Active Vision for Complete Scene Reconstruction and Exploration

Éric Marchand and François Chaumette, *Member*, *IEEE*

Abstract—This paper deals with the 3D structure estimation and exploration of static scenes using active vision. Our method is based on the structure from controlled motion approach that constrains camera motions to obtain an optimal estimation of the 3D structure of a geometrical primitive. Since this approach involves to gaze on the considered primitive, we have developed perceptual strategies able to perform a succession of robust estimations. This leads to a gaze planning strategy that mainly uses a representation of known and unknown areas as a basis for selecting viewpoints. This approach ensures a reconstruction as complete as possible of the scene.

Index Terms—3D reconstruction, scene exploration, purposive and active vision, perception strategies.

————————   ✦   ————————

1 INTRODUCTION

ONE of the main issues in computer vision is to recover the 3D structure of an unknown world from visual data provided by a calibrated camera mounted on the end effector of a robot arm. The system must provide a clear, accurate and complete three-dimensional geometric description of the scene from incomplete and usually noisy images. Most of the approaches proposed to solve this problem are inspired from the Marr paradigm that considers a static or a mobile sensor, but not a controlled one. Unfortunately, this approach appears to be inadequate to solve many problems where appropriate modifications of intrinsic and/or extrinsic parameters of the sensor are necessary. This is why Aloimonos et al. [2], [1], Bajcsy [4], and Ballard [5] have proposed to modify the Marr concept by introducing the **active vision** paradigm. Since the major shortcomings that limit the performance of vision systems are their sensitivity to noise, their low accuracy, and their lack of reactivity, the aim of active vision is generally to elaborate control strategies in order to improve a perception task. Thus, function of the specified task and of the data extracted from the acquired images, an active vision system might be induced to modify its parameters (position, velocity, ocular parameters such as focus or aperture, etc.), but also the way data are processed (region of interest, peculiar image processing, etc.), and the processing resources allocated to the system. Despite some differences, the goal was to show that an active system is more *relevant* to the application (usually because it is goal driven), more robust (because they can handle either uncertainty and/or dynamic environment), and more accurate (because they are able to modify their own configuration).

Our specific concern is to deal with the problem of recovering the 3D spatial structure of scenes composed of cylinders and polyhedral objects, without any knowledge on their dimensions and their localization inside a bounded region (that is assumed to be known). The system we have developed has three main levels; each can be seen as a different cycle of perception-action.

Exploration. The first issue deals with the exploration. The goal is to observe all the parts of the scene and to ensure the completeness

---

• *The authors are with IRISA/INRIA Rennes, Vista Project, Campus de Beaulieu, 35042 Rennes-cedex, France.*
  E-mail: [marchand, chaumett}@irisa.fr.

of the reconstruction (for all the most, a reconstruction as complete as possible) by determining adequate camera viewpoints. This part of our work can be related to active perception as defined by Bajcsy [4]: The position and the orientation of the camera are set in order to increase the knowledge on the 3D structure of the scene. As far as we are concerned, active vision is used to iteratively determine the location of the camera. Knowledge on 3D data previously gathered, and current 2D information are fed back into the exploration process. This leads to a gaze planning strategy that proposes a solution to the next best view problem. It mainly uses a representation of known and unknown region as a basis for selecting viewpoints. We have chosen to handle the "where to look next" question as a function minimization problem. More precisely, we have defined a function to be minimized that integrates the constraints imposed by the system and evaluates the quality of a viewpoint. When an object is observed from a computed viewpoint, the exploration process stops and an incremental reconstruction is performed. Other works on sensor planning have been proposed using a function minimization approach [19], [20]. However, in each case, an a priori knowledge on the structure of the scene was available.

Primitive Reconstruction and Camera Motion Generation. The approach we have chosen to get an accurate three-dimensional reconstruction of an object is based on a continuous structure from motion approach [7]. Very noticeable improvements are obtained in the 3D reconstruction if the camera viewpoint is properly selected and if adequate camera motions are generated: It has been shown in [7] that the considered primitive must remain static at a given position in the image in order to obtain a robust and unbiased estimation (see Fig. 1). The corresponding camera motions can be generated using the visual servoing approach [10]. This aspect of the reconstruction is related to the purposive vision concept [1]: Only motions useful to get an optimal structure estimation are realized. Furthermore, it confirms the point of view of previous works on active vision [2] and on gaze control [5].

Incremental Reconstruction. Since the camera motion is controlled for the estimation of one primitive at a time, this implies to successively gaze on each primitive of the scene. This process is performed using a simple incremental reconstruction of all the primitives observed by the camera. We also call this step *local exploration* owing to the fact that it only uses locally available information.

The remainder of this paper is organized as follows: In Section 2 is presented the incremental reconstruction algorithm. The exploration strategy that ensures the completeness of the reconstruction is described in Section 3. We finally demonstrate with various real-time experiments that the implemented active vision system allows the autonomous and complete reconstruction of scenes with a very good accuracy.

2   INCREMENTAL SCENE EXPLORATION

As already stated, the scene is assumed to be only composed of polyhedral objects and cylinders, so that the contours of all the objects projected in the image plane form a set of segments. The first step in the scene reconstruction process is to obtain the list of these segments. We denote these lists $\omega_{\phi_t} = \left\{ S_i, i = 1 \ldots M \right\}$, where

ϕ_t is the corresponding camera location and viewing direction from which the M segments S_i are observed. For real-time issue, we cannot create a list at each iteration of the estimation process. So, they are created after each reconstruction, and are used for the selection of the next considered segment.

