité

Supposons qu'un point P est reconstruit par un modèle local l_0. Le point P peut être considéré comme peu fiable et rejeté s'il est trop loin des centres des caméras de l_0, ou bien si ces centres et P sont alignés. Ces cas sont possibles même si l_0 est l'un des meilleurs modèles locaux. Donc, une condition de fiabilité est nécessaire pour décider si P peut être inclus dans le modèle global ou pas.

A première vue, la condition de fiabilité pourrait être "P est fiable si $U_{l_0}(P) < U_0$" avec U_0 un seuil. Cependant, le rendu du modèle global aura deux défauts lors de l'étape de visualisation (visite virtuelle de la scène). D'abord, les points de l'avant plan auront une erreur de reprojection plus

FIG. 1 – Le miroir "0-360" avec le Nikon Coolpix 8700.

grande que ceux de l'arrière plan. Ensuite, trop de points de l'arrière plan seront rejetés puisque l'incertitude d'un point croit avec sa profondeur dans les images. Donc, une seconde définition pour la condition de fiabilité pourrait être "P est fiable si $U_{l_0}(P) < U_0 d^\gamma(P, l_0)$ avec $\gamma > 0$ et d la distance moyenne entre P et les centres des caméras de l_0. Maintenant, les deux défauts sont réduits, mais cette définition est heuristique et dépend des seuils p et σ (définis dans la partie 4.1), U_0 et γ.

Une troisième définition est donnée par [4] : "P est fiable s'il y a une paire de caméra (i, j) telle que $\frac{\pi}{2} - \arccos(|d_i.d_j|) < \frac{\pi}{2} - \theta_0$ avec θ_0 un seuil et d_i la direction du rayon qui passe par P et le centre de la i-ème caméra de l_0. Cette définition n'implique qu'un seul seuil θ_0 (une borne inférieure pour un angle entre deux rayons). Elle traite les deux cas "P est trop loin" et "P et les centres des caméras de l_0 sont alignés".

4.4 Du local au global

L'utilisation brutale du critère de sélection pour obtenir un modèle 3D global a une complexité en temps proportionnelle a $(\#L)^2(\#P)$, avec $\#L$ le nombre de modèles locaux dans L et $\#P$ le nombre (supposé constant) de points 3D dans un modèle local. Cette complexité peut être élevée puisque $\#L$ est au moins proportionnel à la longueur de la séquence (parfois plusieurs centaines de points de vues) et $\#P$ a le même ordre de grandeur que le nombre de pixels dans une image catadioptrique (une centaine de milliers ici).

Le temps de calcul est réduit en subdivisant chaque modèle local en petits morceaux plans ("patch"), et en appliquant le test de sélection (ainsi que le test de fiabilité) seulement une fois sur un point 3D représentatif de chaque morceau. Une fois que l'ensemble global des morceaux est sélectionné à partir de tout les modèles locaux, des techniques standards de fusion [2, 18] peuvent être utilisées pour réduire la redondance des morceaux dans l'espace. Dans cet article, chaque morceau est un carré (les carrés sont assemblés en anneaux dans les images catadioptriques) et son point représentatif a la profondeur médiane du morceau. Bien que le système de reconstruction n'inclus pas encore de méthode de fusion, les modèles 3D globaux obtenus actuellement sont convainquants comme cela est montré dans la partie suivante. Deux triangles sont utilisés pour chaque morceau.

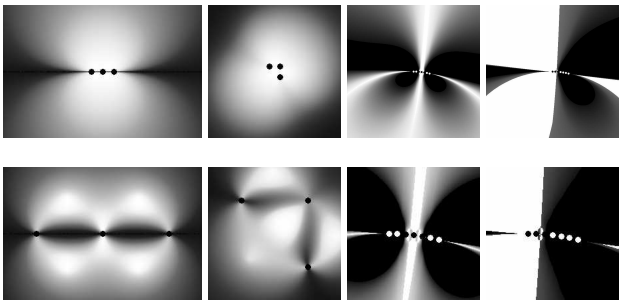

FIG. 2 – Fonctions d'incertitudes virtuelles $P \mapsto U_l(P)$ et $P \mapsto U_l^r(P)$ dans le plan horizontal (les plus petites valeurs sont blanches). Les points de vue du modèle local l sont des points noirs dans ce plan. De gauche à droite : $U_l(P)$ pour 3 points de vue alignés, $U_l(P)$ pour 3 points de vue non alignés, $U_l^r(P)$ pour un modèle local à 3 points de vue au milieu d'une séquence à 7 points de vue, $U_l^r(P)$ pour un modèle local à 3 points de vue en début d'une séquence à 7 points de vue. Les voisinages des points de vue sont zoomés dans la seconde ligne avec normalisation d'histogramme. A droite : la couleur noire est utilisée si $1.5 < U_l^r(P)$, et les points de vue sont blancs si ce ne sont pas ceux de l.

