Tracking (3)

Recognition, Pose and Tracking of Modelled Polyhedral Objects by Multi-Ocular Vision

P. Braud J.-T. Lapresté M. Dhome

LASMEA, URA 1793 of the CNRS, Université Blaise Pascal
63177 Aubière Cedex, FRANCE
dhome@le-eva.univ-bpclermont.fr

Abstract. We developped a fully automated algorithmic chain for the tracking of polyhedral objects with no manual intervention. It uses a multi-cameras calibrated system and the 3D model of the observed object.
The initial phase of the tracking is done according to an automatic location process using graph theoretical methods. The originality of the approach resides mainly in the fact that compound structures (triple junction and planar faces with four vertices) are used to construct the graphs describing scene and model. The association graph construction and the search of maximal cliques are greatly simplified in this way. The final solution is selected among the maximal cliques by a prediction-verification scheme.
During the tracking process, it is noticeable that our model based approach does not use triangulation although the basis of the multi-ocular system is available. The knowledge of calibration parameters (extrinsic as well as intrinsic) of the cameras enables to express the various equations related to each images shot in one common reference system. The aim of this paper is to prove that model based methods are not bound to monocular schemes but can be used in various multi-ocular situations in which they can improve the overall robustness.

KeyWords : **3D model - multi-cameras - graph theory - prediction-verification - localisation without triangulation - tracking**

1 Introduction

Model based recognition technics has received a growing attention from many artificial vision researchers; one can consult [CD86] for a detailled state of the art. Various recognition systems exist that can be classified by the kind of primitives used (2D or 3D) to describe the model and extracted from the images. The more frequently used methods are based on the partionning of the parameters space [Ols94], on the traversal of a tree of matches between model and images primitives [Gri89], or on the direct search of compatibles transforms between model and images [HU90].

These various approaches share the property of using an objects models database. Our aim is quite different, as it consists in the computation of the pose of a polyhedral object which is known to pertain to the scene.

The recognition proposed algorithm follows S. Pollard, J. Porrill, J. Mayhew and J. Frisby [PPMF87] and is based on a research of constraints between 3D primitives extracted from the images and of the model. The originality of the approach resides mainly in the fact that compound structures (triple junction and planar faces with four vertices) are used to construct the graphs describing scene and model. The association graph construction and the search of maximal cliques are greatly simplified in this way.

Actually, the paper intends to deal with the more general problem of temporal tracking of polyhedral objects along multi-ocular images sequences. The automatic initial pose estimate of the 3D object allows to begin the tracking process with no manual intervention.

We, thus propose a new approach to localization and tracking in a multi-cameras system: this approach does not involve triangulation. Our aim is to prove that model based methods are not bound to monocular schemes but can be used in various multi-ocular situations in which they can improve the overall robustness.

To solve the general problem of inverse perspective, various searchers have used the model based paradigm in monocular vision [Kan81] [RBPD81] [FB81] [HCLL89] [Bar85] [SK86] [Hor87] [DRLR89] [Rei91] [Low85].

To compensate the loss of information due to the perspective projection, it is necessary to provide a priori knowledges about what is seen. If a model is available it is then necessary to match primitives arising from the images treatments to parts of the model and write down the projection equations relating these primitives. The methods can be classified using the kind of primitives involved (points, segments, elliptical contours, limbs, ...) and the projection chosen to modelize the image formation process (orthographic or perspective projection for example).

When more than one camera are at disposal, generally, a 3D reconstruction of the observed object is made by triangulation. During a tracking process, the reconstructed model is filtered along the images sequence [Aya89]. Few articles, in such cicumstances, assume the model knowledge. M. Anderson and P. Nordlund [AN93] improve incrementally the reconstructed model and match it with the known 3D model to increase the pose accuracy. Our method is similar, but we show that the step of triangulation is not necessary.

This work is based on a previous paper where it was proven that the localization and tracking of a rigid body by monocular vision can be performed using images segments and model ridges as primitives [DDLR93]. The pose estimate was done using a Lowe-like algorithm ([Low85]), coupled with a constant velocity Kalman filter. We had shown the robustness of the process even if the actual velocity was far from the constant model wired in the Kalman filter.