Another list is used. It contains all the untreated segments previously observed (treated segments are suppressed by back-

projection), and the camera positions ϕ_k from which they have been observed. More precisely, it is defined as:

$$\Omega_{\mathcal{T}_{t_1}^{t_2}} = \left\{ (S_i, \phi_k), i = 1 \ldots N_k, k \in \left[t_1, t_2 \right] \right\}$$

where $\mathcal{T}_{t_1}^{t_2} = \left\{ \phi_{t_1}, \phi_{t_1+1}, \ldots, \phi_{t_2} \right\}$ is the set of all viewpoints between t_1 and t_2, S_i is a 2D segment associated with an unestimated primitive, and ϕ_k the camera location from which it has been observed.

Using these two sets of segments, it is possible to define an incremental reconstruction strategy able to successively consider all the observed segments:

Step 0: Initialization. We consider that the camera is located in ϕ_0 and ω_{ϕ_0} is acquired. We extract from ω_{ϕ_0} a segment S to be estimated.

Step 1: Active 3D estimation and 3D map creation. Let us consider now that the camera is located in ϕ_t. An estimation based on S is performed, including a recognition process [21], [15] (Does this segment correspond to a 3D segment or to a cylinder?), the structure estimation process [7] and the estimation of the primitive length [15]. The obtained parameters (structure and location) of the primitive are introduced into the 3D global map of the scene. We then remove from $\Omega_{\mathcal{T}_0^t}$ all the 2D segments corresponding to this estimated primitive.

Step 2: Segment lists generation. After the active estimation, because of the camera motion implied by this process (see Fig. 1), the camera is located in ϕ_{t+1}. A new local set of segments $\omega_{\phi_{t+1}}$ corresponding to this position is constructed and merged with $\Omega_{\mathcal{T}_0^t}$.

Step 3: Segment selection. Three different cases may occur:

1) In the case where several untreated segments are in the current list $\omega_{\phi_{t+1}}$, a choice is performed in order to select the next segment S. An **active estimation** (step 1) based on this segment is then performed.
2) If all the segments of $\omega_{\phi_{t+1}}$ have been considered and if at least one of the 2D segments previously observed have not been considered (i.e., $\omega_{\phi_{t+1}}$ empty and $\Omega_{\mathcal{T}_0^{t+1}}$ not empty), we look in $\Omega_{\mathcal{T}_0^{t+1}}$ for an untreated segment S (backtracking).

   Then, the camera moves to the position ϕ_k from which it had been observed (thus, $\phi_{t+1} := \phi_k$). An **active estimation** (step 1) is then performed.
3) Finally, if $\Omega_{\mathcal{T}_0^{t+1}}$ is empty (i.e., all the 2D segments observed from any previous camera positions have been considered), a new viewpoint must be found. A **global exploration**, which is described in the next section, is thus necessary.

Since this exploration strategy is local, it avoids computing explicitly new viewpoints. Furthermore, this algorithm can treat the composition of simple primitives such as polygons [15]. However, more complex combinations raise new problems: an object can be occluded by another one (or by itself). Finally, some objects may not have been observed from the different viewpoints. Exploration
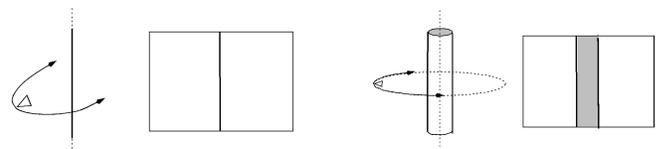

Fig. 1. Optimal camera motion and resulting image in the cases of a straight line and a cylinder.
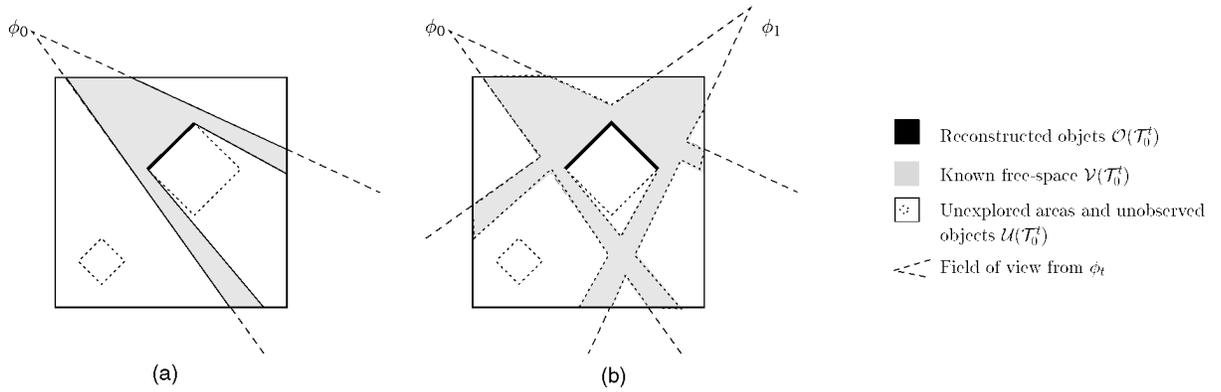
Fig. 2. The scene representation: Reconstructed objects $O\left(\mathcal{T}_0^t\right)$ and free space $\mathcal{V}\left(\mathcal{T}_0^t\right)$ for the discrete trajectory (a) $\mathcal{T}_0^0 = \left\{\phi_0\right\}$ and for (b) $\mathcal{T}_0^1 = \left\{\phi_0, \phi_1\right\}$.
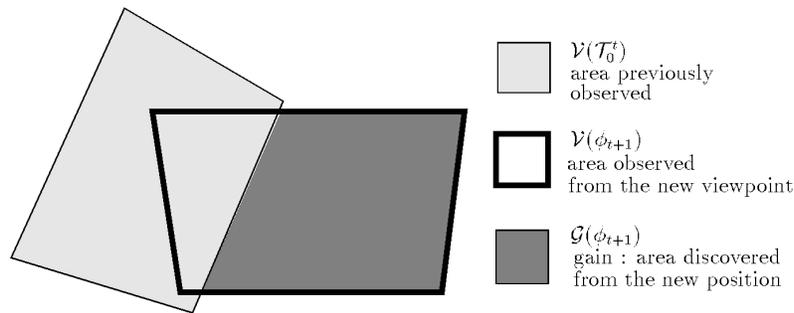

Fig. 3. Quality of a new position (2D projection).

probes are thus necessary to make sure that the whole scene has been reconstructed.