5 Expérimentations

Le contexte expérimental et des expériences synthétiques sur la sélection de modèles 3D locaux sont donnés dans les parties 5.1 et 5.2. Ensuite, les parties 5.3 et 5.4 présentent des résultats obtenus avec le système de modélisation 3D appliqué à des séquences d'images réelles.

5.1 Contexte

L'utilisateur se déplace le long d'une trajectoire sur le sol avec la caméra catadioptrique montée sur un monopode et tenue à la main, en alternant un pas en avant et la prise d'une photo. La caméra est approximativement équiangulaire (le miroir "0-360" avec l'appareil photo Nikon Coolpix 8700, en figure 1), non-centrale, et on suppose que les axes de symétrie du miroir et de l'appareil photo sont les mêmes. Le fabriquant du miroir donne le champ de vue $\alpha_{up} = 37.5$ et $\alpha_{down} = 152.5$ degrés. Ces angles ne sont pas connus exactement puisqu'ils dépendent de la position relative entre le miroir et l'appareil photo. Les images ont pour dimensions 1632×1224 pixels.

5.2 Sélection de modèle 3D locaux

Cette partie présente les expériences sur la sélection de modèles 3D locaux (partie 4) pour des mouvements d'une caméra synthétique.

Premièrement, la fonction d'incertitude virtuelle $U_l(P)$ est montrée à gauche de la figure 2 pour deux modèles locaux. On s'y attendait, $U_l(P)$ augmente si le point P se rapproche de la droite contenant tout les centres/poses de caméra de l (si elle existe) ou si P s'éloigne de tout ces centres. Le premier cas (modèle local dont les 3 centres de caméra sont alignés) arrive souvent si la trajectoire de la ca-

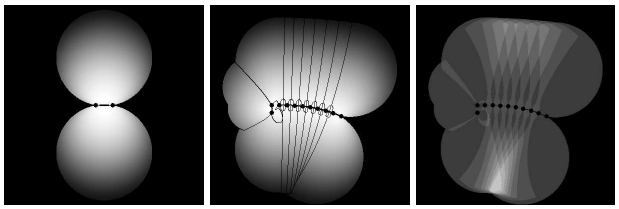

FIG. 3 – A gauche : fonction d'incertitude $P \mapsto U_l(P)$ d'un modèle local à deux points de vue. Au milieu et droite : tout les modèles locaux à trois points de vue consécutifs sont considérés dans une séquence à 11 points de vue. Au milieu : la fonction d'incertitude $P \mapsto U_{l(P)}(P)$ avec $l(P)$ le meilleur modèle local (celui qui minimise $l \mapsto U_l(P)$). Les régions de points P qui ont le même $l(P)$ sont délimités par des lignes noires. A droite : le nombre de modèles locaux l acceptés par la sélection des modèles locaux ($\epsilon = 0.1$) est illustrée par un niveau de gris (gris foncé : 1 modèle local, blanc : 9 modèles locaux). Dans tous les cas, les pixels sont noirs si le point P correspondant ne satisfait pas la condition de fiabilité avec $\theta_0 = 10^o$.

méra est une courbe régulière sur le sol. Puisque ce modèle local reconstruit des points proches de la trajectoire avec beaucoup d'incertitude, ces points apparaîtrons très bruités dans une image rendue (lors de la visite virtuelle) même dans le contexte favorable d'un point de vue au voisinage de la trajectoire initiale. Sélectionner des points dans le modèle global avec une incertitude faible est une mauvaise solution car les points distants seront aussi rejetés. Une grande incertitude est plus acceptable pour les points distants que pour les points proches si les points de vue lors de l'étape de rendu est dans le voisinage de la trajectoire initiale. Le second cas (modèle local dont les 3 centres ne sont pas alignés) est plus favorable car P ne se retrouve jamais aligné avec tous les centres de l.

