Int. Workshop on Algebraic Frames for the Perception-Action Cycle,
AFPAC'97, LNCS 1315, pages 290–305, Kiel, Germany, September 1997.

Visual perception strategies for 3D reconstruction

Éric Marchand and François Chaumette

IRISA - INRIA Rennes
Campus de Beaulieu
35 042 Rennes cedex, France

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 a robust estimation of 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 complex scenes. 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 complete 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.

1 Active Vision to handle the perception action cycles

Most of the approaches proposed to solve vision problems are inspired from the Marr paradigm 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 Aloimonos, Bajcsy, or Ballard (among others) have proposed to modify the Marr concept. They proposed a new paradigm named **active vision**. 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 might 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, either the processing resources allocated to the system [20].

However, this is a general definition of active vision and the different authors who had introduced this concept had different motivations. What is usually called active vision can be divided into four main classes : the *active vision* introduced by Aloimonos [3] is a mathematical analysis of complex problems such as stability, linearity and uniqueness of solutions ; the goal of *active perception* as defined by Bajcsy [4] is to elaborate control strategies for setting sensor parameters in order to improve the knowledge of the environment. Thus, active vision is defined as an intelligent data acquisition process ; the *animate vision* [5] 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 a second hand to facilitate the computational process ; dealing with *purposive vision* [2], the goal 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 they can handle either uncertainty and/or dynamic environment) and more accurate (because they are able to modify their 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 information is used to control the camera motion, and camera motions 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.*, images handled 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 this goal are usually local. The relation between the global modeling of the goal and this set of local sub-model (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. However 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 be handled in real time using continuous schemes. In that case, information must be seen as an infinite flow of data acquired by the sensor. In an 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.).

Let us now consider how this methodology has been applied to the scene reconstruction and exploration problem. Our concern is to deal with the problem of recovering the 3D spatial structure of a whole scene without any knowledge on 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 system is described using a hierarchical parallel automata (see Fig. 1). It has three main **perception-action cycles**. The main one is the exploration-reconstruction cycle which ends only when the reconstruction is complete. This cycle deals with global information and the resulting camera motions are discrete. However, when an object is observed, the system enters in the second cycle which is the incremental reconstruction loop. The main goal of this level is to bridge the gap between a local modeling of the scene and a global one. The latter cycle deals with the active reconstruction itself which is here intrinsically based on a local/continuous approach. There, for each observed segment, a recognition task is performed in order to determine the nature of the primitive (cylinder or segment). Then, if a cylinder has been recognized, an estimation of its parameters based on its two limbs is performed in order to get a more robust reconstruction. Finally in both cases, we have to compute the length of the primitive. In parallel to the reconstruction tasks, due to the camera motion, occlusions and manipulator joint limits avoidance tasks are realized. Let us now examine the various issues raised by this reconstruction problem.

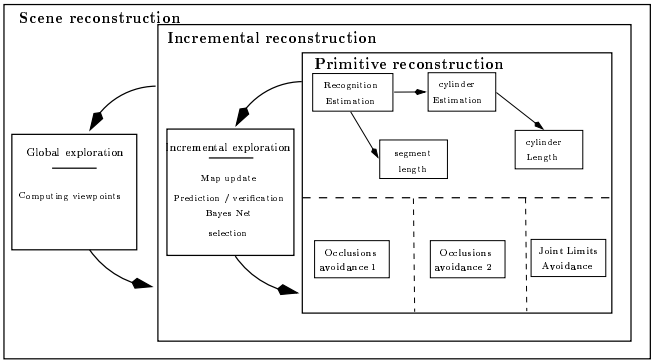


Fig. 1. Hierarchical parallel automata describing the whole reconstruction process

Exploration - Complete reconstruction. The first issue deals with the exploration. 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 [9,21,8,22,23]. As far as we are concerned, in the high

level perception strategies of our reconstruction scheme, active vision is used to determine the camera position which provides the maximum of new information (Section 4). 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. We have chosen to handle this problem as a function minimization problem. We define a function to be minimized which integrates the constraints imposed by the system and evaluates the quality of the viewpoint. When an object is observed, the exploration process ends and its reconstruction is realized.

