Using Singular Displacements for Uncalibrated Monocular Visual Systems

T. Viéville and D. Lingrand

INRIA, Sophia, BP93, 06902 Valbonne, France.
tel : +33 93 65 76 88 fax : +33 93 65 78 45
email : vthierry@sophia.inria.fr

Abstract. In the present paper, we review and complete the equations and the formalism which allow to achieve a minimal parameterization of the retinal displacement for a monocular visual system without calibration.
Considering the emergence of active visual systems for which we **can not** consider that the calibration parameters are either known or fixed, we develop an alternative strategy using the fact that certain class of special displacements induces enough equations to evaluate the calibration parameters, so that we can recover the affine or Euclidean structure of the scene when needed.
A synthesis of what can be recovered for singular displacements in terms of camera calibration, scene geometry and kinematics is proposed. We give, for the different levels of calibration, an exhaustive list of the geometric and kinematic information which can be recovered. Following a strategy based on special kind of displacements, such as fixed axis rotations or pure translations for instance, we describe how to detect this particular classes of displacement.

Key-Words Structure and Motion, Singular Displacements, Self-Calibration

1 Introduction

The analysis of motion in the case of an uncalibrated monocular image sequence has already been developed by several authors, considering point and/or line correspondences or correspondences between planar patches and using either a discrete or a continuous representation of the rigid displacement between two or more frames.

These studies are motivated by the fact that *we must not consider an active visual system is calibrated* [6]. However, it has been demonstrated that, in the general case, it is not possible to self-calibrate the camera when zooming or modifying the intrinsic calibration parameters.

Considering this fact, the key idea of the present study is that **several singular displacements induce enough equations to evaluate the calibration parameters.**

For instance, fixed axis rotations of known angles or pure rotations [5] allow to estimate the calibration parameters, their uncertainty and, for a given kind of displacement, which parameters are optimally estimated, so that active visual strategies can be developed. On the other hand, pure translations do not allow to calibrate the Euclidean geometry of the scene [8], but its affine geometry [11].

Collecting all this information and considering a suitable statistical framework as in [1], it is then possible to infer which kind of displacement will increase at most the information (usually represented by the inverse of a covariance matrix) on the scene geometry, object kinematics and calibration parameters.

This is the goal of this paper.

In order to attain this objective, we are first going to review the theory of motion when no calibration: equations, parameterization of motion, etc...

We then are going to propose a synthesis of what can be recovered in terms of scene geometry and kinematics when calibration is not given as an input: describe the different forms of calibration, the different levels of calibration and give an exhaustive list of the different geometric and kinematic information to be recovered, depending the chosen geometry.

2 Reviewing the theory of motion when no calibration.

Notations. We write vectors and matrices using bold letters, matrices being written with capital letters. The duals of vectors are represented as the transpose of a vector and scalars in italic. The notation $\mathbf{x} \wedge \mathbf{y} = \tilde{\mathbf{x}}\mathbf{y}$ corresponds to the cross-product, the dot-product being written as $\mathbf{x}^T\mathbf{y}$. $\tilde{\mathbf{x}}$ is a 3×3 skew-symmetric matrix[1] . The identity matrix is written $\mathbf{I}$. Geometric objects such as points, lines, planes are written with capital letters in 3D, and small letters in 2D. We represent the components of a matrix or a vector using superscripts from 0 to 2, e.g.: $\mathbf{x} = (x^0, x^1, x^2)^T$.
We write $\mathbf{a} \equiv \mathbf{b}$ if $\mathbf{a}$ is equal to $\mathbf{b}$ up to a scale factor, i.e. $\exists k, \mathbf{a} = k\ \mathbf{b}$.

2.1 Setting the equations

Camera model and frame of reference. We use *the standard pinhole model* for a camera, assuming the camera performs a perfect perspective transform with center C (the camera optic center) at a distance f (the focal length) of the retinal plane. The pinhole model can still be used for a zoom lens if the object-to-image distance is not considered as fixed.