However such a method has limits: object occultations along the sequence can lead to divergence of the tracking. [Gen92] has shown that more than one camera can help if the object is not simultaneously lost by all cameras... but at each iteration he only makes use of one of the images to minimize computation times. We have shown (which is intuitively obvious) in [BDLD94] that using

simultaneously several cameras improves pose estimation as well as robustness.

The paper organization is the following: section 2 describe the recognition algorithm. The multi-ocular localization algorithm is the subject of section 3. The tracking is treated in section 4. Finally we present experimental results in section 5 and then conclusions.

2 Recognition

N. Daucher, M. Dhome, J.-T. Lapresté et G. Rives [DDLR93] have developped a method for localization and tracking of polyhedral objects by monocular vision using a surfacic 3D model. As in every tracking system, it is necessary to know the initial pose of the object. Up to now, this knowledge was obtained from a manual match of image segments and model ridges. From a multi-cameras system and also using a surfacic 3D model, we have build a recognition algorithm that can be summarized by the 6 following steps:

- polygonal segmentation of the images (to be used further);
- corner detection in the images from the N cameras at subpixel precision;
- stereo matching by cameras pairs of the detected corners,
- 3D reconstruction of the matched corners using the cameras parameters. At this step we have 3D points and 2D connection between these points obtained from images segments (from the polygonal segmentation);
- 3D matching between these reconstructed corners and the model corners;
- pose computation using the matchings deduced between the images and the model.

We shall not discuss here of the polygonal segmentation which can now be considered as a classic.

2.1 Corners extraction

We detect corners in two phases: coarse and loose then fine and severe.

1. the method of C. Harris et M. Stephens [HS88] has been implemented because it offers a good trade off between accuracy and simplicity. Corners are extracted at pixel accuracy and many false ones are present.
2. the model based method of R. Deriche et T. Blaszka [DB93] is then used to refine or suppress the previously detected corners. After this step, in each image we possess a list of angular points and triple junctions.

2.2 Stereo matching

We thus produce matches between images pairs with the algorithm from Z. Zhang, R. Deriche, O. Faugeras et Q.-T. Luong [ZDFL94]. They have developped a robust matching scheme based on the computation of the epipolar geometry. In fact, in our case we know the epipolar geometry and it is only the final part of their method we will borrow.

2.3 3D reconstruction

If for instance one has a three cameras system, the outputs of the previous step must be treated to combine the pair of image results. It is necessary to group matches sharing one primitive. To disambiguate, we use the trinocular epipolar constraint.

The reconstruction will be obtained by a least square minimization of the distances between all the lines that have been grouped together as defining the same spatial primitive.

Then, an algorithm relying on previous polygonal segmentations is able to etablish "physical" links between the reconstructed points.

2.4 3D matching

The recognition problem consists to establish the list of the matches between vertices of the model and corners of the reconstructed scene. The major drawback of such an approach is the combinatorial explosion. Thus, to limit the complexity, we have chosen to define new elaborated structures both in the scene and the model: triple junctions and four edges planar faces.

From the graph of the model and the graph of the scene constructed from the elaborated structures, we establish the association graph $\mathcal{A}$.

At last, when $\mathcal{A}$ is constructed it remains to find the maximal cliques. Their exists many algorithms [Fau93] [Sko88] which solve this problem, and they are all based on the extension of an initial clique.

2.5 Localization

We are ready know to choose in the cliques list the best one. It is not necessarily the largest one because false matches can still occur. To choose the best solution we use a prediction-verification scheme.

From the matches a multi-ocular localization is performed computing the model pose compatible with the matches and the associated covariance matrix. The verification algorithm tests the exactness of the pose: an automatic matching scheme between image segments and projection of model ridges controled by the Mahalanobis distance is used. Finally we retain the pose which lead to the maximal number of such matches.

3 Localization

3.1 General formulation

The multi-ocular localization algorithm is an extension of a monocular one described in a paper by N. Daucher, M. Dhome, J.-T. Lapresté et G. Rives [DDLR93].

