Robotica (2001) volume 19, pp. 331–342. Printed in the United Kingdom © 2001 Cambridge University Press

Indoor robot motion based on monocular images

D. Ortín & J.M.M. Montiel*

Dpto. Informática e Ingeniería de Sistemas, María de Luna 3, E-50015 Zaragoza (Spain). josemari@posta.unizar.es 182616@cepsz.unizar.es*

(Received in Final Form: September 19, 2000)

SUMMARY

The estimation of the 2D relative motion of an indoor robot using monocular vision is presented. The camera calibration is known, and its motion is limited to be planar. These constraints are included in the robust regression of epipolar geometry from point matches. Motion is derived from the epipolar geometry. A sequence of 54 real images is used to test the algorithm. Accurate motion, both in rotation and translation angles of 0.4 and 1.7 deg, is successfully derived.

KEYWORDS: Monocular vision; Robust epipolar geometry; Indoor robot; Accurate motion.

I. INTRODUCTION

The goal of this work is to compute the motion of an indoor mobile robot using monocular vision. To achieve it, some of the recent results in geometric computer vision research have been used. Our aim is to adapt the general computer vision algorithms to solve this particular task in the robotics field.

To compute the camera motion between two images, a two stages approach can be applied: First, a set of point matches between both images is computed. Then, the camera motion is computed from those matches. The previous strategy can be traced back to the origins of photogrammetry, and has been successfully applied along the last century by manually computing the matches.[1] The lack of robustness against spurious matches of the previous approach is well known. This lack of robustness impeded the availability of a reliable automatically computed motion using only images.

One of the key factors for the lack of reliability of the motion derivation comes from the lack of reliability of the automatic point matching computation. In fact, the robust algorithms proposed recently give as a joint result both reliable matchings and motion computation.[2,3]

The theoretical support to produce nearly spurious-free matches comes from robust statistics. A compendium of these techniques can be found in references [4,5]. The basic idea is to fit a model with the matches, considering as outliers those matches inconsistent with the fitted model. Robust statistics techniques have been used in perception for fitting models[6] and for registering range images.[7,8] It is also a valid technique for primitive fitting in contour detection.

To apply a robust method, it is necessary to have a model to fit. Computer vision research has done a great progress in producing geometrically based models to encode the constraints between different views observing the same 3D rigid scene. The applied models are very general. On one hand, they are uncalibrated, that is, they do not need knowledge of camera calibration. On the other hand, the camera motion is also unknown. Despite its generality, good experimental performance with real images is achieved. For two images,[2] the uncalibrated epipolar geometry, represented by the fundamental matrix,[9] encodes the rigidity between two views. Torr[3] uses uncalibrated three view relations by means of the trifocal tensor. It encodes the rigidity among three views.[10,11]

The problem of overparameterisation is well known. In fact, in reference [12] the importance of model selection is stressed in order to successfully automatically compute matches and motion from image sequences. For an indoor mobile robot with a fixed camera, the number of parameters of the motion model can be reduced by imposing the knowledge of planar motion and known camera calibration. The parameter reduction improves the results both in performance and in computing time.

The work presented in this paper is a robust regression of the epipolar geometry between two images taken with a calibrated camera on board a mobile robot. Motion is computed from epipolar geometry. Epipolar geometry is computed from point matches. In order to reduce the number of model parameters as much as possible, all the available knowledge has been considered. So, the camera calibration and the planarity of the motion were included in the model. In order to validate it experimentally, the proposed algorithm has been applied to a 54 image sequence and the results are compared with respect to the ground true motion. The mean error in the computed rotation is 0.4 deg. and for the computed translation direction it is 1.7 deg.

The epipolar geometry is only meaningful when the camera translation between two views is non-zero (not pure rotation). Otherwise, it is not well defined. In the case of a pure rotating camera, a homography model between the two views can be fitted. It is also shown how, when motion is close to a pure rotation, the translation is unreliably computed. However, in such a case, a homography can be fitted to the matches. This problem of model selection has been addressed generally in reference [13]. In this paper we

* Corresponding author.

https://doi.org/10.1017/S0263574700003143 Published online by Cambridge University Press

have focused on determining the models to fit, including the planarity and the calibration, which reduces the number of independent parameters. The proposed models have been tested over a real image sequence.

Section II states the problem to be addressed. Next, Section II is devoted to presenting the relation between two images of the same 3D scene, when the camera motion is planar and the camera calibration is known. Section IV, for a complete review, includes a summary of robust regression; it also details how to apply the robust regression to fit the two view relations. Next Section V details with how to compute the motion from the two view relations. Section VI is devoted to the experimental verification with real images of the proposed algorithm. Finally, Sections VII and VIII present the conclusions and the discussion.

II. PROBLEM STATEMENT

Given two locations for a mobile robot, their relative motion next can be computed following the following procedure. First, an image is taken at each position. Second, putative point matches are selected by correlation techniques between the images. Third, they are used to perform a robust regression of the fundamental matrix. The intrinsic parameters of the camera and the planarity of motion are used to further constrain the epipolar geometry. Finally, both rotation and the direction of translation are estimated from fundamental parameters.

The previous algorithm was applied to every single movement described in a 54 step evaluation trajectory (see Section VI-A). The automatically computed motion was compared with the ground true one available for the sequence.

Some modifications have been applied with respect to the classical 3D uncalibrated algorithm, in order to improve its performance and computing time:

(i) Epipolar geometry is computed after correcting the radial distortion of the lens.[14]

(ii) Planar motion and camera calibration are included in the model, reducing the parameters to adjust.

(iii) Neither back projection nor optimal motion and structure estimation are computed.

As additional constraints are included in the model, the number of independent parameters to compute epipolar geometry is reduced to two. This also reduces the number of subsamples required for robust regression, speeding up the process and improving its reliability.

III. TWO VIEW RELATIONS

This section is devoted to presenting the constraints that relate the point matches between two views. First, the relations for planar motion and known calibration are presented. Two cases are detailed: general planar motion and pure planar rotation. The section ends by detailing how the constraints between images can be computed from image matches.

A. Epipolar geometry for a calibrated camera with planar motion

Given two images acquired from different viewpoints, the epipolar constraint is the relation between their points in terms of only geometric criteria. It forces optical centres, spatial point and its projections on the images to be on the same ‘epipolar plane’ (see Figure 1). In mathematical terms both projections, $\mathbf{x}$ and $\mathbf{x}'$, are related with the essential matrix $\mathbf{E}$ by

$$\mathbf{x}^T \mathbf{E}\mathbf{x}' = 0$$

where $\mathbf{x} = (x_i, y_i, 1)^T$ and $\mathbf{x}' = (x_i', y_i', 1)^T$ are the homogeneous image coordinates in the first and the second cameras. These coordinates should be referred to the normalised retina.[15]

As the camera is supposed to be moving on the horizontal plane (see Figure 2), the relative rotation and translation in the first camera reference are:

$$\mathbf{R} = \begin{bmatrix} \cos(\varphi) & 0 & \sin(\varphi) \\ 0 & 1 & 0 \\ -\sin(\varphi) & 0 & \cos(\varphi) \end{bmatrix} \quad \text{and} \quad \mathbf{t} = \begin{bmatrix} \sin(\theta) \\ 0 \\ \cos(\theta) \end{bmatrix}$$

where φ is the rotation angle, and θ is the direction of translation, shown in Figure 2. As only monocular image matches are used as input data, translation can be recovered only up to a scale factor. Because of that, only the translation direction is considered.

