Visual data fusion for objects localization by active vision

Grégory Flandin, François Chaumette

Visual Data Fusion for Objects Localization by Active Vision

Grégory Flandin and François Chaumette

IRISA/INRIA Rennes
Campus de Beaulieu
35042 Rennes cedex, France
Francois.Chaumette@irisa.fr
http://www.irisa.fr/vista

Abstract. Visual sensors provide exclusively uncertain and partial knowledge of a scene. In this article, we present a suitable scene knowledge representation that makes integration and fusion of new, uncertain and partial sensor measures possible. It is based on a mixture of stochastic and set membership models. We consider that, for a large class of applications, an approximated representation is sufficient to build a preliminary map of the scene. Our approximation mainly results in ellipsoidal calculus by means of a normal assumption for stochastic laws and ellipsoidal over or inner bounding for uniform laws. These approximations allow us to build an efficient estimation process integrating visual data on line. Based on this estimation scheme, optimal exploratory motions of the camera can be automatically determined. Real time experimental results validating our approach are finally given.

1 Overview

Whatever the application, a main issue for automated visual systems is to model the environment: it must have a suitable knowledge representation in order to perform efficiently the assigned task. In the context of robot vision, most papers deal with 3D reconstruction and focus on modeling accuracy. Classically, this is done either considering geometric objects (in that case, techniques are based on primitive reconstruction [6,12]) or representing the shapes thanks to meshes [3] or using an exhausting voxel representation of the scene (see [16] for a recent survey of methods involving sample representations), eventually reducing the complexity by means of hierarchical techniques like octrees. But, for several kinds of applications such as path planning, obstacle avoidance or exploration, only a preliminary 3D map of the scene is sufficient. Moreover, very fast algorithms are requested to perform on line computation. This motivated our work about coarse model estimation. This notion was previously studied by Marr [13]. Coarse models are strongly related to the assigned task so that we need to introduce an other concept concerning the kind of object we treat of. It is not restrictive at all to consider that any scene can be partitioned in coherent groups or parts of objects. The term coherent should be understood in the sense that

A. Heyden et al. (Eds.): ECCV 2002, LNCS 2353, pp. 312–326, 2002.

the group or part can be significantly described by a single including volume according to the assigned task. Typically, a set of close objects is coherent when observed from a sufficiently far distance but is incoherent when observed closely enough to treat each object as a coherent one. Even an isolated object may appear incoherent when constituted of inhomogeneous parts. Each part should be described individually when necessary. By misuse of language we will simply call "object" each coherent group or part of physical objects. This paper is dedicated to the description, the estimation and the refinement of objects' including volume. This volume is defined by an ellipsoidal envelope. Thus, if describing a strongly concave physical object by an ellipsoid appears obviously incoherent for the assigned task, this object will need to be partitioned in coherent parts. However, this point is out the scope of this paper. Besides, the image processing comes down to the extraction of an ellipse including the segmented mask of the object in the image. As a consequence, our technique can theoretically apply to any kind of object provided that it can be segmented. In practice, to deal with general scenes with no constraint on the object aspect, we make the only assumption that there is a depth discontinuity at the frontier of the objects so that a motion segmentation algorithm will give the mask of the objects.

The first part of our work focuses on the description of the including volume of an object (center and envelope). The method we developed stems for the class of state estimation techniques. Typically, the problem of parameter and state estimation is approached assuming a probabilistic description of uncertainty. In order to be compared and fused, observations are expressed in a common parameter space using uncertain geometry [2,5]. But in cases where either we do not know the associated distribution or it is not intrinsically stochastic, an interesting alternative approach is to consider unknown but bounded errors. This approach, also termed set membership error description, has been pioneered by the work of Witsenhausen and Schweppe [20,15]. But, in this method, the observation update needs the calculus of sets intersection. A computationally inexpensive way to solve the problem is to assume that error is bounded by known ellipsoids [11]. Mixing probability and set membership theories in a unified stochastic framework, we will take advantage of both representations in order to model the center and envelope of objects. This model is all the more interesting that it enables, for each point of the scene, the calculation of its probability to belong to a given object.

Once a suitable model is available, a common issue is to wonder which movements of the camera will optimally build or refine this model. In a general case, this is referred to optimal sensor planning [17]. When a coarse or fine reconstruction of the scene or objects is in view, we will speak about exploration. It is said autonomous when the scene is totally or partially unknown. In this context, previous works have adopted different points of view. In [4], Connolly describes two algorithms based on the determination of next best views. The views are represented by range images of the scene and the best one tends to eliminate the largest unseen volume. In [19], Whaite and Ferrie model the scene by superquadrics. The exploration strategy is based on uncertainty minimization

and the sensor is a laser range finder. Kutulakos, Dyer and Lumelsky [9] exploit the notion of the occlusion boundary that are the points separating the visible from the occluded parts of an object. Lacroix and Chatila [10] developed motion and perception strategies in unknown outdoor environments by means of either a laser range finder or stereo cameras. A search algorithm provides an optimal path among a graph. This path is analyzed afterwards to deduce the perception tasks to perform. Let us mention the recent work of Arbel and Ferrie [1] about view point selection. Even if it deals with the quite different problem of object recognition, the approach is interesting. It is based on the off-line calculation of an entropy map which is related to ambiguity of recognition. Marchand and Chaumette [12] use controlled motion of a single camera to explore geometrical objects such as polygons and cylinders. The viewpoint selection is achieved minimizing a cost function. In our case where location is modeled by a gaussian distribution and shape by an ellipsoid, the exploration concept must be seen as a way to improve localization (see Figure 1-a). The exploration is optimal when the convergence rate of the estimated location is the best we can do. The strategy we develop thus consists in reducing uncertainty of the distribution associated with the observed object using visual data. The gaussian modeling of uncertainty and a linearization of the visual acquisition process allow us to build analytical solutions to optimal exploration.

