An autonomous active vision system for complete and accurate 3D scene reconstruction

E. Marchand, François Chaumette

International Journal of Computer Vision, ??, 1–23 (1999)

An Autonomous Active Vision System for Complete and Accurate 3D Scene Reconstruction

ÉRIC MARCHAND AND FRANÇOIS CHAUMETTE
IRISA / INRIA Rennes, Campus de Beaulieu, F-35042 Rennes Cedex, France
Eric.Marchand@irisa.fr, Francois.Chaumette@irisa.fr

Received January 12, 1998; Revised September 15, 1998

Abstract. We propose in this paper an active vision approach for performing the 3D reconstruction of static scenes. The perception-action cycles are handled at various levels: from the definition of perception strategies for scene exploration downto the automatic generation of camera motions using visual servoing. To perform the reconstruction, we use a structure from controlled motion method which allows an optimal estimation of geometrical primitive parameters. As this method is based on particular camera motions, perceptual strategies able to appropriately perform a succession of such individual primitive reconstructions are proposed in order to recover the complete spatial structure of the scene. Two algorithms are proposed to ensure the exploration of the scene. The former is an incremental reconstruction algorithm based on the use of a prediction/verification scheme managed using decision theory and Bayes nets. It allows the visual system to get a high level description of the observed part of the scene. The latter, based on the computation of new viewpoints ensures the complete reconstruction of the scene. Experiments carried out on a robotic cell have demonstrated the validity of our approach.

1. Active visual 3D perception

Most of the approaches proposed to solve vision problems are inspired by the Marr paradigm (Marr, 1982) which considers a sensor, static or mobile, but not controlled. 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 Bajcsy (Bajcsy, 1988), Aloimonos (Aloimonos et al., 1987, Aloimonos, 1990), or Ballard (Ballard, 1991) have proposed to modify the Marr concept. They proposed a new paradigm usually named **active vision**. Active vision techniques are issued from an attempt to simulate the human visual system. Dealing with human vision, head motions, eyes saccades and movements, the eyes adaptation to the lighting variations, are important in the perception process ; therefore active vision aims at simulate this power of adaptation. Since the major shortcomings which 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 strategies for adaptively setting camera parameters (position, velocity,...) in order to improve the perception task. Thus, function of the specified task and of the data extracted from the acquired image, an active vision system may be induced to modify its parameters (position, orientation, ocular parameters such as focus or aperture) but also the way data are processed (region of interest, peculiar image processing, etc). It controls either the sensor parameters or the processing resources allocated to the system (Swain and Striker, 1993).

In this first section, our purpose is not to provide an extensive review of ongoing works on active vision, but to point out the motivation of the

first authors involved in this research area and to describe the methodology we have followed for the conceptualization, design and implementation of our active vision system. The different authors who have introduced the active vision concept had different motivations. What is usually called active vision can be divided into four main classes: the *active perception* as defined by Bajcsy (Bajcsy, 1988) aims at elaborating control strategies for setting sensor parameters in order to improve the knowledge of the environment. The *active vision* introduced by Aloimonos (Aloimonos et al., 1987) is a mathematical analysis of complex problems such as stability, linearity and uniqueness of solutions. The goal of active vision is then defined as an intelligent data acquisition process. The *animate vision* (Ballard, 1991) is based on the analysis of human perception. Animate vision mainly uses binocular camera heads. Its goal is on one hand to solve the *gaze control* problem, and on the other hand to facilitate the computational process. Closely related to Bajcsy's active perception, the goal of *purposive vision* (Aloimonos, 1990) is to acquire and extract from the environment only the information needed to ensure the realization of a given task. Actions irrelevant to the specified problem will not be executed.

Despite these differences, the goal of the active vision community is to show that an active system is more *relevant* to the application (usually because it is goal driven), more robust (because it can handle either uncertainty and/or dynamic environment) and more accurate (because it is able to modify its own configuration). From our point of view, we think that these different approaches are closely related. The methodology used in this paper to define efficient exploration and reconstruction strategies is based on the three following relations:

- the perception-action cycle. The main point of the proposed approach is the relation between the motion of the camera and the information acquired during this motion. Visual data is used to control the camera motions, which are used to acquire information. We see this feedback loop as a fundamental characteristic of an active vision system. At this level, real time implementation (*i.e.*, to handle images at video rate) is a fundamental issue to allow an efficient feedback between perception and action.
- the relation between *global* and *local*. A task is usually defined in a global way (by the goal). However, data available to ensure the goal is usually local. The relation between the global modeling of the task and the set of local submodels (closely related to the parameters and the location of the camera) must be studied in order to ensure the execution of the nominal task. Describing a task as a scheduling of elementary tasks is a fundamental step to describe and implement such systems. Therefore, efficient techniques are necessary to link the local and global models.
- the relation between *continuous* and *discrete*. This aspect of the problem is closely related to the previous one. In one hand, the local elementary tasks can usually be handled in real time using continuous schemes (such as, for example, the control laws used to defined the camera motions). In that case, information must be seen as an infinite flow of data acquired by the sensor. In the other hand, the scheduling of these different tasks may require sensor planning strategies and therefore discrete camera motions. In that case, we manipulate discrete information (logic, temporal, etc.).

Active/purposive vision does not usually require an exact reconstruction of the scene. It has even been proposed to avoid this reconstruction and generally uses a qualitative representation of the scene. However, 3D reconstruction can be considered as a problem on its own, useful for various applications such as navigating tasks in clustered environment. Therefore, we will show how the scene reconstruction and exploration problem can be addressed in a purposive way. In that case, the animate vision, active vision, active perception and purposive vision are closely related.

Overview of the 3D reconstruction problem

Our concern is to deal with the problem of recovering the 3D spatial structure of a whole scene without any knowledge on the number, the localization, and the dimension of the different geometrical primitives of the scene (assumed to be

composed of polygons, cylinders and segments). The autonomous system we propose deals with various issues from the automatic generation of camera motion using image-based visual servoing to sensor planning to ensure a reconstruction as complete as possible of the scene.

The whole reconstruction/exploration process has three main **perception-action cycles** (*i.e.*, three levels):

- The first one is the exploration cycle. Its goal is to discover the objects which have not been yet observed by the camera. At this level, we have developed perceptual strategies able to determine the successive camera locations (next best view problem) ensuring the completeness of the exploration (for all the most, a reconstruction as complete as possible). This part of our work can be related to active perception as defined by Bajcsy (Bajcsy, 1988): the position and the orientation of the camera are set in order to increase the knowledge on the scene.
- From each viewpoint, all the objects observed by the camera have to be reconstructed (*i.e.*, the system has to estimate the parameters which describe the structure of each geometrical primitive composing the various objects of the scene). To obtain as accurate results as possible, we have chosen to use a continuous structure from controlled motion approach (Chaumette et al. 1996). It is based on the analysis of the motion of the object in the images sequence and on the measure of the corresponding camera velocity. Very noticeable improvements are obtained in the parameters estimation if the camera viewpoint is properly selected and if adequate camera motions are generated. Following this way, the work described in (Chaumette et al. 1996) confirms the point of view of previous work on active vision (Aloimonos et al., 1987) and on gaze control (Ballard, 1991). Indeed, it has been shown that the primitive must remain static at a given position in the image during the camera motion. These motions can be automatically generated using the visual servoing approach (Hutchinson et al., 1996). This aspect of the reconstruction can be related to the purposive vision concept (Aloimonos, 1990): only useful motions are generated.
- between the high level and the low level previously described, the system enters in an intermediate cycle which is the incremental reconstruction loop, also called local exploration. The main goal of this level is to bridge the gap between a set of local sub-models (a set of independent primitives) and a global model of the scene (composed of objects). It is composed of two processes widely interdependent. The proposed strategy is depicted on Figure 1. The first process deals with a simple incremental reconstruction algorithm (Figure 1a). It contains besides the internal perception-action cycle (described in the previous paragraph), which ensures the reconstruction of a single primitive, and a second cycle which ensures the detection, the successive selection, and finally the reconstruction of all the observed primitives. However, the model of the scene we get at this level is quite incomplete and contains only sparse primitives. Thus we use a second process which copes with these problems (Figure 1b). It proposes a partial solution to the occlusion problem and allows to obtain a high level description of the scene. This approach is based on a prediction/verification scheme managed using a probabilistic approach based on Bayes nets. These nets allow us to emit hypotheses on the existence and on the localization of new segments, and, then, to propose the execution of an action able to verify or to invalidate these hypotheses. Finally, with respect to the result of the verification step, it produces a new 3D model of the scene.

The remainder of this paper is organized as follows: Section 2 and Section 3 describe the internal cycle. Most of the work described in these sections has already been published (Espiau et al. 1992, Chaumette et al. 1996), however it is important to recall, even briefly, the main ideas that leads to the primitive reconstruction. Section 2 deals with image-based visual servoing. Section 3 is devoted to the local aspect of our reconstruction scheme and describes the structure from motion framework based on an active vision

