De la perception visuelle à la reconnaissance des formes, AFCET, juin 1995

Localisation de formes simples par vision active

François Chaumette, Eric Marchand

IRISA / INRIA Rennes
Campus Universitaire de Beaulieu - 35042 Rennes-cedex
E-mail : chaumett@irisa.irisa.fr

Résumé

Cet article traite de la reconstruction tridimensionnelle de scènes constituées d'objets simples (droites, polygones, cylindres) à l'aide des informations visuelles fournies par une caméra mobile. Le principe de vision active utilisé ici consiste à focaliser successivement la caméra sur les différents éléments de la scène et à controler en temps réel et par asservissement visuel le mouvement de la caméra afin d'obtenir une localisation 3D optimale. Dans le but d'assurer la complétude de la zone reconstruite, des algorithmes d'exploration locale et globale sont également présentés.

1 Introduction

Depuis quelques années, de nombreux travaux menés en vision artificielle se sont fixés pour objectif la réalisation de systèmes capables d'accéder à la géométrie spatiale d'une scène à partir de son observation par une ou plusieurs caméras mobiles [1][6][9][15][18]. Ces systèmes ont pour objectif de fournir une description géométrique claire et complète de la scène à partir d'une séquence d'images souvent bruitées et difficilement exploitables. L'étude proposée ici tente d'apporter sa contribution au problème de la reconstruction d'environnements assez restreints (objets statiques, hypothèses sur la nature des objets constituant la scène,...) en utilisant le concept de la vision active [2][3]. La vision active consiste à élaborer des stratégies de contrôle des paramètres de la caméra (position, vitesse,...) de manière à améliorer la connaissance de l'environnement.

Dans cet article, nous utilisons la vision active à deux niveaux différents : un **niveau local** où les mouvements de la caméra sont contraints de manière à optimiser la qualité des résultats de reconstruction d'une primitive 3D, et un **niveau global** pour explorer les zones de la scène non encore observées. Plus précisément, le niveau local consiste à contrôler par asservissement visuel les mouvements de la caméra de manière à obtenir une estimation précise et robuste de primitives géométriques paramétrables telles les points, les droites, les cylindres, les sphères, etc. Cependant, cette approche ne permet de reconstruire qu'une seule primitive à la fois et ne permet pas de s'assurer de la complétude de la zone reconstruite. De plus, une connaissance *a priori* sur la nature de la primitive à reconstruire est nécessaire afin de générer les mouvements adéquats de la caméra. Il faut donc s'abstraire de ces contraintes en définissant des stratégies de perception qui permettent l'acquisition d'une carte précise et complète de la scène. Ce problème d'exploration correspond à un **niveau global** dans le cadre de la perception active. Plusieurs approches ont été proposées afin de résoudre le problème du calcul automatique de point de vue [7] [13] [14] [16] [17]. Dans notre cas, nous devons obtenir une carte de l'environnement sans connaissance *a priori* sur le nombre, la position et la dimension des objets de la scène (composée par hypothèse de cylindres, de polygones et de segments).

L'article est structuré de la manière suivante : la première partie traite du problème de l'estimation des paramètres 3D d'une primitive géométrique basée sur l'utilisation d'une caméra mobile commandable. La seconde partie est dédiée aux aspects globaux de notre processus de reconstruction et présente des stratégies autonomes d'exploration.

2 Reconstruction de primitives 3D

La méthode utilisée pour reconstruire les primitives 3D est détaillée dans [4] et [5]. Elle permet d'obtenir une estimation précise et robuste des paramètres 3D d'une primitive géométrique à partir de l'analyse d'une séquence d'images acquises par une caméra en mouvement.