All coordinates are related to an affine frame of reference $\mathcal{R} = (C, \mathbf{x}, \mathbf{y}, \mathbf{z})$ *attached to the retina*, $\mathbf{z}$ being aligned with the optical axis, $\mathbf{x}$ and $\mathbf{y}$ being aligned with the horizontal and vertical axe in the image. The retinal plane is thus perpendicular to the optical axis Cz,as shown in figure 1.

[1] Remember that a 3×3 skew-symmetric matrix has 3 parameters and can always be represented by the crossproduct of a vector, i.e. is of the form $\tilde{\mathbf{x}}$ for some $\mathbf{x}$.

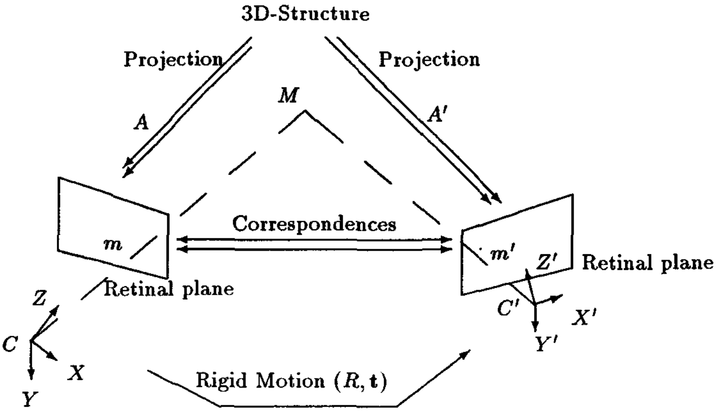

Fig. 1. Elements used in the definition of the problem

Using points as primitives. We represent a 3D-point M by the vector $\mathbf{M} = \mathbf{CM} = (X, Y, Z)^T$ using Euclidean coordinates. Points in the retina, with pixel coordinates coordinates (u, v) will be represented as homogeneous 3-D vectors: $\lambda\ \mathbf{m} = \lambda\ \mathbf{Cm} = \lambda\ (u, v, 1)^T$, corresponding to lines of a given direction passing through the optical center (2-D projective space).

Other primitives will be represented using set of points. This will be discussed in the sequel.

A suitable model of the intrinsic parameters of the camera. In this study, *we do not assume the system is calibrated.* However, we are in a specific situation because we have chosen a "canonical" frame attached to the retina. Therefore, we consider only the matrix of the intrinsic parameters (called A-matrix) in the projection and write:

$$Z\,\mathbf{m} = \mathbf{A}\,\mathbf{M} \quad , \quad \mathbf{A} = \begin{pmatrix} f & 0 & u_0 \\ 0 & f & v_0 \\ 0 & 0 & 1 \end{pmatrix} \tag{1}$$

A complete review can be found in [1].

In the present model, (u_0, v_0) is the principal point, and f the focal length; following [8], we assume that **we know the ratio between the horizontal and vertical focal length** and that we assume that **the two retinal coordinates are orthogonal.** It has been shown experimentally that these assumptions are valid for standard cameras [8] and also for high-level visual sensors [10]. Using this simple model will allow us to improve the obtained results.

We also assume that the intrinsic parameters are different for each camera position, as during a zoom. In the consecutive frame $\mathcal{R}' = (C', \mathbf{x}', \mathbf{y}', \mathbf{z}')$ we write:

$$Z' \, \mathbf{m}' = \mathbf{A}'\mathbf{M}' \tag{2}$$

Representation of rigid displacements. We consider motion of rigid objects and the ego-motion of the camera, *in the discrete case.* We thus represent motion through rigid displacements.