Primitive reconstruction and camera motion generation. The approach we have chosen to get an accurate three-dimensional geometric description of a scene is based on a continuous structure from motion approach [7]. Very noticeable improvements can be obtained in the 3D reconstruction if the camera viewpoint is properly selected and if adequate camera motions are generated (Section 2). These motions are generated using the visual servoing approach [11,13].

This approach has many advantages. First of all, visual servoing allows to generate automatically the camera motions defined for an optimal estimation of the primitives. To this purpose, we define a secondary task such as a trajectory tracking in "parallel" with a priority task (*e.g.*, gaze control). Second, as the camera motions are dedicated to the estimation of one primitive, we have only one features to track in the images sequence. Therefore, we are able to perform a real time estimation of the primitives parameters. Furthermore, as we used a continuous structure from motion approach, the motion of the primitive in the image is very small during the estimation since the primitive must remain at a constant position in the image. Therefore, the spatio-temporal matching process is quite straightforward and can be handled in real time. This real time computation of the camera motion allows us to deal on-line with some other constraints such as occlusions and kinematics problems specific to the manipulator.

Incremental reconstruction. However, since the camera motion is controlled for the estimation of one primitive at a time, this implies to successively focus on each primitive of the scene, using a **local exploration** algorithm. The proposed method is based on a prediction/verification scheme (Section 3). Bayes nets [18,19,6,10] seem to be well adapted to manage this process. They allow us to model "expert" reasoning. Furthermore, 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. This algorithm proposes a partial solution to the occlusion problem and allows us to obtain a high level description of the scene. This way, we can bridge the gap between a set of local sub-models (obtained using a continuous method) and a global model of the scene (thus obtained using a discrete method).

The remainder of this paper is organized as follows. Section 2 is devoted to the local aspect of our reconstruction scheme and describes the structure from motion framework based on an active vision paradigm. Section 3 describes 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 4 where the computing viewpoint issue used to ensure a reconstruction as complete as possible of the scene is proposed. Finally, Section 5 presents experiments carried out on a robotic cell which have demonstrated the validity of our approach.

2 3-D structure estimation using active dynamic vision

The measure of the camera motion, which is necessary for the 3D structure estimation, characterizes a domain of research called dynamic vision. The method used here is a continuous approach [1,12] which stems on the measure of the camera velocity and of the motion of the considered primitive in the image. 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 [7].

For most of the geometrical primitives, it is possible to determine the interaction matrix $L_{\underline{P}}^T$ defined by the classical equation [11]:

$$\dot{P} = L_P^T(P, p_l)T_c \tag{1}$$

where $\dot{P}$ is the time variation of P due to the camera motion T_c. The parameters P describe the position of the object in the image while the parameters p_l describe the position of the object limb surface (*i.e.*, for a volumetric primitive, it defines the 3D surface in which the limbs lie).

From the resolution of a linear system derived from (1), it is possible to obtain the parameters p_l [7]. Then, using geometrical constraints related to the considered primitive, we can estimate the parameters $\underline{p}$ which fully define its 3D configuration.

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. An active vision approach is thus necessary to improve the accuracy of the results by generating adequate camera motions. In fact, two main results dealing with this problem have been achieved [7]:

1. A sufficient and general condition that suppresses the discretization error is to constrain the camera motions such that the projection of the primitive must be kept constant in the image and no variation must occur on the limb surface parameters during the camera motion (*i.e.*, $\dot{P} = \dot{p}_l = 0, \forall t$).
2. A more robust estimation with respect to measurement errors is obtained if the location between the camera and the primitive is considered. Indeed, particular positions of the primitive in the image do minimize the influence of the measurements errors. Thus, 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 realized (see Fig. 2).

For example, in the case of a cylinder, it has been shown that the optimal camera motion is such that the cylinder constantly appears as two static centered, vertical or horizontal, straight lines in the image sequence. Visual servoing [11,13],

which is the main point of low level perception-action cycles, is very well qualified to control camera motions in order to satisfy these constraints.

