Proceedings of the 40th IEEE
Conference on Decision and Control
Orlando, Florida USA, December 2001

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 Schweppe [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], 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 [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 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. We also define a rule that makes propagation of a set distribution possible. This rule is applied to the propagation of visual data in Section 3. 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 and Propagating Rule

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

$$\mathcal{P}_c(x) = \underbrace{\mathcal{P}_c(x|x \in \mathcal{O})}_{\text{constant}}.\mathcal{P}(x \in \mathcal{O}) + \underbrace{\mathcal{P}_c(x|x \notin \mathcal{O})}_{0}.\mathcal{P}(x \notin \mathcal{O})$$

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

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

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 Σ is the inverse of the usual covariance called the information matrix. When dealing with partial inobservability, 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 misure of language) by its matrix E. Σ and E are both symmetric, positive and definite.

From these assumptions, the global distribution associated with an object is completely defined by $\bar{c}$, 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 $\mathcal{N}(\bar{c}, \Sigma)$. We call this distribution a set distribution and we denote

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

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.

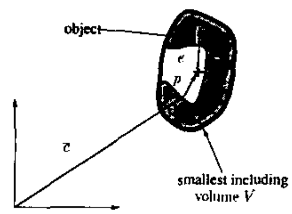


Figure 1: Modeling definitions

Rule 1 (Transformation of a set distribution)
Let c be a random vector following a set distribution $\mathcal{E}(\bar{c}, \Sigma, E)$ and T a diffeomorphism. As a first approximation, the transformed random vector $c' = T(c)$ follows a set distribution $\mathcal{E}'(\bar{c'}, \Sigma', E')$ where

$$\begin{cases} \bar{c'} &= T(\bar{c}) \\ \Sigma' &= J^T \ \Sigma \ J \\ E' &= J^T \ E \ J \end{cases} \quad \text{where} \quad J = \left.\frac{\partial T^{-1}}{\partial c}\right|_{\bar{c}}$$

The proof is achieved approximating T at a first order and applying results of gaussian and ellipsoidal calculus [5]. When T depends on external parameters: $c' = T(c, p_e)$ (where p_e is $\mathcal{N}(\overline{p_e}, P_e)$), we can linearize the transformation around $\bar{c}$ and $\overline{p_e}$ 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 $\mathcal{E}^i(\overline{c^i}, \Sigma^i, E^i)$. 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 ellipsoid in the image. They can be extracted thanks to algorithms like the one proposed in [13]. Σ^i must account for all sources of uncertainty that can occur in the calculus of this ellipse: 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.

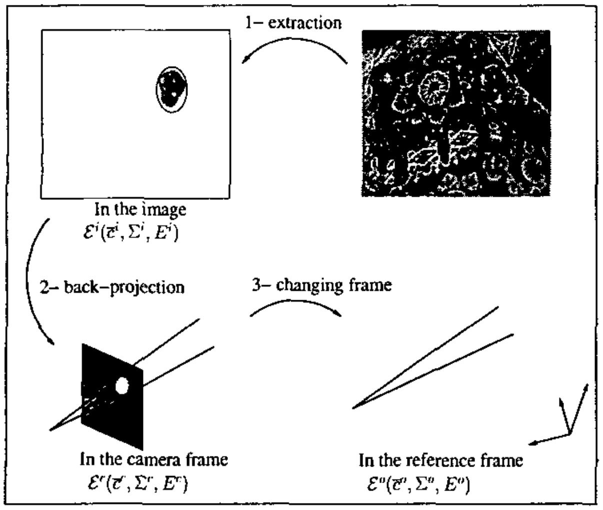


Figure 2: Measure transformations

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. Back-projection strongly depends on the camera configuration. In [5], we study the monocular and binocular configurations. For a lack of place, 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 $\mathcal{E}^o(\overline{c^o}, \Sigma^o, E^o)$. The displacement parameter between $\mathcal{R}$ and $\mathcal{R}_c$ are denoted p_e. We model p_e by a gaussian noise: $p_e = \overline{p_e} + \mathcal{N}(0, P_e)$. If c^o is the coordinate vector of c^c expressed in $\mathcal{R}$, we can write:

$$c^o = \underbrace{\begin{pmatrix} R & t \\ 0 & 1 \end{pmatrix}}_{D} \begin{pmatrix} c^c \\ 1 \end{pmatrix} = T(c^c, p_e)$$

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 diffeomorphism, 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 monocular 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 $\mathcal{E}_0$ that will initialize the knowledge model.