2.1 Reconstruction 3D par vision dynamique

La caméra est modélisée de manière classique par une projection perspective de focale égale à 1. Un point m de coordonnées $\underline{x} = (x, y, z)^T$ se projette donc en M de coordonnées $\underline{X} = (X, Y, 1)^T$ avec :

$$\underline{X} = \frac{1}{z}\,\underline{x} \tag{1}$$

Soit $\mathcal{P}_s$ une primitive géométrique paramétrable décrite par une équation de la forme :

$$h(\underline{x}, \underline{p}) = 0, \forall \underline{x} \in \mathcal{P}_s \tag{2}$$

où h définit la nature de la primitive et $\underline{p}$ sa configuration. Le but de la reconstruction est donc d'estimer la valeur des paramètres $\underline{p}$ afin de pouvoir reconstruire et localiser la primitive h. A partir de cette équation paramétrique, on peut définir en utilisant (1) les deux fonctions suivantes [8]:

$$\begin{cases} g(\underline{X}, \underline{P}) = 0\ , \forall \underline{X} \in \mathcal{P}_i \\ 1/z = \mu(\underline{X}, \underline{p}_0) \end{cases} \tag{3}$$

où $\mathcal{P}_i$ représente la projection dans le plan image de la primitive $\mathcal{P}_s$. Pour des primitives planes, la fonction μ représente le plan dans lequel la primitive se situe. Dans le cas plus général de primitives volumétriques (voir Figure 1), la fonction g représente la projection dans l'image des limbes de la primitive et la fonction μ définit la surface à laquelle appartiennent les limbes.

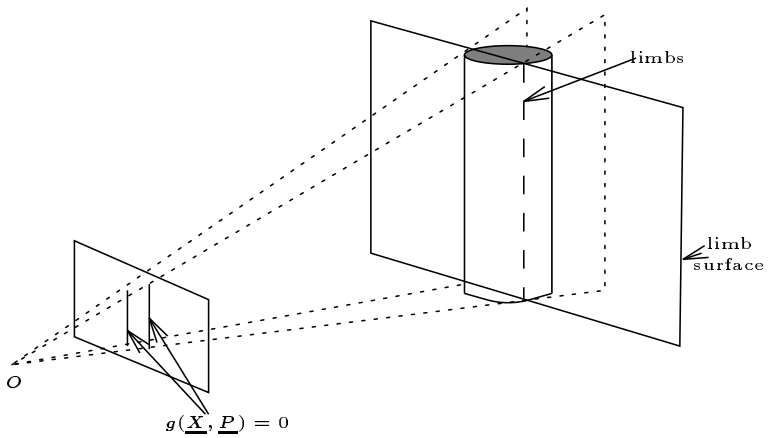


FIG. 1 - *Projection de la primitive dans l'image (g) et surface des limbes (μ) dans la cas d'un cylindre*

Soit $T_c = (V(O), \Omega)^T$ le torseur cinématique de la caméra où $V(O)$ représente la vitesse de translation de la caméra et Ω sa vitesse de rotation. La variation de $\underline{P}$ qui relie le mouvement apparent de la primitive dans l'image au mouvement de la caméra T_c peut être calculée explicitement et s'exprime par [8]:

$$\dot{\underline{P}} = L_{\underline{P}}^T(\underline{P}, \underline{p}_0)\, T_c \tag{4}$$

où $L^T_{\underline{P}}(\underline{P}, \underline{p}_0)$ représente la matrice d'interaction qui caractérise les interactions entre le capteur et la primitive considérée.

L'estimation des paramètres $\underline{p}$ s'effectue en deux étapes: tout d'abord, les paramètres $\underline{p}_0$ caractérisant la surface des limbes sont obtenus à partir de l'équation (4) en utilisant la mesure de $T_c, \underline{P}$ et $\dot{\underline{P}}$: $\underline{p}_0 = \underline{p}_0(T_c, \underline{P}, \dot{\underline{P}})$. Ensuite, les paramètres $\underline{p}$ sont déterminés par l'intersection de la surface des limbes avec le cône de sommet O et de génératrice $g(\underline{X}, \underline{P}) = 0$: $\underline{p} = \underline{p}(\underline{p}_0, \underline{P})$.