The essential matrix $\mathbf{E}$ is defined as the following cross-product: $\mathbf{E} = \mathbf{t} \times \mathbf{R}$.[16] When only planar motion is considered, $\mathbf{E}$ can be defined as follows:

$$\mathbf{E} = \begin{bmatrix} 0 & -\cos(\theta) & 0 \\ \cos(\varphi - \theta) & 0 & \sin(\varphi - \theta) \\ 0 & \sin(\theta) & 0 \end{bmatrix}$$

Normalised coordinates $\mathbf{x}$ are related with pixel coordinates $\mathbf{m}$ by means of the calibration matrix $\mathbf{A}$: $\mathbf{m} = \mathbf{A}\ \mathbf{x}$. It is defined from intrinsic parameters as:

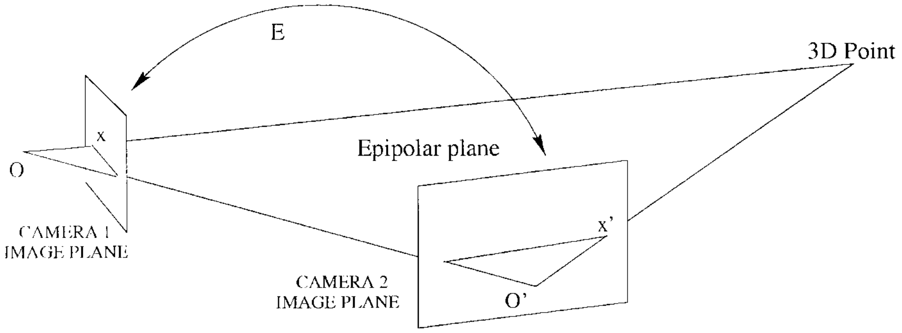

Fig. 1. Epipolar geometry constraint.

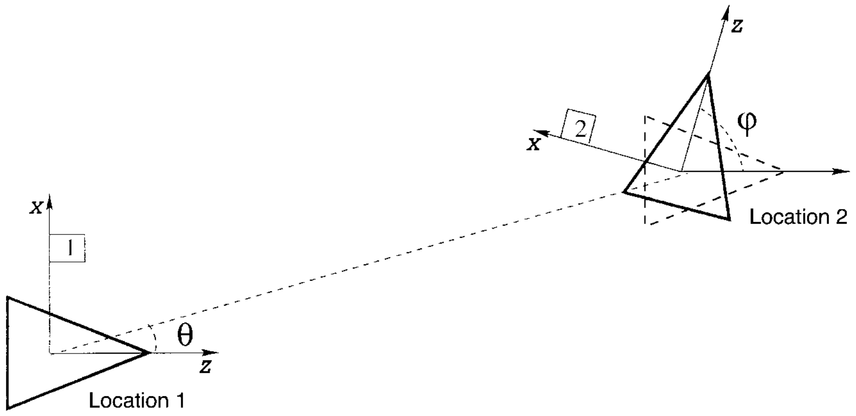

Fig. 2. Motion scheme

$$\mathbf{A}=\begin{bmatrix} k_x & s & x_0 \\ 0 & k_y & y_0 \\ 0 & 0 & 1 \end{bmatrix}$$

where:

- k_x and k_y measure the different pixel sizes in X and Y directions, scaled by focal length.
- x_0 and y_0 give the position of the optical centre of the image, which usually will not be its middle point.
- s measures the X-Y axis angle. In the following, it is assumed to be null.

The epipolar constraint can be also expressed in pixel coordinates by means of the fundamental matrix $\mathbf{F}^9$ as:

$$\mathbf{m}^T\mathbf{F}\mathbf{m}'=0$$

Then:

$$\mathbf{F}=\mathbf{A}^{-T}\mathbf{E}\mathbf{A}^{-1}$$

F can be expressed from calibration and motion parameters as:

$$\mathbf{F}=\frac{1}{k_xk_y}\begin{bmatrix} 0 & -C_\theta & y_0C_\theta \\ C_{(\varphi-\theta)} & 0 & k_xS_{(\varphi-\theta)}-x_0C_{(\varphi-\theta)} \\ -y_0C_{(\varphi-\theta)} & k_xS_\theta+x_0C_\theta & x_0y_0(C_{(\varphi-\theta)}-C_\theta)-k_xy_0(S_{\varphi-\theta}+S_\theta) \end{bmatrix} \tag{1}$$

where:

$$\begin{aligned} C_\theta &= \cos(\theta) \\ S_\theta &= \sin(\theta) \\ C_{(\varphi-\theta)} &= \cos(\varphi-\theta) \\ S_{(\varphi-\theta)} &= \sin(\varphi-\theta) \end{aligned}$$

Planar motion will be derived from this fundamental matrix parameterisation.

B. Epipolar geometry from point matches

As seen in Section III-A, the fundamental matrix **F** is defined from intrinsic camera calibration (matrix **A**) and relative motion (matrices **R** and **t**):

$$\mathbf{F}=\mathbf{A}^{-T}(\mathbf{t}\times\mathbf{R})\mathbf{A}^{-1}$$

Motion can be computed from **A** and **F**. Matrix **A** is derived from camera parameters, and the fundamental matrix can be computed from matches, using

$$\mathbf{m}^T\mathbf{F}\mathbf{m}'=0$$

if $\mathbf{m}=(u,\ v,\ 1)^T$, and $\mathbf{m}'=(u',\ v',\ 1)^T$, then the previous equation can be rearranged as

$$\begin{aligned} &uu'f_{11}+uv'f_{12}+uf_{13}+vu'f_{21}+vv'f_{22} \\ &\quad+vf_{23}+u'f_{31}+v'f_{32}+f_{33}=0 \end{aligned} \tag{2}$$

Given the projections of different 3D points on every camera (**m** and **m′**), they are replaced in equation (2), and fundamental matrix parameters are computed. Reference [17] presents how to compute the fundamental matrix from point matches in the general 3D motion with uncalibrated cameras.

The next sections cover how to include planar motion and camera calibration constraints in the derivation of matrix **F** from point matches.

B.1 **Uncalibrated camera: Linear 6-points algorithm.** For an uncalibrated camera with planar motion, each point match imposes the following constraint:

$$[uv' \quad u \quad vu' \quad v \quad u' \quad v' \quad 1]\begin{bmatrix} f_{12} \\ f_{13} \\ f_{21} \\ f_{23} \\ f_{31} \\ f_{32} \\ f_{33} \end{bmatrix}=0$$

f_{11} and f_{22} have been removed from the unknowns vector as they are null according to equation (1).

As the fundamental matrix is defined up to a scale factor, only six equations (and therefore, six matches) are required to find a linear solution for the uncalibrated planar motion.

The equations to determine **F** can be solved using singular value decomposition of coefficient matrix **M**,[18] as:

$$\mathbf{M}=\mathbf{U}\mathbf{S}\mathbf{V}^T$$

Fundamental matrix parameterisation is the singular vector associated to the minor singular value of **M**.

According to reference [19], rank-two constraints should be imposed in the final solution. So, if $\mathrm{S}=\mathrm{diag}(\sigma_1,\ \sigma_2,\ \sigma_3)$, and $\sigma_1>\sigma_2>\sigma_3$, matrix **M** is projected on **M′**, where

https://doi.org/10.1017/S0263574700003143 Published online by Cambridge University Press

$$\mathbf{M}' = \mathbf{U}\mathbf{S}'\mathbf{V}^T$$

being $\mathbf{S}' = \mathrm{diag}(\sigma_1, \sigma_2, 0)$. This linear estimation can be further refined using nonlinear methods.

B.2 **Calibrated camera: Linear 3-points algorithm.** As our aim is to derive Euclidean motion, it is advantageous to use camera calibration not only to derive motion, but also to compute epipolar geometry parameters (see equation (1)). Then, as only motion parameters are undefined, equation (2) becomes:

$$\frac{1}{k_x k_y}\left[\begin{bmatrix} C_\theta & S_\theta & C_{\varphi-\theta} & S_{\varphi-\theta}\end{bmatrix}\begin{bmatrix} x_0(v'-y_0)-u(v'-y_0) \\ k_x(v'-y_0) \\ u'(v-y_0)-x_0(v-y_0) \\ k_x(v-\mathrm{y}_0)\end{bmatrix}\right]^{\mathrm{T}} = 0 \tag{3}$$

Like in Section III-B.1, it can be solved by singular value decomposition of the coefficient matrix, being $[C_\theta, S_\theta, C_{\varphi-\theta}, S_{\varphi-\theta}]$ the unknown vector.

