DIIS - I3A
Universidad de Zaragoza
C/ María de Luna num. 1
E-50018 Zaragoza
Spain

Internal Report: 2008-V02

Localization and Matching using the Planar Trifocal Tensor with Bearing-only Data [1]

J.J. Guerrero, A.C. Murillo, C. Sagüés

If you want to cite this report, please use the following reference instead:
Localization and Matching using the Planar Trifocal Tensor with Bearing-only Data, J.J. Guerrero, A.C. Murillo, C. Sagüés, *IEEE Transactions on Robotics*, Vol. 24(2): 494-501, 2008.

[1]This work was supported by the projects DPI2006-07928, IST-1-045062-URUS-STP.

Localization and Matching using the Planar Trifocal Tensor with Bearing-only Data.

J.J. Guerrero, A.C. Murillo, C. Sagüés
DIIS - I3A, University of Zaragoza, Maria de Luna 1, 50018 Zaragoza, Spain
Email: {jguerrer, acm, csagues}@unizar.es

Abstract

This paper addresses the robot and landmark localization problem from bearing-only data in three views, simultaneously to the robust association of this data. The localization algorithm is based on the 1D trifocal tensor, which relates linearly the observed data and the robot localization parameters. The aim of this work is to bring this useful geometric construction from computer vision closer to robotic applications. One contribution is the evaluation of two linear approaches of estimating the 1D tensor: the commonly used approach that needs seven bearing-only correspondences and another one which uses only five correspondences plus two calibration constraints. The results in this paper show that the inclusion of these constraints provides a simpler and faster solution and better estimation of robot and landmark locations in the presence of noise. Moreover, a new method that makes use of scene planes and requires only four correspondences is presented. This proposal improves the performance of the two previously mentioned methods in typical man-made scenarios with dominant planes, while it gives similar results in other cases. The three methods are evaluated with simulation tests as well as with experiments that perform automatic real data matching in conventional and omnidirectional images. The results show sufficient accuracy and stability to be used in robotic tasks such as navigation, global localization or initialization of SLAM algorithms.

Keywords: Robot vision, Bearing-only data, Robust matching, 1D Trifocal tensor, global localization, SLAM initialization

I. Introduction

When an unknown scene is observed from multiple unknown positions, a complex but well-known geometric problem appears. The goal is to associate the observations and to recover the robot and landmark locations. Let us focus on the case of 1D bearing-only observations while performing planar robot motion, which is typical for robots working in man-made environments. Robot localization based on bearing-only data has been considered in autonomous guided vehicles using landmarks of known location [1] and also in simultaneous localization and mapping (SLAM), where the initialization and the data matching are difficult problems [2].

This geometric problem is similar to the structure from motion problem studied in computer vision [3], but specialized for 1D observations and 2D locations. In computer vision it is usual to relate different views with an initial matching of relevant features, followed by a robust estimation of a projective transformation or a tensor [4]. These tensors provide general constraints between the bearing-only observations, and allow us to recover the camera localization from them. For example, the fundamental matrix has been extensively used for two-view robust matching [5], since it provides a general constraint for 2D bearing-only observations. Recently, it has been applied to help loop closing methods in SLAM [6].

However, using 1D bearing-only data, two views do not provide a constraint between observations; at least three views are required. The 1D trifocal tensor encodes the three-view constraint between bearing-only observations, and can be used to compute both the robot and landmarks 2D localization in closed form [7]. The 1D trifocal tensor has been used for self calibration of cameras [8] and for radial distortion correction in wide angle cameras [9]. In robotics, it was previously used for initialization of bearing-only SLAM algorithms [10], using conventional cameras and artificial landmarks on a plane.