2.2 Reconstruction 3D par vision active

Les résultats obtenus en utilisant la méthode décrite précédemment sont généralement assez médiocres. En effet, la qualité de l'estimation est très sensible à la nature des mouvements de la caméra. Une solution efficace pour résoudre ce problème consiste à utiliser le formalisme de la vision active. En effet, on peut montrer [4] [5] que pour obtenir une estimation non biaisée, il suffit de contraindre les mouvements de la caméra de sorte que $\dot{\underline{P}} = \dot{\underline{p}}_0 = 0$. Cette contrainte signifie que la caméra doit effectuer une tâche de **fixation** : l'image et la surface des limbes de la primitive doivent rester immobiles le long de la trajectoire effectuée par la caméra. De plus, les effets des erreurs de mesure sur l'estimation dépendent fortement de la position de la primitive dans l'image. C'est pourquoi le mouvement de la caméra doit être contraint de manière à minimiser les effets de ces erreurs de mesure. Afin d'assurer cette minimisation, la primitive doit rester immobile à des positions particulières dans l'image (tâche de **focalisation**). Si l'on considère le cas d'un point par exemple, celui-ci doit apparaître en permanence centré dans l'image, *i.e.* $X = Y = 0, \forall t$.

On peut signaler que les techniques d'asservissement visuel [8] sont parfaitement adaptées pour assurer les contraintes nécessaires à une estimation optimale en réalisant les tâches de fixation et de focalisation. L'asservissement visuel consiste en effet à introduire directement et en boucle fermée les informations extraites de l'image dans une boucle de commande calculant la vitesse adéquate de la caméra.

2.3 Résultats : reconstruction optimale dans le cas d'un cylindre

Les expérimentations présentées dans cet article ont été réalisées sur la cellule de vision robotique de l'IRISA composée d'une caméra CCD montée sur l'effecteur d'un robot à 6 degrés de liberté. Dans un premier temps, nous avons utilisé la méthode présentée pour estimer les paramètres d'un cylindre (cf figure 2.a). De manière à obtenir une estimée robuste et non biaisée, le cylindre doit apparaître centré et vertical ou horizontal dans l'image pendant le processus de reconstruction (cf figure 2.b). Les techniques d'asservissement visuel ont donc été employées pour réaliser cette tâche en temps réel : une itération de la boucle de commande et une estimation sont simultanément effectuées en 100 ms. La figure 3 représente l'erreur entre la valeur estimée du rayon et sa valeur réelle pour chaque itération effectuée. La moyenne obtenue est inférieure à 0.1 mm avec un écart-type inférieur à 0.2 mm, ce qui montre la robustesse, la précision et la stabilité de l'algorithme de reconstruction.

(a)  (b)

FIG. 2 - *Cylindre à reconstruire avant et après la tâche de focalisation*

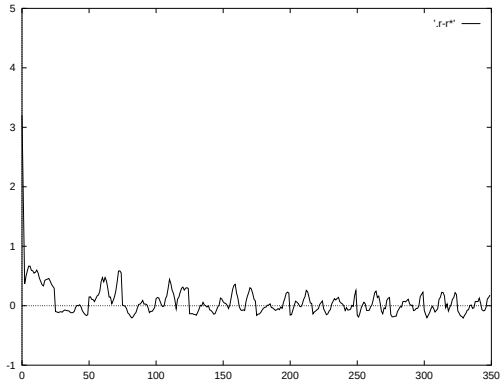


FIG. 3 - *Erreurs successives entre la valeur estimée du rayon et sa valeur réelle (en mm)*

3 Stratégies de perception

Le problème qui nous intéresse maintenant est celui de la reconstruction complète d'une scène contenant plusieurs objets. En effet, l'approche retenue ne permet de reconstruire qu'une seule primitive à la fois. L'objectif de cette partie est donc de définir des stratégies de perception permettant une représentation 3D précise et complète de la zone à reconstruire. De manière schématique, l'approche utilisée consiste à sélectionner automatiquement les informations pertinentes contenues dans les bases de données 2D que l'on peut construire, et à focaliser successivement la caméra sur les différents objets de la scène. Des phases d'exploration sont également nécessaires afin d'assurer la complétude de la reconstruction.

3.1 Reconnaissance de primitives

La seule hypothèse effectuée sur la scène, outre les dimensions d'un volume l'englobant, porte sur le fait qu'elle est constituée uniquement de segments, de polygones et de cylindres. Or, la méthode de reconstruction présentée ci-dessus implique une connaissance *a priori* sur la nature de la primitive observée (segment ou cylindre). Nous avons donc développé une méthode de reconnaissance des primitives basée sur un test statististique au maximum de vraisemblance [11].