3 GLOBAL EXPLORATION

3.1 Related Work

This section deals with the exploration of the scene. Indeed, the goal is to observe all the parts of the scene and to ensure the completeness of the reconstruction (at least, a reconstruction as complete as possible) by determining adequate camera viewpoints. Previous works have been done in order to answer the "where to look next" question.

The problem is different if an a priori knowledge about the scene is available or not. If the complete geometrical description about the scene is known, many approaches about automatic sensor placement are described in [9], [19].

The problem is different if no a priori information about the scene is available i.e., if the sensor is in an unknown environment. It raises the problem of autonomous exploration [6], [8], [16], [20], [23], [22]. In [8], the sensor (a range finder) placement is computed from a local map of the scene that is described by an octree. The proposed solution, called the "planetarium algorithm," gives for all the camera positions on a sphere located around the scene, the viewpoint from which the maximal amount of unexamined region will be visible. The method proposed by Connolly performs an exhaustive search of the best viewpoint and considers only 2 degrees of freedom (such that the range finder is located on a sphere and gazes always at the same point). Wixson [23] describes strategies to search for a known object in a cluttered area. Maver and Bajcsy [16] do not try to optimize a cost function but use explicitly information provided by the analysis of the occlusions to plan the next viewing direction (furthermore, they do not explicitly handle

the completeness problem). Kutulakos et al. [12] presents an approach for exploring a 3D surface, using a mobile monocular camera, which is based on the use of the occlusion boundary. In [22], Whaite and Ferrie present a system that creates a 3D model of the environment using the data gathered by a laser range-finding system through a sequence of exploratory probes. In order to minimize the uncertainty of the parametric forms used to describe the scene, a feedback based on the model uncertainty is used as a basis for selecting viewpoints. We now describe the method used in our system to perform a complete scene exploration.

3.2 Viewpoint Selection

Let us consider a scene composed of a set O of initially unknown objects. At the end of a local exploration process, a subset $O\left(\mathcal{T}_0^t\right) \subseteq O$ has been observed and reconstructed. Thus, we have to determine viewpoints able to bring more information about the scene. By "information," we mean either the observation of a new object (in that case, the local exploration process will be used to estimate its 3D parameters), either the certainty that a given region is object-free. Such viewpoints will be computed using the previously estimated 3D map and the part of the 3D scene which has not been already observed.

Knowing $\mathcal{T}_0^t$, the set of viewpoints from the beginning of the reconstruction process, it is possible to maintain a map of the observed and unexplored regions. The knowledge is thus composed of (see Fig. 2):

- the objects already estimated: $O\left(\mathcal{T}_0^t\right)$;

- the known free space, denoted $\mathcal{V}\left(\mathcal{T}_0^t\right)$. From the current 3D map of the scene, it is possible to compute the region

$\mathcal{V}(\phi)$ observed from a camera position ϕ using a ray-tracing scheme. Thus, we can incrementally determine the region $\mathcal{V}(\mathcal{T}_0^t)$ observed from the beginning of the reconstruction process:

$$\begin{cases} \mathcal{V}(\mathcal{T}_0^0) = \mathcal{V}(\phi_0) \\ \mathcal{V}(\mathcal{T}_0^t) = \mathcal{V}(\mathcal{T}_0^{t-1}) \cup \mathcal{V}(\phi_t) \end{cases} \tag{1}$$

- the unknown region $\mathcal{U}(\mathcal{T}_0^t)$: It is simply computed as:

$$\mathcal{U}(\mathcal{T}_0^t) = \overline{\mathcal{V}(\mathcal{T}_0^t) \cup O(\mathcal{T}_0^t)} \tag{2}$$

We can here emphasize the importance of the incremental reconstruction algorithm which ensures that all the observed objects have been reconstructed. This point is crucial for correctly computing the known free space and unknown region. For instance, the region occluded by an object has of course to be considered as unknown, which necessitates the structure estimation of the occluding object.

A simple strategy able to compute the "next best view" ϕ_{t+1} is to consider the viewpoint that maximizes the volume of the new observed regions [8], [23]. However, such a strategy does not take into account some problems such as the manipulator kinematics constraints or geometric constraints. Furthermore, it is not possible to consider an exhaustive research of the best viewpoint. Indeed this is not realistic, in practice, owing to the size of the configuration space. As in [19], [20] we have thus defined a function to be minimized that integrates the constraints imposed by the robotic system and evaluates the quality of the viewpoint. The function $\mathcal{F}$ to be optimized is taken as a weighted sum of a set of measures that determine the quality or the badness of a viewpoint.

