Vision-based Control Using Probabilistic Geometry for Objects Reconstruction

Grégory Flandin
IRISA - INRIA
Campus de Beaulieu
35042 Rennes Cedex, France

François Chaumette
IRISA - INRIA
Campus de Beaulieu
35042 Rennes Cedex, France

Abstract

We first present a suitable object knowledge representation based on a mixture of stochastic and set membership models and considering an approximation resulting in ellipsoidal calculus by means of a normal assumption for stochastic laws and ellipsoidal over or inner bounding for uniform laws. Then we build an efficient estimation process integrating visual data online and perform online and optimal exploratory motions for the camera. The control schemes are based on the maximization of the a posteriori predicted information.

1. Overview

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) or using an exhausting voxel representation of the scene, eventually reducing the complexity by means of hierarchical techniques. But, for several kinds of applications, only a preliminary 3D map of the scene is sufficient. As a consequence, for a large class of applications, we consider that the knowledge of each object of a scene can come down to the knowledge of its including volume (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 [1, 3, 4]. 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 Schwepppe [15, 11]. 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 [8]. Mixing probability and set membership theories in a unified stochastic framework, we will take advantage of both representations 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 [12]. When a complete reconstruction of the scene 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 [2], 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 [14], White 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 [6] exploit the notion of the occlusion boundary that are the points separating the visible from the occluded parts of an object. Lacroix and Chatila [7] 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. Marchand and Chaumette [9] use active motion of a single camera to explore geometrical objects such as polygons and cylinders. The viewpoint is selected minimizing a cost function.

The strategy we develop in this paper consists in reducing uncertainty of the distribution associated with the observed object using visual data. A gaussian modeling of uncertainty and a linearization of the visual acquisition process allow us to build analytical solutions to optimal
2. Modeling and Propagating Rule

For every object $O$ belonging to a scene $S$ and for every point $x \in S$, we aim at calculating the probability that $x \in O$ denoted $P(x \in O)$. If we consider the coordinates of a point $c \in O$ as a random vector whose distribution is, for every $x \in S$, $P_c(x) = \mathcal{P}_c(x \mid x \in O) \cdot P(x \in O) + \mathcal{P}_c(x \mid x \notin O) \cdot P(x \notin O)$.

$P_c(x \mid x \in O)$ is the probability that a point $c \in O$ is at $x$ knowing that $x \in O$, it is a constant that can be calculated after normalization. Thus, modeling $S$ comes down to finding for each $O$ a suitable distribution to model the density function of $P_c(x)$. To do so, we break down $c$ into the sum of a mean vector $\bar{c}$ and two independent random vectors (see Fig. 1):

$$c = \bar{c} + p + e \quad (1)$$

where $p$ represents the uncertainty on the location of the object and the bounds on the error $e$ define its volume. For computational convenience, we assume that: 1- $p$ follows a normal distribution $\mathcal{N}(0, \Sigma)$ where $\Sigma$ is the inverse of the usual covariance called the information matrix. When dealing with partial observability, an infinity variance along the inobservability axis is thus replaced by a null information. Let us remark that the normal distribution is a quite good approximation of most uncertainty sources and makes the propagation of the law easier. 2- $e$ is uniformly distributed on an ellipsoid denoted (by measure of language) by its matrix $E$. $\Sigma$ and $E$ are both symmetric, positive, and definite.

From these assumptions, the global distribution associated with an object is completely defined by $\bar{c}$, $\Sigma$ and $E$. More precisely, it is the distribution of the sum of independent variables, that is the convolution product of a uniform distribution $U_E$ on $V$ by a normal one $\mathcal{N}(\bar{c}, \Sigma)$. We call this distribution a set distribution and we denote

$$E(\bar{c}, \Sigma, E) = \mathcal{N}(\bar{c}, \Sigma) * U_E$$

We now aim at defining the transformations induced by a change in the parameter space. This will allow us to fuse heterogeneous data and knowledge.