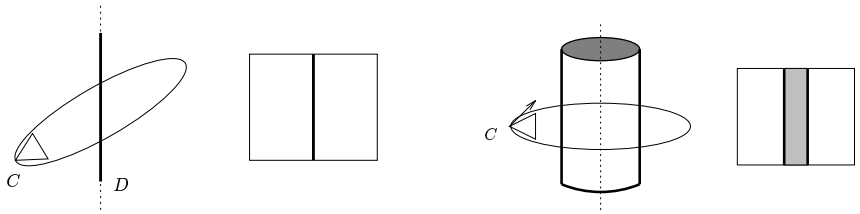


Fig. 2. Optimal camera motion and resulting image in the cases of a straight line and a cylinder

3 A Bayes-Nets Based Prediction / Verification Scheme

The next level of our reconstruction scheme is the incremental reconstruction of the objects observed from a computed viewpoint. In order to obtain as accurate results as possible, we have chosen to perform the reconstruction in sequence. The resulting algorithm [16] allows us to perform an estimation of all the primitives which appear in the camera field of view. However:

- The description of the scene is a low level and local description which only contains a list of 3D segments and cylinders. It might be more interesting to get high level global information such as junctions, polygons, and faces.
- The scene reconstruction is incomplete for two main reasons. First, the projection in the image of some segments have a too small length to make their reconstruction possible. Second, as this algorithm only deals with the observed objects, it has a local perception of the scene. According to this, some objects may not appear in the camera field of view (because of occlusions or because they are located in an unknown and unobserved area).

To cope with these problems, we propose a Bayes Nets based prediction/verification scheme. A Bayes Net [18] 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 conceptor of the application. The advantages of Bayes Nets lies in the ability to reflect the *a priori* knowledge available on the application. This knowledge is reflected 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. The knowledge is also present in the conditional probability tables associated with the variables of the net. These tables reflect the expert reasoning as well as the uncertainty associated with the

observations. Finally, the propagation allows to take each new observation into account. The influence of an observation is propagated to the other variables of the net according to the causality relations.

The goal of our prediction/verification scheme is to determine the relations between reconstructed 3D segments and to infer either the presence of new segments, either the existence of more complex objects. As our reconstruction scheme is incremental, we determine the consequence of the introduction of a new segment S_t in the 3D map of the scene as soon as the structure of S_t is known. Our approach can be decomposed into three steps. For each couple of segments $(S_{t'}, S_t)$, 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.

Prediction Dealing with two segments $S_{t'}$ and S_t, the possible actions are the followings: fuse the segments, create a junction, or add a link (a new segment) between $S_{t'}$ and S_t. Therefore the aim of the prediction step is to create some hypotheses leading to the realization of one (or more) of these actions. 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 ;
- H_4: there are no (or some other) relation between $S_{t'}$ and S_t.

We have a multi-step strategy. First, we compute the belief we have in simple topological relations (proximity ($p(N)$), coplanarity ($p(C)$), collinearity ($p(P)$)) between $S_{t'}$ and S_t. Then, according to these beliefs, it is possible to classify the pair of segments into five classes (see the first raw of the Table in the Fig. 4). Classes are $\mathcal{C}_1$: CNP (coplanar, neighbor and parallel) , $\mathcal{C}_2$: $CN\neg P$, $\mathcal{C}_3$: $C\neg NP$, $\mathcal{C}_4$: $C\neg N\neg P$, and $\mathcal{C}_5$: $\neg C\neg N\neg P$.

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 $P(\mathcal{H}|C)$ where $\mathcal{H}$ is the hypothesis and C the class (see Fig. 4). These tables are defined in an empirical way from 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 guarantee. However, extreme precision is not required. Rather, they must reflect the knowledge we want to transmit to the system.