In Section 2, we precisely describe the model of an object as a mixture of stochastic and set membership models. This model is seen as a probability density called set distribution. In Section 3, we define rules that makes propagation of a set distribution possible. These rules are applied to the propagation of visual data. Multiple images of a same object can then be compared and fused. In Section 4, we describe an estimation process for static objects which is based on camera motion. In the context of exploration, the camera motion has to be defined. With this aim in view, an optimality criterion and two associated exploratory control laws are examined in Section 5.

2 Modeling

In this part, we define a simple and suitable representation of an object Let us consider the 3D coordinates c of a point belonging to an object. We choose to break down c into the sum of a mean vector $\overline{c}$ and two independent random vectors:

$$c = \overline{c} + p + e \tag{1}$$

In (1), $\overline{c}$ represents the location of the center of gravity of the object, p the uncertainty on this location, and the bounds on the error e define the including volume of the object (see Figure 1-b). We distinguish between uncertainty (modeled by p) and error (modeled by e). Indeed, uncertainty belongs to the class of random models whereas error belongs to the class of set membership models (e is uniformly distributed on a bounded set denoted $\mathcal{V}$). We now recall the assumptions made for future developments and whose motivations were given in introduction:

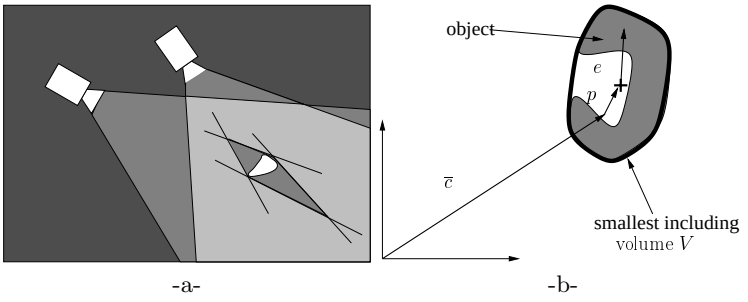

Fig. 1. Modeling principle and definitions

1. we first assume that p follows a zero mean normal distribution whose variance is P. This is a quite good approximation of most of the uncertainty sources such that camera localization and calibration, or image processing. It also makes the propagation of the laws easier. To deal with partial unobservability, we will use the information matrix $\Sigma = P^{-1}$ instead of P. An infinity variance along the inobservability axis is then replaced by a null information. The associated distribution is denoted $\mathcal{N}(0, \Sigma)$.
2. The mathematical representation of the volume $\mathcal{V}$ needs to be simple otherwise its manipulation becomes prohibitive. We decide to approximate $\mathcal{V}$ by ellipsoids and remind that an ellipsoid is completely defined by its quadratic form: its center and a positive definite matrix E. In the sequel and by misuse of language, an ellipsoid will be denoted by its matrix E (since its center is given by $\overline{c}$).

Our work will thus focus on the estimation and refinement of $\overline{c}$, Σ and E.

With a more formal point of view, let us denote $\mathcal{O}$ an object of the scene (denoted $\mathcal{S}$). The coordinates of a point $c \in \mathcal{O}$ can be seen as a random vector whose distribution is, for every $x \in \mathcal{S}$, $\mathcal{P}(c = x)$ denoted $\mathcal{P}_c(x)$. Modeling $\mathcal{S}$ comes down to finding for each $\mathcal{O}$ a suitable distribution to model the density function of $\mathcal{P}_c(x)$. From previous assumptions, the global distribution associated with an object is completely defined by $\overline{c}$, Σ and E. More precisely, it is the distribution of the sum of independent variables, that is the convolution product of a uniform distribution $\mathcal{U}_E$ on $\mathcal{V}$ by a normal one $\mathcal{N}(\overline{c}, \Sigma)$. We call this distribution a set distribution and we denote

$$\mathcal{E}(\overline{c}, \Sigma, E) = \mathcal{N}(\overline{c}, \Sigma) * \mathcal{U}_E$$

– **Remark:** Let us derive the relation between $\mathcal{P}_c(x)$ and the probability $\mathcal{P}(x \in \mathcal{O})$ that a point $x \in \mathcal{S}$ belongs to $\mathcal{O}$:

$$\mathcal{P}_c(x) = \mathcal{P}_c(x|x \in \mathcal{O}).\mathcal{P}(x \in \mathcal{O}) + \mathcal{P}_c(x|x \notin \mathcal{O}).\mathcal{P}(x \notin \mathcal{O})$$

$\mathcal{P}_c(x|x \notin \mathcal{O})$ is the probability that a point $c \in \mathcal{O}$ is at x knowing that $x \notin \mathcal{O}$, it is obviously null. $\mathcal{P}_c(x|x \in \mathcal{O})$ is the probability that a point $c \in \mathcal{O}$

is at x knowing that $x \in \mathcal{O}$. We naturally model it by a uniform law whose value can be calculated after normalization. Indeed, for every $\mathcal{O}$:

$$\int_{\mathcal{S}} \mathcal{P}_c(x|x \in \mathcal{O})dx = \mathcal{P}_c(x|x \in \mathcal{O})Volume(\mathcal{O}) = 1$$

As a consequence $\mathcal{P}_c(x) = \mathcal{P}(x \in \mathcal{O})/Volume(\mathcal{O})$. Thus $\mathcal{P}(x \in \mathcal{O})$ is related to a set distribution up to a scale factor. For every $x \in \mathcal{S}$, its probability to belong to an object can thus be computed. This possibility is very interesting for path planning tasks and to quantify the collision risk.