It means that the tokens in the scene are undergoing a rigid displacement parameterized by a rotation matrix R and a translation vector $\mathbf{t}$:

$$\mathbf{M}' = \mathbf{R}\,\mathbf{M} + \mathbf{t} \tag{3}$$

2.2 Parameterization of motion when no calibration.

The goal of the parameterization of motion is the following: given a set of points in correspondence between two views, i.e. a set of matches $\{m.m'\}$ we want to analyse all constraints which relate the two points i.e find the equations of the form $\forall\{m.m'\}, f(m, m') = 0$. In particular, we would like to predict the location of a point given its correspondent, i.e. a relation of the form $\forall\{m.m'\}, m' = g(m)$. Having such parameterization allows to exact all information available from the retinal displacements, which is measured through the set of matches.

The Qs-representation and the F-matrix. Considering the 2D correspondences between two points m and m' in two different frames, we obtain, combining equations (1),(2) and (3):

$$Z' \, \mathbf{m}' = Z \underbrace{\mathbf{A}'\mathbf{R}\mathbf{A}^{-1}}_{\mathbf{H}_\infty} \mathbf{m} + \underbrace{\mathbf{A}'\mathbf{t}}_{\mathbf{s}} \tag{4}$$

where the Q-matrix $\mathbf{H}_\infty$ corresponds to the "uncalibrated rotational component of the rigid displacement", or more geometrically *the collineation of the plane at infinity,* while the s-vector corresponds to the "uncalibrated translational component of the rigid displacement", also called "focus of expansion" by some authors, and more geometrically *the epipole.* These notations have been introduced in [8] to analyse the motion of points and lines in the general case.

If we eliminate Z and Z' in equation (4) (by taking the cross-product with s and multiplying by $\mathbf{m}'^T$) we obtain:

$$\mathbf{m}'^T \underbrace{[\tilde{\mathbf{s}}\,\mathbf{H}_\infty]}_{\mathbf{F}} \mathbf{m} = 0 \tag{5}$$

The matrix $\mathbf{F} = \tilde{\mathbf{s}}\,\mathbf{H}_\infty$ is the *Fundamental matrix* and is also called the "essential matrix in the uncalibrated case". If we consider that the only information available is related to the retinal correspondences between points, without any knowledge about the depths Z, equation (5) is the only equation that can be derived [8].

Considering a set of matches related by equation (3) the equation (5) is well defined if and only if (i) $\mathbf{s} \neq 0$ and (ii) there is no linear relations between all m' and m. The degenerated cases occur only if the translation is zero, or if all points belong to the same plane [8][2]. This particular case will be analysed in detail.

As discussed in [12] an efficient criterion is the average retinal Euclidean distances between each point and its epipolar line. The following symmetric least-square sum is minimized:

$$\mathbf{F}_\bullet = argmin_{\mathbf{F}} \underbrace{\left[\sum_{\{\mathbf{m}\}} w_{\mathbf{m}} \underbrace{\left[d(\mathbf{m}', \mathbf{F}\mathbf{m})^2 + d(\mathbf{m}, \mathbf{F}^{\mathrm{T}}\mathbf{m}')^2 \right]}_{f_{\mathbf{m}}(\mathbf{F})^2} \right] / \left[2 \sum_{\{\mathbf{m}\}} w_{\mathbf{m}} \right]}_{[\epsilon_{\mathbf{F}}(\mathbf{F})]^2} \tag{6}$$

where $w_{\mathbf{m}}$ is a weighted corresponding to the precision of the match, in fact the inverse of the variance of the precision of the match. The quantity $w_{\mathbf{m}}$ is given in $pixel^{-2}$, while $\epsilon_{\mathbf{F}}(\mathbf{F})$, the *average distance to the epipolar*, is in pixel.

A camera for which F has been computed is called *a weak calibrated camera.*

The vector $\mathbf{s}$ is defined as the basis vector of the kernel of $\mathbf{F}^T$.

The case of a pure rotation, and the planar case. As pointed out previously, in the case of a pure rotation or if the set of points belongs to a unique planar structure, we cannot estimate the F-matrix because all points in one view are related to points in the other view by a relation of the form:

$$\mathbf{m}' \equiv \mathbf{H}\,\mathbf{m} \tag{7}$$

which corresponds to two equations for each match.