The prediction step reasoning can be encoded in a simple Bayes Net (see Fig. 3.a). It is composed of six nodes. Links between these nodes depict the causality relations between the different steps of reasoning and thus its progression. One node is associated to each topological relation, another to the class, and one node is associated to each set of hypotheses. Indeed, two sets of hypotheses are emitted. The first concerns the relation between the closest extremities

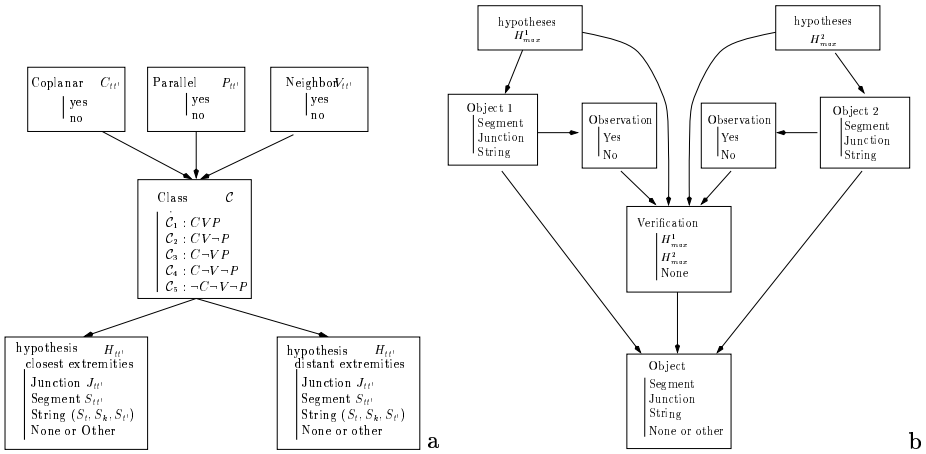
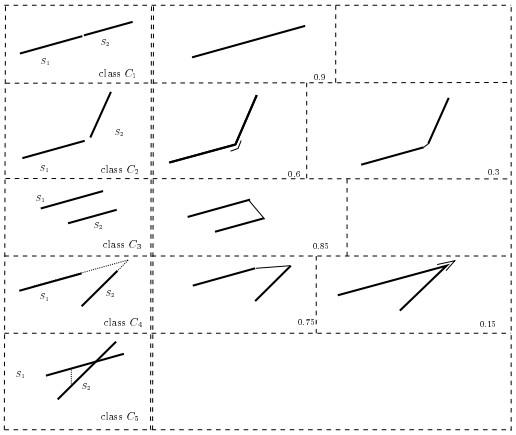


Fig. 3. (a) Prediction net and (b) Verification net



	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

closest 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

distant extremities

Fig. 4. Elementary classes and associated hypothesis (closest extremities) and Conditional probabilities table P(hypotheses | classes) for the closest and distant extremities

of the segments (see Fig. 4) and the second concerns the relation between their distant extremities. In both cases, the same hypotheses can be emitted, though the associated conditional probabilities can be very different. 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.

Verification In order to verify the two selected hypotheses, we use the reasoning encoded in the Bayes Net depicted in Fig. 3.b. We use two similar nets, each associated with one of the two sets of hypotheses (*i.e* close and distant extremities). 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 using the observation node. 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).

The most important node in the verification net is the observation node. Sometimes, the hypotheses can be verified (or invalidated) using direct observation in the images previously acquired. In such cases, the 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 or of a junction is simple. If this junction exists, it has already been observed (because the presence of the two segments has been already verified). Thus, the verification is performed as described above. In the case of 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 the 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 necessary to move the camera to a new viewpoint from which the segment can be observed. Rather than computing explicitly a viewpoint (*e.g.* [9,21]) and researching *off-line* the considered segment, we prefer to turn the camera around a segment which belongs either to the occluding polygon or to a plane to which the considered segment belongs. During this motion, automatically generated by visual servoing [11], an image processing is performed *on-line* to detect the appearance of the researched segment.

Modeling. At this step of the reconstruction process, we have a model of the scene composed of 3D segments, 3D junctions, or even a coplanar string of segments. It is finally quite easy to use this information in order to get 3D polygons. To this end, we use the junction information and the coplanarity information already used in the hypotheses generation (see [14] for further details).

This three-step approach allows us to get a high level and more complete representation of the scene. Section 5 will present experimental results which illustrate the different key points of this algorithm.