Deuxièmement, l'incertitude relative est montrée à droite de la figure 2 pour un modèle local à 3 points de vue dans une séquence d'images à 7 points de vue avec des points de vue presques alignés. Comme cela est dit dans la partie 4, un point P reconstruit par l doit avoir un petit $U_l^r(P)$ pour être accepté dans le modèle global : $U_l^r(P)$ doit être dans $[1, 1+\epsilon]$. Dans le premier cas, le modèle local n'est pas à la fin de la séquence complète. On note qu'une tranche de l'espace 3D (délimitée par deux surfaces à peu près planes au voisinage de la trajectoire) contient des valeurs faibles de $U_l^r(P)$, avec $U_l^r(P) = 1$ au milieu de la tranche. Cette tranche passe par la caméra au centre du modèle local, sa largeur augmente avec la distance avec la caméra au milieu, et est connectée aux deux bouts de la séquence complète des points de vue. Dans le second cas, le modèle local est à l'une des extrémités de la séquence complète et la tranche est quasiment remplacée par un demi-espace. Dans les deux cas, on voit que la sélection de modèles 3D locaux (c'est à dire le seuillage de $U_l^r(P)$) n'est pas suffisante pour décider si un point d'un modèle 3D local doit être accepté ou non dans le modèle global.

FIG. 4 – L'image de référence 211 et sa carte de profondeur (un pixel blanc est sans profondeur).

La figure 3 donne des résultats obtenus en combinant la sélection de modèles 3D locaux et la condition de fiabilité basée sur l'angle entre deux rayons et définie par θ_0 (partie 4.3). Dans tous les cas, les pixels noirs à l'extérieur de la région grise ou blanche sont des points qui ne satisfont pas la condition de fiabilité avec $\theta_0 = 10^o$. A gauche, les deux cercles sont l'ensemble des points P tels que l'angle entre les centres des deux caméras est égal à θ_0. Au milieu et à droite, on considère tout les modèles 3D locaux à 3 points de vue consécutifs dans une séquence à 11 points de vue. La partition de l'espace définie par le meilleur modèle 3D local est donnée au milieu. A droite, on voit le nombre de modèles locaux satisfaisant la sélection de modèles locaux $\epsilon = 0.1$ avec $\theta_0 = 10^o$ pour un point P donné (c'est à dire le nombre maximal de reconstructions possibles pour P dans le modèle global). On note que ce nombre augmente si P s'éloigne de la trajectoire de la caméra.

5.3 Séquence "vieille ville"

Il faut environ 52 minutes pour prendre les 354 photos de la vieille ville. Quelques images sont montrées dans la figure 6. La longueur de la trajectoire est d'environ $(35 \pm 5cm) \times 353 = 122 \pm 17m$ (la longueur exacte des pas entre deux images successives est inconnue). Les rayons des grands et petits cercles dans les images catadioptriques sont 570 et 102 pixels. Un champ de vue rectangulaire de $\frac{\pi}{4} \times \frac{\pi}{4}$ a environ 260×210 pixels dans les directions horizontales et verticales dans ces conditions.

L'estimation de la géométrie (partie 2) est la première étape. 59859 points 3D sont reconstruits automatiquement avec 380947 points dans les images satisfaisant la géométrie. Une vue de dessus du résultat est proposée dans la figure 6. L'erreur finale RMS est de 0.74 pixels, et on considère qu'un point 2D ne respecte pas la géométrie si l'erreur de reprojection associée est supérieure à 2 pixels. Tous les calculs sont effectués avec une calibration imprécise ($\alpha_{up} = 40, \alpha_{down} = 140$ degrés) pour tester la robustesse de la méthode et le résultat du raffinement de calibration. La distorsion radiale $r(\alpha)$ estimée est très proche d'une fonction affine (ce que l'on attendait) et les angles de champ de vue sont améliorés : $\alpha_{up} = 35.5, \alpha_{down} = 152.2$ degrés.