The angles that define motion can be derived by means of a four quadrant arctan. Anyway, as the elements of the unknown vector are not independent from each other, nonlinear optimisation could be applied to recover more accurate estimations. Section III-B.3, next, covers how to deal with it.

B.3 **Calibrated camera: Non-linear 2-points algorithm.** As shown in Section III-B.2, the epipolar constraint can be reduced to a nonlinear equation $f(\theta, \varphi) = 0$. This can be solved by Newton's iterative method. Classical Taylor's expansion was used to approximate equation (3) to a linear function. Then,

$$f(\theta, \varphi) \approx f(\theta_0, \varphi_0) + \mathbf{J}_f(\theta_0, \varphi_0)\begin{bmatrix} \theta-\theta_0 \\ \varphi-\varphi_0 \end{bmatrix}$$

$f(\theta_0, \varphi_0)$ can be computed from equation (3), and the Jacobian, $\mathbf{J}_f(\theta_0, \varphi_0)$, can be written as:

$$\mathbf{J}_f(\theta, \varphi) = \frac{1}{k_x k_y}\left[\begin{bmatrix} -S_\theta & C_\theta & S_{\varphi-\theta} & -C_{\varphi-\theta} \\ 0 & 0 & -S_{\varphi-\theta} & C_{\varphi-\theta}\end{bmatrix}\right.$$

$$\left.\begin{bmatrix} x_0(v'-y_0)-u(v'-y_0) \\ k_x(v'-y_0) \\ u_2(v-y_0)-x_0(v-y_0) \\ k_x(v-\mathrm{y}_0)\end{bmatrix}\right]^{\mathrm{T}}$$

As the epipolar constraint is defined only up to a scale factor, the term $\frac{1}{k_x k_y}$ can be disregarded in both $\mathrm{f}(\theta, \varphi)$ and $\mathbf{J}_f(\theta, \varphi)$ equations.

Experimental results, (see Section VI) show convergence taking the point $\theta = 0$, $\varphi = 0$ as initial seed. Since only two unknowns are determined, two is the minimum number of matches required to compute any solution from scratch.

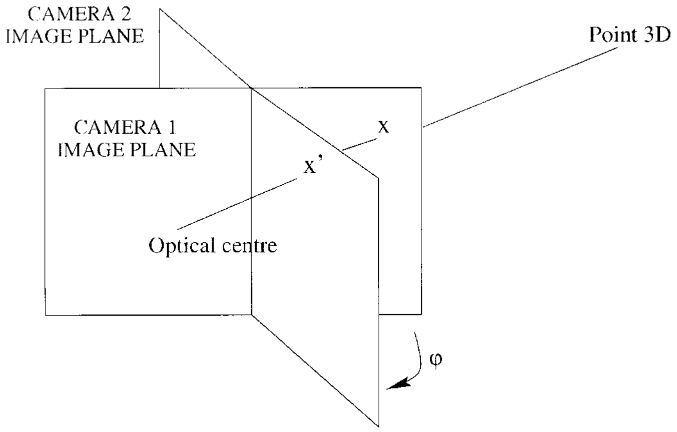

Fig. 3. Rotation homography scheme

C. *Homography for a calibrated camera with plane rotation motion*

When the camera motion is a pure rotation, epipolar geometry is not defined. In this case, an even more restrictive point-to-point geometric constraint, a homography, should be used.

If relative motion is given only by a rotation of the camera (see Figure 3), image coordinates in the first and second images are related by:

$$\mathbf{x} = \mathbf{R}\mathbf{x}'$$

And when rotation is constrained to be on the X-Z axis plane, they are

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} \cos(\varphi) & 0 & \sin(\varphi) \\ 0 & 1 & 0 \\ -\sin(\varphi) & 0 & \cos(\varphi)\end{bmatrix}\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix}$$

As the epipolar constraint, a homography can also relate pixels coordinates. If $\mathbf{A}$ is the camera intrinsic parameter matrix and $\mathbf{x}$ are the normalised coordinates, then pixel coordinates ($\mathbf{m}$) are: $\mathbf{m} = \mathbf{A}\mathbf{x}$. Homography results in:

$$\mathbf{m} = (\mathbf{A}\mathbf{R}_\varphi\mathbf{A}^{-1})\mathbf{m}' = \mathbf{H}\mathbf{m}'$$

And matrix $\mathbf{H}$ can be parameterised as:

$$\mathbf{H} = \frac{1}{k_x}\begin{bmatrix} K_x C_\varphi - x_0 S_\varphi & 0 & S_\varphi(k_x^2 - x_0^2) \\ -y_0 S_\varphi & k_{\mathrm{x}} & k_x y_0(C_\varphi - 1) + x_0 y_0 S_\varphi \\ -S_\varphi & 0 & k_x C_\varphi + x_0 S_\varphi \end{bmatrix} \tag{4}$$

where $C_\varphi = \cos(\varphi)$ and $S_\varphi = \sin(\varphi)$.

D. *Homography from point matches*

The matrix $\mathbf{H}$ that defines a homography can be computed, either from calibration (i.e. from matrixes $\mathbf{A}$ and $\mathbf{R}$) or from point matches

If rotation is constrained to be planar, and camera parameters are known, $\mathbf{H}$, depends only on one parameter, φ. It can be recovered from one point match, solving the following nonlinear equation:

https://doi.org/10.1017/S0263574700003143 Published online by Cambridge University Press

$$\begin{bmatrix} k_x(u-u') & x_0(u+u') - uu' + x_0^2 - k_x^2 \\ k_x(v - y_0) & x_0 v + y_0 u' - vu' - x_0 y_0 \end{bmatrix} \begin{bmatrix} C_\varphi \\ S_\varphi \end{bmatrix}$$

$$- \begin{bmatrix} 0 \\ k_x(v' - y_0) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{5}$$

This equation $(\mathbf{h}(\varphi) = 0)$ is approximated to a linear function, $\mathbf{h}(\varphi) \approx \mathbf{h}(\varphi_0) + J_{\mathrm{h}}(\varphi_0)(\varphi - \varphi_0)$, where $\mathbf{h}(\varphi_0)$ is given by equation (5) and the Jacobian matrix, $J_{\mathrm{h}}(\varphi_0)$, is

$$J_{\mathrm{h}}(\varphi) = \begin{bmatrix} k_x(u-u') & x_0(u+u') - uu' + x_0^2 - k_x^2 \\ k_x(v - y_0) & x_0 v + y_0 u' - vu' - x_0 y_0 \end{bmatrix} \begin{bmatrix} -S_\varphi \\ C_\varphi \end{bmatrix}$$

Every match provides two equations, so an initial seed can be computed linearly (using only one match) by deriving C_φ and S_φ from equation (5).

IV. TWO VIEW RELATIONS ROBUST REGRESSION

Section III-B showed how to compute epipolar geometry parameters from point matches. In practice, matches obtained by correlation techniques are unreliable, and include some spurious rates. Reference [4] shows how just one of them can completely degrade classical least squares techniques. Thus, robust regression must be used in order to compute trustworthy solutions.

A. Robust regression and outlier detection using LMedS

To complete the study, a summary of the Least Median of Squares robust regression (LMedS), algorithm is included in the paper, further details can be found in reference [4], from where this summary has been obtained.

Given n points, LMedS robust regression, solves the non-linear minimisation problem:

$$\min \{\mathrm{median}_i \ (r_i^2)\}$$

where $\{r_i\}$ $i = 1 \ldots n$ are the residual of the n points used to fit the model.

There is no analytical solution for this problem, and it must be solved searching in the space of all possible results. An exhaustive search requires computing all the possible fits using p matches; p is the minimum number of matches required to compute a solution. So the number of fits to compute would be:

$$C_n^p = \frac{(n)!}{(n-p)!}$$

where n is the total number of point putative matches. For instance, if $n = 70$, and $p = 2$, then near 5000 estimations would be required. If $p = 8$, they would be 3.8×10^{14}.