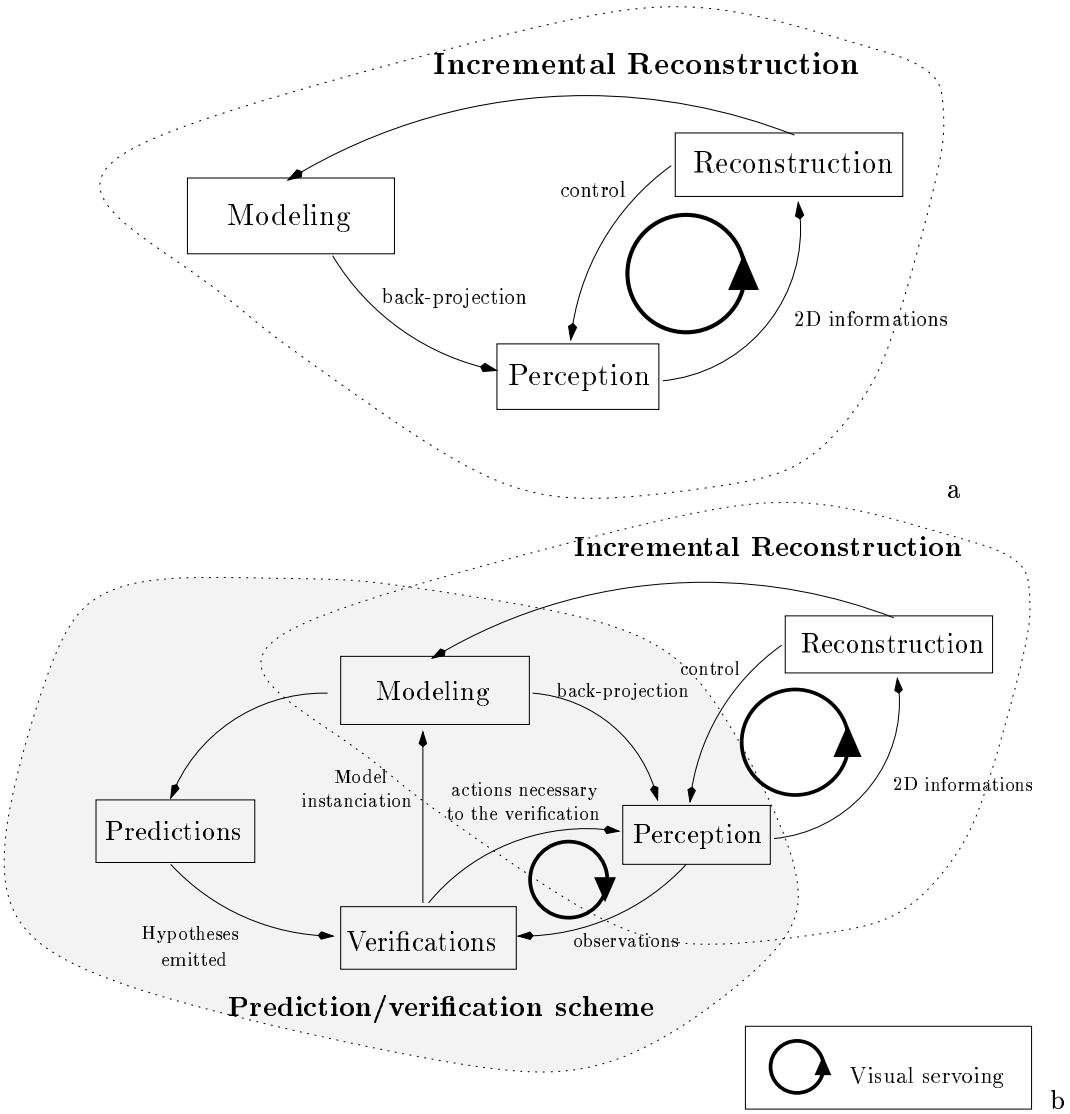

Fig. 1. Overview of the local exploration scheme

paradigm. Section 4 is devoted to the second cycle. It describes the incremental exploration algorithm and the Bayes nets-based prediction / verification scheme used to get a complete description of the observed part of the scene. The last cycle is described in Section 5, where the computing-viewpoint issue used to ensure a reconstruction as complete as possible of the scene is proposed. Finally, Section 6 presents experiments carried out on a robotic cell.

2. Image-based closed-loop control

The automatic generation of camera motion is a key point of the perception action cycle. Two main approaches are currently used in robot control based on visual data (Hutchinson et al., 1996): the *position-based control* which is achieved by computing, from the visual data, the 3D position and orientation of the camera with respect to its environment, and the *image-based visual servoing*, which consists in specifying a task as the regulation in the image of a set of visual features (Espiau et al. 1992, Hutchinson et al., 1996, Hashimoto,

1993). In the remainder of this paper, we will refer to this last approach since it is perfectly suitable for our purpose.

We may choose as visual features in a visual servoing framework the parameters vector (denoted $\mathbf{P}$) which describe the configuration of one or several primitives observed in the image (such as the coordinates of a point, the orientation and distance to origin of a line, the inertial moments of an ellipse, etc). More generally, any differentiable expression obtained from $\mathbf{P}$ can be used (such as the distance between a point and a line, the orientation between two lines, etc).

To ensure the convergence of $\mathbf{P}$ to its desired value, we need to know the interaction matrix $\mathbf{L}$, also called image Jacobian, defined by (Espiau et al. 1992):

$$\dot{\mathbf{P}} = \mathbf{L}(\mathbf{P}, \mathbf{p_r})\mathbf{T} \tag{1}$$

where $\dot{\mathbf{P}}$ is the time variation of $\mathbf{P}$ due to the camera motion $\mathbf{T}$. The parameters $\mathbf{p_r}$ involved in $\mathbf{L}$ represent the depth information between the considered object and the camera frame (see next Section). The interaction matrix plays an essential role in visual servoing. We will see in Section 3 that it is also involved in our 3D structure estimation method.

Using the formalism of the task function approach (Samson et al., 1991), a vision-based task $\mathbf{e}_1$ is defined by:

$$\mathbf{e}_1 = \mathbf{C}(\mathbf{P} - \mathbf{P_d}) \tag{2}$$

where $\mathbf{P_d}$ is the desired value of the selected visual features, $\mathbf{P}$ is their current value (measured from the image at each iteration of the control law), and $\mathbf{C}$, called combination matrix, has to be chosen such that $\mathbf{CL}$ is full rank (see (Espiau et al. 1992) for details). In our case, it can be defined as $\mathbf{C} = \mathbf{W}\mathbf{L}^+(\mathbf{P}, \widehat{\mathbf{p_r}})$ where $\mathbf{L}^+$ is the pseudo-inverse of the matrix $\mathbf{L}$ and $\widehat{\mathbf{p_r}}$ are the estimated values of the 3D parameters $\mathbf{p_r}$ obtained on-line using the 3D structure estimation method presented in the next section. Furthermore, $\mathbf{W}$ is defined as a full rank matrix such that Ker $\mathbf{W}$ = Ker $\mathbf{L}(\mathbf{P}, \widehat{\mathbf{p_r}})$.

If the vision-based task does not constrain all the n available robot degrees of freedom, a secondary task can be performed and we obtain the following task function:

$$\mathbf{e} = \mathbf{W}^+\mathbf{e}_1 + (\mathbf{I} - \mathbf{W}^+\mathbf{W})\mathbf{e}_2 \tag{3}$$

where

- $\mathbf{W}^+$ and $\mathbf{I} - \mathbf{W}^+\mathbf{W}$ are two projection operators which guarantee that the camera motion due to the secondary task is compatible with the regulation of $\mathbf{P}$ to $\mathbf{P_d}$. Indeed, thanks to the choice of matrix $\mathbf{W}$, $(\mathbf{I} - \mathbf{W}^+\mathbf{W})\mathbf{e_2}$ belongs to Ker $\mathbf{L}$, which means that the realization of the secondary task will have no effect on the vision-based task ($\mathbf{L}(\mathbf{I} - \mathbf{W}^+\mathbf{W})\mathbf{e}_2 = 0$). However, if errors are introduced in $\mathbf{L}$ and $\mathbf{W}$ due to errors in the estimation of $\widehat{\mathbf{p_r}}$, $\mathbf{I} - \mathbf{W}^+\mathbf{W}$ no more exactly belongs to Ker $\mathbf{L}$, which will induce perturbations on the visual task due to the secondary task.
- $\mathbf{e_2}$, called secondary task, is the gradient of a cost function h_s to be minimized. This cost function is minimized under the constraint that $\mathbf{e}_1$ is realized. Different secondary tasks have been developed for this application:
  - trajectory tracking which allow the camera to move along a given trajectory. These motions are mainly used in the structure estimation processes (see Section 3). Furthermore, we have also defined motions tied to the estimation of the length of each primitive.
  - joint limits and singularities avoidance (Marchand et al., 1996c). Our solution to this problem uses the robot redundancy with respect to the image constraints. The cost function to be minimized is based on a measure of the robot manipulability in the vicinity of internal or external singularities.

  The cost functions involved in the realization of these tasks are fully described in (Marchand, 1996a).

To assure that $\mathbf{e}$ exponentially decreases and then behaves like a first order decoupled system, we get:

$$\mathbf{T} = -\lambda\mathbf{e} - (\mathbf{I} - \mathbf{W}^+\mathbf{W})\frac{\partial \mathbf{e}_2}{\partial t} \tag{4}$$

where $\lambda > 0$ is a proportional coefficient involved in the exponential convergence of $\mathbf{e}$, and the term $(\mathbf{I} - \mathbf{W}^+\mathbf{W})\frac{\partial \mathbf{e}_2}{\partial t}$ is tied to the generation of a non zero camera motion when the vision-based task is realized.

Let us now examine how visual servoing has been used within a structure estimation problem.

Indeed, it is perfectly suitable to generate camera motion tied to an optimal estimation of primitive parameters.

3. 3D structure estimation using active vision

The measure of the camera motion, which is necessary for 3D structure estimation, characterizes a domain of research called dynamic vision. Approaches for 3D structure recovery may be divided into two main classes: discrete approach (Chien and Aggarwal, 1989, Weng et al., 1990) and the continuous approach (Adiv, 1989, Espiau and Rives, 1987, Xie and Rives, 1989). The former lie on the analysis of the object displacement in the images sequence and on the measure of the camera displacement, whereas the latter is based on the analysis of the motion of the object in the images sequence and on the measure of the corresponding camera velocity. The method used here is a continuous approach. More precisely, we use a "*structure from controlled motion*" method which consists in constraining the camera motion in order to obtain a precise and robust estimation of 3D geometrical primitives such as points, straight lines and cylinders (Chaumette et al. 1996). Simplifying and improving shape estimation by viewpoint control is also reported in (Kutulakos and Dyer, 1994).

The issue is to estimate the parameters $\mathbf{p}$ which fully characterizes a 3D primitive. It can be noticed that part of these parameters ($\mathbf{p_r}$) are involved in the interaction matrix. Thus, from the resolution of a linear system derived from (1), we can obtain the parameters $\mathbf{p_r}$ that describe the position of the rim surface (see Figure 2) Then, measuring the position of the primitive in the image and using geometrical constraints related to the considered primitive, we can estimate the parameters $\mathbf{p}$ which fully define its 3D configuration. We thus have:

$$\widehat{\mathbf{p_r}} = \widehat{\mathbf{p_r}}(\mathbf{P}, \dot{\mathbf{P}}, \mathbf{T}) \quad \text{and} \quad \widehat{\mathbf{p}} = \widehat{\mathbf{p}}(\mathbf{P}, \mathbf{p_r}) \qquad (5)$$

From a geometric point of view, this continuous approach implies solving for the intersection between the rim surface and a generalized cone, defined by its vertex located at the optical center and by the image of the primitive. This approach has been applied to the most representative primitives (*i.e.*, point, straight line, circle, sphere, and cylinder) (Chaumette et al. 1996).

When no particular strategy concerning camera motion is defined, important errors on the 3D structure estimation can be observed. This is due to the fact that the quality of the estimation is very sensitive to the nature of the successive camera motions (Espiau and Rives, 1987). An active vision paradigm is thus necessary to improve the accuracy of the estimation results by generating adequate camera motions. In fact, two main results dealing with this problem have been achieved (Chaumette et al. 1996):

1. A sufficient and general condition that suppresses the discretization error is to constrain the camera motions such that:

$$\dot{\mathbf{P}} = 0 \text{ and } \dot{\mathbf{p}}_\mathbf{r} = 0, \forall t, \qquad (6)$$

   i.e., the projection of the primitive must be motionless in the image, and the equation of the rim must be kept constant in the moving camera frame.
2. A more robust estimation with respect to measurement errors can be obtained if the relation between the camera and the primitive is considered. Some positions of the primitive in the image do minimize the influence of the measurements errors. So, in order to obtain an *optimal estimation*, a gaze control task which constrains the camera motion so that the object remains fixed at its specified position in the image is necessary.