The problem we intend to address here consists to find the extrinsic parameters giving the pose of the viewed object in the acquisition reference system related to the first camera (arbitrarily chosen) and the uncertainties associated to these parameters.

3.2 Criterium definition

We are looking for the 6-dimensional vector $\mathcal{A}_1 = (\alpha_1, \beta_1, \gamma_1, t_{x_1}, t_{y_1}, t_{z_1})^t$ defining the rotation R_1 (by the three Euler angles) and the translation T_1 that minimize the sum of the squared distances of the vertices of the model to the interpretation planes of the images segments.

3.3 Solving the equations

We simply use a classical Newton-Raphson iterative scheme. Starting from an initial guess, to the previous equations we substitute at step k the set of linear equations given by their first order developpement at the current solution. Then we solve the linear system writing down the normal equations.

4 Tracking

4.1 The tracking algorithm

From a multi-cameras system and a 3D object model we have developped a tracking algorithm that can be split in three parts. At each shot k of the images acquisition sequence, we perform succesively:

- An automatic matching between the various images segments and the model ridges. The process is made independantly on each image of the shot looking for informations directly in the grey level data. The search is guided by the prediction of the model pose and its covariance matrix from the previous step.
- The new estimation of the model pose in the global reference system determined from the matchings using the algorithm described in the previous section.
- The prediction of the model pose at the next step of the tracking and the associated covariance matrix obtained by Kalman filtering.

4.2 Automatic matching in grey-level images

At each shot we have the previous pose of the object as well as the predicted new pose. The automatic grey-level match consists in finding independantly for each camera the correspondances between model ridges and image segments.

We have seen previously that a matching could be made using a global polygonal segmentation of the image: this was done in the initialisation phase where no pose estimates were at hand. As the possess the previous pose and the prediction of the new one, we can predict:

- the nature of the grey level transition leading to a segment detection (from the previous image),
- the approximate position in the new image of a comparable signal.

The new location of the segment in the new image will be obtained by correlation with the corresponding signal in the previous image. To be accurate let us say the matching is performed in three phases (independantly on each image):

- We compute the perspective projection of each visible model ridges at computed at step k and predicted for step $k+1$. We also determine with the help of the covariance matrix of the pose parameters, the size of a square window centered on the extremities predicted on the images from step $k+1$.
- In each of these windows we look for the point homologous to the one which is the projection at step k of the same model vertex. This is done by a pyramidal correlation algorithm.
- Between each pair of extremities supposed to belong to the projection of the same model ridge we look for contour points to verify the presence of a ridge projection. The points are extracted by convolution with an ideal model of transition (step or line) choosen from the previous image. Finally the points alignment is tested by a RANSAC paradigm[1] [FB81].

5 Experimentations

5.1 Recognition

The recognition algorithm has been tested on a real images trinocular sequence of a cassette. Figure 1 present images acquired by the three cameras at the beginning of the trinocular sequence. Figure 2 present the final result of the recognition algorithm obtained from this three images. In this sequence, the triangulation basis is weak, so the 3D reconstruction is less accurate. Nevertheless the algorithm converges to the right solution.

In [BDL96], we have shown results of the recognition algorithm obtained from trinocular sequences of others objects.

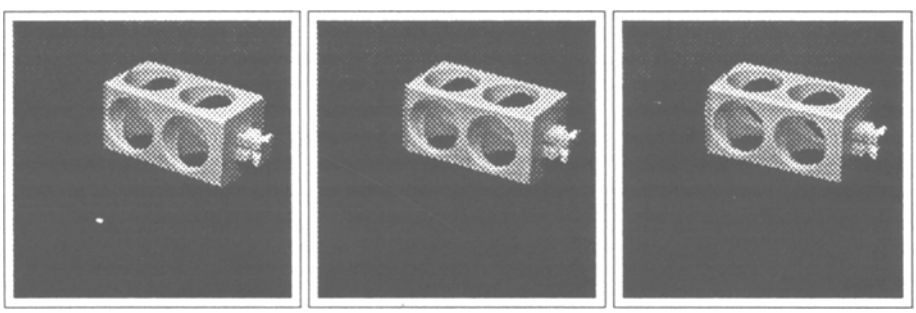

Fig. 1. Left, center and right images: grey level images of the cassette given by cameras #1, #2 and #3 respectively.