Very often this is an unaffordable problem, and only random subsets of data can be analysed. Their number is given by:

$$m = \frac{\log(1-P)}{\log(1-(1-\varepsilon)^p)} \tag{6}$$

where P is the desired probability of finding one subsample without any spurious match, ε is the estimated fraction of outliers and p is the number of points used to compute any solution.

For example, if $P = 90\%$, and ε is supposed to be about the 50%, using $p = 8$ points, the number of required random subsamples is near 600, whereas if only two are required $(p = 2)$ just 9 are enough.

The m different solutions are compared, and the one that minimises the median of squared residuals is selected. As the median does not include the 50% of greater errors, it is not spoiled by outliers, and this solution copes with bad locations and false matches.

When this estimation is selected, every putative match can be classified as outlier or inlier, in function of its residual with respect to the selected model.

$$\begin{cases} \text{If } r_i^2 \leq \left(1.96\left(1.4826\left[1 + \frac{5}{(n-p)}\right]\sqrt{M}\right)\right)^2 \text{ point } i \text{ is classified as match.} \\ \text{Else, it is classified as outlier.} \end{cases} \tag{7}$$

where M is the computed minimal median, n is the total number of point putative matches and p is the minimum number of point matches required to compute any solution.

Finally, to derive an *efficient* solution in presence of noise, all the inliers are used to compute a least squares regression.

B. Algorithm for epipolar geometry robust regression

Given a set of putative matches between two images, the final LMedS algorithm for computing the fundamental matrix when motion is planar and camera parameters are known, results in:

1. Derive the number of m random samples to evaluate, according to equation (6).
2. Select m random samples, with two matches each, to compute m fundamental matrixes (Sec. III-B.3). Use the initial location as an initial seed.
3. Compute the median of the squared residuals for every solution, using the whole set of n matches. Given a fundamental matrix $\mathbf{F}$ and a point match in pixel coordinates: $\{\mathbf{m}, \mathbf{m}'\}$, its corresponding residual is:

$$r^2 = \left(\frac{\mathbf{mFm'}}{(\mathbf{Fm'})_1^2 + (\mathbf{Fm'})_2^2}\right)^2 + \left(\frac{\mathbf{m'F}^T\mathbf{m}}{\mathbf{F}^T\mathbf{m})_1^2 + (\mathbf{F}^T\mathbf{m})_2^2}\right)^2$$

4. Select the fundamental matrix that minimises these medians of squared residuals.
5. Classify matches as inliers or outliers according to equation (7).
6. Use inliers to compute the final fundamental matrix (Sec. III-B.3), using the linear solution (Sec. III-B.2) as an initial seed.

https://doi.org/10.1017/S0263574700003143 Published online by Cambridge University Press

C. Algorithm for rotation homography robust regression

Given a set of putative matches between two images, the final LMedS algorithm for computing the rotation homography when motion is planar and camera parameters are known, results in:

1. Compute the number of solutions m to evaluate (6).
2. Derive m rotation homographies from m different randomly selected points (Sec. III-D).
3. Compute the median of the squared residuals (5) for every homography, using the whole set of n matches. Given a homography matrix $\mathbf{H}$ and a point match given in pixel coordinates: $\{\mathbf{m}, \mathbf{m}'\}$, the corresponding residual is:
$$r^2 = \| \mathbf{m} - \mathbf{H}\mathbf{m}' \|^2 + \| \mathbf{m}' - \mathbf{H}^{-1}\mathbf{m} \|^2$$
4. Choose that solution that minimises the median of residuals.
5. Classify matches as outliers or inliers in function of (7).
6. Compute a final rotation homography with all the inliers (Sec. III-D).

V. MOTION COMPUTATION

Section III-C shows that if the motion is a rotation, the fundamental matrix cannot be defined uniquely. In these cases, the derived direction of the motion will not be reliable. Homographies can be used in order to identify them.

So, given a set of putative matches, LMedS robust regression is applied to compute one homography and classify matches as inliers or outliers (see Section IV). After that, the median residual of the inliers are derived according to next expression:

$$\frac{1}{2}\sqrt{\frac{1}{n_i}\sum_{\text{inliers}} \| \mathbf{m} - \mathbf{H}\mathbf{m}' \|^2 + \| \mathbf{m}' - \mathbf{H}^{-1}\mathbf{m} \|^2}$$

When resulting residuals are small enough, and similar to those acquired using epipolar geometry (the latter will always be smaller, as more parameters are used to fit the same data), motion can be approximated to be a pure rotation. In this case, the azimuth direction would be inaccurate, and translation should not be taken into account.

The automatic model selection is not an easy matter.[13] In this work we only analyse the validity of the epipolar and homography to model two different cases, but not how to select the valid model automatically. Section VI shows how this simplistic low residual criteria can be validated experimentally with real data.

A. Full motion computation algorithm

The resulting algorithm to compute the motion would be the following:

1. Given two images acquired with planar motion, derive putative matches using correlation techniques. The program "image-matching", by Z. Zhang, was used to complete it.
2. Use those matches (after correcting radial distortion) to compute a robust regression of the fundamental matrix, according to Section IV.
3. Derive motion parameters from equations (3) or (5), according to the kind of selected motion.
4. Select motion direction according to the results of structure reconstruction (Section V-B).

B. Structure reconstruction

Motion results have an ambiguity and can be either $[C_\theta, S_\theta, C_{\varphi-\theta}, S_{\varphi-\theta}]$ or $[-C_\theta, -S_\theta, -C_{\varphi-\theta}, -S_{\varphi-\theta}]$ depending on wether the scene is in front of or behind the camera.

The ambiguity can be solved by checking if the scene reconstruction[19] is in front of the camera. 3D points in the initial location reference can be computed as

$$\text{If } \mathbf{A} = \begin{bmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \\ \mathbf{a}_3^T \end{bmatrix} \text{and } \mathbf{B} = \mathbf{A}\mathbf{R}^T$$

$$= \begin{bmatrix} \mathbf{b}_1^T \\ \mathbf{b}_2^T \\ \mathbf{b}_3^T \end{bmatrix} \Rightarrow \begin{bmatrix} \mathbf{a}_1^T - u\mathbf{a}_3^T \\ \mathbf{a}_2^T - v\mathbf{a}_3^T \\ \mathbf{b}_1^T - u'\mathbf{b}_3^T \\ \mathbf{b}_2^T - v'\mathbf{b}_3^T \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$$

$$= \begin{bmatrix} 0 \\ 0 \\ -(u'\mathbf{a}_3 - \mathbf{a}_1)^T\mathbf{R}^T\mathbf{t} \\ -(v'\mathbf{a}_3 - \mathbf{a}_2)^T\mathbf{R}^T\mathbf{t} \end{bmatrix}$$

The median depth ($\text{median}_i\ (Z_i)$) was used to classify the motion, to ensure that the structure was also robust to outliers.

C. Pure rotation computation algorithm

When the motion can be approximated by a rotation, the algorithm to compute it is:

1. Find putative matches between initial and final images by correlation techniques.
2. Correct their radial distortion, and use it to compute a robust regression of a rotation homography, according to Section IV.
3. Derive motion parameters from equation (5).

VI. EXPERIMENTAL RESULTS

A. Experiments description

In order to evaluate the studied algorithm against real data, it is applied over the track shown in Figure 4. This is the 2D motion of a mobile robot in an indoor environment, described in reference [20]. Images, camera calibration and ground-true motion are available. The images correspond to a real, complex environment, with symmetries in the scene, light reflection and moving objects, such as people.

https://doi.org/10.1017/S0263574700003143 Published online by Cambridge University Press

The camera, a B/W Pulnix TM-6EX, was fixed horizontally by means of a spirit level. Its intrinsic parameters are shown in Table I. The image size was 512 × 512 pixels.