There, if the matrix $\mathbf{F}$ is undefined, we still can estimate the matrix $\mathbf{H}$ as in [8], using $\mathbf{H} = \mathbf{I}$ as initial value.

Following the same method as for the F-matrix, an efficient criterion is to minimize the residual disparity again, as in [8] and obtain $\mathbf{H}$ through:

$$\mathbf{H}_\bullet = argmin_{\mathbf{H}} \underbrace{\left[\sum_{\{\mathbf{m}\}} w_{\mathbf{m}} \underbrace{||\mathbf{m}' - \frac{\mathbf{H}\,\mathbf{m}}{((\mathbf{h}^2)^T\,\mathbf{m})}||^2}_{f_{\mathbf{m}}(\mathbf{H})^2} \right] / \left[\sum_{\{\mathbf{m}\}} w_{\mathbf{m}} \right]}_{[\epsilon_{\mathbf{H}}(\mathbf{H})]^2} \tag{8}$$

[2] From algebraic point of view, equation (5) has singular solutions if and only if there exist a linear relation between $\mathbf{m}$ and $\mathbf{m}'$, i.e. a relation of the form $\mathbf{m}' = \mathbf{H}\,\mathbf{m}$. This situation corresponds to the case were the points are related by a collineation, i.e. correspond to a planar structure as reviewed in the sequel.

where we write $\mathbf{H} = (\mathbf{h}^0, \mathbf{h}^1, \mathbf{h}^2)$ in order to have a compact notation[3].

The error $\epsilon_{\mathbf{H}}(\mathbf{H})$, given in pixel, will be called *residual disparity after motion reduction* in the sequel.

Reciprocally, as soon as the points belongs to at least two planes, we can defined a F-matrix [3]. This degenerated situation thus only exists in the case of an unique plane.

3 Using specific displacements for motion analysis.

Let us now discuss situations for which the F-matrix or the H-matrix have a particular form. Considering a robotic system, it is very often that a displacement is not a general displacement but a constrained motion such as a pure translation, a fixed axis rotation, etc... as illustrated in figure 2.

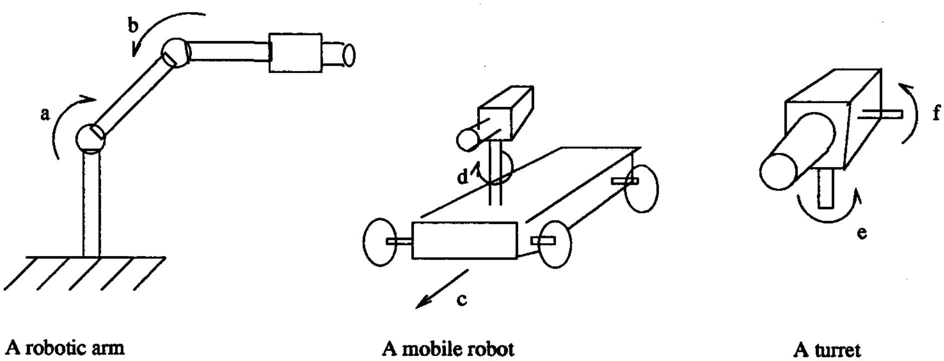

Fig. 2. Examples of robotic mechanism which generates pure translations, pure rotations or fixed axis rotations. *If, on the robotic arm, (a) and (b) have opposite values a pure translation occurs. Applying the same command (c) on both wheels of a mobile robot induces also a translation. A motion of (a) alone, or (b) alone, on the robotic arm induces a fixed axis rotation. A displacement (c) applying the opposite commands on both wheels of a mobile robot also induces a fixed axis rotation. Turrets for camera can induce pure rotations in pan (e) or tilt (f) around the optical center.*