3.2 Bases de données

Une analyse des images acquises par la caméra permet la création de bases de données 2D contenant la projection des objets de la scène dans l'image (dans notre cas, la liste des segments contenus dans une image). Notons ces bases de données ω_{ϕ_t}, où ϕ_t représente la position de la caméra. Une autre base de données, notée Ω_Φ, est également utilisée. Ω_Φ contient tous les segments qui n'ont pas encore fait l'objet d'une reconstruction, ainsi que la position de la caméra depuis laquelle ils ont été observés

3.3 Modélisation incrémentale de l'espace libre pour l'exploration de scène

Depuis un point de vue ϕ_t, il est possible de calculer la zone observée $V(\phi_t)$ en utilisant les informations 3D déjà recueillies. Notons $\mathcal{V}(\Phi)$ la zone de l'espace observé par la caméra depuis le début du processus de reconstruction (*i.e.* les primitives 3D et l'espace libre connus). On a:

$$\mathcal{V}(\Phi) = \bigcup_{i=1}^{t} V(\phi_t)$$

En utilisant les informations visuelles (*i.e.* les bases de données ω_ϕ et Ω_Φ) et la carte partielle de l'environnement $\mathcal{V}(\Phi)$, nous pouvons définir des stratégies de positionnement de la caméra qui assureront une reconstruction complète de la scène. Ce processus d'exploration est constitué de deux niveaux distincts :

- une **exploration locale** est réalisée quand un segment correspondant à une nouvelle primitive apparaît dans le champ de vision de la caméra ou a été précédemment observé depuis une autre position de la caméra. Dans ce cas, un calcul explicite de point de vue n'est pas nécessaire.
- Par contre, dans les autres cas, quand tous les segments précédemment observés ont été reconstruits, une stratégie plus complexe doit être mise en œuvre de manière à se focaliser sur des zones de la scène n'ayant pas encore été observées. Nous parlerons alors d'**exploration globale**.

3.4 Exploration locale

Afin de minimiser le déplacement de la caméra, nous avons mis en œuvre une stratégie *locale* utilisant de façon explicite la connaissance courante sur la scène [12]. La démarche s'appuie sur les hypothèses suivantes : la scène est constituée d'objets 3D liés par des relations topologiques. La projection d'un objet 3D dans l'image peut être représentée par un graphe où les nœuds sont les jonctions multiples et les arcs les contours. Chaque arc (segment 2D) de ce graphe est valué en fonction de sa position dans l'image et de la connaissance éventuelle que l'on a sur la primitive 3D correspondante. Si plusieurs arcs de ce graphe se révèlent être la projection d'objets non reconstruits, une sélection minimisant le déplacement de la caméra est effectuée pour déterminer le prochain segment à traiter. Dans le cas où tous les segments d'une image correspondent à des primitives déjà reconstruites, nous recherchons, dans la base de données globale Ω_Φ, un segment non traité. La caméra se déplace jusqu'à la position depuis laquelle il avait été observé (phase de retour arrière), et une reconstruction de ce segment est effectuée. Si cette base de donnée Ω_Φ est vide, une exploration globale est alors nécessaire.

La stratégie *locale* que nous avons développée assure une reconstruction efficace de toute primitive ayant été observée lors du processus de reconstruction. Elle ne fait pas appel à un calcul explicite de nouveaux points de vue et minimise localement le déplacement de la caméra. Cependant, elle ne donne pas l'assurance d'une reconstruction complète de la scène. Pour résoudre ce problème, des stratégies *globales* doivent être mises en œuvre. Quand toutes les primitives observées pendant les phases d'exploration locale ont été reconstruites, de nouveaux points de vue doivent être calculés de manière à obtenir un maximum d'informations suplémentaires sur la scène.

