Structure from Motion using an Active Vision Paradigm

F. Chaumette , S. Boukir

IRISA / INRIA Rennes
Campus de Beaulieu, 35042 Rennes-cedex, France

Abstract

A method for the reconstruction and localization of geometrical primitives using active dynamic vision is presented in this paper. Our approach is based on the use of the interaction matrix related to the visual data describing a primitive. Next, active vision is considered by computing adequate camera motions with a control law in closed-loop with respect to visual data. Simulation results on the localization of a sphere are presented and show that active vision can to a large extent improve the accuracy of the structure estimation.

1 Introduction

A recent expansion of computer vision and image analysis is related to the estimation of 3D structure from image sequences. Using information extracted from an image sequence and the measure of successive positions of a moving camera (for example mounted on a mobile robot) aims at providing a clear and complete description of the camera environment. A great deal of effort has been concentrated on two main approaches:
- the discrete approach where images are acquired at discrete events and where camera motions may be large between two successive image acquisitions. This approach is based on the assumption that selected features generally persist from image to image and is confronted with the difficult inter-image correspondence problem [7] [4] [15].
- the continuous approach where images are acquired at the video rate [8] [16]. In this case, the emphasis is placed on the evaluation of the optical flow field [6] [13] which is generally noisy and partially incorrect especially near occlusion or motion boundaries. Thus, the main problem encountered by this approach lies on its sensitivity to noise.

The observability of the camera motion, which is necessary for structure estimation, characterizes a recent research domain called "dynamic vision". Furthermore, if together with this observability the motion can also be controlled using vision data, dynamic vision becomes "active vision" [1] [2] [12] [3]. The purpose of active vision is to manipulate the constraints underlying the observed phenomena, to improve the quality of perceptual results.

The aim of this work is to investigate the problem of recovering spatial structure using active vision. In this paper, the problem of 3D reconstruction and localization of geometrical primitives is examined under two different aspects:
- modelling aspect which consists in establishing a general method for three-dimensional reconstruction by dynamic vision;
- optimization aspect which consists in improving the performance of the proposed method by active vision.

2 Structure from motion using dynamic vision

2.1 Modelling

Let us consider a geometrical primitive $\mathcal{P}_s$ of the scene. Its configuration is specified by an equation of the type:

$$h(\underline{x}, \underline{p}) = 0 \ , \quad \forall \underline{x} \in \mathcal{P}_s \tag{1}$$

where $\underline{x} = (x \ y \ z)^T$ are the coordinates of a point in the scene, h defines the kind of the primitive and the value of $\underline{p}$ stands for its corresponding configuration.

Let us now denote $\mathcal{P}_i$ as the projection of $\mathcal{P}_s$ onto the image plane. The configuration of $\mathcal{P}_i$ can be written in the form:

$$g(\underline{X}, \underline{P}) = 0 \ , \quad \forall \underline{X} \in \mathcal{P}_i \tag{2}$$

where $\underline{X} = (X \ Y)^T$ are the coordinates of a point in the image plane, g defines the kind of the image primitive and the value of $\underline{P}$ represents its configuration.

Furthermore, we can define a function μ, which links the points belonging to $\mathcal{P}_s$ with the points belonging to $\mathcal{P}_i$, such that [9]:

$$z = \mu(\underline{X}, \underline{p}_0) \tag{3}$$

where the parameters $\underline{p}_0$ depend on $\underline{p}$.

Remark: For a plane primitive, the function μ represents the plane in which the primitive lies. For a three-dimensional primitive (sphere, cylinder, torus,...), the function $g(\underline{X}, \underline{P})$ is the limb equation (we only consider the contour of $\mathcal{P}_i$) and the matching between 3D points and contour points provides the function $\mu(\underline{X}, \underline{p}_0)$ which is thus called the limb surface.