3.2.1 Quality of a New Position

The quality of a new position ϕ_{t+1} is defined by the volume of the unknown regions that appear in the field of view of the camera. The new observed region $\mathcal{G}(\phi_{t+1})$ is given by:

$$\mathcal{G}(\phi_{t+1}) = \mathcal{V}(\phi_{t+1}) - \mathcal{V}(\phi_{t+1}) \cap \mathcal{V}(\mathcal{T}_0^t) \tag{3}$$

where $\mathcal{V}(\phi_{t+1})$ defines the part of the scene observed from the position ϕ_{t+1} and $\mathcal{V}(\phi_{t+1}) \cap \mathcal{V}(\mathcal{T}_0^t)$ defines the subpart of $\mathcal{V}(\phi_{t+1})$ that has been already observed (see Fig. 3). If the position ϕ_{t+1} does not give any wage of information (i.e., $\mathcal{G}(\phi_{t+1}) = \varnothing$), we must reject this position. The measure related to the quality of position ϕ_{t+1} is thus given by:

$$g(\phi_{t+1}) = \begin{cases} \infty & \text{if } \mathcal{G}(\phi_{t+1}) = \varnothing \\ 1 - \dfrac{volume(\mathcal{G}(\phi_{t+1}))}{volume(\mathcal{V}(\phi_{t+1}))} & \text{else} \end{cases}. \tag{4}$$

Remark: In fact, $\mathcal{G}(\phi)$ defines the maximum volume of unknown region that can be expected using only the current knowledge on the 3D scene. If a new object appears in the camera field of view, the new observed region $\mathcal{G}'(\phi)$ is in fact smaller than the expected one ($\mathcal{G}'(\phi) \subseteq \mathcal{G}(\phi)$). However, this is not a problem as our goal is to discover new objects *or* to be sure that this part of the scene is object free. In both cases, the goal is reached.

3.2.2 Displacement Cost

A term reflecting the cost of the camera displacement between two viewpoints ϕ_t and ϕ_{t+1} is introduced in the cost function $\mathcal{F}$ in order to reduce the total camera displacement and to obtain a smoother trajectory. It is defined using the following relation:

$$C(\phi_t, \phi_{t+1}) = \frac{1}{N_{dof}} \sum_{i=1}^{N_{dof}} \beta_i \frac{\left| q_{i_t} - q_{i_{t+1}} \right|}{\left| Q_{i_{Max}} - Q_{i_{Min}} \right|} \tag{5}$$

where:

- N_{dof} is the number of robot degrees of freedom ;
- q_i is the position of the robot joint i and $\left| Q_{i_{Max}} - Q_{i_{Min}} \right|$ gives the distance between the joint limits on axis i ;
- β_i are weights setting the relative importance of an axis with respect to the others ($\beta_i \in [0, 1]$). For instance, rotational motions may be preferred to translational ones.

3.2.3 Additional Constraints

Additional constraints are associated to camera locations. The goal of these constraints is:

- to avoid unreachable viewpoints. This binary test $\mathcal{A}$ returns an infinite value when the position is unreachable. A position is unreachable if it is not in the operational space of the manipulator, or if this position is located in an unknown region (leading to a collision risk).
- to avoid positions in the vicinity of the robot joint limits. When a new primitive is observed from the computed viewpoint ϕ_{t+1}, an optimal estimation of its parameters is performed. This estimation requires camera motions performed by visual servoing which can not be realized if the robot encounters a joint limit. The measure $\mathcal{B}$ related to this constraint (and similar to those proposed in [13] or [17]) is optimal (equal to 0) if the camera is located at the middle of the extension of each axis of the robot:

$$\mathcal{B}(\phi) = \frac{1}{N_{dof}} \sum_{i=1}^{N_{dof}} \frac{4 \left(q_i - \frac{Q_{i_{Max}} + Q_{i_{Min}}}{2} \right)^2}{\left(Q_{i_{Max}} - Q_{i_{Min}} \right)^2} \tag{6}$$

3.2.4 Cost Function Optimization

The function $\mathcal{F}(\phi_{t+1})$ to be minimized is defined as a weighted sum of the different measures:

$$\mathcal{F}(\phi_{t+1}) = \mathcal{A}(\phi_{t+1}) + \alpha_1 g(\phi_{t+1}) + \alpha_2 C(\phi_t, \phi_{t+1}) + \alpha_3 \mathcal{B}(\phi_{t+1}) \tag{7}$$

Here, as each measure belong to $[0, 1] \cup \infty$, the weights can be predetermined in order to reflect the relative importance of the different measures. For example, the wage of information given by a new position is more important than the cost of the camera displacement. We have defined a priority order of the coefficients α_i such that $\alpha_1 > \alpha_2 > \alpha_3$. More precisely, we have fixed $\alpha_1 = 0.6$, $\alpha_2 = 0.3$, and $\alpha_3 = 0.1$.

As already stated, we have assumed that the scene is located inside a bounded volume. We have limited the set of potential viewpoints (position and viewing direction) to the ball that includes this volume. Let us note that the real set of potential viewpoints is far smaller than this volume. However, this is detected only by the reachability constraint that limits ("on-line") significantly its size. It is then limited by positions unreachable owing to kinematic constraints (joint limits) or because they are located in unknown regions.