Analysing the sequence of images, we found that there were 11 steps (17–19, 29–33, 37–38 and 40–44) in which the observed scene was different in initial and final images. In these cases, the overlap was less than 20%, and epipolar geometry could not be computed. Therefore, they are not considered in the next discussions; results are referred to the other 42 cases.

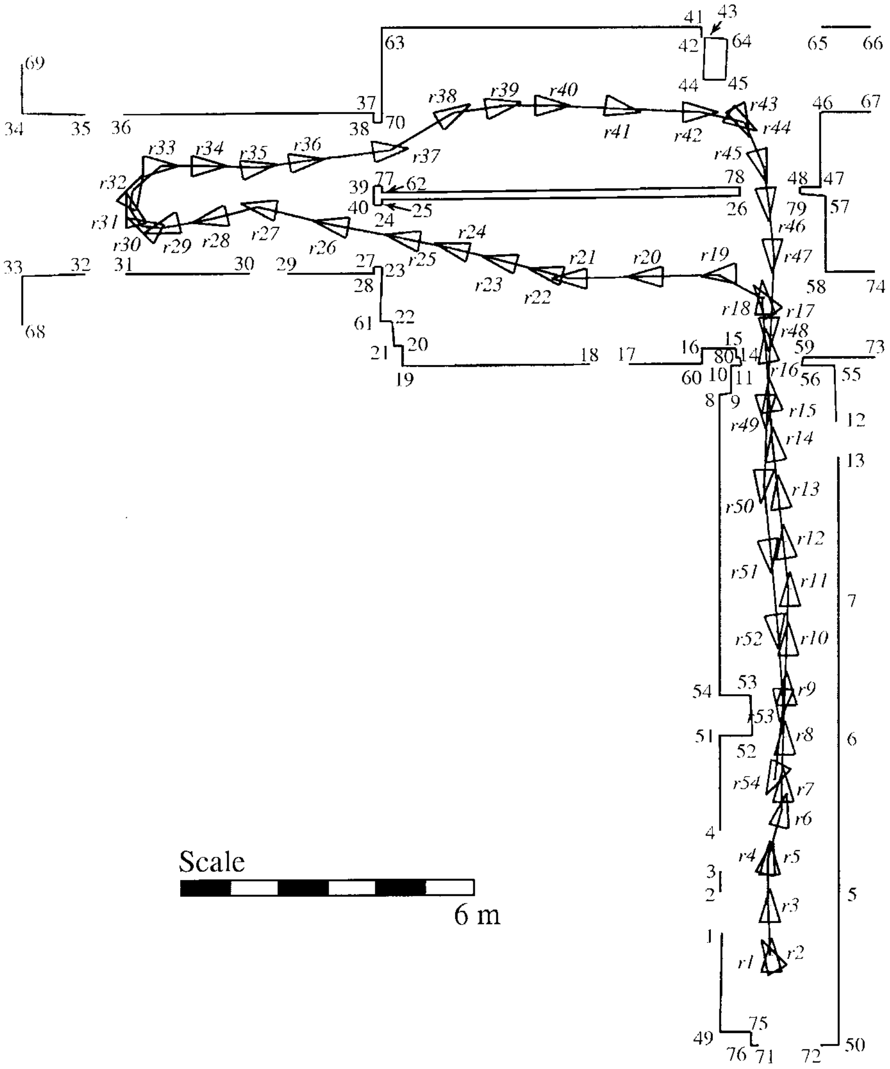

Fig. 4. Robot ground-true trajectory

Table I. Camera intrinsic parameters.

N_{CX}	Number of columns in the camera CCD sensor	752.0
N_{FGX}	Number of columns in the frame grabber	512.0
d_X	Size in X direction of a CCD sel (mm)	0.0086
d_Y	Size in Y direction of a CCD sel (mm)	0.0083
d'_X	Size in X direction of a frame grabber pixel (mm)	0.0126313
d'_Y	Size in Y direction of a frame grabber pixel (mm)	0.0083
C_X	X coordinate of the image centre (pixels)	257.476
C_Y	Y coordinate of the image center (pixels)	252.378
s_X	Scale correction factor in X direction	1.03563
f	Effective focal length (mm)	6.14495
k	Radial distortion factor (mm^{-2})	0.00374955

https://doi.org/10.1017/S0263574700003143 Published online by Cambridge University Press

Although it is clear that no computational improvement can cope with this lack of matches, choosing appropriate lenses and reducing the image acquisition time can limit its effects.

In the following, the derived motion, robust matching results, and computing time are analysed.

B. Motion results

Figure 5 shows the resulting errors between ground-true and computed motion. As remarked in Section VI-A, steps 17–19, 29–33, 37–38 and 40–44 are not displayed or taken into account.

If we focus on step 53, its putative matches are shown in Figure 6. Due to the scene symmetry all of them are wrong, and robust estimation cannot cope with this image pair. Without considering the step 53–54, the mean rotation error is 0.4 deg. and the median rotation error is 0.3 deg. (Figure 5).

Regarding the nearly pure rotations, on steps 1 and 4, a homography could also have been fitted to the matches.

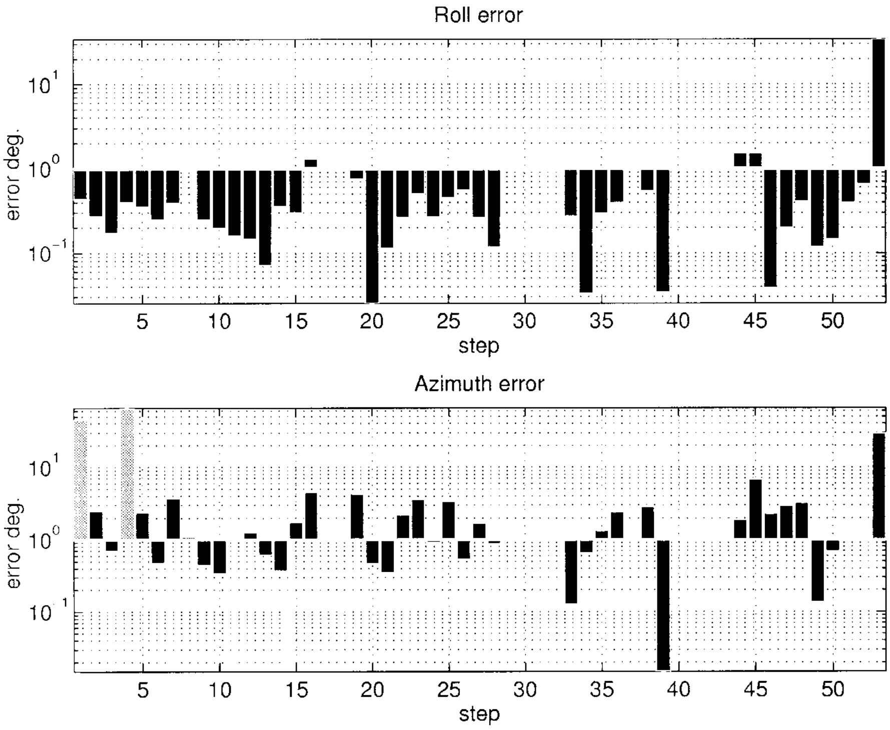

Fig. 5. Computed motion errors with respect to the ground true solution. Errors in grey correspond to pure rotations. Errors in steps 17–19, 29–33, 37–38 and 40–44 are not considered (Section VI-A)