**Rule 1 (Transformation of a set distribution)**

Let $c$ be a random vector following a set distribution $E(\bar{c}, \Sigma, E)$ and $T$ a diffeomorphism. As a first approximation, the transformed random vector $c' = T(c)$ follows a set distribution $E'(\bar{c}', \Sigma', E')$ where

$$\begin{cases}
\bar{c}' = T(\bar{c}) \\
\Sigma' = JT \Sigma J \\
E' = JT E J
\end{cases}$$

where $J = \frac{\partial T^{-1}}{\partial \bar{c}} |_{\bar{c}}$

The proof is achieved approximating $T$ at first order and applying results of gaussian and ellipsoidal calculus [5]. When $T$ depends on external parameters: $c' = T(c, p_c)$ (where $p_c$ is $\mathcal{N}(\bar{p}_c, P_c)$), we can linearize the transformation around $\bar{c}$ and $\bar{p}_c$ to take into account other sources of uncertainty such as the location of the camera (see [5] for details). Besides, we saw that $T$ must be a diffeomorphism. When it is not the case, we need a rule dedicated to projection on a subspace. This rule can be found in [5].

3. Propagation of visual data

Thanks to rule 1, we can infer a lot of transformations specialized to the propagation of visual data. We identify three stages in the chain of visual observation (see Fig. 2):

1. In the image, the measure is a 2D set distribution $E'(\bar{c'}, \Sigma', E')$. 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 $(\bar{c'}, E')$ represents the center and matrix of the smallest outer ellipsoid in the image. They can be extracted thanks to algorithms like the one proposed in [13]. $\Sigma'$ must account for all sources of uncertainty that can occur in the calculus of this ellipsoid: errors on camera intrinsic parameters estimation and inaccuracy of image processing. Let us notice that we implicitly consider that the projected center of an ellipsoid is the center of the projected ellipse. This is theoretically wrong but the difference is very small, more especially as the ellipsoid is centered in the image which will be the case in practice thanks to a visual servoing control scheme.
2. The process transforming the 2D visual data in a 3D observation is called back-projection. The associated 3D set distribution is denoted $E^c(\mathbf{c}^c, \Sigma^c, E^c)$. 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. Back-projection strongly depends on the camera configuration. For a lack of space, we do not present the corresponding rules here.

3. At last, we must express every observation in a common frame called the reference frame $\mathcal{R}$. The associated observation is denoted $E^c(\mathbf{c}^c, \Sigma^c, E^c)$. The displacement parameter between $\mathcal{R}$ and $\mathcal{R}_e$ are denoted $p_e$. We model $p_e$ by a gaussian noise: $p_e = \overline{p}_e + \mathcal{N}(0, P_e)$. If $\mathbf{c}^c$ is the coordinate vector of $\mathbf{c}^c$ expressed in $\mathcal{R}$, we can write:

$$\mathbf{c} = \begin{pmatrix} R & t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{c}^c \\ 1 \end{pmatrix} = T\mathbf{c}$$

where $D$ is related to $p_e$. For a six degrees of freedom robot, $p_e$ accounts for the position of joints and the effector/camera transformation. As a consequence of rule 1 and since $T$ is a dixemorphism, we can infer a changing frame rule which is detailed in [5].

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 monocamera case, they are obtained by two successive positions of the camera. Using the equations of binocular back-projection, we can estimate the parameters of the distribution $E\theta_0$ that will initialize the knowledge model.

Of course, only two images cannot 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 $E_k(\overline{E}_k, \Sigma_k, E_k)$. At time $k + 1$, the observation likelihood is given by $E_k^{(k+1)}(\overline{E}_k^{(k+1)}, \Sigma_k^{(k+1)}, E_k^{(k+1)})$. We estimate separately the uncertainty parameters and the error bounds:

**Uncertainty distribution:** This is the gaussian estimation case. We can show that the a posteriori distribution is $\mathcal{N}(\overline{E}_{k+1}, \Sigma_{k+1})$ where:

$$\overline{E}_{k+1} = \overline{E}_k + \Sigma_{k+1}^{-1}(\Sigma_k \overline{E}_k + \Sigma_{k+1} \overline{E}_k^{(k+1)})$$

This is a mean of previous knowledge and new observation respectively weighted by the confidence (inverse of covariance) we have in them. Besides:

$$\Sigma_{k+1} = \Sigma_k + \Sigma_{k+1}^{(k+1)}$$

is the variance of the error on this estimation.