Owing to the size of the set of potential viewpoints, it seems out of reach to perform an exhaustive research of the best viewpoint. Therefore, to find a good viewpoint (if not the best), we have decided to use a fast deterministic relaxation scheme corresponding to a modified version of the ICM algorithm (see [15] for
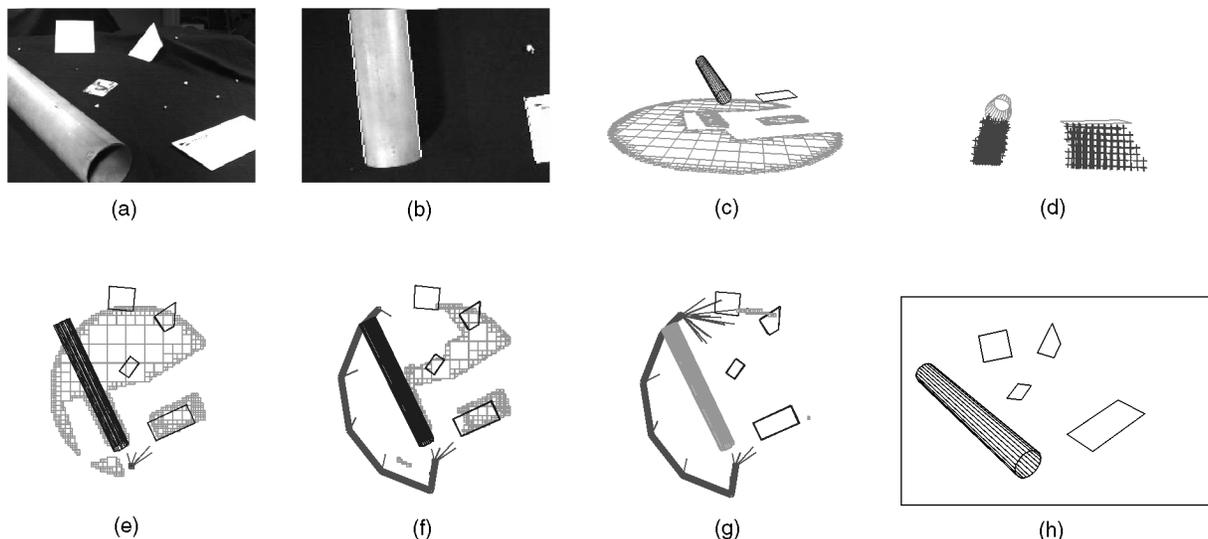
Fig. 4. (a) Birds' eye-view of the scene. (b) First image acquired. (c) After the incremental reconstruction process: reconstructed scene and projection of unknown region onto the ground plane. (d) Reconstructed scene and volumetric representation of the occluded region. (e), (g) Different steps of the global exploration process (camera trajectory, 3D model of the final reconstructed scene and top view of the unknown region), (h) 3D model of the reconstructed scene.
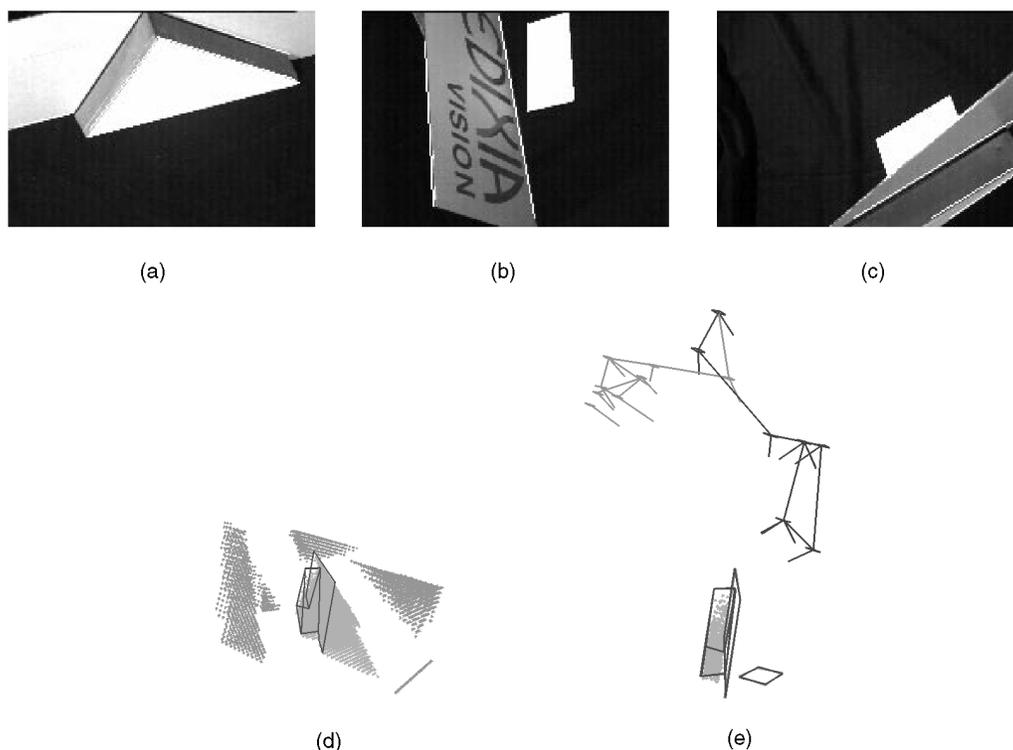

Fig. 5. Scene with a polyhedron object. (a) First image acquired. (b) A bird-view of the same scene from another viewpoint. (c) Image acquired during the exploration: A new object is observed. (d) Reconstructed object and unobserved regions before the exploration. (e) Camera trajectory, reconstructed objects, and residual unobserved regions (inside the polyhedron) after the exploration.

details). A viewpoint is described by a vector with six parameters (x, y, z, θ_x, θ_y, θ_z). First, $\mathcal{F}(\phi)$ is minimized using large variation steps of the parameters. When the minimum is found, the process is iterated with smaller variation steps.

We have chosen to represent explored and unexplored regions using a simple representation in term of voxels. This representation is easy to implement and it is very simple and fast to compute the regions observed from a given viewpoint using a ray-tracing algorithm [3]. Octrees are obviously more efficient

in term of memory, but they appear to be more time consuming for our purpose.