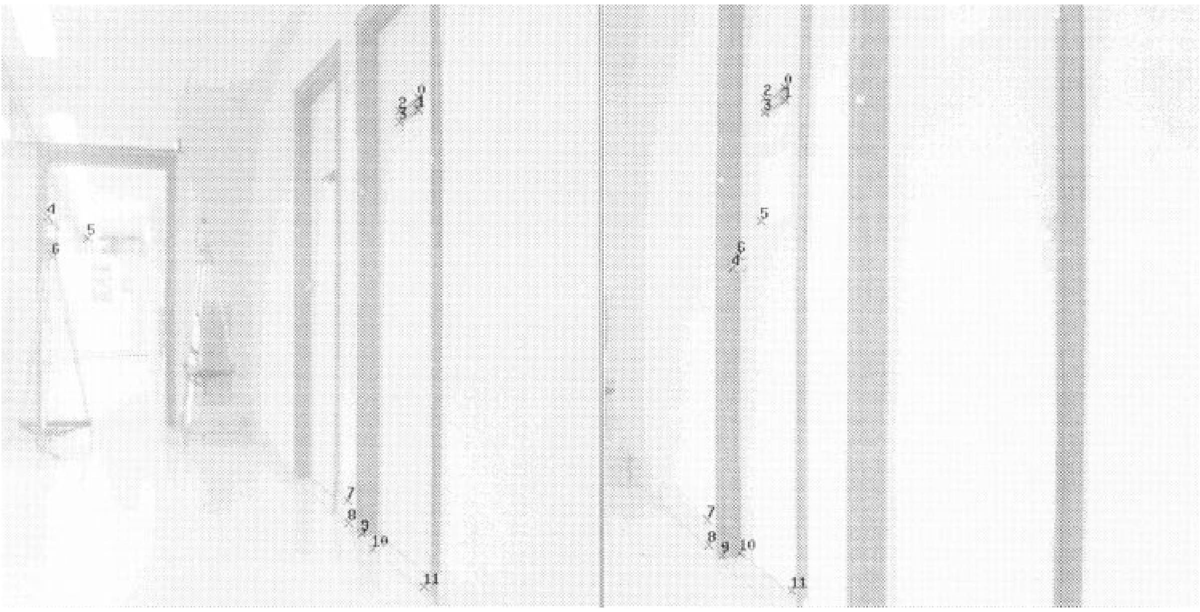

Fig. 6. Initial putative matches in step 53–54

Figure 7 shows the RMS residual for the robust fit of a homography to all the image pairs. It is clear that only in steps 1 and 4 have low residuals, below 2 or 3 pixels. In these cases, the parameterisation of pure rotation considering planarity and camera calibration is valid.

The larger errors in the translation direction are:

- Steps 1 and 4, which correspond to nearly pure rotations, so the computed translation directions are unreliable.
- Step 53 fails because initial putative matches contain a 100% of mismatches.

As steps 1 and 4 can be modelled as pure rotations, they are excluded from the mean translation error computation; step 53 is also excluded. Then the mean azimuth error is 1.7 deg, and the median azimuth error is 1.1 deg. So, the proposed method provides a reliable way of computing accurate motions, using only vision-based techniques.

C. Matching results

Another way of measuring the quality of this technique is analysing the results of matching reliability.

As mentioned, the matching process is composed of two steps. First, putative matches are selected by correlation. Then, outliers are detected using robust statistics.

The total number of automatically computed matches is denoted as AM. A correct match is one pair of image points corresponding to the same 3D feature. Then, matching errors can be classified as:

False negative (FN) Initial putative match which is correct, but that is removed by robust regression. The number of false negative errors in an image pair is denoted as (FN).

Coherent false positive (CFP) A pair of image points corresponding to different 3D points that is not detected by the robust regression because it is coherent with the ground-true epipolar geometry. Therefore, it cannot be detected (see, for example, matches 5 to 14 in Figure 10). The number of coherent false positive matches in an image pair is denoted as (CFP).

Incoherent false positive (IFP) A pair of image points corresponding to different 3D points not detected by robust regression but that is not coherent with the ground-true epipolar geometry. Ideally, they should be detected as an outlier. The number of incoherent false positive matches in an image pair is denoted by (IFP).

False negative ones simply produce a lack of accuracy as fewer points than possible are used to compute the epipolar parameters. Coherent false positives would produce right motion results, because they are coherent with epipolar constraint, and therefore, with camera motion. On the other hand, its corresponding 3D location will be incorrect, as they do not correspond to any physical point. The third ones are properly outliers, that according to reference [4] could spoil regression.

In order to evaluate these misclassifications, every putative match in the sequence has been analysed and classified by a human operator. Then, the automatically selected matches have been classified, and the fraction of every kind of fault has been computed. For every image pair, the following ratios were computed: FN/AM, CFP/AM and IFP/AM. Figure 8 shows these ratios *vs.* the number of image pairs in the evaluation trajectory. The fraction of faults is referred to the whole set of matches of that step that have that fault ratio. For instance, 50% of pairs have neither false negative mismatches nor incoherent false positive ones. And 90% of pairs have less than 5% of false negative (FN) or incoherent false positive mismatches (IFP), and less than 25% of coherent false positive matches (CFP).

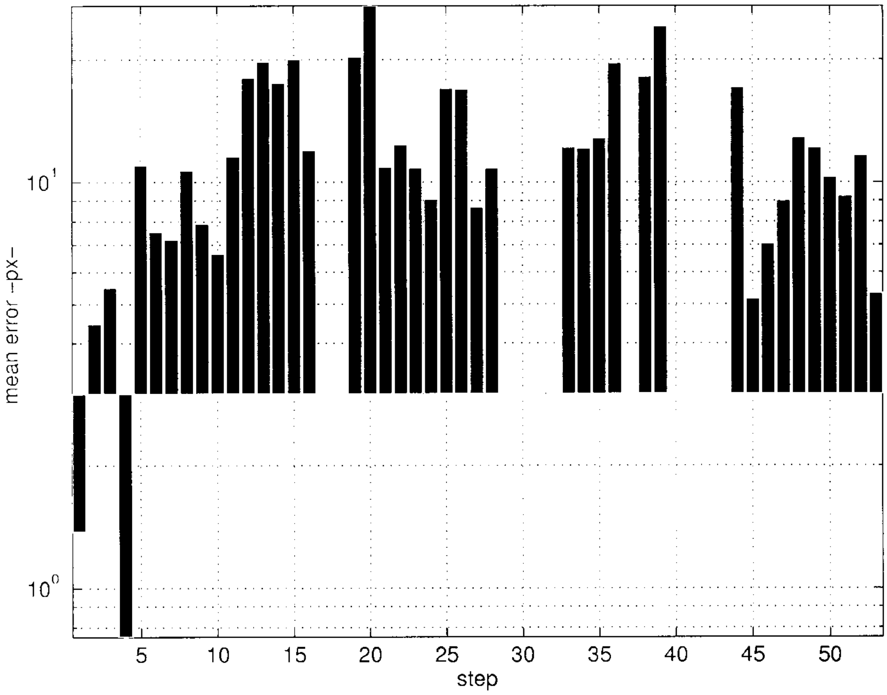

Fig. 7. RMS error for inliers in the robust fit of the homography. Notice how they are lower for steps 1 and 4, which correspond to nearly pure rotations

https://doi.org/10.1017/S0263574700003143 Published online by Cambridge University Press

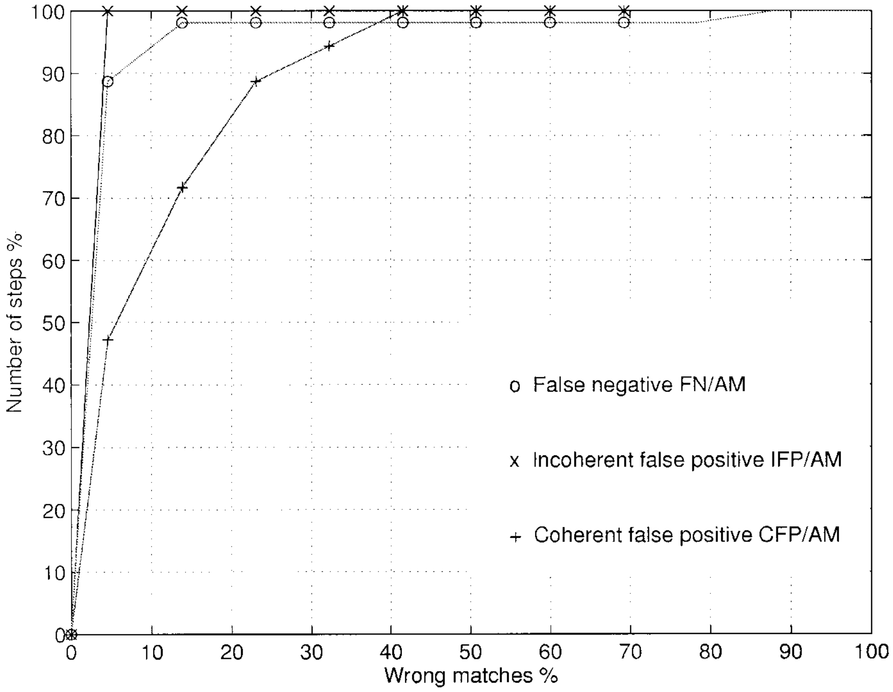

Fig. 8. Accumulated distributions of misclassified matches. Errors are represented as the ratio of misclassified matches to the total number of putative matches in the corresponding image pair