3.5 Exploration globale : calcul de points de vue

Nous devons déterminer les points de vue de la caméra permettant d'amener de nouvelles primitives dans son champ de vision. De tels points de vue sont calculés en utilisant la connaissance courante sur la géométrie spatiale de la scène et une représentation des zones de l'espace déjà observées. En utilisant ces connaissances, nous avons défini une fonction d'énergie à minimiser $\mathcal{F}(\phi_{t+1})$ qui intégre le gain $\mathcal{G}(\phi_{t+1})$ attendu de la nouvelle position (ϕ_{t+1}) (voir figure 4), le coût de déplacement $\mathcal{C}(\phi_t, \phi_{t+1})$ depuis la position courante (ϕ_t) et l'évitement $\mathcal{B}(\phi_{t+1})$ des butées articulaires du robot porteur de la caméra.

La fonction d'énergie $\mathcal{F}(\phi_{t+1})$ est définie par la somme pondérée des ces fonctions à valeur dans $\mathbb{R}^+$:

$$\mathcal{F}(\phi_{t+1}) = \alpha_1 \mathcal{B}(\phi) + \alpha_2 \mathcal{G}(\phi_{t+1}) + \alpha_3 \mathcal{C}(\phi_t, \phi_{t+1}) \tag{5}$$

La détermination des coefficients α_i dans un problème d'optimisation de ce type est un problème non trivial [10]. Nous nous sommes contentés de choisir ces coefficients de manière empirique. Cependant,

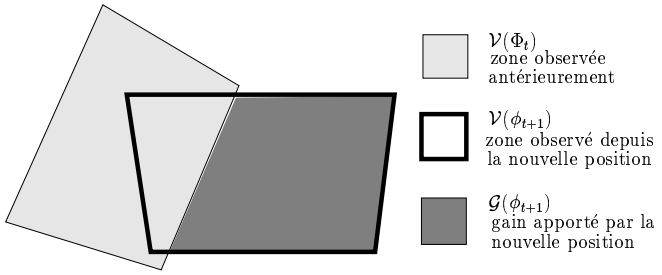


FIG. 4 - *Mesure du gain apporté par une nouvelle position*

leur valeur fixe l'ordre de priorité associé à chacun des critères. L'accessibilité de la nouvelle position étant bien sûr prioritaire et la découverte de nouvelle zone à explorer étant notre objectif, nous avons choisi $\alpha_1 > \alpha_2 > \alpha_3$.

Chaque position $\phi \in SE_3$ peut *a priori* être solution de ce problème d'optimisation. Cependant, de manière à contraindre le problème, nous autorisons la caméra à se déplacer sur la surface d'une sphère englobant la scène. La position de la caméra peut alors être décrite par un vecteur à 5 paramètres $(\theta, \varphi, \Omega_x, \Omega_y, \Omega_z)$ où θ et φ représentent la latitude et la longitude de la caméra sur la sphère et où Ω_x, Ω_y et Ω_z représentent l'orientation de la caméra.

Pour minimiser $\mathcal{F}(\phi)$, nous avons choisi d'utiliser une méthode déterministe classique de type gradient conjugués avec un pas de descente à deux niveaux : nous utilisons tout d'abord des incréments importants afin de déterminer la région de l'espace des paramètres où l'optimum de la fonction $\mathcal{F}(\phi)$ est probablement situé. Puis, nous itérons le processus depuis cette nouvelle position avec un incrément plus faible. Contrairement aux méthodes stochastiques de type recuit simulé, nous ne pouvons assurer que la convergence s'effectue vers le minimum global de la fonction. Cependant, le gain en temps de calcul est très important et les expériences ont montré qu'un optimum correct est toujours atteint en un faible nombre d'itérations. De plus, l'intérêt de trouver un minimum global de la fonction d'énergie ne nous a pas paru fondamental dans la mesure où une position apportant un complément d'informations important est trouvée.

3.6 Résultats : exploration de la scène

Signalons tout d'abord que la mise en œuvre des différents algorithmes présentés dans cet article a été réalisée à l'aide d'un automate hiérarchique parallèle (voir figure 5) capable de gérer et d'enchaîner l'ensemble des étapes nécessaires à la reconstruction d'une scène. Le formalisme des systèmes à évènements discrets que nous avons employé autorise le parallèlisme, la préemption et le séquencement de tâches. Cet automate sélectionne et gère les actions à effectuer en fonction des événements 2D perçus dans l'image, des connaissances acquises et répertoriées au fur et à mesure de la reconstruction, ainsi que de la détection de la fin d'une tâche d'asservissement visuel [12].