Let $T = (V_x \ V_y \ V_z \ \Omega_x \ \Omega_y \ \Omega_z)^T$ be the camera velocity where V_x, V_y, V_z and $\Omega_x, \Omega_y, \Omega_z$ are respectively its translational and rotational components. The time variation of the parameters $\underline{P}$, which links the motion

0-8186-2910-X/92 $3.00 © 1992 IEEE

in the image with a camera motion T, can be explicitly derived [9] and we obtain:

$$\dot{\underline{P}} = L_{\underline{P}}^{T}(\underline{P}, \underline{p}_0)\, T \tag{4}$$

where $L_{\underline{P}}^{T}$, called interaction matrix related to $\underline{P}$, fully characterizes the interaction between the camera and the considered primitive.

We are now able to present a general method to localize a geometrical primitive by dynamic vision (i.e. to compute the value of $\underline{p}$ using the measure along an image sequence of the camera velocity T and of the image parameters $\underline{P}$ and $\dot{\underline{P}}$). It consists of the following steps:

- Step 1: Using the implicit function theorem, we obtain from (4):

$$\underline{p_0} = \underline{p_0}\,(T, \underline{P}, \dot{\underline{P}}) \tag{5}$$

- Step 2: Next, knowing $g(\underline{X}, \underline{P})$ and $\mu(\underline{X}, \underline{p}_0)$ and using geometrical constraints related to the considered primitive, we can achieve the parameters $\underline{p}$ which completely define its configuration:

$$\underline{p} = \underline{p}\,(\underline{P}, \underline{p_0}) \tag{6}$$

From a geometrical point of view, this continuous approach remains to determine the intersection between a generalized cone (defined by its vertex located at the optical center and the function $g(\underline{X}, \underline{P})$) and the limb surface given by the function μ. On the other hand, the discrete approach is based on the intersection between two generalized cones (one for each camera position) and therefore seems more complicated for achieving explicit results in the case of complex primitives.

This approach has already been used in the particular case of points and segments [8] [16]. In [5] to which we refer the reader for more details, we have used the two steps described above to recover the structure of non-polyhedral primitives such as circles, spheres and cylinders. In the next section, we give the obtained results for one of these primitives: the sphere.

2.2 Case of a sphere

We have chosen a position between the camera and the sphere as shown in figure 1. The sphere and the camera are represented in the left-hand part of this figure. The resulting image seen by the camera is depicted in the right-hand side of the same figure.

From this position, we apply a motion to the camera (in the present case $V_x = V_y = V_z = 0.5$ cm/s and $\Omega_x = \Omega_y = \Omega_z = 3$ dg/s) and, at each iteration, we apply the method described above in order to obtain the parameters which define the sphere (i.e. the coordinates x_0, y_0, z_0 of its center and its radius r). The results are shown in figure 2: in the left-hand part, the value of x_0, y_0, z_0 and r (in cm) computed at each iteration and expressed in a reference frame (the values should thus be constant since the sphere is motionless) and, in the right-hand part, the error between these computed values and the real ones.

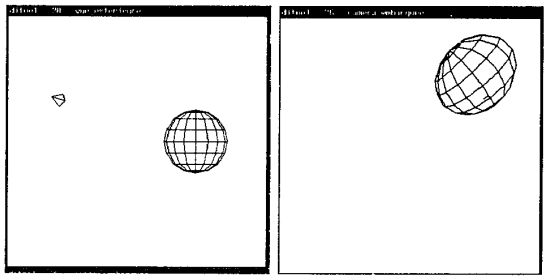

Figure 1: Position between the camera and the sphere

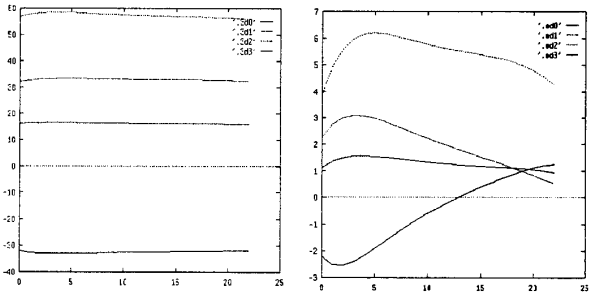

Figure 2: Reconstruction of a sphere by dynamic vision

In that case where none strategy for the camera motion is defined, we observe important errors. We will see in the next section that these results can be greatly improved by using active vision.