For example, in the particular case of a cylinder, it can be shown that the optimal camera motion is such that the tracked cylinder rims constantly appear as static, centered, vertical or horizontal straight lines in the image sequence (see Figure 3). The visual servoing approach is very well qualified to control camera motions in order to satisfy these constraints. Indeed, as the parameters $\mathbf{p_r}$ of the primitive are estimated on-line, these parameters are feedback into the visual servoing process. This allows to update, at each iteration (frame) of the estimation/visual servoing process, the interaction matrix $\mathbf{L}$, as well as the projection operators $\mathbf{W}^+$ and $\mathbf{I} - \mathbf{W}^+\mathbf{W}$. This adaptive behavior ensures a simultaneous convergence of the visual task and

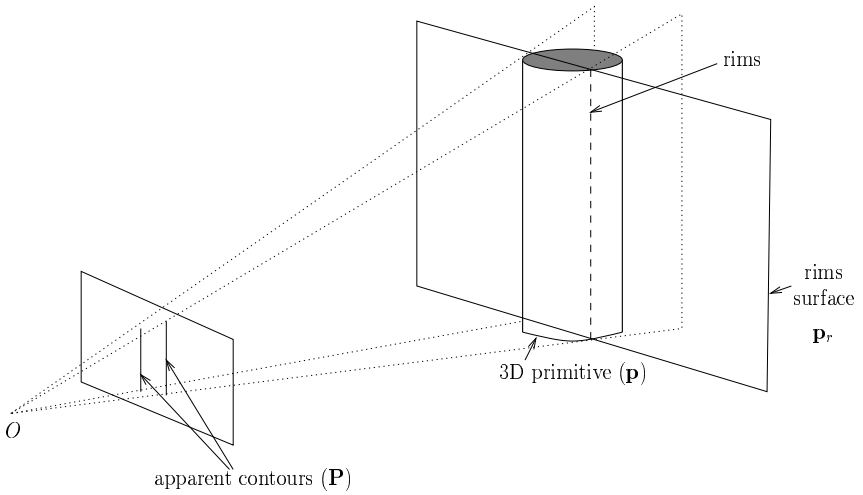

Fig. 2. Projection in the image ($\mathbf{P}$, *i.e.*, the apparent contour) of the primitive ($\mathbf{p}$) and rim surface ($\mathbf{p_r}$) in the case of a cylinder

the structure estimation scheme. Indeed, perturbations introduced in $\mathbf{I}-\mathbf{W}^+\mathbf{W}$ progressively disappear thanks to the improvement in the estimation, which improves the achievement of the visual task, and thus the estimated value of $\mathbf{p_r}$.

The main advantage of this method leads in the accuracy and robustness of the estimation results. As already stated, specific camera motions are automatically generated to obtain an optimal estimation of 3D parameters (*e.g.*, the position of a point can be computed with a precision of 1 mm, for a distance point/camera of 1 m, and the error on the estimation of the radius of a cylinder is less than 0.5 mm for a similar range – see Section 6 for more details). However this reconstruction scheme has some drawbacks. First, the primitive (cylinder or straight line) is assumed to have an infinite length. A specific process has thus been defined to compute the length and the spatial position of the primitive along its axis (Marchand and Chaumette 1996b). To achieve this task, visual servoing is used to observe the extremities of the primitive at a given position in the image. Second, since the structure estimation method is specific to each kind of primitive, a preliminary recognition process is required. This is done using a statistical test (Marchand and Chaumette 1996b). Finally, as the reconstruction scheme involves fixating at and gazing on the different primitives of the scene, this can be done on only one primitive at a time. Hence, reconstruction has to be performed in sequence. Next sections are devoted to this high level issue: the complete reconstruction and exploration of a scene composed of several objects.

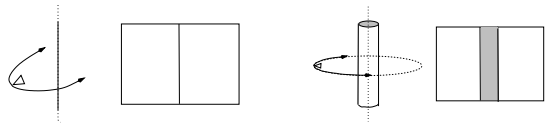

Fig. 3. Optimal camera motion and resulting image in the cases of a straight line and a cylinder

4. Toward a global representation of the scene

4.1. 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 2D segments. The first step in the scene reconstruction process is to obtain the list of these segments. We denote these lists $\omega_{\phi_t} = \{S_i, i = 1 \dots M\}$, where ϕ_t is the corresponding camera location from which the M segments S_i are observed. In fact, we consider in ω_{ϕ_t} only the segments which do not correspond to a reconstructed primitive. This is done by a simple matching process between all the observed 2D segments and the back-projection of the current 3D map of the scene. Finally, only the segments which are long enough to be accurately located us-

ing the structure from controlled motion method are considered.

Another list, denoted $\Omega_{\mathcal{T}_{t_1}^{t_2}}$, is also used. It contains all the unmatched segments previously observed, and the camera positions ϕ_k from which they have been observed. More precisely, we have:

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

where $\mathcal{T}_{t_1}^{t_2} = \{\phi_{t_1}, \phi_{t_1+1}, \dots, \phi_{t_2}\}$ is the set of all viewpoints between t_1 and t_2, S_i is an unmatched 2D segment, 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. The system automatically extracts from ω_{ϕ_0} a segment S_i to be considered.
Step 1: Active 3D estimation and 3D map creation. The parameters of the 3D primitive corresponding to S_i are estimated using the structure estimation process described in the previous section. The reconstructed primitive is introduced into the 3D map of the scene. In order to integrate this new object in the global model of the scene, the Bayes nets-based prediction/verification scheme which will be described in the next subsection is used at this step. We will come back at the end of this section on our motivations to introduce this scheme. We then remove from $\Omega_{\mathcal{T}_0^t}$ all the 2D segments corresponding to the reconstructed primitive.
Step 2: Local and global 2D lists generation. After the active estimation, 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 can occur:

a) In the case where several segments are in the current list $\omega_{\phi_{t+1}}$, a choice is performed in order to select the next segment S_i to be taken into account. We iterate the steps 1, 2 and 3 until one of the segments present in the current list $\omega_{\phi_{t+1}}$ has not been reconstructed.
b) **Backtracking.** If all the segments of $\omega_{\phi_{t+1}}$ have been considered and if at least one of the objects previously observed has not been reconstructed (*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 the couple (S_i, ϕ_k) for which the distance between ϕ_{t+1} and ϕ_k is minimal. Then, the camera moves back to position ϕ_k and the parameters of the primitive corresponding to S_i are estimated (step 1).
c) Finally, if $\Omega_{\mathcal{T}_0^{t+1}}$ is empty (*i.e.*, all the 2D segments observed from any previous camera positions have been taken into account), new viewpoints must be found in order to ensure the completeness of the reconstruction. A **global exploration**, which will be described in Section 5, is thus necessary.

Discussion. The proposed algorithm allows to perform an estimation of the structure of all the primitives which appear in the camera field of view. Although we have just mentioned the prediction/verification scheme (step 1), it is one of the key features of our algorithm. Indeed, if it is not considered, the incremental reconstruction scheme raises the following problems:

- The description of the scene is a low level and local description and contains only a list of 3D segments and cylinders. It might be more interesting to get high level global information such as junctions, polygons, or faces.
- The scene reconstruction is incomplete for two main reasons:
  - The projection in the image of some segments had a too small length to make their reconstruction possible.
  - As this algorithm estimates only the observed primitives, it has only a local perception of the scene. According to this, some objects may never appear in the camera field of view (because of occlusions or because they are located in an unknown and unobserved area). So they will never be reconstructed.

For example, let us consider the object depicted on Figure 4. The model obtained using the simple incremental reconstruction algorithm presented above is given in Figure 4.c. It outlines the problems recalled above: the 3D model is composed of four segments which *a priori* come apart,

a small number of segments has not been taken into account because of their small size, and two long segments have not been estimated (because they were always occulted). The method proposed now allows the system to complete this model as shown on Figure 4.d.

4.2. A Bayes-nets Based Prediction / Verification Scheme

Our goal here is to improve the incremental scene reconstruction in order to obtain a more complete and high level representation of the scene. As already stated, our approach is based on Bayes nets prediction/verification scheme. It has been applied here to the reconstruction of polyhedral scenes. Similar strategies could be extended to the reconstruction of different kind of scenes such as cylinders network.

4.2.1. Bayes nets In our application, measurement errors appear either in the 3D data acquired using the structure from motion approach or in the extraction of segments in images. Mainly, the consequence of this uncertainty is the confrontation of different possible alternatives for guiding the reconstruction and the exploration of the scene. The goal of decision theory is to provide well defined and mathematical approaches for making a decision in presence of uncertainty. Different approaches have been proposed and have been already used in computer vision: Dempster Shafer theory (Hutchinson and Kak, 1989) or hidden Markov models (Rimey and Brown, 1991). Among these different approaches, Bayes nets (Pearl, 1988) seem to be well adapted to our problem. They allow us to model "expert" reasoning. They are adapted to the automatic generation of action while performing this reasoning. Thus we can directly introduce perception strategies within the scene interpretation process. Using Bayes nets in active vision is more recent. Most discriminant works have been proposed by Rimey and Brown (Rimey and Brown, 1994) with the TEA-1 system (selective perception for visual search), Buxton and Gong (Buxton and Gong, 1995) (traffic analysis), or Djian and Rives (Djian et al., 1995) (for object recognition).

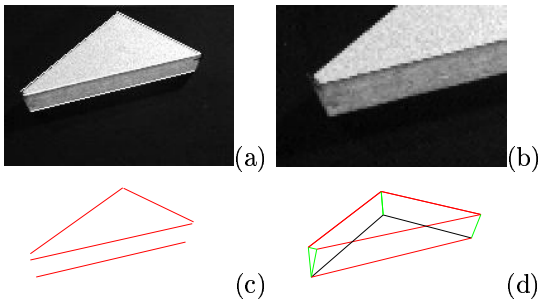

Fig. 4. Polyhedral scene: (a) view of the scene, (b) other view of the same object (note the little segment), (c) 3D model obtained using the incremental algorithm, (d) 3D model obtained using the prediction/verification scheme.

Bayes nets allow representation of joint probabilities distributions of a set of variables using a set of *a priori* knowledge on the relations between these variables. A Bayes net is a directed acyclic graph where nodes represent the discrete random variables and where links between nodes represent the causality between the variables. Such a net can be used to represent the knowledge available on a particular domain. The graph structure and the *a priori* knowledge introduced in the graph (as conditional probability tables) must be defined by the designer of the application.

The advantages of Bayes nets lies in the ability to reflect the *a priori* knowledge available on the application. The knowledge is reflected at two levels:

- in the structure of the net through the nature and the number of nodes (variables), the different states of these variables and the relations (links) between these variables.
- in the conditional probability tables, associated with each variable of the net, which reflect the expert reasoning. These tables also model the uncertainty associated with the observations.

Finally, the propagation allows each new observation to be taken into account. The influence of an observation is propagated to the other variables of the net according to the causality relations. Describing this complex process is not the goal of this paper. A simple overview of the propagation algorithms is proposed in (Krause and Clark, 1993) and a detailed description in (Neapolitan, 1990).

4.2.2. Overview of our approach Dealing with our problem, the information available on the scene is composed by a set $\mathcal{S}(\mathcal{T}_0^{t-1})$ of 3D segments. $\mathcal{S}(\mathcal{T}_0^{t-1})$ is a subset of $\mathcal{O}(\mathcal{T}_0^{t-1})$ which represents all the known objects of the scene (*i.e.* 3D segments but also junctions, polygons, etc.). The goal is to determine the relations between segments and to infer either the presence of new segments, or the existence of more complex objects. As our reconstruction is incremental, we have to determine the consequence of the introduction of a new segment S_t in $\mathcal{S}$. Therefore, this module is used each time that a new segment is introduced in $\mathcal{S}$.

Our approach can be decomposed into three steps. For each couple of segments $(S_{t'}, S_t), t' \in [0, t-1]$, we propose hypotheses on the relation between these two segments. Then, we verify if these hypotheses match the observations. Finally, the system proposes a new model of the scene resulting from the integration of the new segment.