La seconde étape est le calcul de tous les modèles 3D lo-

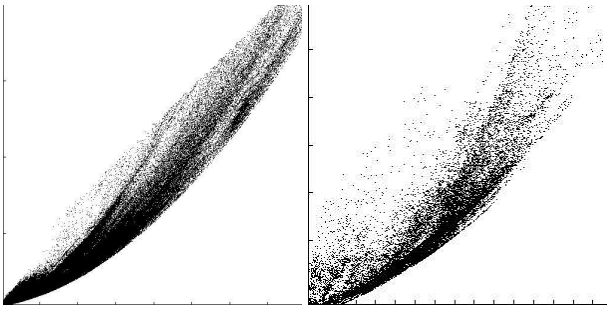

FIG. 5 – Résultats quantitatifs pour l'incertitude des modèles globaux ("vieille ville" à gauche et "petite salle" à droite). Chaque morceau a des coordonnées $(d_l(P), U_l(P))$ avec P le point représentatif du morceau, $d_l(P)$ et $U_l(P)$ la profondeur moyenne et l'incertitude ($p = 90\%$) de P par le modèle 3D local l qui reconstruit P. Les intervalles (cm) sont $[103, 890] \times [0.63, 39.9]$ (à gauche) et $[55.6, 212] \times [0.72, 7.56]$ (à droite), et excluent 0.5% des morceaux.

caux (partie 3). Un modèle local est construit pour chaque triplé d'images successives dans la séquence, et comporte en moyenne 10000 morceaux/patches qui partitionnent l'image de référence. La figure 4 montre la carte de profondeur et l'image de référence d'un des modèles 3D locaux. Enfin, le modèle 3D global est obtenu par sélection des morceaux de tout les modèles 3D locaux en utilisant la sélection des modèles 3D locaux (partie 4.2) et la condition de fiabilité (partie 4.3). Un morceau de modèle 3D local est accepté si son point 3D représentatif (introduit dans la partie 4.4) vérifie (1) la condition de fiabilité angulaire avec $\theta_0 = 5^o$ et (2) la sélection de modèles 3D locaux avec $\epsilon = 0.1$. Seulement 24% de tous les morceaux (819767 patches) sont retenus dans le modèle global.

La figure 7 montre trois vues de ce modèle 3D. Il n'est pas difficile pour le lecteur de mettre en correspondance la vue globale oblique du modèle 3D (en haut de la figure 7) et la vue de dessus de l'estimation de la géométrie (figure 6). De plus, la vidéo jointe (accessible sur ma page web) est composée de deux parties. La première partie montre toutes les images originales. Des images panoramiques sont aussi données pour faciliter la compréhension de la scène, mais elles ne sont pas utilisées par la méthode. On note un changement d'éclairage majeur sur l'image numéro 295. La seconde partie est une visite virtuelle de la scène complète dans un voisinage des points de vue originaux. Les résultats sont acceptables dans les zones texturées en dépit de la faible résolution (environ 210-260 pixels pour un angle dans l'espace de $\frac{\pi}{4}$). Actuellement, le principal problème est que plusieurs parties de murs ou du sol ne sont pas assez texturées pour être appariées et reconstruites. Un second problème est causé par les morceaux qui ont tous une forme (un carré de 8×8 pixels dans les images) et une position régulière dans les images. Les approximations des contours d'occultation qui en résultent sont parfois très grossières.

Des résultats quantitatifs pour les incertitudes à 90% des points sont donnés dans la figure 5 pour le modèle global. Dans $99,5\%$ des cas, l'incertitude croit de 0.6 cm a 40 cm lorsque la profondeur des points croit de 1 m à 9 m.

5.4 Séquence "petite salle"

La séquence "petite salle" est composée de 18 images d'intérieur, telles que la longueur du pas entre chaque prise de vue est d'environ 15 cm (la valeur exacte est inconnue). Les résultats sont : 2974 points 3D reconstruits, 15096 points dans les images satisfaisant la géométrie avec RMS=0.73 pixels, $\alpha_{up} = 37.8$ et $\alpha_{down} = 152.7$ degrés (avec $\alpha_{up} = 40$ et $\alpha_{down} = 140$ degrés en valeurs initiales).

Seulement 26% des morceaux (35709 morceaux) sont retenus dans le modèle global après la sélection des modèles 3D locaux ($\epsilon = 0.1$) et le test de fiabilité ($\theta_0 = 5^o$). La figure 5 donne des incertitudes quantitatives pour le modèle 3D global et la figure 8 montre plusieurs vues du modèle global. Les principaux objets sont faciles à reconnaître. Comme cela a été dit pour la séquence "vieille ville", plusieurs parties de la salle ne sont pas reconstruites à cause du manque de texture et la forme des morceaux n'est pas adaptée aux contours d'occultation.