It is clear that robust regression process can detect and reject most incoherent false matches, and properly classify true ones. The fraction of coherent false positives is larger, as they are hardly identifiable by epipolar constraints. Figure 9 shows the ratio between residuals with respect to the ground true solution and the estimated standard deviation used to classify matches. Most errors are due to CFP and few ones are IFP, both of them with small residuals.

Figure 10 illustrates the matching results. It shows a pair of images, and the epipolar geometry derived from ground-true motion. As the robot advances towards the end of the corridor, the epipole (the intersection of epipolar lines) is close to the vanishing point. Then, any features lying on these epipolar lines are easily mismatched. Mismatches 5 to 14 are hardly recognisable, as they lie on the right epipolar line; they are CFP. Some other points (see matches 0 to 2) are right matches rejected by the robust fit. They are FN

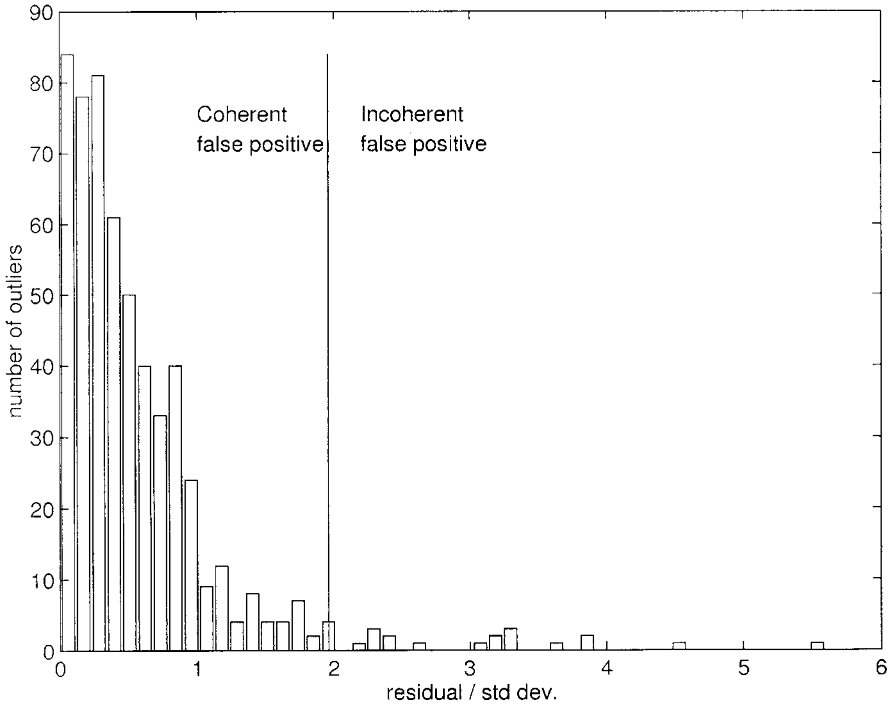

Fig. 9. Distance of CFP and IFP with respect to the ground true geometry. Distance is represented as the ratio of distance in pixels to standard deviation computed in Section IV-A

https://doi.org/10.1017/S0263574700003143 Published online by Cambridge University Press

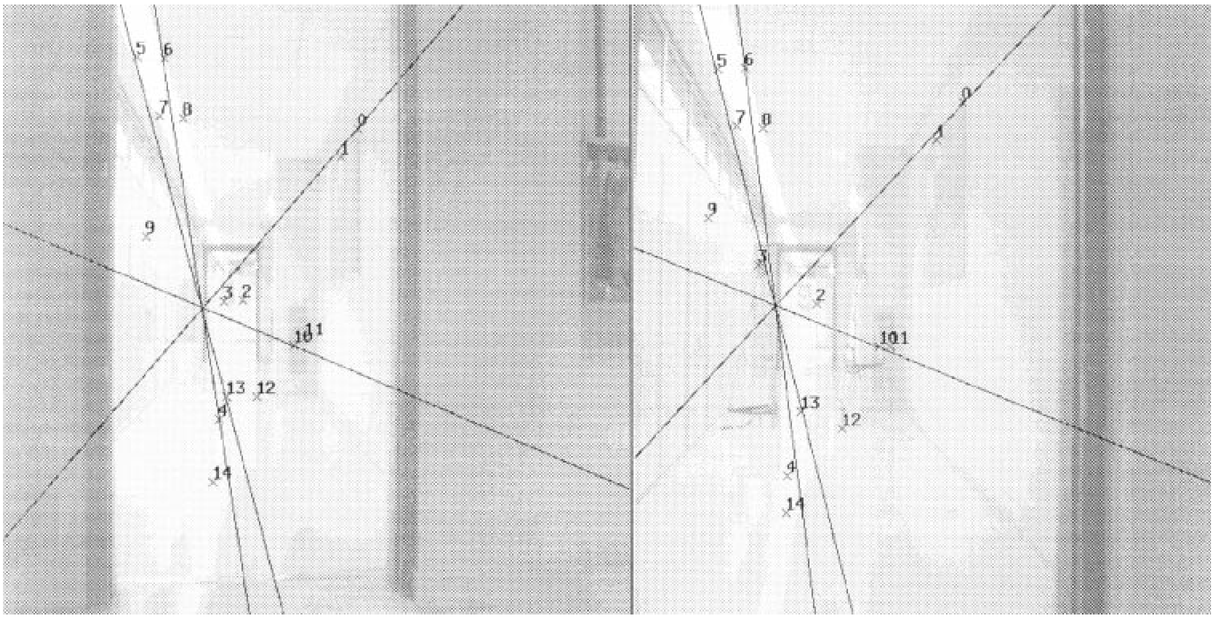

Fig. 10. Examples of misclassified matches: False negatives 0–2 (regarded as outliers), coherent false positive errors 5–14 and two incoherent false positive, 3 and 4 not removed by the robust fit

mismatches. Finally, points 3 and 4 are examples of IFP errors.

D. Computing time

Computing times for every step in the evaluation trajectory are displayed in Figure 11. This shows the accumulated computation time *vs.* the number of steps for epipolar geometry and rotation homography models. As one can see, the mean time is about 0.9 sec for epipolar geometry (a), and for instance, the 90% of steps need less than 1.25 sec to compute the motion from matches. The final time is always under 1.5 sec.

Similarly, Figure 11 shows the computing time for the robust fit of the homography and motion derivation. The mean time is 0.4 sec. and for all the steps it is lower than 0.7 sec. As only one parameter is fitted, the computing time is smaller than in the previous case (a).

The algorithm was computed using MATLAB functions, on a 450 MHz Pentium III PC.

It is important to realize how the previous knowledge about motion and camera parameters is essential to reduce the number of required subsamples. So, as the number of parameters to adjust goes down from seven (general motion) to two (the planar motion of a calibrated camera), the number of randomly tentative solutions to compute is reduced, thus speeding up the process. The inclusion of additional constraints in the epipolar geometry computation also provides more reliable solutions, as overparameterisation is avoided.