La scène considérée est composée d'un cylindre et de plusieurs polygones disposés dans des plans différents. La figure 6 montre une vue extérieure de la scène et des différent objets qui la composent.

Exploration locale La figure 7 représente les images acquises avant chaque reconstruction optimale. À chacune de ces images est associée la base de données 2D correspondante. Les lignes noires correspondent aux éléments de la base de données qui n'ont pas encore été traités et les lignes pointillées représentent les segments correspondant à des primitives 3D déjà reconstruites.

La figure 7.a montre l'image acquise depuis la position ϕ_0 de la caméra. Aucune reconstruction n'a encore été effectuée. Notons que la scène complète n'est pas visible depuis cette position de la caméra

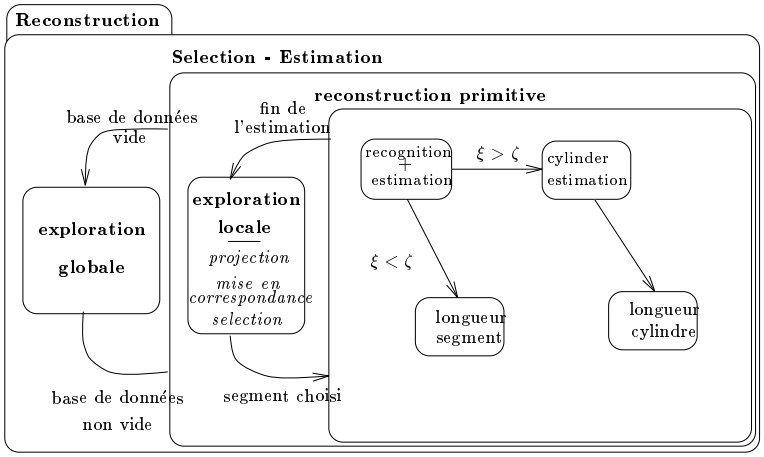


FIG. 5 - *Automate hiérarchique parallèle*

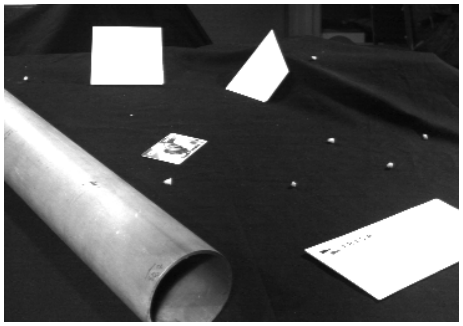
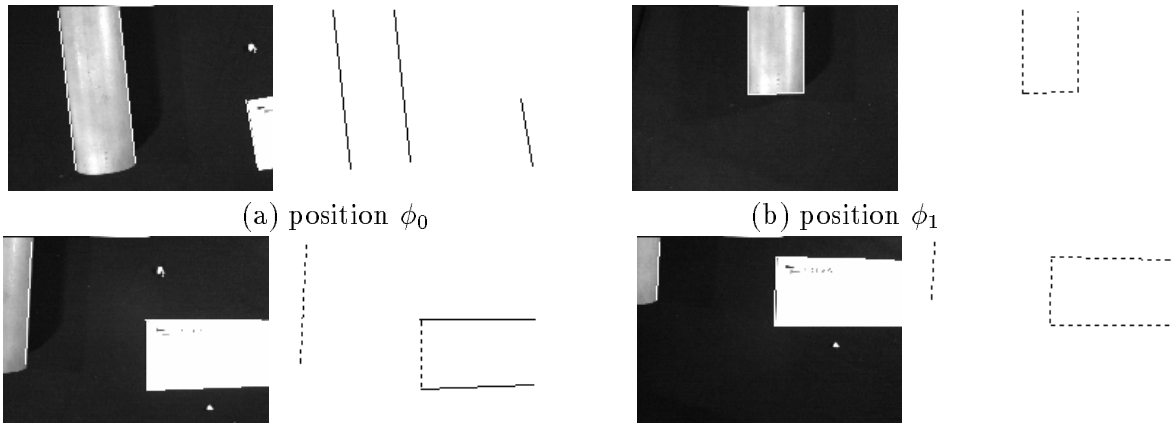