Of course, 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 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 $\mathcal{N}(\overline{c_{k+1}}, \Sigma_{k+1})$ where:

$$\overline{c_{k+1}} = (\Sigma_k + \Sigma^o_{k+1})^{-1}(\Sigma_k \overline{c_k} + \Sigma^o_{k+1}\overline{c^o_{k+1}})$$

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

$$\Sigma_{k+1} = \Sigma_k + \Sigma^o_{k+1}$$

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^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 [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 $E^- \subset E \subset E^+$ 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 $\mathcal{E}_k(\overline{c_k}, \Sigma_k, E_k)$. 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 with assurance, the predicted a priori information is simply the propagation of Σ_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 is not resctrictive 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 $k+1$:

$$\widehat{\Sigma_{k+1}} = \begin{pmatrix} 2A - \frac{BB^T}{c} & B \\ B^T & c \end{pmatrix} \quad (2)$$

5.2. 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:

$$V_{z\,k} = -\frac{\lambda z_k}{2 S_k}(S_k - S^*)$$

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 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

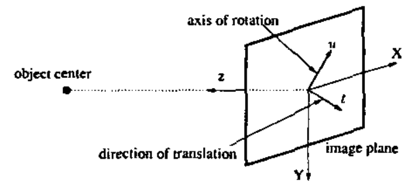


Figure 3: Definition of the exploratory control law

$\alpha \geq 0$ around the unit vector $u = (u_x, u_y, 0)$. At a first order approximation ($\alpha << 1$), the associated matrix is

$$R \simeq I + \alpha \begin{pmatrix} 0 & 0 & u_y \\ 0 & 0 & -u_x \\ -u_y & u_x & 0 \end{pmatrix}$$

The optimal exploratory motion is defined by $(u_x, u_y) = argmax \quad tr\widehat{\Sigma_{k+1}}$. If we denote V and Δ the eigenvector matrix and the diagonal matrix of Σ_k, we show that (see [5]):

$$tr(\widehat{\Sigma_{k+1}}) \approx constant + 2\alpha\gamma(\Delta, V)\Delta\beta^T$$

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 $2\alpha\gamma(\Delta, V)\Delta\beta^T$.

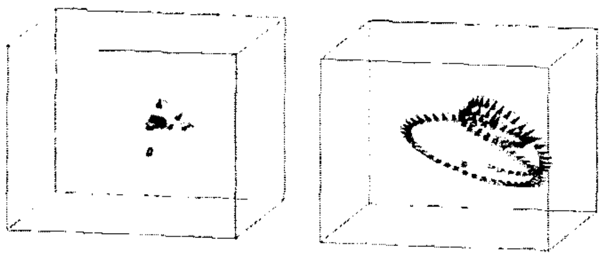
In [5], we notice that the study is correct if and only if v_z is not an eigenvector of Δ. We will see, in the simulations, that when v_z is an eigenvector of Δ, the camera is in a local minimum. Besides, when Δ is spherical (i.e. when each eigen-value equals the other) $tr[\widehat{\Sigma_{k+1}}]$ is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

5.2.2 Best view point exploration

Now, instead of locally optimizing $\widehat{\Sigma_{k+1}}$, the best view point 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, we can show that the next best view point is located on the eigen vector of Σ_k associated with the biggest eigenvalue that is the most informative direction. The motion vector $(V_x, V_y)_k$ 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.



-a- -b-

Figure 4: Simulated trajectory for -a- the locally optimal exploration, -b- the best view point exploration

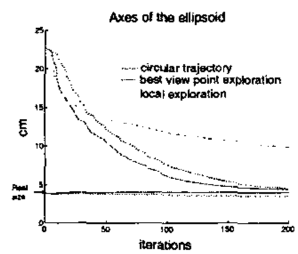


Figure 5: Comparison between circular trajectory and best view point strategy

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. 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 minimum. This is due to the biggest slope maximization induced by this technique. Even if further studies should be done about this issue, let us note that local minima correspond to points where v_z is an eigen-vector of Δ (see Section 5.2.1). We can 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 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 exploration process ran over a fixed number of iterations. It would be interesting to identify a suitable stopping criterion. 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 estimation algorithm imagined by Odobez and Bouthemy [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 degrees 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 shapes is given on Figure 6-c. Figure 6-d is the projection of this 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. 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 minimum 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. Several propagating rules have been infered 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

[1] N. Ayache. *Artificial Vision for Mobile Robots*. The MIT Press, Cambridge, MA, 1991.