This paper focusses on the use of the 1D trifocal tensor for matching 1D bearing-only data as well as computing robot and landmarks localization. In previous related works [10], a minimum of seven matches were used to compute this tensor. In the calibrated case, two constraints for the 1D tensor were presented in [7]. We evaluate here two linear methods to estimate the tensor; first the commonly used one that needs 7 matches, and a second one that needs a minimum of 5 matches and includes these two constraints. We also present and evaluate a third method to estimate the 1D trifocal tensor, which only needs four matches if three of them are collinear in the 2D scene. The key issue in this third method is to take advantage of typical structure in man-made environments (e.g. vertical planar walls). An interesting result from the experiments is that both methods that include the calibration constraints perform better, particularly for real-time robotic applications: they provide smaller errors in localization estimation and significantly reduce the computational cost of the robust data matching process. Notice that these best performing methods have not yet been applied previously.

The 1D trifocal tensor can be used with bearing-only data from any source, for instance from vision sensors. Omnidirectional images are the most intuitive situation where we can obtain it, e.g. extracting radial lines in the image and using its bearing [11]. This kind of image has received increasing attention from robotics researchers lately, due to its many well known advantages for robotics [12], [13], [14]. In this paper we present examples using two different image features: SIFT [15] which is very

This work was supported by projects MEC DPI2006-07928 and IST-1-045062-URUS-STP.
Part of this work was presented in IEEE ICRA 2006 and in IEEE IROS 2006.

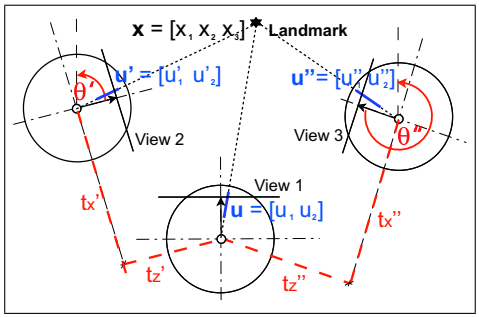


Fig. 1. The aim is to obtain the relative location of the robot (θ', θ'', $\mathbf{t}' = [t'_x t'_z]$, $\mathbf{t}'' = [t''_x t''_z]$) and the position of the landmarks $\mathbf{x}$, from triplets of bearing-only observations $\mathbf{u}$, $\mathbf{u}'$, $\mathbf{u}''$, that are automatically matched.

effective in widely separated images, and vertical lines from natural landmarks projected in conventional and omnidirectional images.

From our point of view, some results in vision research are difficult to be used in robotic applications, probably due to the current divergence of computer vision and robotics communities. Here, we try to experiment and make advanced results accessible to robotic researchers. Whereas most of the mathematics can be recovered from computer vision papers, its particularization to the 1D bearing-only observations with planar sensor motion and the extensive experimentation presented here are useful in robotics. Besides, the two reduced methods using 5 and 4 matches make the 1D tensor more suitable for robotics. We believe that this work would be interesting for researchers using omnidirectional vision systems as well as for people working in bearing-only localization and mapping, navigation or visual servoing. Two-view relations like the fundamental matrix or the homography have been extensively used in these systems [16], [17], but the use of multiple views, like image triplets are still poorly studied despite its attractive properties.

II. Planar Trifocal Tensor

Let us assume a robot moving on a plane and observing an unknown scene, where 1D bearing-only observations are acquired from different robot locations. If v is the number of views and l is the number of observed landmarks, we have vl equations. Three motion parameters are needed for each robot location (except for the first one, where we locate the origin) and two parameters (up to a global scale factor) for each landmark. If the number of views v is 2, the problem is unsolvable even with an infinite number of landmarks, since $vl < 3(v-1) + 2l - 1$. Three is the minimum number of views needed to solve it, with at least 5 observed landmarks.

As in [10], we are interested in obtaining a linear solution to the bearing-only localization and mapping problem (Fig. 1) that does not suffer from local minima. The first issue is to represent the bearing measurements obtained by the robot using a projective formulation. Thus, we can easily convert a bearing measurement α to its projective formulation in a 1D virtual retina as $\mathbf{u} = (\sin\alpha, \cos\alpha)^T$.