4 Global exploration - complete scene reconstruction

Since it is not possible to ensure that the model of the scene issued from the local exploration process is complete, we present now the last perception-action cycle which includes the two previous ones and ensures a reconstruction as complete as possible. 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.

Knowing the set of viewpoints since the beginning of the reconstruction process, it is possible to maintain a map of the observed and unexplored areas using a ray tracing scheme. The knowledge is composed by: the objects already reconstructed $\mathcal{O}$, the known free space $\mathcal{V}$, and the unknown area $\mathcal{U}$. Using this knowledge, we have defined a gaze planning strategy which 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. We have chosen to handle the "where to look next" problem as a function minimization problem. Such a function $\mathcal{F}(\phi)$ has to integrate the constraints imposed by the system and to evaluate the quality of a viewpoint in order to select the next camera viewpoint ϕ_{t+1} which corresponds to its minimal value. The cost function is minimized using a fast deterministic relaxation scheme corresponding to a modified version of the ICM algorithm. The camera viewpoints are constrained inside an hemisphere located around the scene, 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.

The function $\mathcal{F}$ is taken as a weighted sum of the following measures:
Quality of a new position. The quality of a new position ϕ_{t+1} is function of 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:

$$\mathcal{G}(\phi_{t+1}) = \mathcal{V}(\phi_{t+1}) - \mathcal{V}(\phi_{t+1}) \cap \mathcal{V} \tag{2}$$

where $\mathcal{V}(\phi_{t+1})$ defines the part of the scene observed from the position ϕ_{t+1} and $\mathcal{V}(\phi_{t+1}) \cap \mathcal{V}$ defines the sub-part of $\mathcal{V}(\phi_{t+1})$ which has been already observed (see Fig. 5).
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 (see [16]).
Reachability Constraints. To avoid unreachable viewpoints, we use a binary test which 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 area (leading to a collision risk).

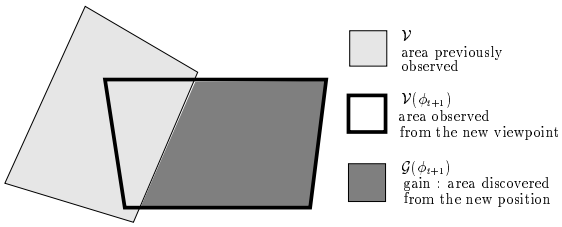


Fig. 5. Quality of a new position (2D projection).

If a new object is observed from the selected viewpoint, its reconstruction is performed. On the other hand, if the observed space is free of new objects, a next viewpoint has to be computed. This process is iterated until the end of the exploration. In theory, it must end when all the space has been observed, *i.e.*, if $\mathcal{U} = \emptyset$. However, this condition is usually unreachable. Ensuring the completeness of the reconstruction is not always possible. Some areas may remain 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. Thus the exploration process is said to be 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.

5 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. Describing the complete implementation of our system 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 in the vision/control part 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 safety 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 algorithm, we have implemented the control/estimation algorithm and the task controller using SIGNAL [17]. SIGNAL is a real-time synchronous data-flow language adapted to implementation of vision-based tasks such as visual servoing and estimation. Dealing with the high level PAC, we have used SIGNALGTi, an extension that introduces intervals of time. SIGNALGTiprovides 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).

Structure from controlled motion. As already stated, we are interested in the reconstruction of cylinders and segments. We here presents the results obtained for the structure estimation of a cylinder. Fig. 6.a represents the initial image acquired by the camera and the selected cylinder. Fig. 6.b contains the image acquired by the camera after the convergence of the visual gazing task.

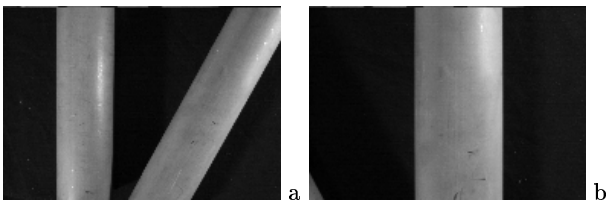


Fig. 6. Position of the cylinder in the image before (a) and after (b) the focusing task

Fig. 7 describes the evolution of the estimation of the parameters of the cylinder displayed in Fig. 6. Fig. 7.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 is determined with an accuracy less than 0.5 mm whereas the camera is one meter away from the cylinder (and even less than 0.1 mm with good lighting conditions). Fig. 7.b reports the error between the estimated value of the radius and its true value (*i.e.*, $r_i - r*$) using the two limbs-based estimation. As far as depth is concerned, its standard deviation is less than 2.5 mm (that is 0.25%).