On note qu'un seul modèle 3D local au centre de la séquence ne suffit pas à reconstruire tout les objets, et que plus de modèles 3D locaux sont bienvenus pour réduire les incertitudes.

6 Conclusion

Une méthode de modélisation 3D automatique à partir d'une séquence d'images catadioptrique est proposée. D'abord, des photos sont prises avec une caméra catadioptrique approximativement équiangulaire. Deuxièmement, la géométrie de la séquence (incluant quelques paramètres intrinsèques) est estimée avec un modèle de caméra central. Troisièmement, plusieurs modèles 3D locaux (sur trois vues) sont reconstruits tout le long de la séquence. La reprojection des images sur un cube virtuel (au lieu d'un cylindre) permet l'utilisation de méthodes de stéréo dense classiques et facilite la reconstruction du sol. Enfin, un modèle 3D global est obtenu en appliquant une sélection sur les modèles 3D locaux : chaque modèle 3D local est partitionné en petits morceaux, et un morceau est rejeté si un autre modèle local est disponible pour reconstruire le morceau avec moins d'incertitude. De tels modèles 3D n'avaient pas été obtenus jusqu'à présent avec une caméra catadioptrique, car la mise en correspondance est toujours difficile en pratique et la dernière étape a été largement ignorée dans les travaux précédents.

Plusieurs améliorations sont possibles et incluent : une meilleure utilisation de la visibilité dans la sélection des modèles 3D locaux, l'amélioration de la mise en correspondance dans les zones faiblement texturées, et plus d'investigations sur le choix des modèles locaux à calculer (réduire leur nombre, intégrer plusieurs baselines pour une même image de référence).

Références

[1] R. Bunschoten and B. Krose. Robust scene reconstruction from an omnidirectional vision system. *IEEE Transactions on Robotics and Automation*, pages 351–357, 2003.

[2] B. Curless and M. Levoy. A volumetric method for building complex models from range images. *SIGGRAPH*, 30, 1996.

[3] K. Daniilidis. The page of omnidirectional vision. www.cis.upenn.edu/˜ kostas/omni.html.

[4] P. Doubek and T. Svoboda. Reliable 3d reconstruction from a few catadioptric images. In *OMNIVIS'02*.

[5] S. Fleck, F. Busch, P. Biber, W. Strasser, and H. Andreasson. Omnidirectional 3d modeling on a mobile robot using graph cuts. In *IEEE ICRA'05*.

[6] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. CUP, 2000.

[7] S. Kang and R. Szeliski. 3-d scene data recovery using omnidirectional multibaseline stereo. *IJCV*, 25(2), 1997.

[8] V. Kolmogorov and R. Zabih. Computing visual correspondances with occlusions using graph-cuts. In *ICCV'01*.

[9] M. Lhuillier and L. Quan. Match propagation for image-based modeling and rendering. *IEEE PAMI*, 24(8), 2002.

[10] M. Lhuillier and L. Quan. A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE PAMI*, 27(3), 2005.

[11] M. Lhuillier. Automatic structure and motion using a catadioptric system. *CVIU*, doi : 10.1016/j.cviu2007.05.004 (à paraitre).

[12] M. Lhuillier. Toward Flexible 3D Modeling using a Catadioptric Camera. In *CVPR'07*.

[13] L. McMillan and G. Bishop. Plenoptic modelling : an image-based rendering system. In *SIGGRAPH'95*.

[14] B. Micusik and T. Pajdla. Structure from motion with wide circular field of view cameras. *IEEE PAMI*, 28(7), 2006.

[15] D. Nister. *Automatic Dense Reconstruction from Uncalibrated Video Sequence.* PhD thesis, Royal Institute of Technology KTH, Stockholm, Sweden, 2001.

[16] M. Okutomi and T. Kanade. A multiple-baseline stereo. *IEEE PAMI*, 15(4), 1993.

[17] M. Pollefeys, L. V. Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch. Visual modeling with a hand-held camera. *IJCV*, 59(3), 2004.

[18] M. Soucy and D. Laurendeau. A general surface approach to the integration of a set of range views. *IEEE PAMI*, 17(4), 1995.

[19] D. Strelow, J. Mischler, S. Singh, and H. Herman. Extending shape-from-motion estimation to noncentral omnidirectional camera. In *IEEE/RSJ IROS'01*.