A trifocal tensor relates linearly three different views independently of the observed scene. It allows us to uncouple the mapping and the robot localization problems, as it provides a general constraint for data association when neither robot nor landmark locations are known. When we have landmarks in a 2D scene, observed as a $\mathcal{P}^1$ projective space, the 1D trifocal tensor is used. Let us name the homogeneous representation of a feature in a 2D space as $\mathbf{x} = [x_1, x_2, x_3]^T$ and the representation of its observation in the $\mathcal{P}^1$ projective space as $\mathbf{u} = [u_1, u_2]^T$. This projection can be expressed by a 2×3 matrix $\mathbf{M}$ for the three sensor locations, as:

$$\lambda\mathbf{u} = \mathbf{M}\mathbf{x}, \qquad \lambda'\mathbf{u}' = \mathbf{M}'\mathbf{x}, \qquad \lambda''\mathbf{u}'' = \mathbf{M}''\mathbf{x}, \tag{1}$$

where λ, λ' and λ'' are scale factors.

If we suppose all the 2D features in a common reference frame placed in the first robot location, the projection matrixes relating the scene and the image features are $\mathbf{M} = [\mathbf{I}|\mathbf{0}]$, $\mathbf{M}' = [\mathbf{R}'|\mathbf{t}']$ and $\mathbf{M}'' = [\mathbf{R}''|\mathbf{t}'']$ for the first, second and third location respectively. Here, $\mathbf{R}' = \begin{bmatrix} \cos\theta' & \sin\theta' \\ -\sin\theta' & \cos\theta' \end{bmatrix}$ and $\mathbf{R}'' = \begin{bmatrix} \cos\theta'' & \sin\theta'' \\ -\sin\theta'' & \cos\theta'' \end{bmatrix}$ are the rotations and $\mathbf{t}' = [t'_x, t'_z]^T$ and $\mathbf{t}'' = [t''_x, t''_z]^T$ are the translations (Fig. 1).

Removing $\mathbf{x}$, λ, λ' and λ'' from the projection equations (1), a constraint for the three observations of a landmark appears. It is written as the trifocal constraint [8]:

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} T_{ijk} u_i u'_j u''_k = 0, \tag{2}$$

where T_{ijk} $(i, j, k = 1, 2)$ are the eight elements of the $2 \times 2 \times 2$ tensor whose components are the 3×3 minors of the 6×3 matrix $[\mathbf{M}^T\ \mathbf{M}'^T\ \mathbf{M}''^T]^T$ in such a way that to obtain $T_{ijk} = [\bar{i}\bar{j}\bar{k}]$ we take the $\bar{i}$ row of $\mathbf{M}$, the $\bar{j}$ row of $\mathbf{M}'$ and the $\bar{k}$ row of $\mathbf{M}''$, meaning $\bar{\cdot}$ a mapping from [1,2] to [2,-1].

A. Robot and Landmark localization from the trifocal tensor

The camera and landmark locations can be computed in closed form from the 1D trifocal tensor. The motion parameters can be related to the components of the tensor by developing the elements of the projection matrixes $(\mathbf{M}, \mathbf{M}', \mathbf{M}'')$. In this work, two methods to recover the robot and landmarks localization from those relationships were used: the algorithm presented in [10], which is based on the decomposition of the tensor into two intrinsic homographies [18], and the method developed in [7]. Both methods give us almost identical results, therefore, no explicit comparison will be shown. They both provide two symmetric solutions for the location parameters, defined up to a scale for the translations. This two-fold ambiguity [7] is one of the drawbacks of using only three 1D views to solve this problem. Both solutions of the camera and landmark localization problem are coherent with the triplets of 1D bearing-only observations. The consistency in the visibility of the landmarks can help in the search of the right solution. A fourth view, when available, can be used to disambiguate between the two solutions [10]. However, in practice some additional information like approximate camera location or orientation (e.g. from odometry or reference information) can be easily used. If the image triplets are nearly collinear the selection of the correct location is more sensitive to noise, but the consequences of a bad selection are smaller because both solutions are close. More details about the two-fold ambiguity can be seen in [7]. Once the relative location of the camera has been estimated, the location of the landmarks can be obtained by solving the projection eq. (1) for each landmark [10].