3.3 Completeness of the Reconstruction

3.3.1 Termination Criteria

In theory, the reconstruction must end when all the space has been observed, i.e., if at instant t: $\mathcal{U}\left(\mathcal{T}_0^t\right) = \varnothing$. However, this condition is usually unreachable. Ensuring the completeness of the recon-
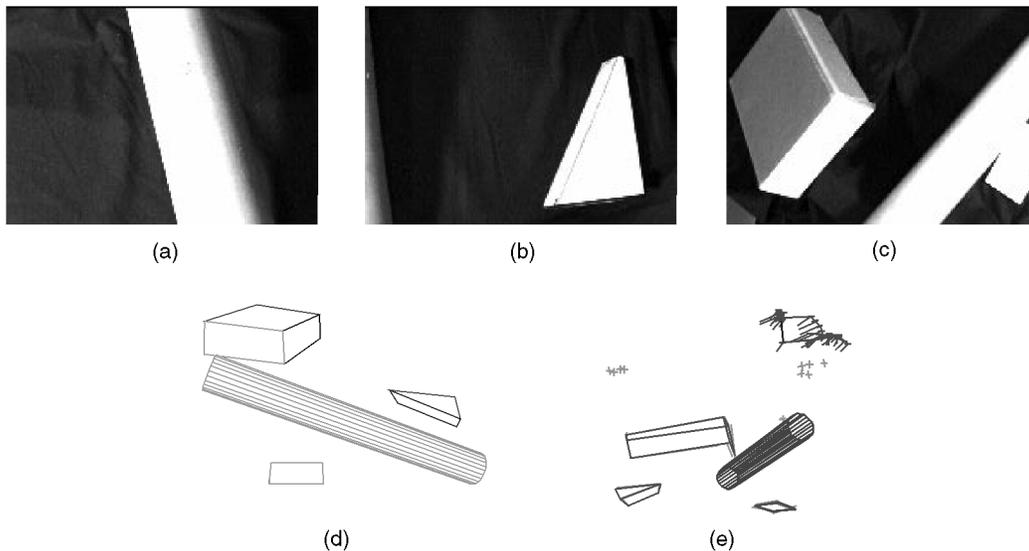
Fig. 6. Scene with cluttered regions. (a) First image. (b) After an exploration step (observation of a new object). (c) After another one step (two new objects appear in the field of view, one was previously occluded by the cylinder). (d) Reconstructed scene. (e) Reconstructed scene, camera trajectory and residual unobserved region.
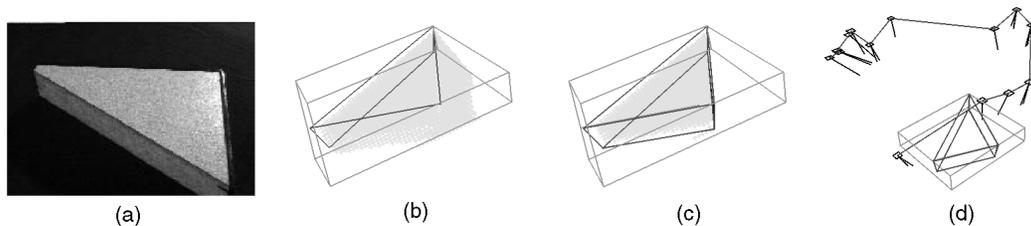

Fig. 7. Polyhedron reconstruction. (a) First image acquired. (b) Residual unobserved regions after a first reconstruction/exploration process (the parallelepiped represents the volume on which the camera gazes to complete the reconstruction). (c) Model computed at the end of the reconstruction process. (d) Camera trajectory.

struction is not always possible. Some regions may be observed only from a set of viewpoints unreachable by the camera. Furthermore, owing to the objects topology, some regions may be unobserved whatever the camera position. Thus, the following termination condition may be preferred:

$$\begin{cases} \mathcal{V}\left(\mathcal{T}_0^t\right) \cup \mathcal{V}(\phi_{t+1}) = \mathcal{V}\left(\mathcal{T}_0^t\right) \\ O\left(\mathcal{T}_0^t\right) \cup O(\phi_{t+1}) \cup O\left(\mathcal{T}_0^t\right) \end{cases}. \tag{8}$$

This means that the exploration process is as complete as possible if, for the new computed viewpoint, the camera looks at a known part of the scene. However, the use of a deterministic minimization algorithm prevents the camera from covering all the reachable viewpoints. Thus, small parts of the scene usually remain unobserved. Second, the marginal gain

$$\frac{volume\left(\mathcal{G}(\phi_{t+1})\right)}{volume\left(\mathcal{V}(\phi_{t+1})\right)}$$

decreases rapidly when the number of viewpoints increases. Thus, even if the whole scene is observable, the observation of the last residual regions requires a large number of viewpoints. For these different reasons, we decide to stop the exploration when a sufficient part of the scene has been observed (typically 95 percent). However, it is necessary to verify that the remaining unobserved regions do not contain any object (and if one object is found, to perform its reconstruction). That is the goal of the following algorithm.

3.3.2 Gazing on the Regions of Interest

In a first step, we compute a segmentation of the residual regions considering the polyhedron incorporating a set of connected unobserved regions (defined here by small voxels). As the goal is to define small regions, if the volume of the computed polyhedrons is too large, they are subdivided. In a second step, considering sequentially each subscene, an exploration process restricted to each of these subscenes is performed. Finally, the exploration ends when, for each created sub-scenes, $\mathcal{U}\left(\mathcal{T}_0^t\right) = \varnothing$ or when the criterion (8) is achieved (note that this process can be performed recursively). This strategy allows to decrease the number of viewpoints while increasing the part of observed regions and ensuring a reconstruction as complete as possible (leading generally to 99 percent of the whole scene).

4 EXPERIMENTAL RESULTS

The results presented in this section mainly deal with the global exploration algorithm. We will not give any result dealing with the reconstruction of each primitive using the structure from controlled motion algorithm. We refer the reader to [7] for these results. A process that allows to group 3D line segments into planar polygons and polyhedron is described in [15].