In order to apply our model to the particular case of visual data, we identify three stages in the chain of visual observation (see Figure 2):

1. In the image, the measure is a 2D set distribution $\mathcal{E}^i(\overline{c^i}, \Sigma^i, E^i)$ where $\overline{c^i}$ and E^i represent the center and matrix of the smallest outer ellipse including the projection of the object in the image. This projection must be extracted by segmentation algorithms (see Section 6). Besides Σ^i must account for all sources of uncertainty such that camera intrinsic parameters or image processing (see Section 3.2).
2. The process transforming the 2D visual data in a 3D observation is called back-projection. The associated 3D set distribution is denoted $\mathcal{E}^c(\overline{c^c}, \Sigma^c, E^c)$. This leads us to distinguish between the **measure** related to the 2D information in the image and the **observation** which is the associated back-projection.
3. At last, we must express every observation in a common frame called the reference frame ($\mathcal{R}$). The associated observation is denoted $\mathcal{E}^o(\overline{c^o}, \Sigma^o, E^o)$.

Thanks to these successive transformations, visual measurements can be compared and fused. In order to define the influence of such transformations on a set distribution, the next section is dedicated to the description of propagating rules.

3 Propagating Rules

In Section 3.1, we define general rules applying to any kind of transformation. In Section 3.2, we specialize them to the case of visual transformations. Every rule is given without proof. More details on the demonstrations are given in [8].

3.1 General Transformations

Rule 1 (Transformation of a set distribution)
Let c be a random vector following a set distribution $\mathcal{E}(\overline{c}, \Sigma, E)$. As a first order approximation, if we denote $J = \frac{\partial T^{-1}}{\partial c}\Big|_{\overline{c}}$ the jacobian of T^{-1}, the transformed random vector $c' = T(c)$ follows a set distribution $\mathcal{E}'(\overline{c'}, \Sigma', E')$ where

$$\overline{c'} = T(\overline{c}), \quad \Sigma' = J^T \quad \Sigma \quad J \quad \text{and} \quad E' = J^T \quad E \quad J$$

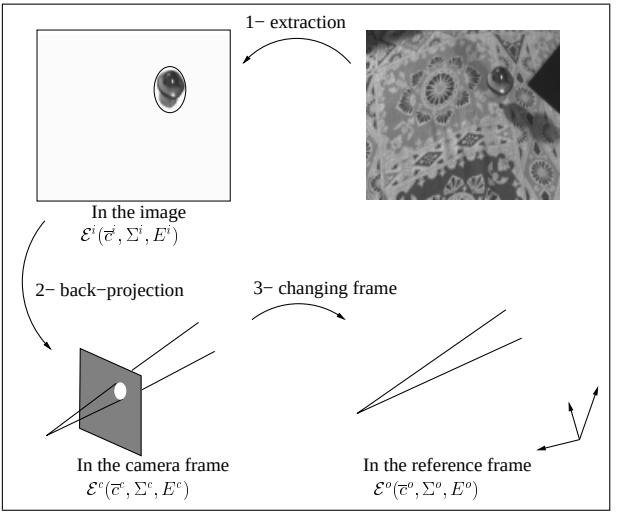

Fig. 2. Successive transformations of measurements.

In this rule, T^{-1} denotes the inverse transformation of T which implicitly requires T to be a diffeomorphism. When it is not the case (by instance for perspective projection from 3D points to 2D image points), we need a rule dedicated to projection on a subspace:

Rule 2 (Projection of a set distribution on a subspace)
Let $c = (c1, c2)^T$ *be a random vector following a set distribution*

$$\mathcal{E}(\begin{pmatrix} \overline{c_1} \\ \overline{c_2} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix}, \begin{pmatrix} E_{11} & E_{12} \\ E_{12}^T & E_{22} \end{pmatrix})$$

the projected random vector $c1$ *follows a set distribution* $\mathcal{E}'(\overline{c'}, \Sigma', E')$ *where*

$$\overline{c'} = \overline{c_1}, \quad \Sigma' = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T \quad \textit{and} \quad E' = E_{11} - E_{12}E_{22}^{-1}E_{12}^T$$

When T depends on external parameters p_e ($c' = T(c, p_e)$ where p_e is $\mathcal{N}(\overline{p_e}, P_e)$), we can linearize the transformation around $\overline{c}$ and $\overline{p_e}$ to take into account other sources of uncertainty such as the location of the camera or calibration errors.