III. Simultaneous Robust Matching and 1D Trifocal Tensor Computation

To recover the structure and motion parameters from the 1D trifocal tensor, the latter must be estimated in calibrated coordinates. The first part of this section explains two methods to estimate this tensor, and the second part shows how to perform the robust matching simultaneously to the tensor estimation.

A. Two ways of obtaining a tensor in calibrated coordinates

To compute the tensor we can use eq (2), that relates the feature coordinates in the three views and the tensor elements

$$\begin{array}{c}[u_1u'_1u''_1 \;\; u_1u'_1u''_2 \;\; u_1u'_2u''_1 \;\; u_1u'_2u''_2 \\ u_2u'_1u''_1 \;\; u_2u'_1u''_2 \;\; u_2u'_2u''_1 \;\; u_2u'_2u''_2]\mathbf{T} = 0,\end{array} \tag{3}$$

where $\mathbf{T} = [T_{111}\; T_{112}\; T_{121}\; T_{122}\; T_{211}\; T_{212}\; T_{221}\; T_{222}]^T$ is a vector with the elements of the trifocal tensor. In a general case, $\mathbf{T}$ has eight parameters defined up to an overall scale factor. As each observation gives one equation (3), at least seven matches are required to compute it. Seven triplets of corresponding features in calibrated coordinates can be used to construct a 7×8 matrix $\mathbf{A}$. The solution of this system of equations corresponds to the eigenvector associated with the smallest eigenvalue of the matrix $\mathbf{A}^T\mathbf{A}$, that can be obtained from singular value decomposition (SVD) of matrix $\mathbf{A}$ [4].

When the tensor is not computed from features expressed in calibrated coordinates, we must perform a second step to bring the tensor to the calibrated case. For example, if image coordinates in pixels are used, the transformation of these coordinates from the image $\mathbf{u}^p$ to a calibrated retina $\mathbf{u}^c$ is linearly expressed as: $\mathbf{u}^p = \mathbf{K}\mathbf{u}^c$, where $\mathbf{K}$ is obtained from the camera internal parameters $\mathbf{K} = \begin{bmatrix} f & u_0 \\ 0 & 1 \end{bmatrix}$, being f the focal length and u_0 the position of the principal point, both expressed in pixels. Then, if the tensor $\mathbf{T}^p$ is computed from image coordinates, the tensor in calibrated coordinates $\mathbf{T}^c$ can be obtained from $\mathbf{T}^p$ and the calibration matrix $\mathbf{K}$ as

$$T^c_{lmn} = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} T^p_{ijk}K(i,l)K(j,m)K(k,n), \tag{4}$$

where $K(i,l)$ is the element in row i, column l of matrix $\mathbf{K}$.

On the other hand, the camera motion parameters can be related to the trifocal tensor elements by expanding the elements of the projection matrixes $(\mathbf{M}, \mathbf{M}', \mathbf{M}'')$as:

$$\begin{array}{ll} T_{111} = t'_z\sin\theta'' - t''_z\sin\theta'; & T_{211} = -t'_z\cos\theta'' + t''_z\cos\theta' \\ T_{112} = t'_z\cos\theta'' + t''_x\sin\theta'; & T_{212} = t'_z\sin\theta'' - t''_x\cos\theta' \\ T_{121} = -t'_x\sin\theta'' - t''_z\cos\theta'; & T_{221} = t'_x\cos\theta'' - t''_z\sin\theta' \\ T_{122} = -t'_x\cos\theta'' + t''_x\cos\theta'; & T_{222} = -t'_x\sin\theta'' + t''_x\sin\theta'. \end{array} \tag{5}$$