The first results (see Fig. 4) illustrate the capability of our algorithm to discover new objects located at very different locations inside the workspace. Fig. 4a shows an external view of the scene (never observed by the camera) and Fig. 4b shows the first image acquired by the camera. Figs. 4c and 4d show the primitives reconstructed at the end of the first local exploration process. As all

the primitives have not been observed, a global exploration process is thus necessary. Thirty viewpoints (Figs. 4e to 4g) are necessary to ensure that 99 percent of the scene is observed. (On the camera trajectory (Fig. 4e to Fig. 4g), only discrete viewpoints ϕ are represented. In fact, only the pointing direction seems to change because we do not show on this graph the camera motions used in the structure estimation itself. As this motion corresponds to two circle arcs of the same length (see Fig. 1), the position at the end of the estimation process is almost the same that at the beginning.)

This scene was quite simple, despite the fact that the objects are located at different locations. It is composed only of simple polygons and does not feature complex objects, self-occlusions, cluttered regions, or concave objects. We give afterward the results of the reconstruction of different scenes that feature this kind of problems.

Scene with a polyhedron object. The following result depicts the experiment of the reconstruction of a scene composed of a concave polyhedron (in fact two convex polyhedrons placed side by side forming a concave one) and a flat rectangle (see Fig. 5). Fig. 5a shows the first image acquired by the camera. Using the incremental reconstruction scheme, a large part of the scene can be reconstructed. However, large parts of the scene remain unobserved (44 percent—see Fig. 5d) and the rectangle and one segment of the concave polyhedron have not been yet reconstructed. After the computation of seven new viewpoints using the global exploration scheme (no new primitive appears in these views), the rectangle behind the polyhedron is detected and reconstructed (see Fig. 5c). A gazing process of the unobserved regions is finally performed after their segmentation. This leads to the exploration of 99.2 percent of the scene. As the 0.8 percent remaining unknown regions are detected to be inside the polyhedron, the completeness of the exploration is ensured.

Scene with cluttered regions. The next scene (see Fig. 6) features various polyhedral objects and a cylinder located in different planes. It can be considered as a lightly cluttered scene (the number of the objects is not very important but the scene features a large amount of occlusions and self-occlusions). In a first time, only the cylinder is observed and reconstructed. The exploration allows the observation of every objects of the scene. This includes some primitives located below the cylinder.

Seeing behind occlusions. By considering one of the polyhedron presented in the previous scene in another configuration, it may not be fully reconstructed after a first reconstruction/exploration process (see Fig. 7). Indeed some regions, which are occluded, have never been observed and some segments have not been reconstructed. This can be due to a failure in the optimization process (local minima). Therefore, as explained in Section 3.3.2, the system segments the unobserved region (see the parallelepiped located around the object on Fig. 7b) and the same exploration process is used, restricted on this region. Occluded regions too small to be detected in a first time can thus be appropriately handled using this method.

5 DISCUSSION

Image processing. The scenes considered in this paper are quite simple. First, the images are not noisy; second, we have restricted the problem to polygonal and cylindrical shape. The main reason for the use of simple images is a real-time issue. Let us recall that during the reconstruction of a primitive, the camera motion is computed in real-time with respect to acquired images [7]. For example, the reconstruction of a segment, in order that the camera achieves a motion of sufficient amplitude, involves the acquisition of around 200 images. Robust real-time tracking algorithms in noisy environments are not yet available (recent work such as

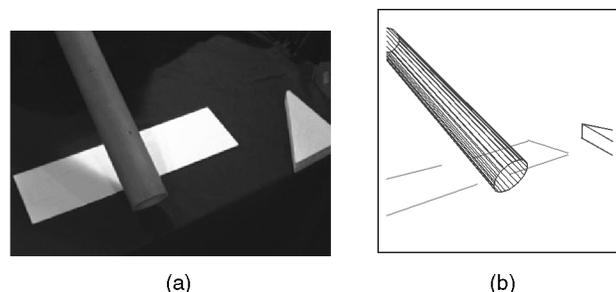(a)   (b)

Fig. 8. Sometimes it fails. (a) Bird-view of the scene. (b) (Incomplete) Reconstructed scene.

XVision [11] try to cope with these problems). Therefore, we have restricted ourselves to simple and well-contrasted images. Dealing with more complex shapes, their reconstruction is one perspective of this work.

Sometimes it fails... Other problems can be due to the lightning conditions, which influence the image processing, and to the shadows that can be observed (indeed shadow on a plane can be reconstructed as a line segment). Dealing with the exploration, we said that the reconstruction is as complete as possible. Indeed, they are two types of failures in our system: the inability to find regions of space that need exploration (e.g., due to the small set of viewpoints where those parts of the scene may be visible) and the inability to reach viewpoints that are known to make new parts of the scene visible. The latter type is benign, since the system itself is able to establish its own failure. This last kinds of problem occurred during the reconstruction of the scene presented in Fig. 8. A small amount of scene located below the cylinder has never been observed (the robot cannot reach the adequate viewpoint); this implies an incomplete reconstruction since some portions of the rectangle are never observed and therefore not reconstructed.

6 CONCLUSION

In this paper, we have proposed a method for 3D environment perception using a sequence of images acquired by a mobile camera. Since the method used for reconstruction is based on particular camera motions, perceptual strategies able to appropriately perform a succession of individual primitive reconstruction have been proposed. An important feature of our approach is its ability to easily determine the next primitive to be estimated without any knowledge or assumption on the number, the localization and the spatial relation between objects. Experiments carried out on a robotic cell have proved the validity of our approach, but have also shown its limitations: the constraints on the camera motion, which are necessary to obtain precise results, imply the sequencing of visual estimations. Furthermore, the necessity to process images in real-time does not allow us to consider very complex scenes. Related work have been done to improve this reconstruction/exploration process:

- Dealing with the local exploration process, we have developed a prediction/verification scheme based on the use of Bayesian networks which allows the system to predict the position of new segments to avoid explicit reconstruction of each primitive [15].
- The technique that is proposed in the previous paragraphs to solve the "next best view" problem is a greedy algorithm. We do no try to consider the whole trajectory in order to reduce either the number of viewpoints, or the global distance performed by the camera. Therefore, we have also proposed a method in order to reduce the number of viewpoints needed to explore the scene [14].