The previously analysed times are necessary to derive the motion from matches. The putative matches computation

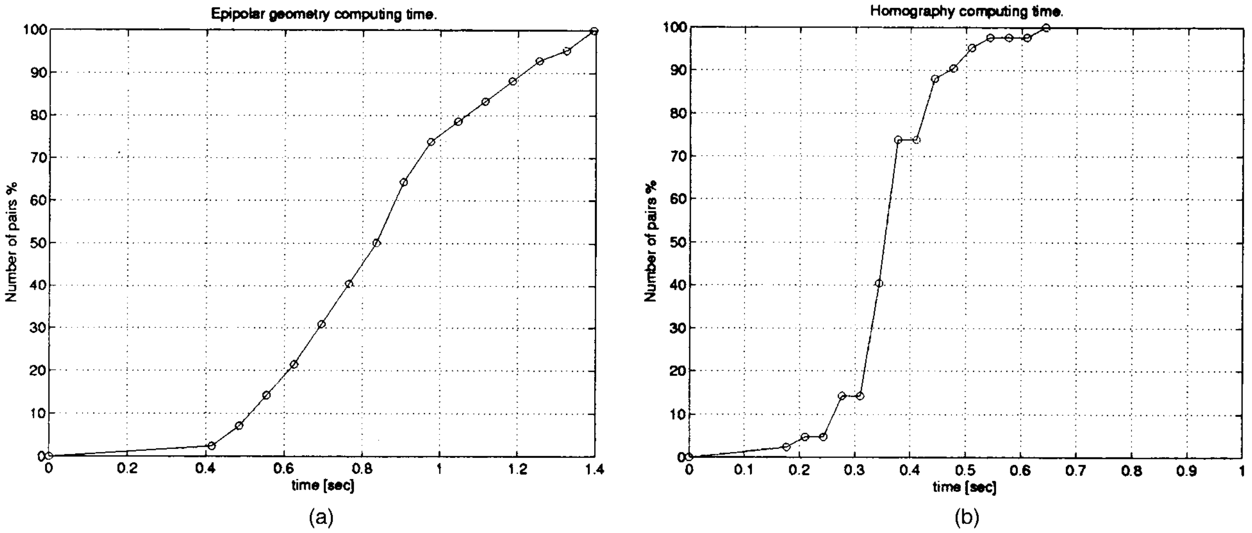

Fig. 11. Accumulated robust regression and motion derivation computing time distribution. (a) Epipolar geometry (b) Rotation homography

https://doi.org/10.1017/S0263574700003143 Published online by Cambridge University Press

time should also be taken into account. This mean time is about 3 seconds per image pair.

VII. CONCLUSIONS

Using epipolar geometry robust regression, the planar motion of a robot carrying one vision camera can be computed accurately. Constraining the trajectory of the camera to be plane and including its intrinsic parameters to derive motion, the minimum number of required matches is only two.

The robust fit of only two parameters (compared with seven in the general uncalibrated case) reduces the computation overhead, and avoids overparameterisation problems.

The experimental trajectory was composed of 54 image pairs. The results after processing have been:

- 11 image pairs cannot be processed because the overlapping is less than 20%. These 11 pairs were removed from the experiment.
- One image pair gave a completely wrong result because of the initial 100% spurious putative matches.
- 2 image pairs where the translation module was less than 10 cm were successfully processed fitting a homography.
- 41 image pairs yield results, with the following performance:

Orientation error: mean 0.4 deg. median 0.3 deg.

Direction of translation: mean 1.7 deg. median 1.1 deg. (the two pure rotation steps have not been considered in these mean values)

Computing time: 3 sec. for putative matches and 0.9 sec. to derive general motion.

No precision device was used to keep the camera horizontal. It was archived by means of a spirit level, used to fixate it to the robot before the experiment was carried out.

VIII. DISCUSSION AND FUTURE WORK

This work has pointed out the capability of some recent computer vision techniques to assist robotic tasks. We consider that this is a promising direction where more results should be exploited in order to improve the robot perception of its environment.

We have shown the importance of considering the degeneration of epipolar geometry when the camera motion is nearly a pure rotation. We have also described the one-parameter model that should be fitted in such a case.

Future work will be directed towards automatic model selection between homography and epipolar geometry.

Acknowledgements

This work has been partially supported by the Diputacin General de Aragn under the project P29/98 and by the Spanish Goverment “Acciones Integradas” reference HB 1998-0047.

The authors wish to express their gratitude to Zhengyou Zhang, the autor of the public available program “image-matching”, used to compute the putative matches for our experiments.

(http://www-sop.inria.fr/robotvis/personnel/zzhang/softwares.html)

References

1. K. Kraus and P. Waldhäusl, *Photogrammetry*, volume 1. Dümmler, 1993.
2. 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”, *Artificial Intelligence Journal*, **78**, 87–119 (October 1995).
3. P.H.S Torr and A. Zisserman. “Robust parameterisation and computation of the trifocal tensor”, *Image and Vision Computing*, **15**, 591–605, 1997.
4. P.J. Rousseeuw and A.M. Leroy, *Roboust Regression and Outlier Detection* (Wiley, 1987).
5. P.J. Huber, *Robust Statistics* (Wiley, 1981).
6. R.C. Bolles and M.A. Fischler. “A ransac-based approach to model fitting and its applications to finding cylinders in range data.” *Int. Joint. Conf. Artificial Intelligence*, (1981), pp. 637–643.
7. G. Blais and M.D. Levine. “Registering multiview range data to create 3d computer objects”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **17**(8), 820–824 (August 1995).
8. J. Luck, C. Little and W. Hoff, “Registration of range data using a hybrid simulated annealing and iterative closest point algorithm,” *IEEE Int. Conf. on Robotics and Automation*, San Francisco, USA (2000), pp. 3739–3744..
9. O. Faugeras, *Three-Dimensional Computer Vision* (The MIT Press, Cambridge, Massachusetts, USA, 1993).
10. A. Shashua, “Algebraic functions for recognition”, *IEEE Trans. Pattern Analysis and Machine Intelligence*, **17**(8), 779–789 (1995).
11. R.I. Hartley, “Lines and points in three views and the trifocal tensor”, *Int. J. Computer Vision*, **22**(2), 125–140 (1997).
12. P. Torr, A.W. Fitzgibbon and A. Zisserman, “Maintaining multiple motion model hypotheses over many views to recover matching and structure.” *IEEE Int. Conf. on Computer Vision*, India (January 1998), pp. 485–491.
13. P. H. S. Torr, A. Zisserman and S. Maybank. “Robust detection of degenerate configurations for the fundamental matrix”, *Computer Vision and Image Understanding*, **71**(3), 312–333 (September 1998).
14. R.Y. Tsai. “A versatile camera calibration technique for high accuracy 3d machine vision metrology using Off-the-Shelf tv cameras and lenses”, *IEEE Journal of Robotics and Automation*, **RA-3**(4), 323–344 (August 1987).
15. G. Xu and Z. Zhang, *Epipolar Geometry in Stereo, Motion and Object Recognition: A Unified Approach* (Kluwer Academic Publishers, 1996).
16. J. Weng, T.S. Huang, and N. Ahuja, *Motion and Structure from Image Sequences* (Springer-Verlag, Heidelberg, 1993).
17. R.I. Hartley, “In defence of the eight-point algorithm”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **19**(6), 580–593 (June 1997).
18. G.H. Golub and C.H Van Loan, *Matrix Computations* (Johns Hopkins University Press, 1989).
19. Z. Zhang, “A new multistage approach to motion and structure estimation: From essential parameters to Euclidean motion via fundamental matrix,” *Technical report* (INRIA, Shophia-Antipolis, 1996).
20. J.A. Castellanos, J.M. Martínez, J. Neira, and J.D. Tardós. “Experiments in multisensor mobile robot localisation and map building.” *3rd IFAC Symposium on Intelligent Autonomous Vehicles*, Madrid, Spain (March 1998).

https://doi.org/10.1017/S0263574700003143 Published online by Cambridge University Press