[1] RANdom SAmple Consensus

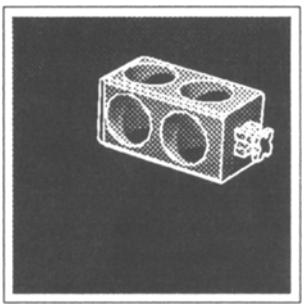

Fig. 2. Final solution obtained by the recognition algorithm on the cassette: superimposition of the grey level image from camera #1 and of the model projection.

5.2 Tracking

During the contract ESTEC #8019/88/NL/PP with the European Spatial Agency, SAGEM provided a trinocular sequence of real images and the CAO model of an "Orbital Replacement Unit" (ORU). The tracking algorithm has been tested on this trinocular sequence. Process has been initiated with the result of the recognition algorithm. The results can be seen in figures 3

In [BDLD94], we have shown by comparing experiments on trinocular and monocular sequences that:

- the multi-ocular approach was more robust,
- The pose computation was more accurate in the multi-cameras case.

In fact, we have shown that the monocular approach was robust if the Kalman filtering had properly modelled the object kinematics. If the object is occulted along some images and the movement has not been properly modelled yet, when the visual information is restored, the pose prediction can be wrong and imply a divergence of the tracking.

Of course the multi-cameras approach is more robust and does not break down in the case of an occultation occuring on one of the cameras. This property could be used in real application to request from some of the cameras a transitory task as (re)calibration.

Speaking of accuracy, the more sensible parameter in the monocular case is obviously z: the depth. Comparing the results, we have found an increased accuracy in trinocular localization.

6 Conclusion

We have developped an algorithm of temporal tracking of polyhedral objects with no human intervention (see [Bra96] for more details). The major lack in existing methods is the initialisation step : we have adressed it.

The initial phase of the tracking is done according to an automatic location process based on a multi-cameras system and using graph theoretical methods.