3.2 Specialized Rules

Measure in the Image. First of all, the projection of each object in the image must be extracted. We will see in Section 6 how we achieve this task in practice. Then $(\overline{c^i}, E^i)$ represents the center and matrix of the smallest outer ellipse in the image. They can be extracted thanks to algorithms like the one proposed in [18]. Σ^i must account for all sources of uncertainty that may occur in the calculus of this ellipse: errors on camera intrinsic parameters and inaccuracy of image

processing. If we denote c^i the coordinates in meter, c^p the coordinates in pixels and p_{int} the vector of intrinsic parameters, the pixel/meter transformation is $c^i = T(c^p, p_{int})$. If we assume that c^p is $\mathcal{N}(\overline{c^p}, P^p)$ and p_{int} is $\mathcal{N}(\overline{p_{int}}, P_{int})$ then we can write:

$$P^i = J_{c^p} P^p J_{c^p}^T + J_{p_{int}} P_{int} J_{p_{int}}^T$$

where $J_{c^p} = \frac{\partial T}{\partial c^p}\Big|_{\overline{c^p}, \overline{p_{int}}}$ and $J_{p_{int}} = \frac{\partial T}{\partial p_{int}}\Big|_{\overline{c^p}, \overline{p_{int}}}$ The gaussian assumption we make for intrinsic parameters can appear not very realistic since it looks nearer to a bias. But we will attach importance to the choice of a sufficiently large P_{int} so that the gaussian model can take small bias into account. We will model P_{int} and P^p by diagonal matrices, considering that the uncertainties are not correlated. Since P^p is related to image processing uncertainty (mainly due to segmentation), an estimate is not easy to derive. We will fix its diagonal values to sufficiently large variances that will account for most segmentation uncertainties.

Back-projection. Back-projection strongly depends on the camera configuration. In the **monocular configuration**, because of partial inobservability, $\mathcal{E}^c$ is degenerated (see Figure 2). Let us denote $c^c = (x^c, y^c, z^c)^T$. To account for the inobservability of z^c, we increase the random vector c^i adding artificially the independent measure z^c whose distribution is $\mathcal{E}(\overline{z^c}, 0, 0)$. In other words, we allocate a value for z^c with a null confidence. Then we show:

Rule 3 (Back-projection of a set distribution)
The back-projection c^c of c^i follows a set distribution:

$$\begin{cases} \overline{c^c} = (\overline{z^c}\overline{X^i}, \overline{z^c}\overline{Y^i}, \overline{z^c})^T \\ \Sigma^c = J^T \begin{pmatrix} \Sigma^i & 0 \\ 0 & 0 \end{pmatrix} J \\ E^c = J^T \begin{pmatrix} E^i & 0 \\ 0 & 0 \end{pmatrix} J \end{cases} \quad \text{where } J = \frac{\partial T^{-1}}{\partial c}\Big|_{\overline{c}} = \begin{pmatrix} 1/\overline{z^c} & 0 & -\overline{X^i}/\overline{z^c} \\ 0 & 1/\overline{z^c} & -\overline{Y^i}/\overline{z^c} \\ 0 & 0 & 1 \end{pmatrix}$$

In this rule, $\overline{z^c}$ is a priori unknown but we will see in Section 4 that it can be fixed by the previous estimate. Contrarily, the perspective projection rule is:

Rule 4 (Perspective projection of a set distribution)
If we denote

$$E^c = \begin{pmatrix} E_{11}^c & E_{12}^c \\ {E_{12}^c}^T & E_{22}^c \end{pmatrix} \text{ and } \Sigma^c = \begin{pmatrix} \Sigma_{11}^c & \Sigma_{12}^c \\ {\Sigma_{12}^c}^T & \Sigma_{22}^c \end{pmatrix}$$

where E_{11}^c and Σ_{11}^c are 2x2 matrices, E_{12}^c and Σ_{12}^c are 2x1 matrices and E_{22}^c and Σ_{22}^c are scalar then the perspective projection c^i of c^c follows a set distribution:

$$\begin{cases} \overline{c^i} = (\overline{x^c}/\overline{z^c}, \overline{y^c}/\overline{z^c})^T \\ \Sigma^i = J^T[\Sigma_{11}^c - \Sigma_{12}^c {\Sigma_{22}^c}^{-1} {\Sigma_{12}^c}^T]J \\ E^i = J^T[E_{11}^c - E_{12}^c {E_{22}^c}^{-1} {E_{12}^c}^T]J \end{cases} \quad \text{where } J = \begin{pmatrix} \overline{z^c} & 0 \\ 0 & \overline{z^c} \end{pmatrix}$$

The **binocular configuration** can be achieved either considering a stereo-vision system or using simultaneously an eye-in-hand and an eye-to-hand system. For a lack of place, the associated back-projection rule is not detailed here. The complete formulation is given in [8].

Changing the frame. $\mathcal{E}^c$ is expressed in the camera frame $\mathcal{R}_c$. Yet all the observations must be compared in the same frame $\mathcal{R}$ also called the reference frame. The displacement parameter between $\mathcal{R}$ and $\mathcal{R}_c$ are denoted p_e. To take the uncertainty on the camera localization into account, we model p_e by a gaussian noise: $p_e = \overline{p_e} + \mathcal{N}(0, P_e)$. As previously, the associated rule is not given explicitly here since a complete formulation can be found in [8].

4 Estimation Process

We now describe how the set distribution of an object can be estimated and refined using camera motion. At the first step, two images of the same object are available. In the monocular case, they are obtained by two successive positions of the camera. Using the equations related to the back-projection in the binocular configuration, we can estimate the parameters of the distribution $\mathcal{E}_0$ that will initialize the knowledge model.

4.1 Fusing New Images

As shown in Figure 1-a, only two images can not provide a good estimation neither for the object volume nor for its location, especially when the view points are close. In the exploration context, a sequence of several images is available ; we must be able to take them into account in an efficient and robust way. At time k, the known a priori distribution is $\mathcal{E}_k(\overline{c_k}, \Sigma_k, E_k)$. At time $k+1$, the observation likelihood is given by $\mathcal{E}^o_{k+1}(\overline{c^o_{k+1}}, \Sigma^o_{k+1}, E^o_{k+1})$. We estimate independently the uncertainty parameters and the error bounds.