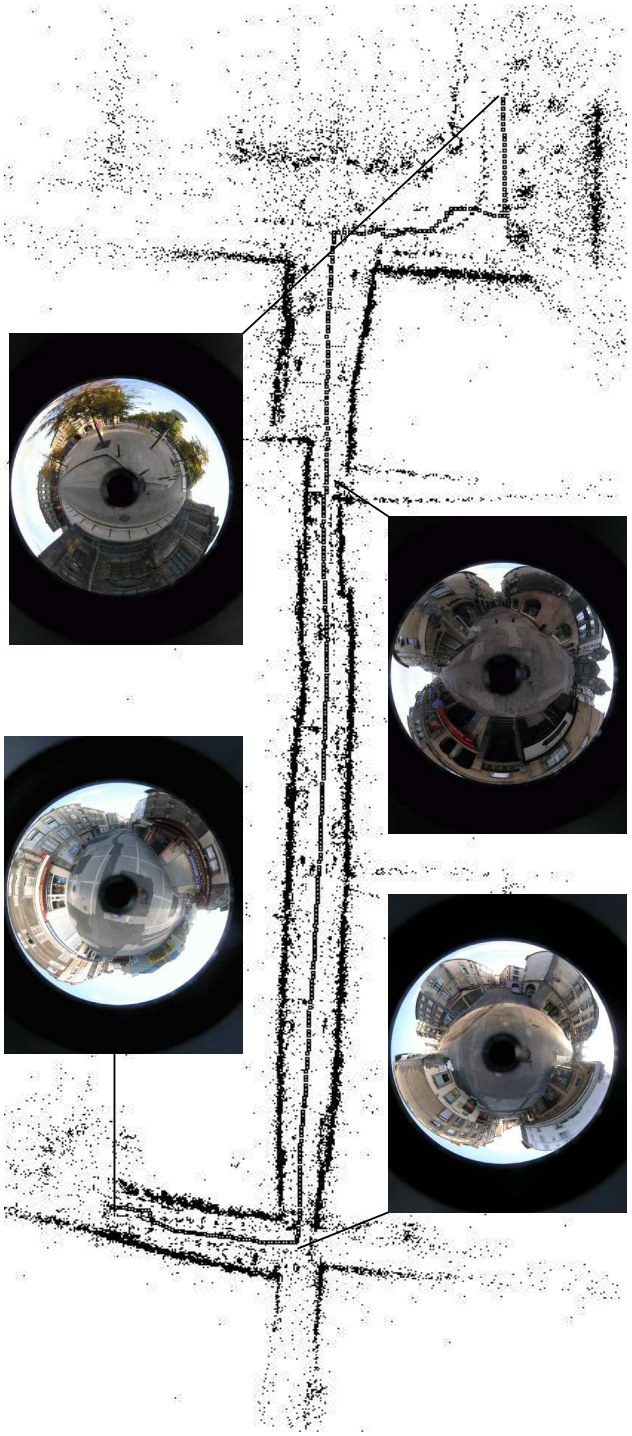

FIG. 6 – Quelques images de la séquence "vieille ville" et une vue de dessus de la géométrie estimée, incluant 354 poses de la caméra (carrés noirs) et 59859 points 3D (points noirs).