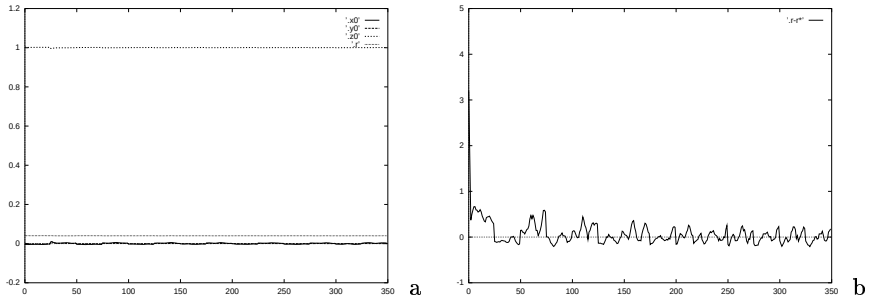


Fig. 7. Estimation of the parameters of a cylinder in the camera frame (a) estimated position of a point on the axis (x_0, y_0, z_0) and radius (r) (in mm) (b) error between the real and estimated radius of the cylinder (in mm)

The experiment of the cylinder structure estimation 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 $\widehat{r} - r^*$ 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 particulary robust, stable and accurate.

From a local to a global description of the scene. We present in this section the reconstruction results obtained for a polyhedral object (see Fig. 8.a). This scene allows us to illustrate the interests of the proposed method. The result of the scene reconstruction using the simple incremental reconstruction process is depicted on Fig. 8.b. As already stated, as they are too small, some of the vertices of the polyhedron have not been reconstructed. Furthermore, due to the local approach used in that process, others remain occluded and thus non reconstructed. We now focus on two aspects of the Bayes Nets prediction verification scheme.

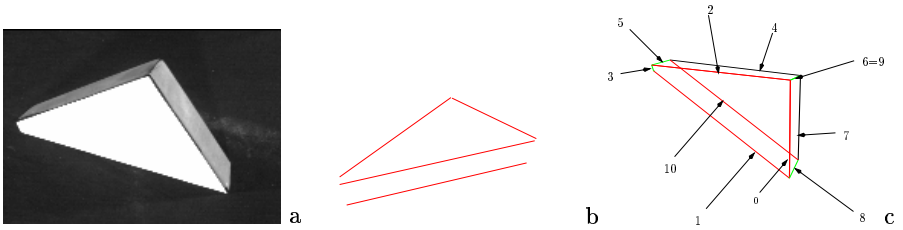


Fig. 8. Polyhedral scene: (a) view of the scene (b) model of the "polyhedron" scene acquired using the incremental algorithm (c) model of the same scene acquired using the prediction/verification scheme and numbering of the reconstructed segments in the order of their introduction in the map of the scene

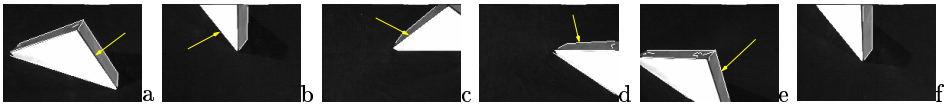


Fig. 9. Polyhedral scene : arrows point at the next primitive to be estimated

Consider that segments S_0 and S_1 have been already estimated and that S_2 has just been reconstructed (see Fig. 9.abc), the system considers the relation between S_2 and S_0 and between S_2 and S_1. Dealing with S_2 and S_0, the system concludes easily to the presence of a junction between them. Dealing with the couple (S_1,S_2), there is around 1cm between their closest extremities. The belief for S_2 and S_1 to be neighbor is 61% and to be 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, it is likely that there exists a junction with a 46% belief and a a segment between them with a 41% belief. 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 Fig. 8.a) 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 is verified with a 95% belief. Indeed, 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.