Uncertainty distribution. This is the gaussian estimation case. We can show that the a posteriori distribution is $\mathcal{N}(\overline{c_{k+1}}, \Sigma_{k+1})$ where $\Sigma_{k+1} = \Sigma_k + \Sigma^o_{k+1}$ and $\overline{c_{k+1}} = (\Sigma_k + \Sigma^o_{k+1})^{-1}(\Sigma_k \overline{c_k} + \Sigma^o_{k+1}\overline{c^o_{k+1}})$.

Error bounds. The new bound on the error is given by the intersection between two ellipsoids (E_k and E^o_{k+1}) supposed to be centered at the origin. This intersection is not an ellipsoid itself. We thus need to approximate it. Two types of approximation can be performed: an outer approximation E^+ or an inner approximation E^- (see [8]). Because it is very pessimistic, the use of E^+ is more robust to measurement errors than the use of E^- but the convergence rate of E^+ is very low, depending on the sample rate. The use of a medium approximation $E^- \subset E \subset E^+$ is worth considering. For future experiments, we choose a weighted mean between E^+ and E^-.

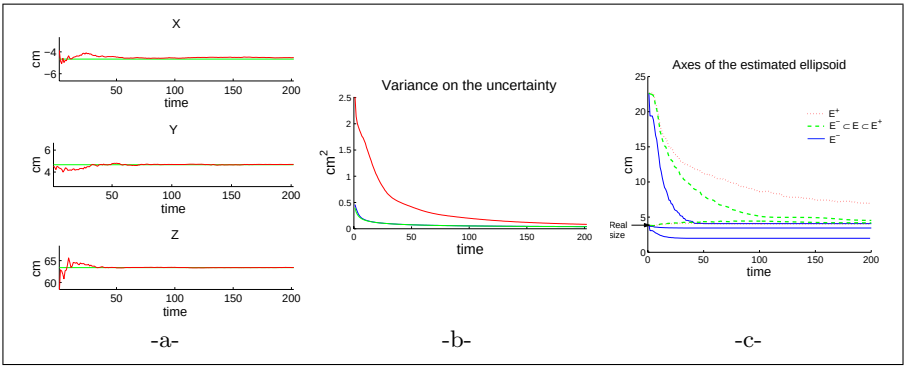

Fig. 3. Simulated reconstruction. a- Estimation of the location. b- Variance on the uncertainty. c- Axes of the estimated ellipsoid.

4.2 Simulation

The previous estimation process has been simulated in order to analyze its convergence rate and precision. The results we present were obtained in the context of a single camera. Simulations were computed as follows: after the initialization stage, the virtual camera is moving with a constant speed (3cm per iteration) along a circular trajectory. At the center of this trajectory is placed the object: a virtual sphere with known position ($X = -4.7cm$, $Y = 4.7cm$ and $Z = 63cm$) and radius (4 cm). The uncertainty on the camera location is a normal unbiased additive noise with a standard deviation outweighing 10 cm for translation and 5 deg for rotations. The covariance on center estimation and intrinsic parameters were chosen as follows:

$$P^p = \begin{pmatrix} 4\,0 & \\ 0\,4 & \end{pmatrix} \; P_{int} = \begin{pmatrix} 4\,0 & 0 & 0 \\ 0\,4 & 0 & 0 \\ 0\,0 & 10^{-7} & 0 \\ 0\,0 & 0 & 10^{-7} \end{pmatrix}$$

It corrsponds to a two pixels standard deviation on the extracted center of the object, to a two pixels standard deviation on the principal point and to a 0.3mm standard deviation on the pixel size at a one meter focal length. Successive observations refine the estimated location of the sphere and its volume. Figure 3-a shows convergence of the estimated location to the real position, ensuring final accuracy. On the figure 3-b describing the variance on the uncertainty, we notice that the axis initially parallel to the optical axis is badly estimated at the initialization.

We have simulated the convergence of the axes for E^+, E^- and $E = 0.98E^+ + 0.02E^-$. We note that the final accuracy and convergence rate strongly depend on the used combination of inner and outer estimation (see Figure 3-c). The outer approximation is converging very slowly whereas the inner approximation under-estimates the including volume of the object. The chosen E seems to be a good compromise between accuracy and convergence speed.

5 Exploration Process

We now want to identify a control law that automatically generates exploratory movements of the camera. The principle of this command is to minimize the uncertainty of the predicted a posteriori knowledge for the next iteration. We could imagine a command based on the reduction of the estimated including ellipsoid. But this strategy seems not very judicious since the real shape of the object is a priori unknown.