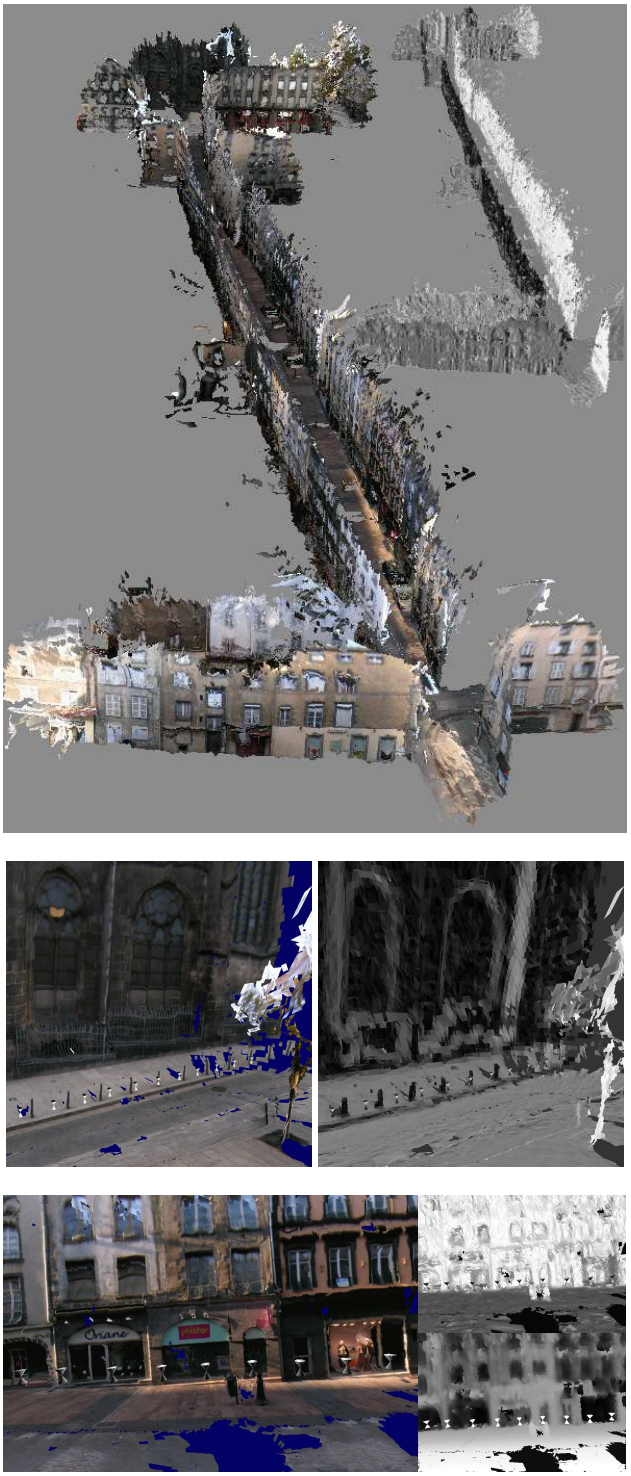

FIG. 7 – Trois vues de la “vieille ville” avec placage de texture (incluant les poses de la caméra), obtenue par sélection des modèles 3D locaux et la condition de fiabilité. Les orientations des morceaux/patches (ou cartes de profondeur) sont aussi donnés en niveaux de gris.

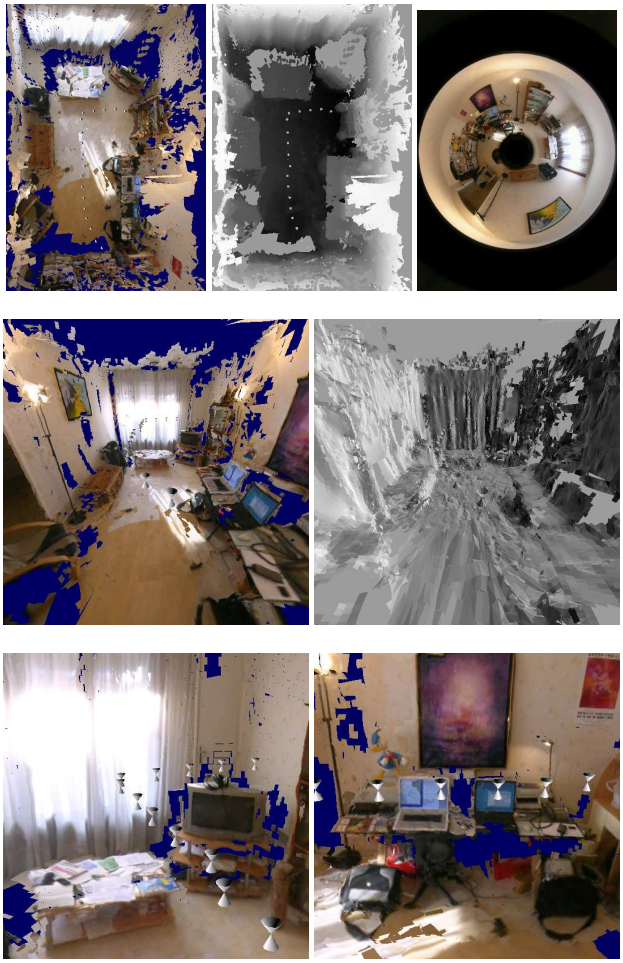

FIG. 8 – Quatre vues (avec placage de texture) du modèle 3D global de la “petite salle”, obtenu avec la sélection de modèles 3D locaux et la condition de fiabilité. Une carte de profondeur et les orientations de morceaux/patches sont donnés en haut et au milieu. Une image de la séquence est aussi montrée.