[2] C. Connolly. The determination of next best views. In *Proc. IEEE Int. Conf. on Robotics and Automation*, volume 2, pages 432–435, St Louis, Missouri, March 1985.

[3] H. F. Durrant-Whyte. *Integration, Coordination, and Control of Multi-Sensor Robot Systems*. Kluwer Academic Publishers, Boston, 1987.

[4] H. F. Durrant-Whyte. Uncertain geometry in robotics. *IEEE Journal of Robotics and Automation*, 4(1):23–31, 1988.

[5] G. Flandin and F. Chaumette. Visual data fusion for complex objects localization. Technical Report 1394, IRISA-INRIA, April 2001. *ftp://ftp.irisa.fr/techreports/2001/PI-1394.ps.gz*

[6] K. N. Kutulakos, C. R. Dyer, and V. J. Lumelsky. Provable strategies for vision-guided exploration in three dimensions. In *IEEE Int. Conf. Robotics and Automation*, pages 1365–1372, Los Alamitos, CA, 1994.

[7] 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.

[8] 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.

[9] 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.

[10] J.M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Jal of Vis. Comm. and Im. Repr.*, 6(4):348–365, 1995.

[11] F. C. Schweppe. Recursive state estimation: unknown but bounded errors and system inputs. *IEEE Trans. on Automatic Control*, AC-13:22–28, 1968.

[12] 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.

[13] E. Welzl. Smallest enclosing disks (balls and ellipsoids). *Lecture Notes in Computer Science*, 555:359–370, 1991.

[14] P. Whaite and F. P. Ferrie. Autonomous exploration: Driven by uncertainty. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 339–346, Los Alamitos, CA, USA, June 1994. IEEE Computer Society Press.

[15] H. S. Witsenhausen. Sets of possible states of linear systems given perturbed observations. *IEEE Transactions on Automatic Control*, AC-13:556–558, 1968.

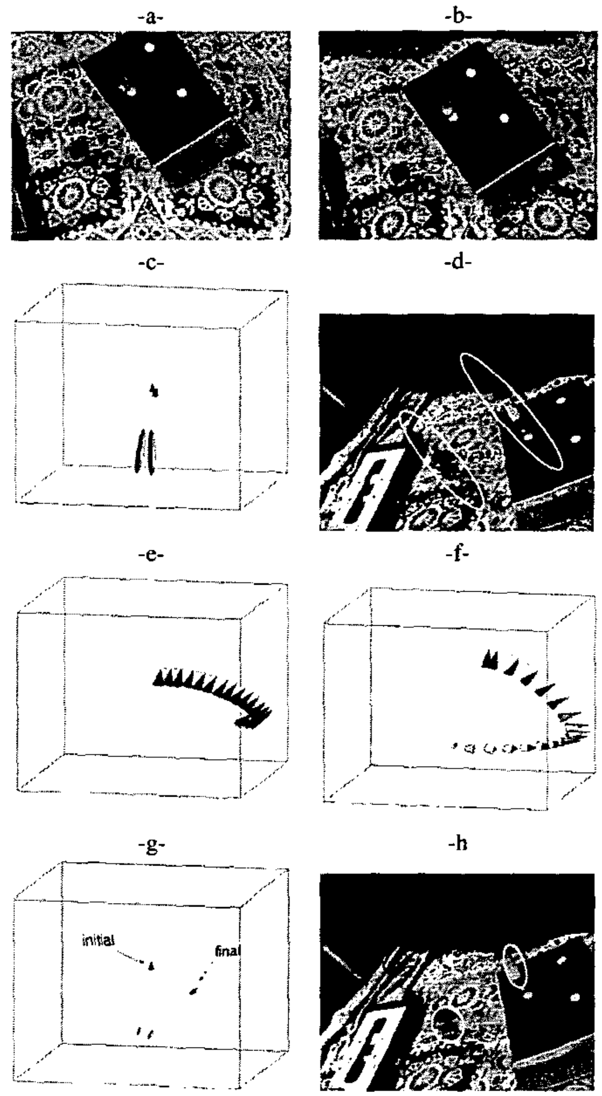


Figure 6: Experimental results