FIG. 6 - *Vue extérieure de la scène*



(a) position ϕ_0 (b) position ϕ_1

(c) position ϕ_2 (d) position ϕ_6

FIG. 7 - *Exploration locale de la scène (images acquises et bases de données)*

puisque trois segments seulement sont visibles depuis cette position. Le premier segment extrait de la base de données ω_{ϕ_0} est celui correspondant au limbe de droite du cylindre. Après la phase de reconnaissance et de reconstruction du cylindre, la caméra est positionnée en ϕ_1 (image 7.b). Les segments de la base de données ω_{ϕ_1} ont tous été reconstruits. Après consultation de la base de données globale Ω_Φ, on constate qu'un segment a été observé depuis la position ϕ_0 et n'a pas encore été traité. La caméra se déplace donc en ϕ_0 et se focalise sur le segment retenu. Après la reconstruction de ce segment, la caméra est positionnée en ϕ_2 (image 7.c). Deux segments correspondant à des primitives non reconstruites apparaissent dans la base de données ω_{ϕ_2}. Le segment le plus proche du centre de l'image est sélectionné et reconstruit. Ce processus est renouvelé jusqu'à ce que toutes les primitives observées pendant cette phase d'exploration locale aient été estimées (image 7.d correspondant à la position ϕ_6 de la caméra). Notons que des primitives qui n'apparaissaient pas dans le champ de vision initial de la caméra ont été découvertes et reconstruites. La scène estimée à l'issue de cette étape est présentée sur la figure 8.

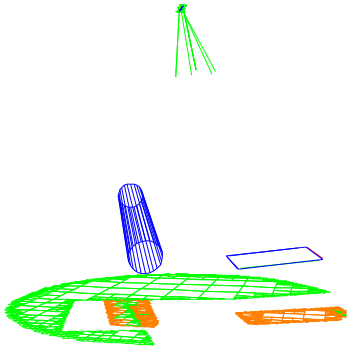


FIG. 8 - *Résultat de l'exploration locale (positions de la caméra, scène reconstruite et projection sur le sol de la zone non observée)*

Exploration globale La figure 9.a montre une visualisation 3D de la scène reconstruite à l'issue des différentes phases d'exploration globale nécessaires à une reconstruction complète. La trajectoire parcourue par la caméra pendant cette exploration est présentée sur la figure 9.b. On peut noter le faible nombre de positions requises ainsi que la continuité de la trajectoire effectuée en raison de la prise en compte dans la fonction d'énergie du coût du déplacement de la caméra.

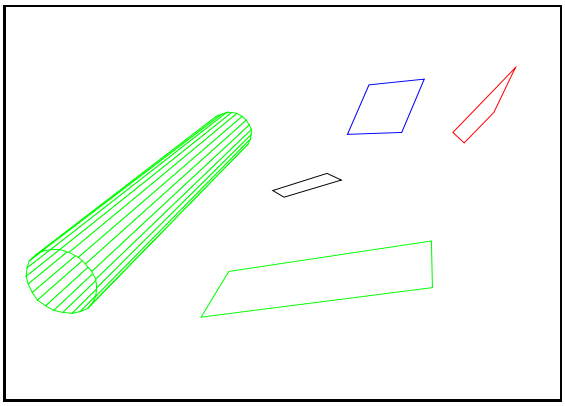
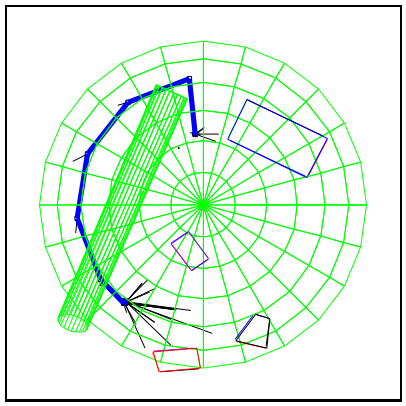
(a)   (b)