Let us now consider a second interesting case. When segment S_7 is reconstructed, within relations with other segments, the system proposes the creation of a junction with S_4 and the creation of a segment between their two distant extremities. Such a segment has never been observed (and could not have been observed according to the current knowledge on the scene and on the camera trajectory). Therefore, as described in the previous section, the camera focuses on S_7, and turns around it (see Fig. 10). During this motion, automatically generated by visual servoing, observers are looking for a moving segment located at its expected position in the images. The discovered segment is then reconstructed and introduced in the scene model (see Fig. 10.c).

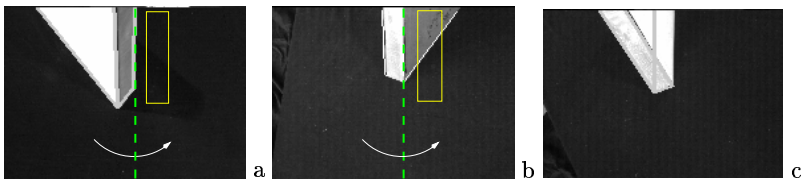


Fig. 10. Verification of a hypothesis: (a) rotation around S_7 (b) S_{10} is discovered and (c) reconstructed

Scene exploration - computing viewpoints. The example reported here (see Fig. 11.a) deals with a scene composed of a cylinder and four polygons which lie in different planes. In Fig. 11.b is displayed the initial image acquired by the camera. Only the cylinder and a polygon are reconstructed during the first local incremental reconstruction process (see Fig. 11.c). Fig. 12 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. Fig. 12.a corresponds to the camera position ϕ_6 obtained just after the local exploration process. The first camera displacements allows to reduce significantly the unknown areas (see Fig. 12.b). A new object is then detected. A local exploration process is performed. It ends at position ϕ_{24} (Fig. 12.c). At this step, the two polygons on the "top" of the scene have been reconstructed. A new global exploration is then performed. After a last exploration process, the last polygon is observed and reconstructed (Fig. 12.d). At this step, 99% of the space has been observed, which ensures that the reconstruction of the scene is

complete. Fig. 11.d shows the final 3D model of the scene (to be compared to Fig. 11).

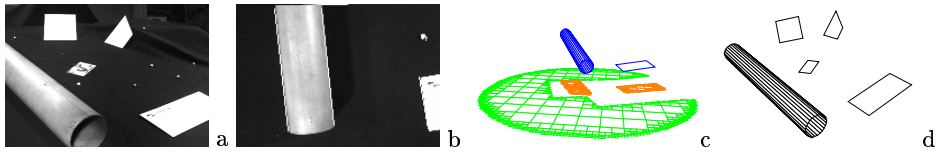


Fig. 11. (a) External view and (b) first view of the scene and results of the first local exploration/incremental reconstruction process : (c) reconstructed scene and projection on a virtual plane of the unknown area (d) 3D model of the reconstructed scene.

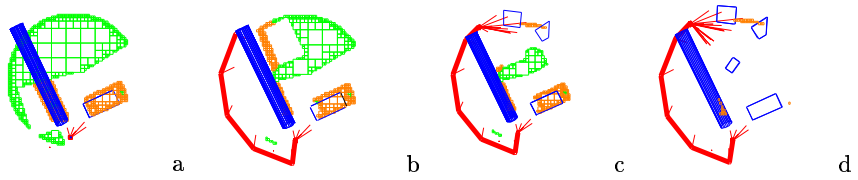


Fig. 12. Different steps of the global exploration process (camera trajectory, 3D model of the reconstructed scene and projection on an virtual plane of the unknown area).

6 Conclusion

We have proposed an active vision approach to the 3D reconstruction of static scenes composed of cylinders and polyedral objects. 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. As the structure from controlled motion approach used to perform primitives estimation is based on particular camera motions, perceptual strategies able to appropriately perform a succession of such individual primitive reconstructions have been proposed in order to recover the complete spatial structure of complex scenes. 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. To this purpose, an algorithm has been proposed to ensure the incremental reconstruction and exploration of the scene. It is based on a computing viewpoints algorithm and the use of a prediction/verification scheme managed using decision theory and Bayes Nets. Finally, experiments have proved the validity of our approach (accurate, stable and robust results with efficient exploration algorithms).