**Error bounds:** The new bound on the error is given by the intersection between two ellipsoids ($E_k^{+}$ and $E_{k+1}^{-}$) 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 [5]). 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 $E^- \subseteq E \subseteq E^+$ is worth considering. For future experiments, we chose a simple weighted mean between $E^+$ and $E^-$. Simulations concerning the previous estimation process can be found in [5].

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.

5.1. Predicted a posteriori information

At time $k$, we have deduced, from the estimation process, the knowledge $E_k(\overline{E}_k, \Sigma_k, E_k)$. 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 with assurance, the predicted a priori information is simply the propagation of \( \Sigma_k \) through a changing frame \((R, t)\). If we assume that the motion is perfectly known, thanks to rule 1, the associated information is \( R^T \Sigma_k R \). In the absence of real measurement, the predicted observation is the propagation of the predicted a priori knowledge through projection and back-projection. Let us denote

\[
R^T \Sigma_k R = \begin{pmatrix} A & B \\ B^T & c \end{pmatrix}
\]

where \( A \) is \((2,2), B \) is \((2,1)\) and \( c \) a scalar. If we assume that the object is centered in the image (this is not restrictive since we want to impose the visibility of the object during the exploration) then, thanks to the previous rules and the estimation process of Section 4, we can deduce the predicted a posteriori knowledge in the camera frame at time \( k + 1 \):

\[
\hat{\Sigma}_{k+1} = \begin{pmatrix} 2A - BB^T/B^T c & B \\ B & c \end{pmatrix}
\]

(2)

5.2. Exploratory control law

Motion parameters \((R, t)\) must be calculated in such a way that \( \hat{\Sigma}_{k+1} \) 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 \( \hat{\Sigma}_{k+1} \).

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:

\[
V_{z_k} = \frac{\lambda_{z_k}}{2S_k} (S_k - S^*)
\]

where \( S_k \) is the current surface while \( S^* \) is the desired one. An other consequence is that the direction of translational motion \( t \) is related to the axis of rotation by the equality \( t = z \wedge u \), where \( z \) is the unit vector normal to the image plane (see Fig. 3). 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.

5.2.1 Locally optimal exploration

In that part the camera motion locally optimizes the increase of \( \Sigma_k \) and the criterion is the trace of \( \hat{\Sigma}_{k+1} \). At time \( k + 1 \), the camera will have rotated with an angle

\[
\alpha \geq 0 \text{ around the unit vector } u = (u_x, u_y, 0).
\]

At a first order approximation (\( \alpha \ll 1 \)), the associated matrix is

\[
R \approx I + \alpha \begin{pmatrix} 0 & 0 & u_y \\ 0 & 0 & -u_x \\ -u_y & u_x & 0 \end{pmatrix}
\]

The optimal exploratory motion is defined by \( (u_x, u_y) = \arg \max_{(u_x, u_y)} \text{ tr}[\hat{\Sigma}_{k+1}] \). If we denote \( V \) and \( \Delta \) the eigenvector matrix and the diagonal matrix of \( \Sigma_k \), we show that (see [5]):

\[
\text{tr}[\hat{\Sigma}_{k+1}] \approx \text{constant} + 2\alpha \gamma(\Delta, V) \Delta \beta^T
\]

where \( \gamma(\Delta, V) \) is a matrix valued vector. As a consequence, the optimal control is given by the projection of \( \Delta \gamma(\Delta, V) \) on the image plane which maximizes \( 2\alpha \gamma(\Delta, V) \Delta \beta^T \).