In eq. (5), we can find the following two linear constraints for the elements of the tensor expressed in calibrated coordinates:

$$\begin{array}{c} -T_{111} + T_{122} + T_{212} + T_{221} = 0 \\ T_{112} + T_{121} + T_{211} - T_{222} = 0. \end{array} \tag{6}$$

In [7] it is shown that a tensor $\mathbf{T}$ is expressed in calibrated coordinates if and only if it satisfies eq. (6). These calibration constraints are linear with the elements of the 1D trifocal tensor, which allows us to linearly obtain this tensor from only five matches expressed in calibrated coordinates. Notice that instead of using eq. (3), the following eq. (7) is applied now to each triplet of corresponding features.

$$\begin{array}{l} T_{111}(u_1u'_1u''_1 + u_2u'_2u''_1) + T_{112}(u_1u'_1u''_2 + u_2u'_2u''_2) + \\ T_{121}(u_1u'_2u''_1 + u_2u'_2u''_2) + T_{122}(u_1u'_2u''_2 - u_2u'_2u''_1) + \\ T_{211}(u_2u'_1u''_1 + u_2u'_2u''_2) + T_{212}(u_2u'_1u''_2 - u_2u'_2u''_1) = 0 \end{array} \tag{7}$$

To sum up, there are two ways to estimate the tensor:

- To compute the tensor using (3), making a search in a 7 degrees of freedom (d.o.f.) space of solutions. This is the classical solution, used for example in [10] or [9], named in the rest of the paper TT7 for short.
- To compute a tensor from data in calibrated coordinates $\mathbf{u}^c$ using (7), which imposes the two linear calibration constraints (6). So, a search in a 5 d.o.f. space of solutions is made. Let us name this method as TT5 from now on.

The second option simplifies the computations using subsets of 5 matches instead of the 7 matches needed in the first one. These are easier to obtain and allow us to reduce the number of iterations required for the robust computation, as it will be explained in next subsection. Besides, the TT5 gives more accurate motion estimation in presence of noise, according to the results from the experimental section (V).

B. Robust Matching and 1D Trifocal Tensor Computation

The computation of $\mathbf{T}$ can be carried out using SVD as explained before. With more matches than the minimum number required, that procedure gives the least squares solution, which assumes that all measurements can be interpreted with the same model. This is very sensitive to outliers, so robust estimation methods are necessary to avoid outliers in the process. From the existing robust estimation methods, we use RANSAC [19], which carries out a search in the space of solutions obtained from subsets of minimum number of matches. This robust estimation allows us to obtain simultaneously the tensor estimation and a robust set of correspondences. It consists of the following steps:

- After extracting relevant features in the three views, the automatic matching process first obtains a putative set of matches (*basic matching*) based on the appearance of the features in the image.
- Afterwards, the geometrical constraint imposed by the tensor is included to obtain a *robust matching* set using a RANSAC voting approach. This robust estimation efficiently rejects the outliers from the basic matching.
- Optionally, the tensor constraint can be used to grow the final set of matches, obtaining new ones with weaker appearance-based similarity but fitting the geometric constraint.

The RANSAC voting approach needs an error value to measure how good a certain match fits the tensor constraint. This fitting error is computed as the average difference between the observed coordinates and the coordinates estimated by the tensor. Given the coordinates of a certain feature in two images (e.g. **u'** and **u"**) and the components of the trifocal tensor, the estimation ($\widehat{\ }$) of the coordinates $\mathbf{u} = (u_1, u_2)$ in a third view is obtained as follows:

$$\widehat{\frac{u_1}{u_2}} = -\frac{T_{211}u'_1u''_1+T_{212}u'_1u''_2+T_{221}u'_2u''_1+T_{222}u'_2u''_2}{T_{111}u'_1u''_1+T_{112}u'_1u''_2+T_{121}u'_2u''_1+T_{122}u'_2u''_2}. \tag{8}$$