4.2.3. Prediction Dealing with two segments $S_{t'}$ and S_t, the possible actions are the following: fuse the segments, create a junction, or add a new segment between $S_{t'}$ and S_t. Therefore the goal is to create some hypotheses leading to the realization of one (or more) of these actions. We now describe this process.

The hypotheses are directly linked to the actions:

- H_1: there is a junction between $S_{t'}$ and S_t ;
- H_2: there are one or two segments between $S_{t'}$ and S_t ;
- H_3: $S_{t'}$ and S_t are identical ; and
- H_4: there is no relation(or some other relations different from H_1, H_2 or H_3) between $S_{t'}$ and S_t.

We have a multi-step strategy. First, we look for simple topological relations (proximity, coplanarity, collinearity) between $S_{t'}$ and S_t. Then, we have defined five distinct classes that represent particular combinations of the previous relations. Knowing these classes, we infer the hypotheses. This reasoning can be encoded in a simple Bayes net (see Figure 5). It is composed of six nodes corresponding to different steps in the reasoning. One node is associated to each topological relation, another to the class, and two nodes are associated to each set of hypotheses. Links between these nodes depict the causality relations between the different steps of reasoning and thus its progression.

More precisely, according to the belief we have in the three topological relations (coplanarity $p(C_{tt'})$, proximity $p(N_{tt'})$ and parallelism $p(P_{tt'})$), it is possible to classify the pair of segments into five classes (see the first column of the Table in the Figure 6). Segments can belong to classes $\mathcal{C}_1$ (they are coplanar, neighbor and parallel, CNP), $\mathcal{C}_2$ ($CN\neg P$), $\mathcal{C}_3$ ($C\neg NP$), $\mathcal{C}_4$ ($C\neg N\neg P$), or $\mathcal{C}_5$ ($\neg C\neg N\neg P$).

Using the belief we have in the belonging of the couple of segments to each class, the system can infer the belief in each possible hypothesis. We have defined decision strategies which are able to determine the best hypothesis according to the available knowledge. These strategies are coded in conditional probability tables $P(H|C)$ (where H is the hypothesis and C the class).

In order to emit an hypothesis, it is necessary to propose a set of elementary considerations about topological relationship that we usually find in a group of segments. These considerations often reflect the truth, though they provide no guarantees. However, we can use them at the basis of the hypotheses generation strategies.

To illustrate this point, let us take the example of two coplanar and neighboring segments (class $\mathcal{C}_2$, see second line of Figure 6). The best hypothesis we can do in that case is that there is a junction between these two segments. However, according to the uncertainty associated with the 3D position of these segments, it is also possible, with a lower belief, to predict the presence of a little segment linking the closest extremities. The other hypotheses (H_1 and H_4) must not be rejected, but we have a very low belief in their achievement. This kind of reasoning can be encoded in the conditional probabilities table associated to the class $\mathcal{C}_2$. This table is defined in an empirical way, as extreme precision is not required. Rather, it must reflect the knowledge we want to transmit to the system.

Two sets of hypotheses are emitted. The first concerns the relation between the closest extremities of the segments and the second concerns the relation between their distant extremities. In both cases, the same hypotheses can be emitted, though

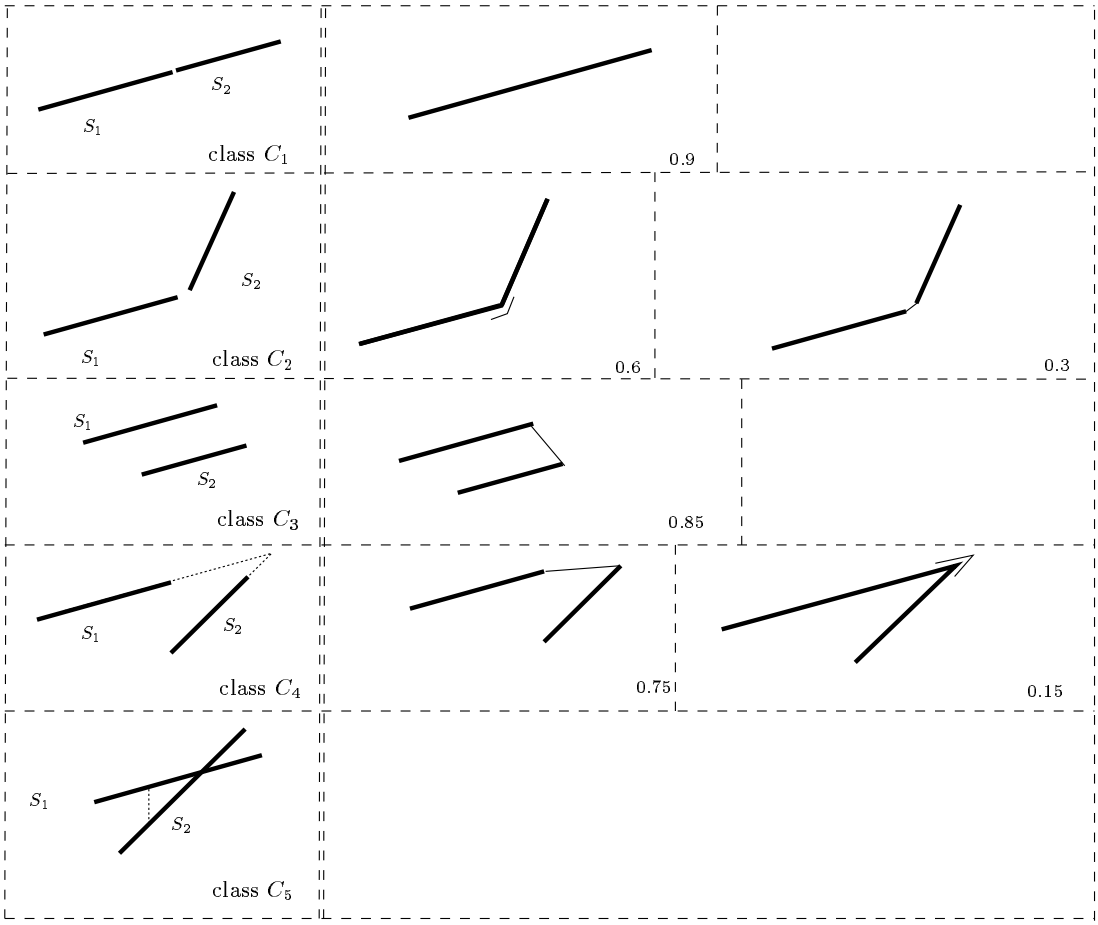

Fig. 5. Elementary classes and associated hypothesis (closest extremities)

the associated conditional probabilities can be very different (see Table 1 and 2).

As already stated, the hypothesis with the higher belief is not always the correct one, and this is the reason why we will always consider for each case (closest and distant extremities) the two hypotheses with the highest belief (H^1_{max} and H^2_{max}). These two hypotheses are then verified or invalidated. Two hypotheses are sufficient in our case according to the proposed conditional probabilities tables. In a different problem, more hypotheses may have to be checked. A similar approach in a pattern recognition problem has been proposed in (Djian et al., 1995).

4.2.4. Verification In order to verify the two selected hypotheses, we use the reasoning encoded in the Bayes net depicted in Figure 7. Considering the two hypotheses, we first define the nature (segment, junction, string) and the position of the created object associated with each hypothesis. Then, we compute the belief in the existence of this object. This step (observation) is the most important. Sometimes the hypotheses can be verified (or invalidated) using direct observation in the images previously acquired, though

Table 1. Conditional probabilities table $P(H \mid C)$ for the closest extremities

	Classes				
hypotheses	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_4$	$\mathcal{C}_5$
H_1 fusion	0.90	0.025	0.05	0.05	0.025
H_2 junction	0.025	0.60	0.05	0.15	0.025
H_3 string	0.025	0.30	0.85	0.75	0.025
H_4 other	0.05	0.075	0.05	0.05	0.925

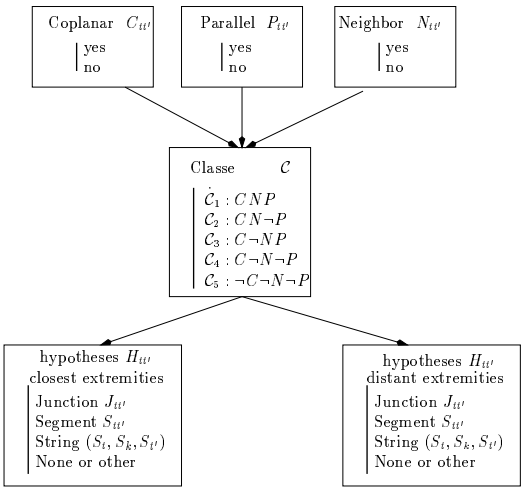

Fig. 6. Hypotheses net

this is not always possible. In such cases, it is necessary to move the camera in order to get the confirmation or the invalidation of the hypothesis. Finally, knowing the belief in each hypothesis and the belief in the related observation, it is possible to determine the most probable hypothesis (or to reject both).

Observation - Verification actions. As already stated, the most important node in the verification net is the *observation* node. This validation is performed using the 3D information associated with the hypotheses and the 2D observation. We perform a back-projection of the 3D objects in each image previously acquired by the camera and we try to associate this projection to the observed data in more than one image (to avoid false matching). For each possible matching, we compute the belief granted to this matching.

The case of a single segment (*i.e.*, that correspond to merge two segments) or of a junction is simple. If this junction (or this segment) exists

Table 2. Conditional probabilities table $P(H \mid C)$ for the distant extremities

	Class				
hypotheses	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_4$	$\mathcal{C}_5$
H_1 fusion	0.025	0.025	0.025	0.025	0.025
H_2 junction	0.025	0.025	0.025	0.025	0.025
H_3 string	0.025	0.5	0.7	0.5	0.025
H_4 other	0.925	0.45	0.25	0.45	0.925

it has already been observed (because the presence of the two segments, from which the hypotheses has been emitted, has been already verified). Thus, the verification is immediate. The case of a string is more interesting. In a string, with three segments, the presence of two of them is certain (they have been used to predict the presence of the third). However the last one has not been yet reconstructed (most of the time) and its presence is not validated. When no matching is found in images previously acquired, it is necessary to know why. The first possibility is that the segment under consideration does not exist, the second is that it is occluded by another object. In the latter case, it is necessary to move the camera to a new viewpoint from which the segment can be observed.

Many approaches can be used to compute this viewpoint (*e.g.* (Cowan and Kovesi 1988, Tarabanis et al., 1995b)). Cowan and Kovesi proposed to compute the set of viewpoints from which the primitive is observable. The method they proposed is simple. However, the resulting set of viewpoints can be huge and the choice of one position from the set requires the introduction of new constraints. Thus, computing the viewpoint can be time consuming.