[3] The relation $\mathbf{m}' = \frac{\mathbf{H}\,\mathbf{m}}{((\mathbf{h}^2)^T\,\mathbf{m})}$ is a vectorial form for :

$$\begin{cases} u' = \frac{H^{00}\,u + H^{01}\,v + H^{02}}{H^{20}\,u + H^{21}\,v + H^{22}} \\ v' = \frac{H^{10}\,u + H^{11}\,v + H^{12}}{H^{20}\,u + H^{21}\,v + H^{22}} \\ 1 = 1 \end{cases}$$

Moreover we can make the following assumptions about this kind of hardware [6]:

- The displacements are reproducible.
- We can measure the angles of the rotation. For technical reasons we do not assume the same thing for translations (it is not sure that we can estimate the norm of the linear translation of a zoom for instance [6] while the precision of the translation of a mobile robot is not very high).
- All extrinsic parameters are unknown, and equations are expected to provide unstable estimates of them [1].

These particular constraints are far from having negative properties. On the contrary, they induce additional equations which help solving the reconstruction or calibration problem.

Furthermore, the estimation of the displacement are easier in these cases, because we have to evaluate less parameters.

However, the system must be able to recognize if the displacement corresponds to such a particular case, so that we must characterize the situation in each case.

Finally, the different class of displacements might have several implications on the perception strategy, which is also to be discussed.

The parameterization of all these different kind of displacements have been given in [7] and will not be reported here.

4 Defining a hierarchical motion module

4.1 Combining different models of displacements

Following the previous discussion, when we estimate a rigid displacement, we consider several cases, depending on the nature of the displacement. Collecting all constraints proposed in [7], we can describe the following set of models:

Considering a rigid structure, the following class of displacements can be identified, N is the number of parameters:

Class of Displacement	Parameterization or constraint	Information Recovered	N
Pure rotation	$\mathbf{F} = 0$	$\mathbf{H}_\infty, \mathbf{t} = 0$	0
Z-axis pure translation	$\mathbf{F} = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	$\mathbf{H}_\infty = \mathbf{R} = \mathbf{I}, \mathbf{t} \equiv \mathbf{z}$	0
Pure retinal translation	$\mathbf{F} = \begin{pmatrix} 0 & 0 & a \\ 0 & 0 & b \\ -a & -b & 0 \end{pmatrix}, \; \lvert\lvert \mathbf{F} \rvert\rvert = 1$	$\mathbf{H}_\infty = \mathbf{R} = \mathbf{I}, \; \mathbf{t} \equiv \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \\ 0 \end{pmatrix}$, $\theta = atan(\frac{s1}{s0})$	1
Pure translation	$\mathbf{F} = \begin{pmatrix} 0 & c & a \\ -c & 0 & b \\ -a & -b & 0 \end{pmatrix}, \; \lvert\lvert \mathbf{F} \rvert\rvert = 1$	$\mathbf{H}_\infty = \mathbf{R} = \mathbf{I}$	2
Retinal displacement	$\mathbf{F} = \begin{pmatrix} 0 & 0 & a \\ 0 & 0 & b \\ c & d & e \end{pmatrix}, \; \lvert\lvert \mathbf{F} \rvert\rvert = 1$	$\mathbf{R}, \mathbf{t}/\lvert\lvert \mathbf{t} \rvert\rvert, eq(\mathbf{a})$	4
Zoom displacement	$\mathbf{F} = \begin{pmatrix} 0 & f & a \\ -f & 0 & b \\ c & d & e \end{pmatrix}, \; \begin{matrix} cb - ad = 0 \\ \lvert\lvert \mathbf{F} \rvert\rvert = 1 \end{matrix}$	$\mathbf{R} = \mathbf{I}, \mathbf{t}/\lvert\lvert \mathbf{t} \rvert\rvert, eq(\mathbf{a})$	4
Fixed axis rotation	$det(\mathbf{F} + \mathbf{F}^T) = 0, \; det(\mathbf{F}) = 0, \; \lvert\lvert \mathbf{F} \rvert\rvert = 1$	$eq(\mathbf{a})$	6
Retinal translation	$det(\mathbf{F}) = 0, s^2 = 0, \lvert\lvert \mathbf{F} \rvert\rvert = 1$	$\mathbf{t} \equiv \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \\ 0 \end{pmatrix}, \; eq(\mathbf{a})(Kruppa)$, $\theta = atan(\frac{s1}{s0})$	6
General rigid displacement	$det(\mathbf{F}) = 0, \lvert\lvert \mathbf{F} \rvert\rvert = 1$	$eq(\mathbf{a})(Kruppa)$	7