At time k, we have deduced, from the estimation process, the knowledge $\mathcal{E}_k(\overline{c_k}, \Sigma_k, E_k)$. For notational convenience, it is expressed in the current camera frame instead of the so called reference frame. If, at time $k+1$, the predicted camera motion is (R,t), we can deduce the corresponding predicted a priori information, the predicted observation and finally the predicted a posteriori information. Since the object is known to be static, the predicted a priori information is simply the propagation of Σ_k through a changing frame (R,t). If the motion was perfectly known, thanks to the changing frame rule, the associated information would be $R^T \Sigma_k R$. In the absence of real measurement, the predicted observation is the propagation of the predicted a priori knowledge through projection and back-projection. Thanks to Rule 4, the predicted measure information is

$$\widehat{\Sigma^i_{k+1}} = J_0^T (A - \frac{BB^T}{\gamma}) J_0 \text{ where } J_0 = \begin{pmatrix} z_{k+1} & 0 \\ 0 & z_{k+1} \end{pmatrix} \text{ and } R^T \Sigma_k R = \begin{pmatrix} A & B \\ B^T & \gamma \end{pmatrix}$$

A is a 2x2 matrix, B is a 2x1 matrix and γ a scalar. Thanks to Rule 3, $\widehat{\Sigma^i_{k+1}}$ corresponds to the following 3D information:

$$\widehat{\Sigma^o_{k+1}} = J_1^T \begin{pmatrix} J_0^T (A - \frac{BB^T}{\gamma}) J_0 & 0 \\ 0 & 0 \end{pmatrix} J_1 \text{ where } J_1 = \begin{pmatrix} 1/z_{k+1} & 0 & -X^i_{k+1}/z_{k+1} \\ 0 & 1/z_{k+1} & -Y^i_{k+1}/z_{k+1} \\ 0 & 0 & 1 \end{pmatrix}$$

X^i_{k+1} and Y^i_{k+1} are the predicted coordinates of $\overline{c^i}$ at time $k+1$. In practice, we use a visual servoing control scheme such that the explored object is centered in the image $(\forall k, X^i_k = Y^i_k = 0)$. This is a first step to impose the visibility of the object during the exploration process. In practice, a tracking rotational motion is calculated as the one presented in [7]. Then, combining the predicted observation with the predicted a priori knowledge (see Section 4.1), we deduce the predicted a posteriori knowledge in the camera frame at time $k+1$:

$$\widehat{\Sigma_{k+1}} = \begin{pmatrix} 2A - \frac{BB^T}{\gamma} & B \\ B^T & \gamma \end{pmatrix} \tag{2}$$

5.1 Exploratory Control Law

Motion parameters (R,t) must be calculated in such a way that $\widehat{\Sigma_{k+1}}$ is maximal in one sense. In order to introduce the idea of isotropy concerning the whole view point directions, we will attach importance to the sphericity of $\widehat{\Sigma_{k+1}}$.

We can show, in equation (2), that the depth z_{k+1} from the camera to the object does not influence the predicted information matrix. This is due to the linear approximation we made in Rule 1. As a first consequence, the optimal translation can be calculated in the image plane so that we can use the remaining degree of freedom to regulate the projected surface. An other consequence is that the direction of translational motion t is related to the axis of rotation u by the equality $t = z \wedge u$ where z is the unit vector normal to the image plane. As a consequence, we can define the exploratory control law either using u or using t. We now examine and compare two types of exploratory motions.

Locally optimal exploration. In that part the camera motion locally optimizes the increase of Σ_k and the criterion is the trace of $\widehat{\Sigma_{k+1}}$. At time $k+1$, the camera will have rotated with an angle $\alpha \geq 0$ around the unit vector $u = (u_x, u_y, 0)$. The locally optimal (LO) motion is defined by $(u_x, u_y) = argmax \quad tr[\widehat{\Sigma_{k+1}}]$. We show in [8] that (u_x, u_y) is the solution of a linear system which is easily computed. We also noticed that the study is correct if and only if the center of the camera is not located on the direction pointed by an eigenvector of Σ_k. In such a situation, the camera is in a local minimum. Simulations presented in the next section will confirm that.

Best view point exploration. Now, instead of locally optimizing $\widehat{\Sigma_{k+1}}$, the best view point (BVP) motion tends to reach the next best view point: the one which leads to the "biggest spherical" $\widehat{\Sigma_{k+1}}$, that is proportional to the identity matrix. Judging from equation (2), the BVP is the point at which:

$$\begin{aligned} \widehat{\Sigma_{k+1}} = \begin{pmatrix} 2A - \frac{BB^T}{\gamma} & B \\ B^T & \gamma \end{pmatrix} \quad & \text{is diagonal} \\ \Leftrightarrow B = 0 \text{ and } A \text{ diagonal} \Leftrightarrow R^T \Sigma_k R \text{ diagonal} & \\ \Leftrightarrow R \text{ is composed of eigenvectors of } \Sigma_k & \\ \text{and } \gamma \text{ is the biggest eigenvalue of } \Sigma_k & \end{aligned}$$

Both previous results show that the third column vector of R is the eigenvector of Σ_k associated with its maximal eigenvalue. Since the object is centered in the image, the center of the camera must be located on the direction pointed by this eigenvector. In other words, the next BVP is located on the eigenvector of Σ_k associated with the biggest eigenvalue that is the most informative direction.