Rather than computing explicitly a viewpoint and investigating *off-line* the considered segment, the camera rotates around a segment which belongs either to the occluding polygon or to a plane

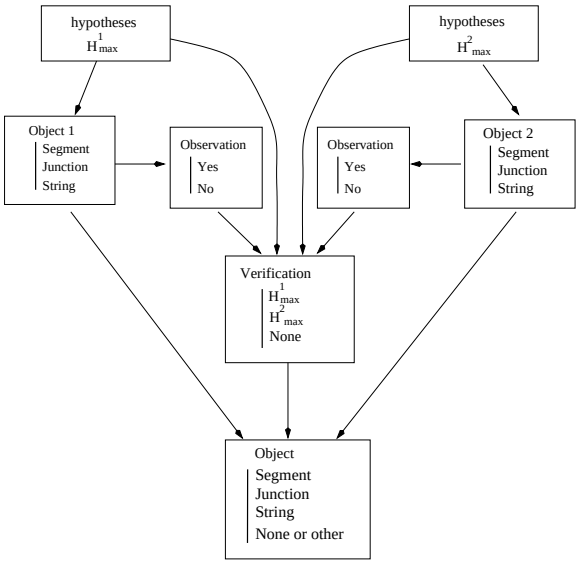

Fig. 7. Verification net

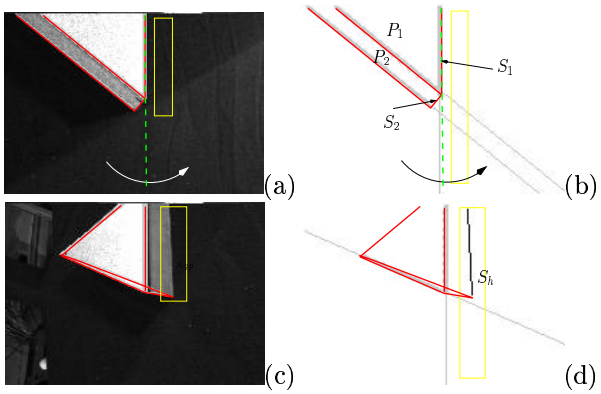

Fig. 8. Hypotheses verification: (a-b) view of the scene from the initial position of the camera, (c-d) view of the scene from the final position of the camera (note the presence of the supposed segment S_h in the observer).

to which the considered segment belongs (according to the way hypotheses are generated, such a segment exists). During this motion, automatically generated by visual servoing, the system looks on-line for the appearance of the segment at its predicted position in the image (according to the current camera position). To detect this appearance, dynamic observers are used. Such observers, as defined for example in (Djian et al., 1995), are small windows in which dedicated image processing is performed. In our case, the observers detect a moving edge using the algorithm proposed in (Bouthemy 1989).

The example shown in Figure 8 describes this strategy. Consider here a polyhedral object where two faces have been already reconstructed. The presence of segments in the plane formed by the segments S_1 and S_2 has been supposed. According to the proposed strategy, the camera gazes on one of these segments (the longest one S_1 is chosen) and turns around it which allows the predicted segment to appear in the observer (see Figure 8.c-d).

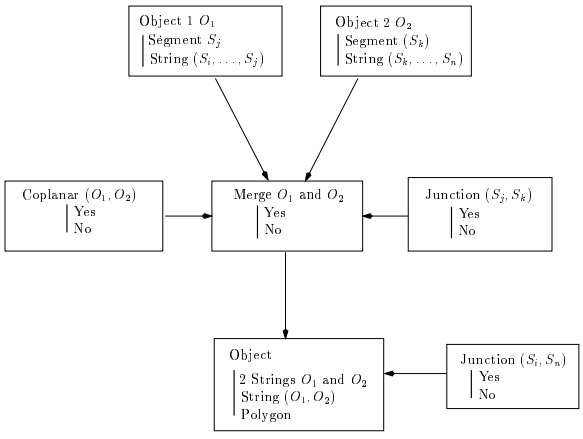

Fig. 9. Modeling net

4.2.5. Modeling The goal now is to use the 3D information (3D segments, 3D junctions, or even a coplanar string of segments) in order to get 3D polygons. To this end, we use the junction information and the coplanarity information already used in the hypotheses generation.

We have created a third Bayes net in order to infer the existence of new strings of coplanar segments as well as polygons. If we consider two objects O_1 and O_2, O_1 and O_2 can be either segments or a string of coplanar segments (the string could have been constructed using this approach in a previous iteration of the process). The action node in the modeling net (see Figure 9) is aimed at linking these two objects according to the belief in the existence of O_1 and O_2, the belief in their coplanarity and the belief in the existence of a junction between the two objects. Then, knowing the confidence in the existence of a new string and the probability of a closed string (another junction), it is possible to compute the probability that the resulting object is one string, a polygon, or composed of two independent strings of coplanar segments.

This three-step approach allows us to get a high level and more complete representation of the scene. Section 6 will present experimental results which illustrate the different key points of this algorithm. However, it is not yet possible to ensure that the model of the scene issued from this process is complete. Therefore, we present now the last perception-action cycle which includes the two previous ones and ensures an exploration as complete as possible.

5. Global exploration - complete scene reconstruction

This last perception-action cycle deals with the exploration of the scene. As already stated, the goal is to determine where the objects are and to ensure the completeness of the reconstruction (for all the most a reconstruction as complete as

possible). Previous works have been done in order to answer the "*where to look next*" question. Differences can be done 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 (Cowan and Kovesi 1988, Tarabanis et al., 1995a). 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 (Connolly, 1985, Maver and Bajcsy, 1993, Triggs and Laugier, 1995, Wixson, 1994, Reed et al., 1997).

As far as we are concerned, in the high level perception strategies of this reconstruction scheme, active vision is used to determine the camera position which provides the maximum of new information. Knowledge on 3D data previously gathered, and current 2D information are feedback into this exploration process. It allows us to determine the next object to be reconstructed or the next camera viewpoint. The "next best view problem" is handled as a function minimization scheme. We define a function to be minimized which integrates the constraints imposed by the system and evaluates the quality of the viewpoint. The resulting gaze planning strategy proposes a solution to the next best view problem that mainly uses a representation of known and unknown areas as a basis for selecting viewpoints.

5.1. Exploration Strategy

Let us consider a scene composed of a set $\mathcal{O}$ of initially unknown primitives. At the end of a local exploration process, a subset $\mathcal{O}(\mathcal{T}_0^t) \subseteq \mathcal{O}$ has been observed and reconstructed. Thus, we have to determine viewpoints able to bring more information about the scene. By *information*, we mean either a new object, either the certainty that a given area 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. If a new object is observed from the computed viewpoint, the local exploration process is used to estimate its structure.

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

- the objects already reconstructed: $\mathcal{O}(\mathcal{T}_0^t)$;
- the known free space, denoted $\mathcal{V}(\mathcal{T}_0^t)$. Indeed, it is possible to compute the area $\mathcal{V}(\phi)$ observed from position ϕ using a ray tracing scheme. Thus, knowing $\mathcal{T}_0^t$, we can determine the area $\mathcal{V}(\mathcal{T}_0^t)$ observed from the beginning of the reconstruction process ($\mathcal{V}(\mathcal{T}_0^t) = \bigcup_{i=0}^{t} \mathcal{V}(\phi_i)$).
- the unknown area $\mathcal{U}(\mathcal{T}_0^t)$. From the location of the reconstructed objects and the known free space, $\mathcal{U}(\mathcal{T}_0^t)$ is computed as:

$$\mathcal{U}(\mathcal{T}_0^t) = \overline{\mathcal{V}(\mathcal{T}_0^t) \cup \mathcal{O}(\mathcal{T}_0^t)} \tag{7}$$

We want to ensure the completeness of the reconstruction. This happens when $\mathcal{U} = \emptyset$. However, this condition is usually unreachable. Ensuring the completeness of the reconstruction is not always possible. Some areas may be observed only from a set of viewpoints unreachable by the camera. Furthermore, due to the objects topology, some areas may be unobserved whatever the camera position. We will see in Section 5.3 how we deal with the termination condition of the following exploration scheme.

5.2. Viewpoint Selection.

A simple strategy able to compute the "next best view" ϕ_{t+1} is to consider the viewpoint which maximizes the volume of the new observed areas (Connolly, 1985, Wixson, 1994). However, such a strategy does not take into account some problems such as the manipulator kinematics constraints or geometric constraints. As in (Tarabanis et al., 1995a, Triggs and Laugier, 1995), we have thus defined a function to be minimized which integrates the constraints imposed by the robotic system and evaluates the quality of the viewpoint. The function $\mathcal{F}$ to be minimized is taken as a weighted sum of a set of measures which determine the quality or the badness of a viewpoint. All the measures belongs to $[0, 1] \cup \infty$.

Quality of a new position The quality of a new position ϕ_{t+1} is defined by the volume of the unknown area which appears in the camera field of view. The new observed area $\mathcal{G}(\phi_{t+1})$ is given by (see Figure 10):

$$\mathcal{G}(\phi_{t+1}) = \mathcal{V}(\phi_{t+1}) - \mathcal{V}(\phi_{t+1}) \cap \mathcal{V}(\mathcal{T}_0^t) \qquad (8)$$

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 sub-part of $\mathcal{V}(\phi_{t+1})$ which has been already observed. If the position ϕ_{t+1} does not give any payoff in terms of information (*i.e.* $\mathcal{G}(\phi_{t+1}) = \emptyset$), we must reject this position. The measure of the quality of the position ϕ_{t+1} can thus be defined by:

$$g(\phi_{t+1}) = \begin{cases} \infty & \text{if } \mathcal{G}(\phi_{t+1}) = \emptyset \\ 1 - \frac{volume(\mathcal{G}(\phi_{t+1}))}{volume(\mathcal{V}(\phi_{t+1}))} & \text{otherwise} \end{cases} \qquad (9)$$

Remark: In fact, $\mathcal{G}(\phi)$ defines the potential volume of unknown area using only the current knowledge on the 3D scene. If a new object appears in the camera field of view, the new observed area is in fact smaller than the expected one ($\mathcal{G}'(\phi_t) \subseteq \mathcal{G}(\phi_t)$) but it is not actually a problem since the main goal of the application is to discover new objects.

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. It is defined using the following relation:

$$\mathcal{D}(\phi_t, \phi_{t+1}) = \frac{1}{N_{dof}} \sum_{i=1}^{N_{ddl}} \beta_i \frac{\mid q_{i_t} - q_{i_{t+1}} \mid}{\mid Q_{i_{Max}} - Q_{i_{Min}} \mid} \qquad (10)$$

where N_{dof} is the number of robot degrees of freedom, q_i is the position of the robot joint i and $\mid Q_{i_{Max}} - Q_{i_{Min}} \mid$ 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.

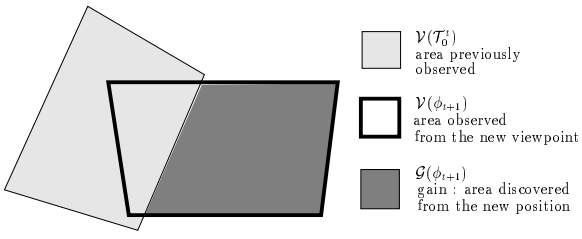

Fig. 10. Quality of a new position (2D projection).

Reachability Constraints. To avoid unreachable viewpoints, we use a binary test which returns an infinite value when the position is unreachable:

$$\mathcal{A}(\phi) = \begin{cases} 0 & \text{if } \phi \text{ is reachable} \\ \infty & \text{else} \end{cases} \qquad (11)$$

A position is unreachable if it is not in the operational space of the manipulator, or if this position is located in an unknown area (leading to a collision risk).