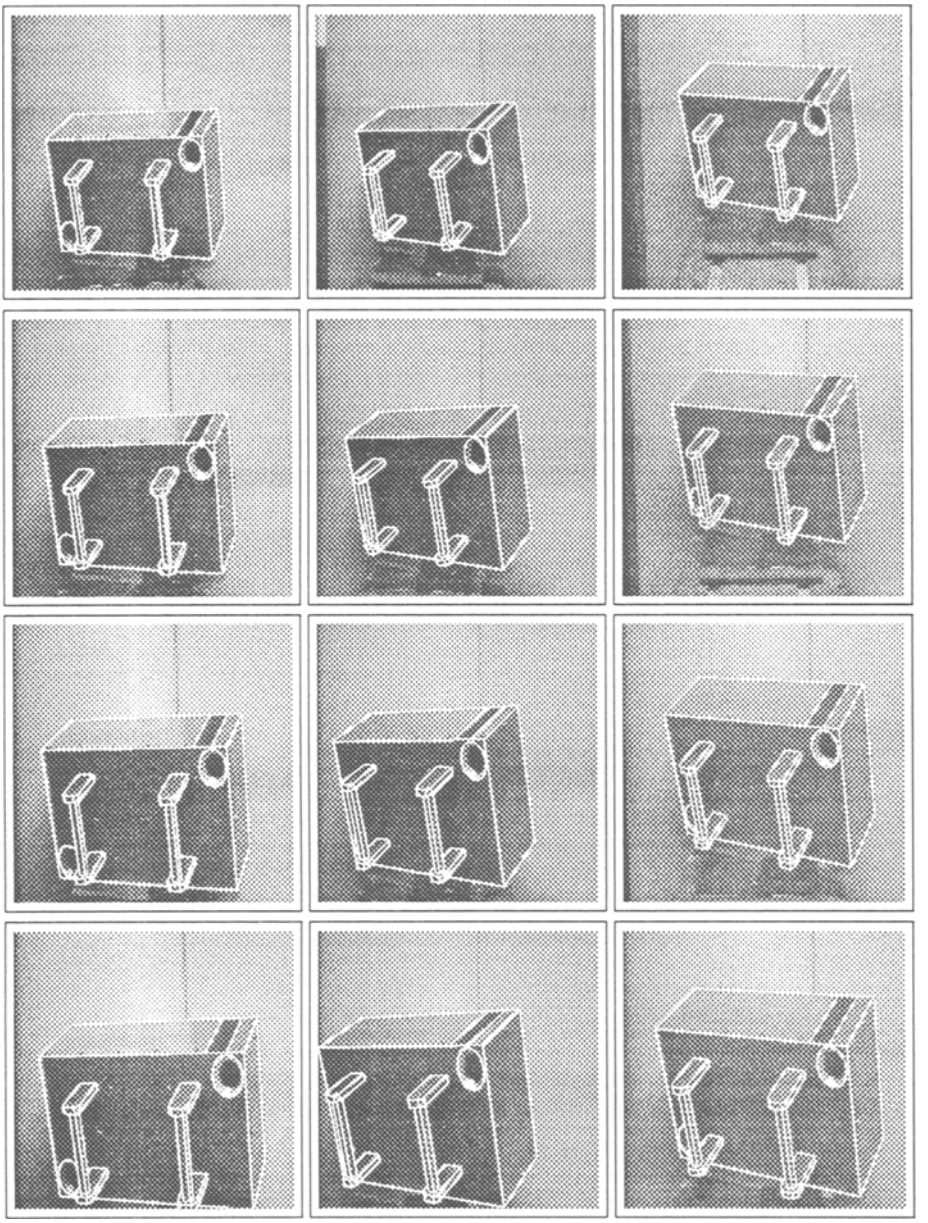

Fig. 3. trinocular tracking of the ORU. Left, center and right are cameras #1, #2 and #3 respectively; from top to bottom : shots 0, 13, 26 and 39 superimposition of the grey level image and the perspective projection of the models ridges.

The originality of the approach resides mainly in the fact that compound structures (triple junction and planar faces with four vertices) are used to construct the graphs describing scene and model. The steps of construction of the association graph and the search of maximal cliques have been greatly simplified in this way.

The algorithm has also proven its robustness facing a poor triangulation basis and thus a less accurate reconstruction.

Moreover we have shown for the tracking process that model based methods that are generally viewed as purely monocular ones can be extended to multi-oculars by merging data without using the triangulation basis of the system (after the initialisation phase).

Up to now, few articles have treated of the model based approach in a multi-camera environment. It appears interesting to have an alternative solution to the stereographic reconstruction. This method allows to disconnect a camera or to assign it to a different task as long as at least one camera goes on providing tracking informations.

References

[AN93] M. Anderson and P. Nordlund. A model based system for localization and tracking. In *Workshop on Computer Vision for Space Applications*, pages 251–260, September 1993.

[Aya89] N. Ayache. *Vision stéréoscopique et perception multisensorielle.* InterEditions, Science Informatique, 1989.

[Bar85] S.T. Barnard. Choosing a basis for perceptual space. *Computer Vision Graphics and Image Processing*, 29(1):87–99, 1985.

[BDL96] P. Braud, M. Dhome, and J.-T. Lapresté. Reconnaissance d'objets polyédriques par vision multi-oculaire. In $10^{ème}$ *Congrès Reconnaissance des Formes et Intelligence Artificielle*, Rennes, France, January 1996.

[BDLD94] P. Braud, M. Dhome, J.T. Lapresté, and N. Daucher. Modelled object pose estimation and tracking by a multi-cameras system. In *Int. Conf. on Computer Vision and Pattern Recognition*, pages 976–979, Seattle, Washington, June 1994.