Finally, let us note that the exploration algorithm has been tested using the structure from controlled motion reconstruction method proposed in [7]. However, it can be used with any kind of reconstruction scheme such as stereovision or laser range finder. Using such reconstruction scheme would allow to consider more complex scenes and shapes and a faster reconstruction (due to a lower number of constraints in the camera motion and the acquisition of a smaller number of images). However, a compromise between accuracy of the reconstruction and performance will have to be done.

ACKNOWLEDGMENTS

REFERENCES

[1] Y. Aloimonos, "Purposive and Qualitative Active Vision," *Proc. IAPR Int'l Conf. Pattern Recognition, ICPR '90*, vol. 1, pp. 346-360, Atlantic City, N.J., June 1990.

[2] Y. Aloimonos, I. Weiss, and A. Bandopadhay, "Active Vision," *Int'l J. Computer Vision*, vol. 1, no. 4, pp. 333-356, Jan. 1987.

[3] J. Amanatides and A. Woo, "A Fast Voxel Transversal Algorithm for Ray Tracing," *Proc. Eurographics*, pp. 3-9, Amsterdam, Aug. 1987.

[4] R. Bajcsy, "Active Perception," *Proc. IEEE*, vol. 76, no. 8, pp. 996-1,005, Aug. 1988.

[5] D.H. Ballard, "Animate Vision," *Artificial Intelligence*, vol. 48, no. 1, pp. 57-86, Feb. 1991.

[6] A. Blake, A. Sisserman, and R. Cipolla, "Visual Exploration of Free-Space," *Active Vision,* A. Blake and A. Yuille, eds. Cambridge, Mass.: MIT Press, 1994.

[7] F. Chaumette, S. Boukir, P. Bouthemy, and D. Juvin, "Structure From Controlled Motion," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 5, pp. 492-504, May 1996.

[8] C. Connolly, "The Determination of Next Best Views," *Proc. IEEE Int'l Conf. Robotics and Automation*, pp. 432-435, St Louis, Mo., Mar. 1985.

[9] C.K. Cowan and P.D. Kovesi, "Automatic Sensor Placement From Vision Task Requirements," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 10, no. 3, pp. 407-416, May 1988.

[10] B. Espiau, F. Chaumette, and P. Rives, "A New Approach to Visual Servoing in Robotics," *IEEE Trans. Robotics and Automation*, vol. 8, no. 3, pp. 313-326, June 1992.

[11] G. Hager and K. Toyama, "The XVision System: A General-Purpose Substrate for Portable Real-Time Vision Applications," *Computer Vision and Image Understanding*, vol. 69, no. 1, pp. 23-37, Jan. 1998.

[12] K.N. Kutulakos, C.R. Dyer, and V. Lumelsky, "Provable Strategies for Vision-Guided Exploration in Three Dimensions," *Proc. IEEE Int'l Conf. Robotics and Automation*, vol. 2, pp. 1,365-1,372, San Diego, Calif., June 1994.

[13] E. Marchand, F. Chaumette, and A. Rizzo, "Using the Task Function Approach to Avoid Robot Joint Limits and Kinematic Singularities in Visual Servoing," *Proc. IEEE/RSJ Int'l Conf. Intelligent Robots and Systems, IROS '96*, vol. 3, pp. 1,083-1,090, Osaka, Japan, Nov. 1996.

[14] E. Marchand and F. Chaumette, "Active Sensor Placement for Complete Scene Reconstruction and Exploration," *Proc. IEEE Int'l Conf. Robotics and Automation*, vol. 1, pp. 743-751, Albuquerque, N.M., Apr. 1997.

[15] E. Marchand and F. Chaumette, "An Autonomous Active Vision System for Complete and Accurate 3D Scene Reconstruction," Technical Report 1156, IRISA, INRIA, Rennes, France, Jan. 1998.

[16] J. Maver and R. Bajcsy, "Occlusions as a Guide for Planning the Next View," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 5, pp. 417-433, May 1993.

[17] B. Nelson and P. Khosla, "Strategies for Increasing the Tracking Region of an Eye-in-Hand System by Singularity and Joint Limits Avoidance," *Int'l J. Robotics Research*, vol. 14, no. 3, pp. 255-269, June 1995.

[18] K. Tarabanis, P.K. Allen, and R. Tsai, "A Survey of Sensor Planning in Computer Vision," *IEEE Trans. Robotics and Automation*, vol. 11, no. 1, pp. 86-104, Feb. 1995.

[19] K. Tarabanis, R. Tsai, and P.K. Allen, "The MVP Sensor Planning System for Robotic Vision Tasks," *IEEE Trans. Robotics and Automation*, vol. 11, no. 1, pp. 72-85, Feb. 1995.

[20] B. Triggs and C. Laugier, "Automatic Camera Placement for Robot Vision," *Proc. IEEE Int'l Conf. Robotics and Automation*, vol. 2, pp. 1,732-1,738, Nagoya, Japan, May 1995.

[21] R. Vaillant, "Using Occluding Contours for 3D Objects Modeling," *Proc. First European Conf. on Computer Vision, ECCV '90*, pp. 454-464, Antibes, France, Apr. 1990.

[22] P. Whaite and F. Ferrie, "Autonomous Exploration: Driven by Uncertainty," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, CVPR'94*, pp. 339-346, Seattle, June 1994.

[23] L.E. Wixson, "Viewpoint Selection for Visual Search," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, CVPR '94*, pp. 800-805, Seattle, June 1994.