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

$$\mathcal{F}(\phi_{t+1}) = \mathcal{A}(\phi) + \alpha_1 g(\phi_{t+1}) + \alpha_2 \mathcal{D}(\phi_t, \phi_{t+1}) \qquad (12)$$

Here, the weights are predetermined in order to reflect the relative importance of the different measures. We think that the payoff in terms of information given by a new position is more important than the cost of the camera displacement. This implies a priority order of the coefficients α_i such that $\alpha_1 > \alpha_2$ (this can be done since these measures belong to $[0, 1]$, or are equal to the infinity, which means an automatic rejection of the viewpoint). More precisely, we have fixed $\alpha_1 = 0.6$ and $\alpha_2 = 0.4$.

We have decided to constrain the camera viewpoints inside an hemisphere located around the scene (assumed to be inside the hemisphere), but only in the region already observed and object-free (in order to avoid collision). At the beginning of the exploration process, as the observed area is null, the camera motion is limited to the surface of the sphere. To minimize $\mathcal{F}(\phi)$, a fast deterministic relaxation scheme (ICM algorithm) is used. Unlike stochastic relaxation methods such as simulated annealing, we cannot ensure that the global minimum of the function is reached. However, our method is not time-consuming and experimental results show that we always get a correct minimum in a low number of iterations. Furthermore, in our problem, finding the global minimum at each iteration of the exploration is not really necessary as long as the new viewpoint discovers a large part of the scene.

5.3. Achieving completeness: gazing on the regions of interest

At the end of the reconstruction process, some residual areas may remain unexplored. It is due to the fact that it is not possible to ensure that 100% of the scene has been observed: the topology of the objects, the kinematic constraints of the manipulator prevent observing the whole space. Thus, small parts of the scene usually remain unobserved. Second, the marginal gain of information decreases rapidly while the number of viewpoints increases. Thus, even if the whole scene is observable, the observation of the last residual areas requires a large number of viewpoints. For these different reasons, we decide to stop the exploration when a subset of the observable space has been really observed (typically, we define a threshold located around 95% of the observable space). However, it is necessary to verify that the remaining unobserved areas do not contain any objects (and if any to perform its reconstruction).

In a first time, we compute a segmentation of the residual areas considering the polyhedron incorporating a set of connected unobserved areas (defined here by small voxels). As the goal is to define small regions, if the volume of the computed polyhedron is too important, they are subdivided. In a second time, considering sequentially each sub-scene, a global exploration algorithm, restricted to the sub-scene, is then performed. Like in the case of the global exploration algorithm we consider that the camera motion is limited inside an hemisphere located around this polyhedron. The knowledge previously acquired is used in order to consider a collision avoidance process. This strategy allows to decrease significantly the number of viewpoints while increasing the part of observed areas.

If this strategy is more efficient, it remains that, as already stated, some parts of the scene can not be observed from any camera viewpoint. Therefore, for each sub-scene, we apply the following termination condition:

$$\forall \phi_{t+1}, \begin{cases} \mathcal{V}(\mathcal{T}_0^t) \cup \mathcal{V}(\phi_{t+1}) = \mathcal{V}(\mathcal{T}_0^t) \\ \mathcal{O}(\mathcal{T}_0^t) \cup \mathcal{O}(\phi_{t+1}) = \mathcal{O}(\mathcal{T}_0^t) \end{cases} \quad (13)$$

This means that the exploration process is as complete as possible if, for all reachable viewpoints, the camera looks at a known part of the scene. We thus can be sure that, at the end of the exploration process, all the areas of the scene are either free-space, either an object which has been reconstructed, either an unobservable area.

6. Experimental results

The whole application presented in this paper has been implemented on an experimental testbed composed of a CCD calibrated camera mounted on the end effector of a six degrees of freedom cartesian robot (see Figure 11).

6.1. Implementation.

Describing the implementation is not the goal of this paper ; however we want to underline the fact that, if it is important to bridge the gap between continuous/local and discrete/global aspects in the vision and control parts of an active vision system, it is also important to consider this gap from a software engineering point of view in order to obtain a safe and correct implementation of such system. As classical asynchronous languages are not really adapted to specify and program either the continuous and the discrete part of our system, we have implemented the control and structure estimation algorithms as well as the task controller (*i.e.*, the manager of the high level perception strategies) using SIGNAL. SIGNAL is a real-time synchronous data-flow language (Le Guernic et al., 1991) adapted to the implementation of vision-based tasks such as visual servoing and estimation (Marchand et al., 1997). Dealing with the high level perception action cycle, we have used SIGNAL *GTi*, an extension that introduces intervals of time, which provides constructs for the specification of hierarchical preemptive tasks executed on these intervals. It allows to consider in an unified framework the various aspects of the perception action cycle: from data-flow task (estimation, visual servoing) to multi-tasking and hierarchical task preemption (perception strategies).

The image processing part is implemented in C and performed on a commercial image processing board (Edixia IA 1000). It consists in tracking the projection of the selected straight line along the image sequence and in determining the (ρ, θ)

parameters describing the position of the line (two in the case of the optimal estimation of the cylinder) in the image. The extraction, maintenance and tracking of the contour segment (in fact a list of edge points) are achieved in 80 ms. The method we have used is described in (Boukir et al., 1998). It is based on a local and robust matching of the moving edge-points constituting the selected line (Bouthemy 1989).

6.2. *Structure from controlled motion.*

As already stated, we are interested in the reconstruction of cylinders and segments. We here present the results obtained for the structure estimation of a cylinder based on the projection of its two rims. Similar results are obtained for straight lines and thus for segments (Chaumette et al. 1996). In order to obtain a non-biased and robust estimation, the two rims of the cylinder must always appear centered and horizontal or vertical in the image sequence during the camera motion, which here consists in turning around the cylinder (see Figure 3.b). Figure 12.a represents the initial image acquired by the camera and the selected cylinder. Figure 12.b contains the image acquired by the camera after the convergence of the visual servoing task.

Figure 13 describes the evolution of the estimation of the parameters of the cylinder displayed in Figure 12. Figure 13.a shows its radius r and the coordinates x_0, y_0, z_0 of a point of its axis. Let us note that the cylinder radius $\widehat{r}$ is determined with an accuracy less than 0.5 mm whereas the camera

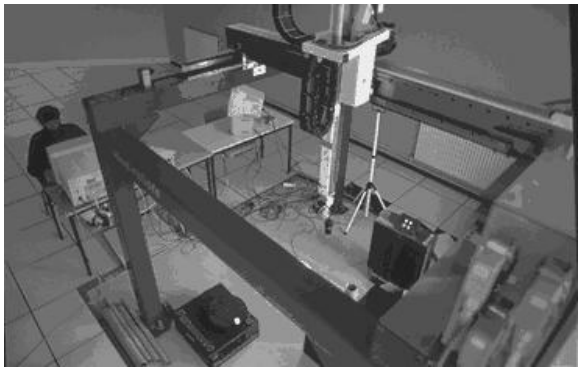

Fig. 11. Experimental cell (camera mounted on a 6 dof AFMA robot)

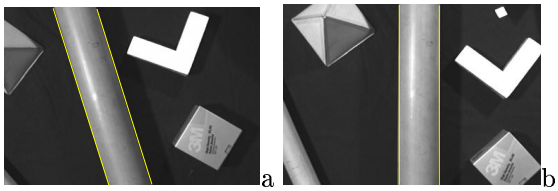

Fig. 12. Position of the cylinder in the image before (a) and after (b) gaze control task

is one meter away from the cylinder (and even less than 0.1 mm with good lighting conditions). Figure 13.b reports the error between the true value of the radius and its estimated value (*i.e.*, $r_i - r*$) using the two rims-based estimation. As far as depth $\widehat{z_0}$ is concerned, the standard deviation σ_{z_0} is less than 2.5 mm (that is 0.25%).

To show experimentally that the active structure from motion approach is really stable and robust, the estimation of the structure of a given cylinder has been carried out fifty times from different initial camera locations. For each one of the 50 experiments, we have computed the estimated radius $\widehat{r}$, and the estimated depth $\widehat{z_0}$. Each time, the measured error $\hat{r} - r^*$ is less than 0.5 mm and the standard deviation of all the estimations (*i.e.*, $\sigma_{\widehat{r}}$) is around 0.02 mm (resp. $\sigma_{\widehat{z_0}} = 0.23$ mm). These results underline the fact that our estimation algorithm is particularly robust, stable and accurate.

6.3. *From a local to a global description of the scene.*

We present in this section the reconstruction results obtained for a polyhedral object (see Figure 4). This object allows us to illustrate the interesting points of the proposed method. Even if the images are well contrasted, this object is quite complex since some segments are too small to be accurately estimated, some are occluded and some have non-trivial geometric characteristics (one of the object angles has been cut (see Figure 4.b)).

Figure 14.a shows the first image of this object acquired by the camera and Figure 14.b to 14.f depict the view of the scene after the reconstruction of each segment. Dashed lines represent the segments previously reconstructed ; others correspond to primitives for which structure is still unknown. Arrows point to the next segment S_i to be reconstructed. Each time, the parameters

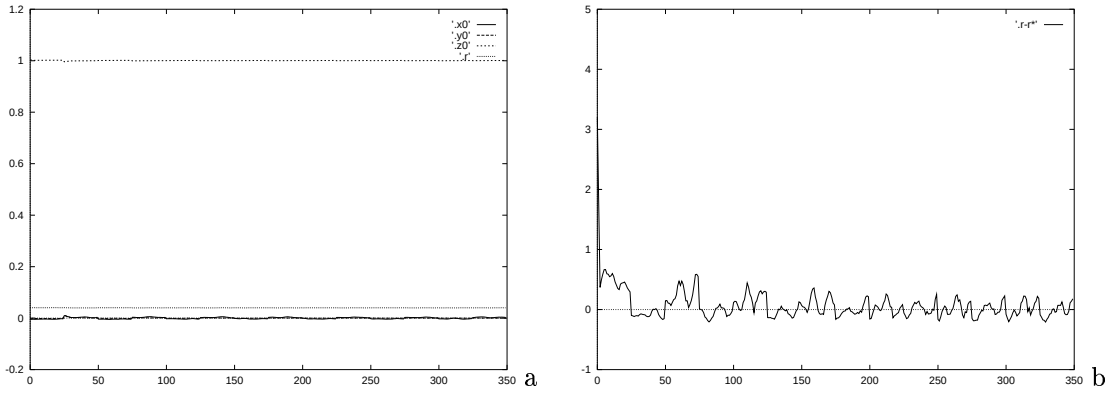

Fig. 13. Estimation of the parameters of a cylinder: (a) estimated position of a point on the axis (x_0, y_0, z_0) and radius r (in mm), (b) error between the real and estimated radius of the cylinder (in mm)

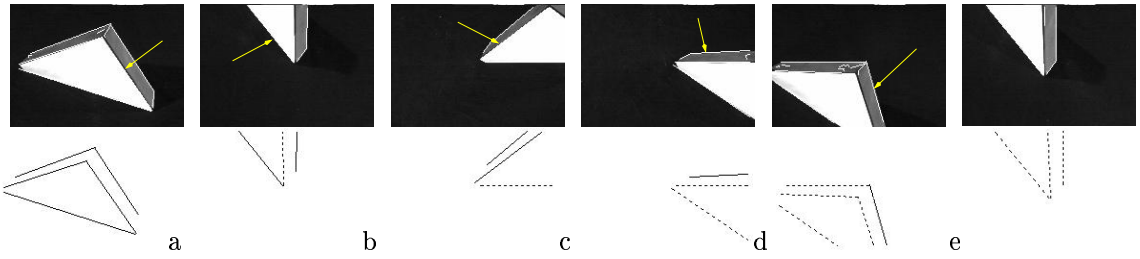

Fig. 14. Polyhedral object: the top line shows the acquired images and the bottom line shows the corresponding lists of segments (dashed lines represent the segments previously reconstructed).

which describe its structure are estimated using the structure from controlled motion approach. This explains that the last reconstructed segment always appears vertical or horizontal centered in the image. Finally, a numbering of the segments in the order of their introduction in the 3D map of the scene is presented on Figure 15.