In [5], we notice that the study is correct if and only if \( v_e \) is not an eigenvector of \( \Delta \). We will see, in the simulations, that when \( v_e \) is an eigenvector of \( \Delta \), the camera is in a local minimum. Besides, when \( \Delta \) is spherical (i.e. when each eigen-value equals the other) \( \text{tr}[\hat{\Sigma}_{k+1}] \) is constant. In that case \((u_x, u_y)\) can be randomly chosen.

5.2.2 Best view point exploration

Now, instead of locally optimizing \( \hat{\Sigma}_{k+1} \), the best view point motion tends to reach the next best view point: the one which leads to the “biggest spherical” \( \hat{\Sigma}_{k+1} \), that is proportional to the identity matrix. Judging from equation 2, we can show that the next best view point is located on the eigenvector of \( \Sigma_k \) associated with the biggest eigenvalue that is the most informative direction. The motion vector \((V_x, V_y)\) must be directed to the intersection between the image plane and the biggest information axis.

5.3. Simulation

Exploratory motions have been simulated so that we can analyze the associated trajectory. Simulations were computed as follows: after the initialization stage, the virtual camera is moving with a constant speed (3cm per iteration) along a trajectory generated thanks to the previous control laws. At the center of this trajectory is placed the object: a virtual sphere with known position and radius.
The uncertainty on the camera location is a norm
unbiased additive noise with a standard deviation outweighing
10 cm for translation and 5 deg for rotations. Successive
observations refine the estimated location of the sphere
and its volume over 200 iterations.

The local exploration (see Fig. 4-a) leads to a local min-
imum. This is due to the biggest slope maximization in-
duced by this technique. Even if further studies should be
done about this issue, let us note that local minima cor-
respond to points where \( v_z \) is an eigen-vector of \( \Delta \) (see
Section 5.2.1). We can go out of such points by slightly
deleasing the object in the image. Applying the cri-
terion
\[
Q = \frac{\text{Initial volume} - \text{Final volume}}{\text{Initial volume} - \text{Real volume}}
\]
the local exploration resulted in a 67% reduction of the
volume for a 200 iterations simulation. The best view
point exploration seems to overpass such local minima
(see Fig. 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 best view point
exploration can be seen on Fig. 5 where we compare the
convergence of the axes when no exploration strategy is
used (circular trajectory) to the case of the locally optimal
exploration and to the case of the best view point strategy.
In the third case, both the convergence rate and the final
accuracy are better.

Let us note that for both previous simulations, the ex-
ploration process ran over a fixed number of iterations. It
would be interesting to identify a suitable stopping cri-
terion. It could deal with completeness of observation or
accuracy of reconstruction.

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 estima-
tion algorithm imagined by Odobez and Boutheemy [10]. It
furnishes a map of points whose motion is not consistent
with dominant motion. In our situation, it corresponds to
the mask of the objects.

We implemented the exploration process on a six de-
grees of freedom robot. 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 6 is an example with two different objects.
At the initialization, two images of the scene are
acquired (see Fig. 6-a and 6-b). The associated estimation
for the two objects is given on Figure 6-c. Figure 6-d is the
projection of this first estimation in the final image. It con-
vincs us of the need to refine this estimation. In a second
step, the camera is autonomously exploring the objects.
We fixed arbitrarily the exploration period to 20 seconds
and the strategy is based on the exploration of one of the
two objects. Both the locally optimal and the best view
point exploration have been tested. The locally optimal
trajectory (see Fig. 6-e) does not encounter a local min-
imum thanks to noise inherent to experimentation. The
best view point trajectory (see Fig. 6-f) is quite similar to
the simulated one even if the experimental time is much
shorter because of robot joint limits. The final estimate
(see Fig. 6-g) has been projected in the final image (see
Fig. 6-h) to show the efficiency of the algorithm.

7. Conclusion

We have defined a model representing each object as an
approximated probabilistic law allowing 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. Sev-
eral 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.

The model defined in this way 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.

Acknowledgments. This study was partly supported by INRIA LARA project and by Brittany County Council.

References