3 Structure from motion using active vision

As a matter of fact, it has been observed that the quality of estimation through a monocular image sequence is very sensitive to the successive motions of the camera [8]. Therefore, the goal in 3D reconstruction by active vision is to find an optimum motion which gives a better estimation of the spatial structure. The problem is stated in terms of the minimization of the errors inherent to the process.

3.1 Suppression of the discretization error

The main error encountered in structure from motion using dynamic vision comes from the discretization error. Indeed, our method is based on the measure of $\dot{\underline{P}}$, i.e. the time variation in the image of the parameters representing the considered primitive. The exact value of $\dot{\underline{P}}$ is generally unrecoverable and the image measurements provide only $\Delta\underline{P}$, the displacement of $\underline{P}$ between two image acquisitions. Using $\Delta\underline{P}/\Delta t$ instead of $\dot{\underline{P}}$ in the method described above generates discretization errors which may be important as seen in the results of the previous section.

On the other hand, if we ensure that $\dot{\underline{P}}\Delta t = \Delta\underline{P}$, the discretization will have no effect. Such a condition

is satisfied if :

$$\underline{\ddot{P}} = \cdots = \underline{P}^{[n]} = 0, \forall t \tag{7}$$

From (4), we have $\underline{\dot{P}} = f(\underline{P}, \underline{p_0}, T)$. Thus :

$$\underline{\ddot{P}} = \frac{\partial f}{\partial \underline{P}}\underline{\dot{P}} + \frac{\partial f}{\partial \underline{p_0}}\underline{\dot{p}_0} + \frac{\partial f}{\partial T}\dot{T} \tag{8}$$

If we consider $\dot{T} = 0$ (by setting T as the mean velocity between two camera positions), then, a sufficient and general condition to ensure (7) that will suppress the discretization error is to constrain the camera motions such that :

$$\underline{\dot{P}} = \underline{\dot{p}_0} = 0, \forall t \tag{9}$$

The visual servoing approach [14] [10] [9] is perfectly suitable to manage such camera motions. Indeed, this approach enables the establishment of control laws in closed-loop with respect to visual data. These control laws automatically compute camera motions ensuring some particular constraints on the camera environment, as $\underline{\dot{P}} = 0$ and $\underline{\dot{p}_0} = 0$ (see [9] and [5] for more details).

Figure 3, which is structured in the same way as figure 2 shows simulation results performed using this strategy. We can see that the discretization errors vanish and, since the process has been started with the same camera and sphere positions (see figure 1), we can observe the very important improvement brought by our active vision strategy with respect to dynamic vision.

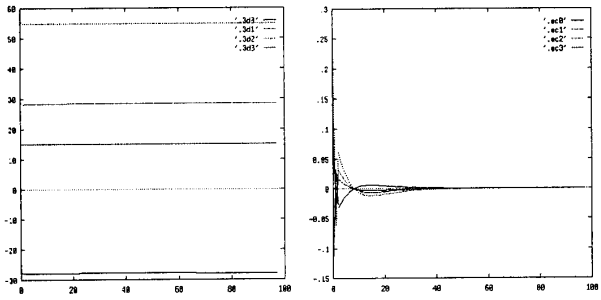

Figure 3: Reconstruction of a sphere by suppressing the discretization error (without noise)

Finally, figure 4 shows similar results when noise is added to image measurements (uniform noise of 3 pixels for a 512×512 image) and to the camera velocities (uniform noise of 2.5% of each velocity component). This is a rough way of simulating a bad calibration and measurement errors of the camera motion and of the acquired visual data. The results presented below show the robustness of the proposed method.