We will not describe the whole process of this object reconstruction using the Bayes nets predic-

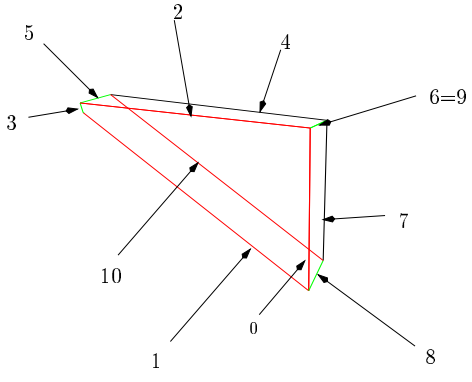

Fig. 15. Reconstructed scene and numbering of the reconstructed segments

tion/verification scheme (see (Marchand, 1996a) for a complete description). Let us only focus on two aspects of this algorithm.

Interest for multiple hypotheses. Suppose that segments S_0 and S_1 have been already reconstructed. Just after the reconstruction of S_2 (see Figure 14.abc), the system considers the relation between S_2 and S_0 and between S_2 and S_1. Dealing with the segments S_2 and S_0, the system easily deduces the presence of a junction between these two segments. This leads to the creation of a string of segments (S_2, S_0, S_1). Let us now consider the case of the couple (S_1,S_2). These two segments are close in the 3D space (there is around 1cm between their closest extremities). The certainty for S_2 and S_1 to be neighbor is 61% and coplanar is 99% ; thus they are likely to belong to the class $\mathcal{C}_2$. According to the strategies encoded in the Hypotheses Bayes net, the two main hypotheses are that a junction exists with a 46% belief and a segment between (41%). The remaining 13% are shared between the two other hypotheses.

After the verification process, and according to the observations, the former hypothesis (junction) is verified with a 60% belief. This high value (even if this hypothesis is false, see Figure 4.b) results from the fact that these two segments are very close in the different images (around 5 pixels). Thus the observations reinforce this hypothesis. However, the latter hypothesis (that it exists a segment between S_2 and S_1) is verified with a 95% belief. A 2D segment is observed at the predicted position in many images. Finally, according to the belief in each hypothesis, to the belief in the observations, a new segment S_3 is added to the model of the scene (with a confidence of 53%, while the confidence in a junction creation is only 37%). This underlines the interest to consider a multi-hypotheses approach. A classical approach might have chosen the first (and wrong) hypothesis.

Verification actions. Let us consider a second interesting case. When segment S_7 has just been reconstructed, the system concludes that a junction with S_4 exists. However, when observing such a pair of segments, a hypothesis corresponding to the creation of a segment between the two distant extremities is also done. This segment does not appear in any of the previously acquired images. However, knowing the position of the previously reconstructed polygons and the previous location of the camera, the system concludes that, due to occlusions, this segment could never have been observed. So, in order to verify this hypothesis, a verification action corresponding to a motion of the camera is made. As described in Section 4.2, the camera gazes on S_7, and turns around it (see Figure 16). During this motion, automatically generated by visual servoing, observers are looking for moving edges located at the computed segment position in the images. The structure of the discovered segment is then estimated and introduced in the scene model (see Figure 16.c).

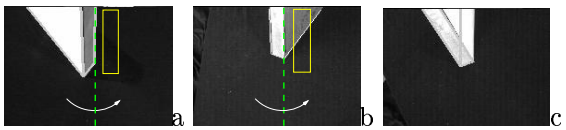

Fig. 16. Verification of a hypothesis: (a) rotation around S_7 (b) S_{10} is discovered and (c) reconstructed

6.4. Scene exploration - computing viewpoints.

Let us consider now a similar but more complex scene (see Figure 17). Using the incremental reconstruction scheme, a large part of the scene can be reconstructed without explicitly computed viewpoints (see Figure 17d-e). However, large unobserved areas (here, more than 44% of the scene), one segment of the concave polyhedron and the little polygon have not been yet reconstructed (see Figure 17f).

In order to ensure the completeness of the reconstruction, the system computes new viewpoints using the algorithm described in Section 6. A few viewpoints are computed before observing the last polygon. However, if no new object appears in these views, they allow the system to check that the corresponding part of the scene is object-free. After 7 viewpoints, the polygon appears in the camera field of view (see Figure 17.b). During the reconstruction of this polygon, the wide segment at the bottom of the big rectangle, which was occluded from the previous viewpoints, is observed (see Figure 17.c) and then reconstructed. Then, the camera gazes on the remaining unobserved area and a new exploration is performed (6 viewpoints) until 99.2 % of the scene is observed. Since the 0.8% remaining residual area are detected to be inside the polyhedron and thus unobservable, the reconstruction is as complete as possible.

Note that more than 6000 images have been automatically acquired and real-time processed in order to get the complete model of that scene. This underlines the robustness of our system.

Other Results. The last example (see Figure 18.a) deals with a scene composed of a cylinder and five polygons which lie in different planes. In Figure 18.b is displayed the initial image acquired by the camera. Only the cylinder and a polygon have been reconstructed during the first local incremental reconstruction process described in Section 4 (see Figure 18.c).

Figure 18.d-g presents the different steps of the global exploration of the scene. Each figure shows the obtained 3D scene, the camera trajectory and the projection on a virtual plane of the unknown areas. Figure 18.d corresponds to the camera position ϕ_6 obtained just after the first incremental re-

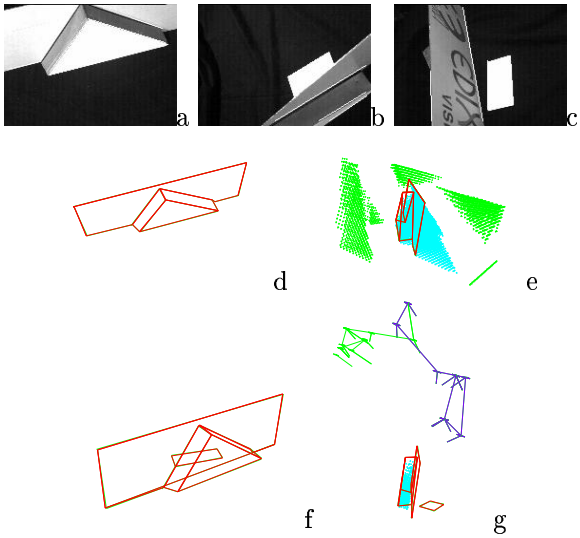

Fig. 17. Polyhedral scene reconstruction: (a) First image of the scene acquired by the camera, (b) detection of a new polygon during the exploration process, (c) detection of the last primitive, (d) Reconstructed model of the scene after the first local exploration process, (e) unobserved and occluded areas (56% of the scene has been already observed), (f) final reconstructed model of the scene, (g) camera trajectory and residual area (99.2% of the scene has been observed).

construction. The first camera displacements significantly reduces the unknown areas. At position ϕ_{13} (see Figure 18.e), a new object is observed and a reconstruction process is thus performed. It ends at position ϕ_{24} (Figure 18.f). At this step, the two polygons on the "top" of the scene have been reconstructed. After a new exploration process, the last polygon is reconstructed and the camera is located in ϕ_{30} (Figure 18.g). At this step, 99% of the space has been observed, which ensures that the reconstruction of the scene is complete. Figure 18.h shows the final 3D model of the scene (to be compared to Figure 18.a) and the camera trajectory.

More results on scene exploration (dealing with the influence of weight α_i in (12), and with the algorithm to gaze on residual areas) can be found in (Marchand and Chaumette 1996b, Marchand, 1996a).

6.5. 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 shapes. 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. 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 at a subsample of video rate (120ms). Robust real-time tracking algorithms in noisy environments are not yet available (recent work such as XVision (Hager and Toyama, 1998) tries to cope with these problems). Therefore, we have restricted ourselves to simple and well contrasted images. We hope in the future to address the interesting issue of the reconstruction of complex shapes.

Encoding strategies. One of the main reported drawback of Bayesian approaches is usually the ability of the conditional probabilities tables to reflect the available knowledge. In our case, we have tested our system with different values in these

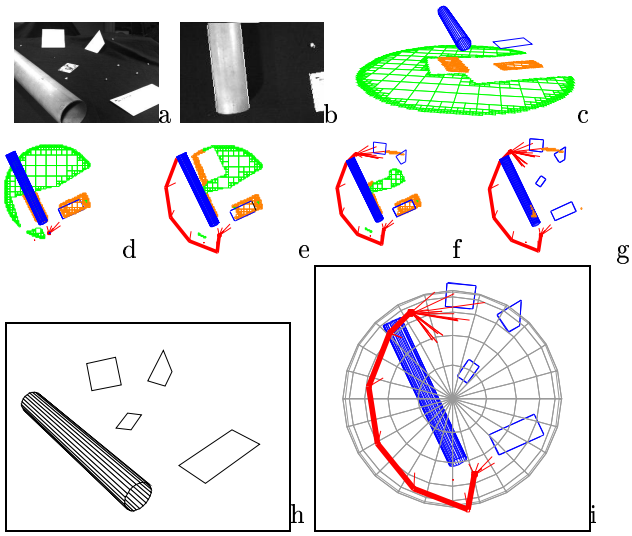

Fig. 18. Cylinder and polygons reconstruction (a) External view, (b) first view of the scene and results of the first incremental reconstruction process, (c) reconstructed scene and projection on a virtual plane of the unknown area, (d-g) different steps of the global exploration process (camera trajectory, 3D model of the reconstructed scene and projection on an virtual plane of the unknown area), (h) 3D model of the reconstructed scene, (i) polar view of the camera trajectory.

tables. Important variations (around ± 0.10) do not greatly affect the behavior of the system. The main reason is that the two most likely hypotheses are always verified. Therefore, using such Bayes nets provides a robust behavior. However, let us note that in many cases, when a more important knowledge about the task is available, a training process is always possible in order to learn the *a priori* probabilities from a set of examples.

Furthermore, in our case, excepting our assumption that the scene is composed only of segments and cylinders, we do not have any *a priori* knowledge on the scene. Others hypotheses may have been introduced in the system (*e.g.*, "the probability for a junction to be a triple junction is high"). However, introducing such hypotheses seems neither essential nor justified, since this kind of hypotheses must be chosen with respect to their relevance to the considered application.

Combinatorics of the system. The last scene (see Figure 18) allows us to illustrate the low combinatorics of our system. In theory, the complexity increases with the number of segments since, for each new segment, the relations with all those already reconstructed have to be considered. However, in practice, only the segments which are not too far away are considered. This simple heuristic allows the system to deal with distant objects independently. Furthermore, only the hypotheses with a belief greater than a given threshold are taken into account (10% typically), which avoids the verification of highly unlikely hypotheses. Finally, we can recall that the knowledge introduced in the Bayes nets implies that there is no relation (or a complex one) between segments which are not coplanar. No verification has thus to be performed for such pairs of segments. Even for scenes composed of several complex objects, all these considerations make the number of effective relations decrease rapidly with respect to the number of segments, as well as the number of verifications performed.