[Bra96] P. Braud. *Reconnaissance, localisation et suivi d'objets polyédriques modélisés par vision multi-oculaire.* PhD thesis, Université Blaise Pascal de Clermont-Ferrand, France, January 1996.

[CD86] R.T. Chin and C.R. Dyer. Model-based recognition in robot vision. *ACM Computing Surveys*, 18(1):67–108, March 1986.

[DB93] R. Deriche and T. Blaszka. Recovering and characterizing image features using an efficient model based approach. In *Int. Conf. on Computer Vision and Pattern Recognition*, pages 530–535, New-York, June 1993.

[DDLR93] N. Daucher, M. Dhome, J.T. Lapresté, and G. Rives. Modelled object pose estimation and tracking by monocular vision. In *British Machine Vision Conference*, volume 1, pages 249–258, October 1993.

[DRLR89] M. Dhome, M. Richetin, J.T. Lapresté, and G. Rives. Determination of the attitude of 3d objets from a single perspective image. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11(12):1265–1278, December 1989.

[Fau93] O.D. Faugeras. *Three Dimensional Computer Vision : A Geometric Viewpoint*, chapter 12. MIT Press, Boston, 1993.

[FB81] M.A. Fischler and R.C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Com. of the ACM*, 24(6):381–395, June 1981.

[Gen92] D.B. Gennery. Tracking of known three-dimensional objects. *International Journal of Computer Vision*, 7(3):243–270, April 1992.

[Gri89] W.E.L. Grimson. On the recognition of parameterized 2d objects. *International Journal of Computer Vision*, 2(4):353–372, 1989.

[HCLL89] R. Horaud, B. Conio, O. Leboulleux, and B. Lacoll. An analytic solution for the perspective 4 points problem. In *Computer Vision Graphics and Image Processing*, 1989.

[Hor87] R. Horaud. New methods for matching 3-d objects with single perspective view. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 9(3):401–412, May 1987.

[HS88] C.G. Harris and M. Stephens. A combined corner and edge detector. In *Fourth Alvey Vision Conference*, pages 189–192, Manchester, United Kingdom, August 1988.

[HU90] D.P. Huttenlocher and S. Ullman. Recognizing solid objects by alignment with an image. *International Journal of Computer Vision*, 5(2):195–212, 1990.

[Kan81] T. Kanade. Recovery of the three dimensional shape of an object from a single view. *Artificial Intelligence, Special Volume on Computer Vision*, 17(1-3), August 1981.

[Low85] D.G. Lowe. *Perceptual Organization and Visual Recognition*, chapter 7. Kluwer Academic Publishers, Boston, 1985.

[Ols94] C.F. Olson. Time and space efficient pose clustering. In *Int. Conf. on Computer Vision and Pattern Recognition*, pages 251–258, Seattle, Washington, June 1994.

[PPMF87] S.B. Pollard, J. Porrill, J. Mayhew, and J. Frisby. Matching geometrical descriptions in three space. *Image and Vision Computing*, 5(2):73–78, 1987.

[RBPD81] P. Rives, P. Bouthemy, B. Prasada, and E. Dubois. Recovering the orientation and position of a rigid body in space from a single view. Technical report, INRS-Télécommunications, 3, place du commerce, Ile-des-soeurs, Verdun, H3E 1H6, Quebec, Canada, 1981.

[Rei91] P. Reis. *Vision Monoculaire pour la Navigation d'un Robot Mobile dans un Univers Partiellement Modélisé.* PhD thesis, Université Blaise Pascal de Clermont-Ferrand, March 1991.

[SK86] T. Shakunaga and H. Kaneko. Perspective angle transform and its application to 3-d configuration recovery. In *Int. Conf. on Computer Vision and Pattern Recognition*, pages 594–601, Miami Beach, Florida, June 1986.

[Sko88] T. Skordas. *Mise en correspondance et reconstruction stéréo utilisant une description structurelle des images.* PhD thesis, Institut National Polytechnique de Grenoble, October 1988.

[ZDFL94] Z. Zhang, R. Deriche, O. Faugeras, and Q.-T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. Technical Report 2273, INRIA Sophia-Antipolis, France, May 1994.