5.2 Simulation

Exploratory motions have been simulated so that we can analyze the associated trajectory. The simulation conditions are the same as described in Section 4.2 except that the trajectory is no longer circular but controlled by either the LO control law or the BVP one. The LO exploration (see Figure 4-a) leads to a local minimum. This is due to the biggest slope maximization induced by this

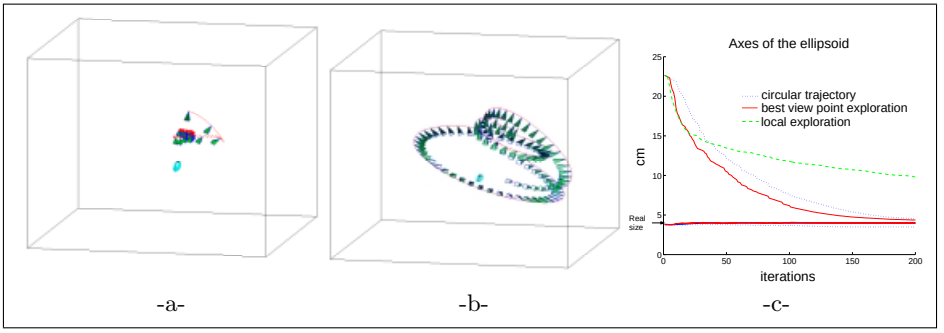

Fig. 4. Simulated exploration. a- Locally optimal exploration. b- Best view point exploration. c- Axes of the ellipsoid.

technique. If needed, we could go out of such points by slightly off-centering the object in the image. Applying the criterion

$$Q = \frac{\text{Initial volume- Final volume}}{\text{Initial volume - Real volume}}$$

the LO exploration resulted in a 67% reduction of the volume for a 200 iterations simulation. The BVP exploration seems to overpass such local minima (see Figure 4-b) and leads to very intuitive trajectories: the camera is spiraling around the object, from the top to its middle plane. For this simulation, the volume reduction was about 99.5%. The gain induced by the BVP exploration can be seen on Figure 4-c where we compare the convergence of the axes when no exploration strategy is used (circular trajectory) to the case of the BVP strategy. In the second case, both the convergence rate and the final accuracy are better. Thanks to its simplicity and local minima avoidance, the BVP strategy appears to be a better solution to exploration.

Finally, a suitable stopping criterion is the trace of the covariance matrix which is related to the uncertainty of the model. The BVP exploration ran until the criterion reached an arbitrarily fixed to 0.002 threshold. Because of the local minimum, this value has not been reached for the LO strategy. The simulation has thus been arbitrarily stopped after 200 iterations in this case.

6 Experimentation

In order to validate the previous study in real situation, we need to extract the mask of the object we explore. In order to deal with general scenes, we want to impose no constraint on the object aspect (color, texture, grey level, ...). With this aim in view, we make the only assumption (not very restrictive in most situations) that there is a depth discontinuity at the frontier of the objects. Then for every translational motion of the camera, the projected motion of each object is distinguishable from the other. A motion segmentation algorithm will

give the mask of the objects. For real time constraints, this algorithm must be fast and robust. We chose the parametric motion estimation algorithm imagined by Odobez and Bouthemy [14]. It furnishes a map of points whose motion is not consistent with dominant motion (background motion). In our situation, it corresponds to the mask of the objects.

We implemented the exploration process on a six degrees of freedom robot. The covariance models P^p and P_{int} were chosen identical to the one in section 4.2. Besides, the uncertainty model on robot joints measure is a normal unbiased additive noise whose standard deviation is outweighing 10 cm for translation and 5 deg for rotations.

The speed of the algorithm (about 150ms per loop including motion segmentation) allows us to estimate the location and volume of several objects in real time. Figure 5 shows an example with two different objects: a mushroom and a stick. At the initialization, two images of the scene are acquired (see Figures 5-a and 5-b). The segmentation of the objects are coarsely approximated by hand for the initialization but it could be done automatically by slightly moving the camera around the initial positions. The associated estimated ellipsoids including the two objects are given on Figure 5-c. Figure 5-d shows the projection of these first estimation in the final image. It convinces us of the need to refine this estimation. In a second step, the camera is autonomously exploring the objects. The strategy is based on the exploration of one of the two objects (the mushroom). Both the LO and the BVP exploration have been tested. The LO trajectory (see Figure 5-e) does not encounter a local minimum thanks to noise inherent to experimentation. The BVP trajectory (see Figure 5-f) is quite similar to the simulated one. The final estimated ellipsoid (see Figure 5-g) was projected in the final image (see Figure 5-h) to show the efficiency of the algorithm. The objective to estimate a coarse including volume for the objects is reached.

7 Conclusion

We have defined a simple and coarse model representing each detected object of a scene. It allows us to calculate for every point of a scene its probability to belong to an object. This model is computationally cheap because it only requires a 3D vector and two 3D symmetric matrices to represent the center, the uncertainty and the volume. Several propagating rules have been inferred from stochastic geometry resulting in an estimation scheme which is fast and robust. Based on this estimation process, we defined and compared two exploration processes which proved to be optimal in one sense.

Our approach stems for the class of coarse model estimation techniques. It differs from previous work in several points. First the model we use is a mixture of stochastic and set membership models. Its description in a unified probabilistic framework makes it robust to different sources of uncertainty (image processing noise, calibration errors, etc.), and furnishes a general frame to coarse estimation and fusion. It allowed us to develop very general rules for model propagation. As a consequence, the method can apply to any kind of sensor.