where $eq(\mathbf{a})$ means that we obtain equations about the intrinsic calibration parameters, these equations being either linear equations or the quartic Kruppa equations, as specified. In these cases, it is not possible to maintain an estimation of all calibration parameters.

In the planar case, we have:

Class of Displacement	Parameterization or constraint	Information Recovered	Number of Parameters
Stationary structure	$\mathbf{H} = \mathbf{I}$	$\mathbf{R} = \mathbf{I}, \mathbf{t} = 0, \mathbf{a} = \mathbf{a}'$	0
Constant retinal displacement	$\mathbf{H} = \begin{pmatrix} 0 & 0 & a \\ 0 & 0 & b \\ 0 & 0 & 1 \end{pmatrix}$	$\mathbf{R} = I, \mathbf{t} \equiv (a, b, 0), \mathbf{a} = \mathbf{a}', \mathbf{n} \equiv \mathbf{z}$	2
Retinal planar zoom	$\mathbf{H} = \begin{pmatrix} c & 0 & a \\ 0 & c & b \\ 0 & 0 & 1 \end{pmatrix}$	eq($\mathbf{a}$)	4
Retinal planar rotation	$\mathbf{H} = \begin{pmatrix} c & d & a \\ -d & c & b \\ 0 & 0 & 1 \end{pmatrix}$	$\mathbf{R}$, eq($\mathbf{a}$)	4
Pure planar retinal translation	$\mathbf{H} = \mathbf{I} + \mathbf{s}\nu^T, s^2 = 1$	$\mathbf{R} = \mathbf{I}, \mathbf{s}/\lvert\lvert \mathbf{s} \rvert\rvert, \nu$	5
Pure planar translation	$\mathbf{H} = \mathbf{I} + \mathbf{s}\nu^T, s^2 = 0$	$\mathbf{R} = \mathbf{I}, \mathbf{s}/\lvert\lvert \mathbf{s} \rvert\rvert, \nu$	5
Retinal planar displacement	$\mathbf{H} = \begin{pmatrix} a & b & c \\ d & e & f \\ 0 & 0 & 1 \end{pmatrix}$	$\mathbf{R}$, $\mathbf{t}/\lvert\lvert \mathbf{t} \rvert\rvert$, $\mathbf{n}$, eq($\mathbf{a}$)	6
General planar displacement	$\mathbf{H} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & 1 \end{pmatrix}$		8

In fact some other variants have also been introduced in order to have alternative models with very few parameters. For instance a model with zero parameters, corresponding to a collineation equal to the identity, i.e. a stationary structure is introduced. This allows to have a simple model assuming that points are not moving.

Furthermore, this set of model has a very interesting structure, i.e. some models are generalizations of others. This allows to take as best model the first model, starting from the bottom, which statistical significance is smaller that every models immediatly higher in the hierarchy.

4.2 Experimental results

An example with real data. We consider a sequence of 16 imagesfor which the displacement is an approximative retinal translation, with some erroneous rotation because of the actual set-up. A retinal displacement is thus expected.