References

1. G. Adiv. Inherent ambiguities in recovering 3D motion and structure from a noisy flow field. *IEEE Trans. on PAMI*, 11(5):477–489, May 1989.
2. Y. Aloimonos. Purposive and qualitative active vision. In *ICPR'90*, pp. 346–360, Atlantic City, June 1990.
3. Y. Aloimonos, I. Weiss, A. Bandopadhay. Active vision. *IJCV*, 1(4):333–356, January 1987.
4. R. Bajcsy. Active perception. *Proc. of the IEEE*, 76(8):996–1005, August 1988.
5. D.H. Ballard. Animate vision. *Artificial Intelligence*, 48(1):57–86, February 1991.
6. H. Buxton, S. Gong. Visual surveillance in a dynamic and uncertain world. *Artificial Intelligence*, 78(1-2):431–459, October 1995.
7. F. Chaumette, S. Boukir, P. Bouthemy, D. Juvin. Structure from controlled motion. *IEEE Trans. on PAMI*, 18(5):492–504, May 1996.
8. C. Connoly. The determination of next best views. In *IEEE Int. Conf. on Robotics and Automation*, pp. 432–435, St Louis, March 1985.
9. C.K. Cowan, P.D. Kovesi. Automatic sensor placement from vision task requirements. *IEEE Trans. on PAMI*, 10(3):407–416, May 1988.
10. D. Djian, P. Probert, and P. Rives. Active sensing using bayes nets. In *Proc. of Int. Conf. on Advanced Robotics, ICAR'95*, pp. 895–902, Sant Feliu de Guixols, Spain, September 1995.
11. B. Espiau, F. Chaumette, P. Rives. A new approach to visual servoing in robotics. *IEEE Trans. on Robotics and Automation*, 8(3):313–326, June 1992.
12. B. Espiau, P. Rives. Closed-loop recursive estimation of 3D features for a mobile vision system. In *IEEE Int. Conf. on Robotics and Automation*, pp. 1436–1443, Raleigh, April 1987.
13. S. Hutchinson, G. Hager, P Corke. A tutorial on visual servo control. *IEEE Trans. on Robotics and Automation* 12(5):651–670, October 1996.
14. E. Marchand. *Stratégies de perception par vision active pour la reconstruction et l'exploration de scènes statiques.* PhD thesis, Université de Rennes 1, June 1996.
15. E. Marchand, F. Chaumette. Controlled camera motions for scene reconstruction and exploration. In *CVPR'96*, pp. 169–176, San Francisco, June 1996.
16. E. Marchand, F. Chaumette, A. Rizzo. Using the task function approach to avoid robot joint limits and kinematic singularities in visual servoing. In *IROS'96*, pp. 1083–1090, Osaka, Japan, November 1996.
17. E. Marchand, E. Rutten, F. Chaumette. 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, Mars 1997.
18. J. Pearl. *Probabilistic reasoning in intelligent systems : Networks of plausible inference.* Morgan Kaufmann Publisher Inc., San Mateo, California, 1988.
19. R.D. Rimey, C. Brown. Control of selective perception using bayes nets and decision theory. *IJCV*, 12(2/3):173–207, April 1994.
20. M.J. Swain, M.A. Stricker. Promising direction in active vision. *International Journal of Computer Vision*, 11(2):109–127, October 1993.
21. K. Tarabanis, P.K. Allen, R. Tsai. A survey of sensor planning in computer vision. *IEEE Trans. on Robotics and Automation*, 11(1):86–104, February 1995.
22. B. Triggs, C. Laugier. Automatic camera placement for robot vision. In *IEEE Int. Conf. on Robotics and Automation*, pp. 1732–1738, Nagoya, Japon, May 1995.
23. L.E. Wixson. Viewpoint selection for visual search. In *CVPR'94*, pp. 800–805, Seattle, June 1994.