FIG. 9 - *Visualisation de la scène reconstruite et trajectoire de la caméra pendant l'exploration globale*

4 Conclusion

Nous avons proposé dans cet article une méthode permettant la reconstruction d'un environnement 3D en utilisant une séquence d'images acquises par une caméra mobile. Nous avons tout d'abord décrit un processus de reconstruction permettant une estimation précise et robuste des paramètres d'une primitive géométrique. Cette méthode étant basée sur des mouvements particuliers de la caméra, des stratégies perceptives permettant d'assurer la complétude de la reconstruction par une succession de reconstructions indépendantes ont été développées. Des expérimentations menées sur une cellule robotique ont démontré la validité de notre approche (résultats de reconstruction précis, robustes et stables, algorithmes d'exploration de la scène simples mais efficaces) mais aussi ses limitations : les contraintes fortes sur les mouvements de la caméra impliquent un séquencement fort des tâches de reconstruction et n'autorisent pas actuellement la recontruction précise de plusieurs primitives en parallèle.

Références

[1] G. Adiv : Inherent ambiguities in recovering 3D motion and structure from a noisy flow field. *IEEE Trans. on PAMI*, 11(5):477–489, Mai 1989.

[2] Y. Aloimonos: Purposive and qualitative active vision. *ICPR*, pp 346–360, New Jersey, 1990.

[3] R. Bajcsy : Active perception. *Proc. IEEE*, 76(8):996–1005, Août 1988.

[4] S. Boukir : Reconstruction 3D d'un environnement statique par vision active. *Thèse de l'Université de Rennes I*, IRISA, Octobre 1993.

[5] F. Chaumette, S. Boukir, P. Bouthemy, D. Juvin : Optimal estimation of 3D structures using visual servoing. *CVPR*, pp 347–354, Seattle, Juillet 1994.

[6] C. Chien, J.K. Aggarwal : Model construction and shape recognition from occluding contour. *IEEE Trans. on PAMI*, 11(4):372–389, Février 1989.

[7] C. Connoly : The determination of next best views. *IEEE Int. Conf. on Robotics and Automation*, pp 432–435, St Louis, Missouri, Mars 1985.

[8] B. Espiau, F. Chaumette, P. Rives : A new approach to visual servoing in robotics. *IEEE Trans. on Robotics and Automation*, 8(3):313–326, Juin 1992.

[9] O. Faugeras : Three-dimensionnal computer vision: a geometric viewpoint. *MIT press*, 1993.

[10] M.A. Gennert, A.L. Yuille : Determining the optimal weights in multiple objective function optimization. *ICCV*, pp 87–94, 1988.

[11] E. Marchand, F. Chaumette : Real time estimation of 3D environment with an active vision system. *Int. Workshop on Intelligent Robotic Systems*, pp 311–318, Grenoble, Juillet 1994.

[12] E. Marchand, F. Chaumette : Active visual 3D perception. *IEEE Workshop on Vision for Robots*, Pittsburgh, USA, Août 1995.

[13] J. Maver, R. Bajcsy : Occlusions as a guide for planning the next view. *IEEE Trans. on PAMI*, 15(5):417–433, Mai 1993.

[14] B. Triggs, C. Laugier : Automatic camera placement for robot vision. *IEEE Int. Conf. on Robotics and Automation*, Nagoya, Japan, 1995.

[15] A.M. Waxman, B.K. Parsi, M. Subbarao : Closed-form solutions to image flow equations for 3D structure and motion. *IJCV*, 1(3):239–258, Octobre 1987.

[16] P. Whaite, F. Ferrie : Autonomous exploration: Driven by uncertainty. *CVPR*, pp 339–346, Seattle, Juillet 1994.

[17] L.E. Wixson : Viewpoint selection for visual search. *CVPR*, pp 800–805, Seattle, Juillet 1994.

[18] M. Xie, P. Rives : Toward dynamic vision. *IEEE Workshop on Interpretation of 3D Scenes*, Austin, Texas, Novembre 1989.