The early-vision module has provided matches between 44 points and the errors are given in table (1).

displacement	number of outliers	residual error
stationary structure	0	17.9633
pure retinal translation	0	17.4948
planar retinal displacement	11	6.52437
retinal displacement (F22 = 0)	8	1.95843
retinal displacement (F22 = 1)	8	0.735489
pure translation	0	14.7264
zoom displacement (s2 = 0; s'2 = 0; s0.s'0 != 0)	0	11.677
zoom displacement linear (s2 = 1; s'2 = 1)	4	0.887803
general rigid displacement	11	0.84458

Table 1. Table of residues for the real scene.

The model is correctly estimated also in this real case, which thus allow us to conclude on the validity of the proposed mechanism.

5 Conclusion

In the present paper we have reviewed and completed the description of a general framework which allows not only to estimate a minimal parameterization of the rigid displacement between two frames, but also to determine several particular cases which occur in practice and have important advantages with respect to the calibration problem. This is true for several standard displacement, except a zoom displacement which seems to be a singular case, for the proposed model.

The statistical framework to implement these equations has been already described in [8] and has been applied here to the estimation of collineations from a minimal parameterization. This paper however generalizes the set of models to general rigid displacements, and proposes a complete analysis of the underlying rigid displacement in each case.

Similar attempts to use degenerated models of parameterization of motion have been already issued in the past [8, 4, 9]. However, we collect here new results about the Euclidean representation associated to each parameterization.

Furthermore, the implementation also integrates two new aspects: (i) clustering data and (ii) testing different models to represent the data.

Finally, this work tries to develop -with a certain degree of completeness- all the different singularities which occur for a rigid displacement and which can be detected without calibration. A practical motion module has been developed and successfully experimented.

A step further, we will use this hierarchical approach to not only parameterize the retinal displacement but also analyse the calibration of the visual system and recover the scene structure. A preliminary study has been issued [2] for retinal displacements.

References

1. O. Faugeras. *Three-dimensional Computer Vision: a geometric viewpoint.* MIT Press, Boston, 1993.
2. D. Lingrand and T. Viéville. Dynamic foveal 3d sensing using affine models. Technical Report RR-2687, INRIA, 1995.
3. T. Luong. *Matrice Fondamentale et Calibration Visuelle sur l'Environnement.* PhD thesis, Université de Paris-Sud, Orsay, 1992.
4. P. Torr, A. Zisserman, and S. Maybank. Robust detection of degenerated configurations for the fundamental matrix. In *5th International Conference on Computer Vision*, pages 1037–1042, 1995.
5. T. Viéville. Autocalibration of visual sensor parameters on a robotic head. *Image and Vision Computing*, 12, 1994.
6. T. Viéville, E. Clergue, R. Enciso, and H. Mathieu. Experimentating with 3D vision on a robotic head. *Robotics and Autonomous Systems*, 14(1), 1995.
7. T. Viéville and D. Lingrand. Using singular displacements for uncalibrated monocular visual systems. Technical Report RR-2678, INRIA, 1995.
8. T. Viéville, C. Zeller, and L. Robert. Using collineations to compute motion and structure in an uncalibrated image sequence. *International Journal of Computer Vision*, 1995. To appear.
9. C. Wiles and M. Brady. Closing the loop on multiple motion. In *5th International Conference on Computer Vision*, pages 308–313, 1995.
10. R. Willson. *Modeling and Calibration of Automated Zoom Lenses.* PhD thesis, Department of Electical and Computer Engineering, Carnegie Mellon University, 1994.
11. C. Zeller and O. Faugeras. Applications of non-metric vision to some visual guided tasks. In *The 12th Int. Conf. on Pattern Recognition*, pages 132–136, 1994.
12. Z. Zhang, R. Deriche, Q.-T. Luong, and O. Faugeras. A robust approach to image matching: Recovery of the epipolar geometry. In *Proc. International Symposium of Young Investigators on Information\Computer\Control*, pages 7–28, Beijing, China, Feb. 1994.

Acknowledgments. We are especially thankful to **O.D. Faugeras** for some powerful ideas which are at the origin of this work. This work is partially achieved under
Esprit Project P8878/REALISE.