3.2 Minimizing the effects of the measurement errors

A more robust estimation can be obtained if we consider particular positions between the camera and

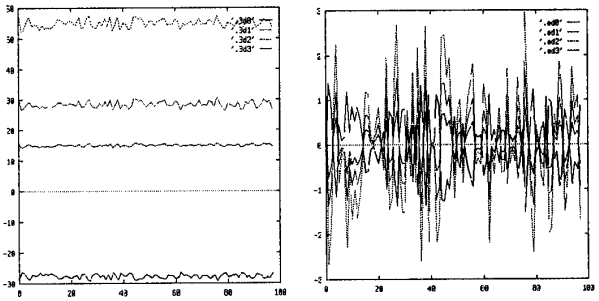

Figure 4: Reconstruction of a sphere by suppressing the discretization error (with noise)

the considered primitive. More precisely, the effect of the different measurement errors, which occur in $\underline{P}$, $\underline{\dot{P}}$ and T, upon the estimation of $\underline{p}$ depends on the measured values of $\underline{P}$, $\underline{\dot{P}}$ (which is constrained to zero in order to suppress the discretization error) and T (which is also constrained to ensure that $\underline{\dot{P}} = \underline{\dot{p}_0} = 0$).

Let us denote p as a parameter of the primitive. If we suppose that errors on $\underline{P}, \underline{\dot{P}}$ and T are not correlated, the uncertainty σ_p on p can be written in the form:

$$\begin{aligned}(\sigma_p)^2 &= \sum_{i=1}^{m}(\frac{\partial p}{\partial P_i})^2(\sigma_{P_i})^2 + \sum_{j=1}^{m}(\frac{\partial p}{\partial \dot{P}_j})^2(\sigma_{\dot{P}_j})^2 \\ &+ \sum_{k=1}^{6}(\frac{\partial p}{\partial T_k})^2(\sigma_{T_k})^2 \end{aligned} \tag{10}$$

We have to minimize each term $p_{a_i} = (\frac{\partial p}{\partial a_i})^2$, $a_i \in \{\underline{P}, \underline{\dot{P}}, T\}$ with respect to the 2D parameters $\underline{P}$. Solving explicitly this problem is complicated. Therefore, we just select some image configurations of the primitive that intuitively should provide a better estimation. Then, we check that these configurations minimize each term (i.e. $\frac{\partial p_{a_i}}{\partial \underline{P}} = 0$ for the selected values of $\underline{P}$). Let us note that we have also verified that all values of $\underline{P}$ do not minimize p_{a_i} [5].

In the particular case of a sphere, the effects of measurement errors are minimized when the image of the sphere is a centered circle (see Figure 5). Once more, the visual servoing approach enables to compute camera motions ensuring this constraint between the camera and the sphere. Figure 6 shows the results obtained after the convergence of the control law. By comparing these results with the previous ones we can notice the important improvement provided by this strategy. Let us note that similar results have been obtained for other primitives like points, lines and cylinders [5].

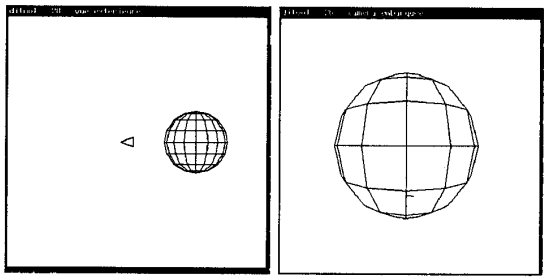

Figure 5: Optimal position between the camera and the sphere

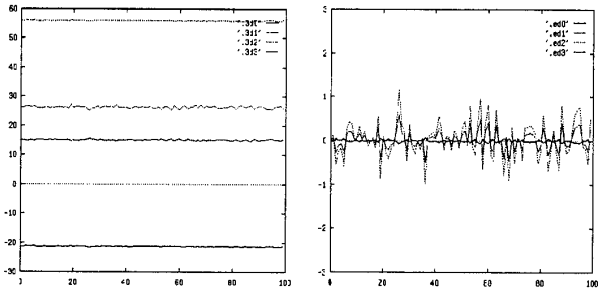

Figure 6: Reconstruction of a sphere by minimizing the effects of the measurement errors (with noise)

4 Conclusion

We have proposed in this paper an original method for 3D structure estimation of geometrical primitives. More specifically, the interaction matrix provides a general method valid for complex primitives such as cylinders, spheres,...