7. Conclusion and Future Works

We have presented a system able to perform the 3D reconstruction of unknown static scenes using a single mobile camera. Although the considered scenes and images are simple, the system is able to perform a complete and very precise reconstruction of these scenes using image acquired and processed at nearly video rate. To achieve this task, we have considered an active vision paradigm. Let us emphasize the fact that this paradigm has been used at each level of our system. Our work confirms earlier ideas related to active vision.

- Reconstruction is done using a "structure from controlled motion" approach. The method we have used relies on the analysis of the motion of the object in the images and on the measure of the camera velocity. In order to optimize the quality of the reconstruction, the camera motion has to be controlled. Visual servoing appears to be a very efficient way to achieve this motion. Unfortunately, the main advantage of this method, accuracy, leads to its main drawback: it requires specific camera motions. Thus, it has been necessary to develop perceptual strategies able to appropriately perform a succession of such individual primitive reconstructions.
- To this purpose, an algorithm has been proposed to ensure the incremental reconstruction of the scene. It is based on the use of a prediction/verification scheme managed using decision theory and Bayes nets. The main goal of this prediction/verification scheme was to bridge the gap between a representation in term of isolated primitives and a representation in term of objects. Sub-goals were to deal with small segments and to propose a partial solution to the occlusion problem.
- The last problem was to ensure the completeness of the reconstruction. To this purpose, we have proposed perception strategies able to automatically compute new viewpoints which give new information about the organization of the scene.

Moreover, the various components of the system have been integrated in a purposive way. Indeed, the task was defined as a global goal while the data used to achieved it were local (*i.e.*, some segments extracted from the image sequence). We have shown that to achieve such a complex goal, various perception-action cycles have to be considered: from the high level scene exploration cycle

downto the low-level visual servoing cycle. It is important to note that if the notion of *feedback* is the key point of these cycles, the information used in and the actions resulting from these processes are very different. Data used to close the loops can be local (*e.g.*, features extracted from the images sequence, joints encoding, etc) or global (*e.g.*, 3D location of objects, computed free space, etc) leading to various possible kind of action such as camera motions (continuous or discrete), images acquisition, etc. We see the integration of this different cycles as a very important issue in our system.

An active reconstruction scheme, as proposed here, may be used for map creation or scene inspection for applications where a full autonomy and a great precision are necessary. Let us examine how the various methods presented in this paper can be used for different applications.

- It is not necessary to emphasize the various applications of visual servoing. We have used visual servoing to perform gazing tasks, to constrain the camera motion along a given trajectory with respect to a geometrical primitive, to deal with mechanical constraints such as joint limits. Other applications are hand-eye coordination (*e.g.*, grasping), dynamic sensor planning, target tracking, etc.
- Dealing with the reconstruction problem itself, we plan to extend the system for the reconstruction of complex cylindrical environment, as can be found in nuclear environment. Furthermore, we also plan to deal with active reconstruction of more complex shapes. For this purpose, other strategies, such as those proposed in (Kutulakos and Dyer, 1994), have to be defined and developed.
- Dealing with the Bayes nets scheme, this approach can be used to guide the camera motion during the reconstruction of far more complex shapes in order to cope with problems such as occlusions or modification in the object topology. Obviously, the Bayes nets will have to be modified for these tasks. As the structure of the nets are assumed to reflect the knowledge on the task, specific nets have to be designed for each application.
  However, if the method can not scale directly to any kind of application, the methodology itself based on the use of three nets (dedicated to the prediction, the verification and the modeling) can be used for various kind of applications such as information gathering, visual search, or object recognition (Djian et al., 1995) (hypothesis on the location of a given feature can be done and verification action can be performed). Furthermore in many cases, the structure of the verification net may remain unmodified and can be used in many applications.
  More generally, this methodology can be applied to various purposes. The approach proposed here is interesting for various reasons. A Bayes net can be seen as a controller to manage a set of simple vision tasks. Furthermore, their ability to deal with uncertainty confers to the system adaptability and robustness.
- The exploration algorithm has been tested using the structure from controlled motion reconstruction method presented in this paper. However, it can be used with any kind of reconstruction scheme such as stereovision or laser range finder. Constraints introduced in the optimization function $\mathcal{F}$ can be changed function of the sensor or robot characteristics (*e.g.,* scanning or tolerance constraints in the case of a laser range finder (Reed et al., 1997), overlap constraints in the case of a mobile robot).

Acknowledgements

This work was partly supported by the MESR (French Ministry of the University and Research) within project VIA (*Vision Intentionnelle et Action*) and under contribution to a student grant.

References

Adiv, G. 1989. Inherent ambiguities in recovering 3D motion and structure from a noisy flow field. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11(5):477–489.

Aloimonos, Y. Weiss, I. and Bandopadhay, A. 1987. Active vision. *International Journal of Computer Vision*, 1(4):333–356.

Aloimonos, Y. 1990, Purposive and qualitative active vision. In *IAPR Int. Conf. on Pattern Recognition, ICPR'90*, Atlantic City, New Jersey, vol. 1, pp. 346–360.

Bajcsy, R. 1988. Active perception. *Proceedings of the IEEE*, 76(8):996–1005.

Ballard, D.H. 1991. Animate vision. *Artificial Intelligence*, 48(1):57–86.

Boukir, S., Bouthemy, P., Chaumette, F. and Juvin, D. 1998. A local method for contour matching and its parallel implementation, *Machine Vision and Application*, 10(5):321–330.

Bouthemy, P. 1989. A maximum likelihood framework for determining moving edges. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11(5):499–511.

Buxton, H. and Gong, S. 1995. Visual surveillance in a dynamic and uncertain world. *Artificial Intelligence*, 78(1-2):431–459.

Chaumette, F., Boukir, S., Bouthemy, P. and Juvin, D. 1996. Structure from controlled motion. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(5):492–504.

Chien, C. and Aggarwal, J.K. 1989. Model construction and shape recognition from occluding contour. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11(4):372–389.

Connolly, C. 1985. The determination of next best views. In *IEEE Int. Conf. on Robotics and Automation*, St Louis, pp. 432–435.

Cowan, C.K. and Kovesi, P.D. 1988. Automatic sensor placement from vision task requirements. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 10(3):407–416.

Djian, D., Probert, P. and Rives, P. 1995. Active sensing using bayes nets. In *Proc. of Int. Conf. on Advanced Robotics, ICAR'95*, San Feliu de Guixols, Spain, pp. 895–902.

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

Espiau, B. and Rives, P. 1987. Closed-loop recursive estimation of 3D features for a mobile vision system. In *IEEE Int. Conf. on Robotics and Automation*, Raleigh, North Carolina, vol. 3, pp. 1436–1443.

Hager, G.D. and Toyama, K. 1998. The XVision system: A general-purpose substrate for portable real-time vision applications. *Computer Vision and Image Understanding*, 69(1):23–37.

Hashimoto, K. editor. 1993. *Visual Servoing : Real Time Control of Robot Manipulators Based on Visual Sensory Feedback*. World Scientific Series in Robotics and Automated Systems, Vol 7, World Scientific Press, Singapor.

Hutchinson, S., Hager, G. and Corke, P. 1996. A tutorial on visual servo control. *IEEE Trans. on Robotics and Automation*, 12(5):651–670.

Hutchinson, S. and Kak, A. 1989. Planning sensing strategies in a robot work cell with multi-sensor capabilities. *IEEE Trans. on Robotics and Automation*, 5(6):765–783.

Krause, P. and Clark, D. 1993. *Representing uncertain knowledge: artificial intelligence approach*. Kluwer Academic Publishers, Oxford, United Kingdom.

Kutulakos, K. and Dyer, C. 1994. Recovering shape by purposive viewpoint adjustment. *International Journal of Computer Vision*, 12(2):113–136.

Le Guernic, P., Le Borgne, M., Gautier, T. and Le Maire, C. 1991. Programming real time application with SIGNAL. *Proceedings of the IEEE*, 79(9):1321–1336.

Marchand, E. 1996a. *Stratégies de perception par vision active pour la reconstruction et l'exploration de scènes statiques*. PhD thesis, Université de Rennes 1, IRISA.

Marchand, E. and Chaumette, F. 1996b. Controlled camera motions for scene reconstruction and exploration. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition, CVPR'96*, San Francisco, California, pp. 169–176.

Marchand, E., Chaumette, F. and Rizzo, A. 1996c. Using the task function approach to avoid robot joint limits and kinematic singularities in visual servoing. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, IROS'96*, Osaka, Japan, vol. 3, pp. 1083–1090.

Marchand, E., Rutten, E. and Chaumette, F. 1997. From data-flow task to multi-tasking: Applying the synchronous approach to active vision in robotics. *IEEE Trans. on Control Systems Technology*, 5(2):200–216.

Marr, D. 1982. *-Vision- A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman, San Francisco, California.

Maver, J. and Bajcsy, R. 1993. Occlusions as a guide for planning the next view. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(5):417–433.

Neapolitan, R. 1990. *Probabilistic Reasoning in Expert Systems*. John Wiley, Chichester.

Pearl, J. 1988. *Probabilistic reasoning in intelligent systems : Networks of plausible inference*. Morgan Kaufmann Publisher Inc., San Mateo, California.

Reed, M.K., Allen, P.K. and Stamos, I. 1997. Automated model acquisition from range images with view planning. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition, CVPR'97*, pp. 72–77, Puerto Rico.

Rimey, R.D. and Brown, C. 1991. Controlling eye movements with hidden markov models. *International Journal of Computer Vision*, 7(1):47–65.

Rimey, R.D. and Brown, C. 1994. Control of selective perception using bayes nets and decision theory. *International Journal of Computer Vision*, 12(2/3):173–207.

Samson, C., Le Borgne, M. and Espiau, B. 1991. *Robot Control: the Task Function Approach*. Clarendon Press, Oxford, United Kingdom.

Swain, M.J. and M.A. Stricker, M.A. 1993. Promising direction in active vision. *International Journal of Computer Vision*, 11(2):109–127.

Tarabanis, K., Allen, P.K. and Tsai, R. 1995a. A survey of sensor planning in computer vision. *IEEE Trans. on Robotics and Automation*, 11(1):86–104.

Tarabanis, K., Tsai, R. and Allen, P.K. 1995a. The MVP sensor planning system for robotic vision tasks. *IEEE Trans. on Robotics and Automation*, 11(1):72–85.

Triggs, B. and Laugier, C. 1995. Automatic camera placement for robot vision. In *IEEE Int. Conf. on Robotics and Automation*, Nagoya, Japan, vol. 2, pp. 1732–1738.

Weng, J., Huang, T.S. and Ahuja, N. 1990. Estimation and structure from line matches: Performance obtained and beyond. In *IAPR Int. Conf. on Pattern Recognition, ICPR'90*, Atlantic City, New Jersey, vol. 1, pp. 168–172.

Wixson, L. 1994. Viewpoint selection for visual search. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition, CVPR'94*, Seattle, Washington, pp. 800–805.

Xie, M. and Rives, P. 1989. Toward dynamic vision. In *IEEE Workshop on Interpretation of 3D Scenes*, Austin, Texas, pp. 91–99.