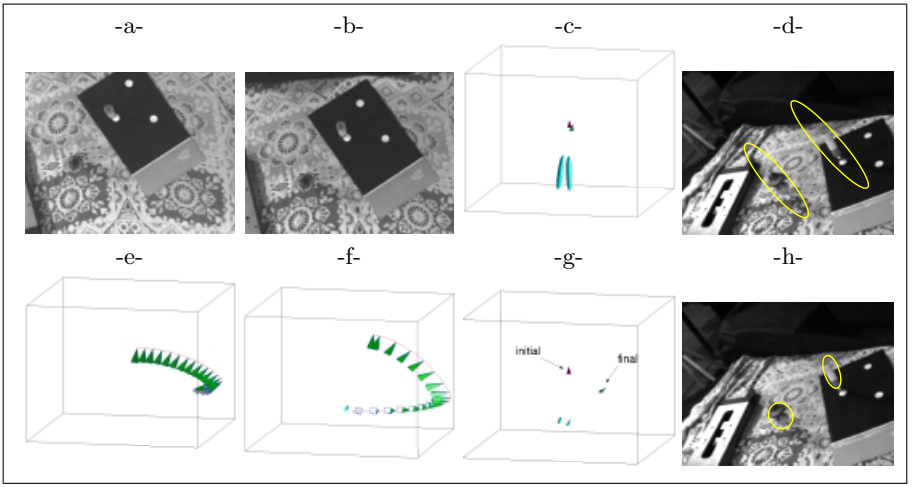

Fig. 5. Two objects reconstruction. a,b-Initial images. c- First estimation of the including ellipsoids. d- Associated projection in the last image. e- Locally optimal exploration. f- Best view point exploration. g- Final reconstruction for the best view point strategy. h- Associated projection in the final image.

Second we derived an analytical solution to optimal exploration so that the calculation of an exploration trajectory which reduces the current uncertainty on the model is solvable on-line. Even if it is obviously not based on accurate reconstruction, the answer we give to which camera motion will improve the coarse model is well-posed. At last, we claim that a main interest of our work is its applicability to real-time processes thanks to the fastness of the algorithms due to adequacy of the model to the assigned task.

As we previously said, the problem of strongly concave objects is worth considering and constitutes an interesting outlook. Judging from our study, reconstruction of such objects is really feasible at the only condition that we are able to partition the object in coherent parts and to match them along the sequence.

At last, the model we defined and the associated tools we developed constitute a good basis to build higher level tasks. We focused on the exploration of objects appearing entirely in the field of view of the camera. Our future work will be dedicated to the research of all the objects of a scene.

References

1. T. Arbel and F. Ferrie. Viewpoint selection by navigation through entropy maps. *7th IEEE Int. Conf. on Computer Vision*, Vol. I, pages 248–254, Los Alamitos, September 1999.
2. N. Ayache. *Artificial Vision for Mobile Robots.* The MIT Press, Cambridge, MA, 1991.
3. J. D. Boissonnat. Representing 2D and 3D shapes with the Delaunay triangulation. *7th Int. Conf. on Pattern Recognition*, pages 745–748, Montreal, Canada, 1984.

4. C. Connolly. The determination of next best views. In *IEEE Int. Conf. on Robotics and Automation*, Vol. 2, pages 432–435, St Louis, Missouri, March 1995.
5. H. F. Durrant-Whyte. *Integration, Coordination, and Control of Multi-Sensor Robot Systems*. Kluwer Academic Publishers, Boston, 1987.
6. O. D. Faugeras, F. Lustman, and G. Toscani. Motion and structure from motion from point and line matches. In *First Int. Conf. on Computer Vision*, pages 25–34, Washington DC, 1987.
7. G. Flandin, F. Chaumette, E. Marchand. Eye-in-hand / eye-to-hand cooperation for visual servoing. *IEEE Int. Conf. Robotics and Automation*, Vol. 3, pages 2741-2746, San Fransisco, April 2000.
8. G. Flandin, F. Chaumette. Visual Data Fusion: Application to Objects Localization and Exploration. *IRISA Research report*, No 1394, April 2001.
9. K. N. Kutulakos, C. R. Dyer, and V. J. Lumelsky. Provable strategies for vision-guided exploration in three dimensions. *IEEE Int. Conf. Robotics and Automation*, pages 1365–1372, Los Alamitos, CA, 1994.
10. S. Lacroix and R. Chatila. *Motion and perception strategies for outdoor mobile robot navigation in unknown environments*. Lecture Notes in Control and Information Sciences, 223. Springer-Verlag, New York, 1997.
11. D. G. Maksarov and J. P. Norton. State bounding with ellipsoidal set description of the uncertainty. *Int. Journal on Control*, 65(5):847–866, 1996.
12. E. Marchand and F. Chaumette. Active vision for complete scene reconstruction and exploration. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(1):65–72, January 1999.
13. D. Marr and K. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. Royal Soc. London Bulletin*, pages 269–294, 1977.
14. J.M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, 6(4):348–365, 1995.
15. F. C. Schweppe. Recursive state estimation: unknown but bounded errors and system inputs. *IEEE Trans. on Automatic Control*, AC-13:22–28, 1968.
16. G. Slabaugh, B. Culbertson, T. Malzbender, and R. Schafer. A survey of methods for volumetric scene reconstruction from photographs. *Technical report, Center for Signal and Image Processing*, Georgia Institute of Technology, 2001.
17. K. A. Tarabanis, P. K. Allen, and R. Y. Tsai. A survey of sensor planning in computer vision. *IEEE Trans. on Robotics and Automation*, 11(1):86–104, February 1995.
18. E. Welzl. Smallest enclosing disks (balls and ellipsoids). *Lecture Notes in Computer Science*, 555:359–370, 1991.
19. P. Whaite and F. P. Ferrie. Autonomous exploration: Driven by uncertainty. *Int. Conf. on Computer Vision and Pattern Recognition*, pages 339–346, Los Alamitos, CA, June 1994.
20. H. S. Witsenhausen. Sets of possible states of linear systems given perturbed observations. *IEEE Trans. on Automatic Control*, AC-13:556–558, 1968.