Since the nature of the camera motion affects the accuracy of the results, we have focused our efforts on this critical aspect of dynamic vision. We have shown that, by cleverly positioning the camera using a control law in closed-loop with respect to visual data, noticeable improvements have been obtained in the reconstruction and localization of the primitives. Our approach has been validated in simulations and turned out to be a powerful and effective one. The final results confirmed the point of view of previous works about the promising future of active vision [3] [1] [2] [12]. Finally, let us report that actual work is devoted to the implementation of the algorithms described in this paper in an experimental cell in order to obtain results from real images.

Acknowledgements

This work is done in collaboration with CEA-LETI DEIN-SIR under contract 1 91 C 244 00 31315 01 1.

References

[1] J. Aloimonos, A. Bandyopadhay: *Active Vision*, 1^{st} Int. Conf. on Computer Vision, pp. 35-54, London, England, June 1987.

[2] R. Bajcsy: *Active Perception*, Proc. of the IEEE, Vol 76, n. 8, pp. 996-1005, August 1988.

[3] A. Bandyopadhay, B. Chandra, D.H. Ballard: *Ego-motion using Active Vision*, Int. Conf on CVPR 86, pp. 498-503, Miami Beach, Florida, USA, June 1986.

[4] B. Boukarri: *Reconstruction 3D récursive de scènes structurées au moyen d'une caméra mobile. Application à la robotique*, PhD-Thesis, Orsay University, CEA, France, October 1989.

[5] S. Boukir, F. Chaumette: *Reconstruction 3D de primitives géométriques par vision active*, IRISA Research report, France, May 1992.

[6] P. Bouthemy: *A Maximum Likelihood Framework for Determining Moving Edges*, IEEE Trans. on PAMI, Vol 11, n. 5, pp. 499-511, May 1989.

[7] C.H. Chien, J.K. Aggarwal: *Model Construction and Shape Recognition from Occluding Contours*, IEEE Trans. on PAMI, Vol 11, n. 4, pp. 372-389, April 1989.

[8] B. Espiau, P. Rives: *Closed-Loop Recursive Estimation of 3D Features for a Mobile Vision System*, IEEE Int. Conf. on Robotics and Automation, Raleigh, North Carolina, USA, Vol. 3, pp. 1436-1443, April 1987.

[9] B. Espiau, F. Chaumette, P. Rives: *A new approach to visual servoing in robotics*, to appear in IEEE Trans. on Robotics and Automation, 1992.

[10] J.T. Feddema, O.R. Mitchell : *Vision-Guided Servoing with Feature-Based Trajectory Generation*, IEEE Trans. on Robotics and Automation, Vol. 5, n. 5, pp. 691-700, October 1989.

[11] K. Price: *Multi-frame Feature-Based Motion Analysis*, 10^{th} Int. Conf. on Pattern Recognition, Vol 1, pp. 114-117, June 1990.

[12] G. Sandini, M. Tistarelli: *Active Tracking Strategy for Monocular Depth Inference over Multiple Frames*, IEEE Trans. on PAMI, Vol 12, n. 1, pp. 13-27, January 1990.

[13] B.G. Schunck: *Image Flow Segmentation and Estimation by Constraint Line Clustering*, IEEE Trans. on PAMI, Vol 11, n. 10, pp. 1010-1027, October 1989.

[14] L.E. Weiss, A.C. Sanderson : *Dynamic Sensor-Based Control of Robots with Visual Feedback*, IEEE Journal of Robotics and Automation, Vol 3, n. 5, pp. 404-417, October 1987.

[15] J. Weng, T.S. Huang, N. Ahuja: *Estimation and Structure from Line Matches: Performance Obtained and Beyond*, 10^{th} Int. Conf on Pattern Recognition, Vol 1, pp. 168-172, June 1990.

[16] M.Xie, P. Rives, C. Labit : *Reconstruction of 3D polyhedral scene by a moving camera.* Congrès PIXIM 89, Paris, France, September 1989.