Similarly for each view, the total error is computed as the mean of the three squared errors $(\frac{u_1}{u_2} - \widehat{\frac{u_1}{u_2}})^2$, $(\frac{u'_1}{u'_2} - \widehat{\frac{u'_1}{u'_2}})^2$, $(\frac{u''_1}{u''_2} - \widehat{\frac{u''_1}{u''_2}})^2$, which are expressed in pixels. If angle units are required, the reprojection error can be obtained from the expression $(\arctan\frac{u_1}{u_2} - \arctan\widehat{\frac{u_1}{u_2}})^2$. Both residues are more convenient than the residue obtained from (3), which has only an algebraic meaning and unknown units.

In practice, there is an agreement between the computational cost of the search in the space of solutions, and the probability of failure $(1-P)$. A random selection of m subsets of n matches ends up with a good solution if all the matches are correct in at least one of the subsets. Assuming a ratio ε of outliers, to obtain a good result with a probability P the number of subsets to explore is $m = \frac{\log(1-P)}{\log(1-(1-\varepsilon)^n)}$. For random sampling methods, such as RANSAC , it is important to solve the model in closed form because a narrow space of solutions must be evaluated. It is also important to use as small number of matches as possible to reduce the probability of a mismatch being included in the random subsets. So, there is a great advantage for TT5 with regard to TT7 (both models are linear with respect to observations), because of the smaller number of matches required. For instance, let us assume a situation with a 40% of outliers. If estimating TT5, the RANSAC method needs to perform 57 iterations to get a result with 99% of probability of being correct. However, if we use TT7, the RANSAC algorithm will need 163 iterations for the same level of confidence, which is almost three times more.

IV. TENSOR COMPUTATION WITH A PLANE IN THE SCENE

When the robot moves in man-made environments, there are many vertical planar surfaces in the scene, which can help in the computation of the trifocal tensor. The implicit coplanarity of the observations gives new constraints for the tensor, which simplify the situation and reduce the number of required matches. These constraints have been presented for three views of a 3D scene, with the 2D trifocal tensor [20], [3]. There, it is shown a relation between the tensor $\mathbf{T}$, defined between three views, and two homographies $\mathbf{H}'$ (from image 1 to 2) and $\mathbf{H}''$ (from image 1 to 3) corresponding to the same scene plane. Here, we develop these constraints for 1D bearing-only observations of a 2D scene.

Two 1D homographies $\mathbf{H}'$ and $\mathbf{H}''$, corresponding to a line in the 2D scene, relate the bearing-only observations in three views ($\mathbf{u}$, $\mathbf{u}'$ and $\mathbf{u}''$) of a certain landmark as:

$$\mathbf{u}' = \mathbf{H}'\mathbf{u}\ , \quad \mathbf{u}'' = \mathbf{H}''\mathbf{u}. \tag{9}$$

On the other hand, the constraint imposed by the 1D trifocal tensor (2) can be reordered as,

$$u_1(T_{111}u'_1u''_1 + T_{112}u'_1u''_2 + T_{121}u'_2u''_1 + T_{122}u'_2u''_2)+ \\ u_2(T_{211}u'_1u''_1 + T_{212}u'_1u''_2 + T_{221}u'_2u''_1 + T_{222}u'_2u''_2) = 0.$$

If we name $\mathbf{T}_1 = \begin{bmatrix} T_{111} & T_{112} \\ T_{121} & T_{122} \end{bmatrix}$ and $\mathbf{T}_2 = \begin{bmatrix} T_{211} & T_{212} \\ T_{221} & T_{222} \end{bmatrix}$, the trifocal equation can be written as:

$$[u_1 u_2] \begin{bmatrix} \mathbf{u}'^T\mathbf{T}_1\mathbf{u}'' \\ \mathbf{u}'^T\mathbf{T}_2\mathbf{u}'' \end{bmatrix} = 0. \tag{10}$$

Substituting with (9) and naming $\mathbf{B}_1 = \mathbf{H}'^T\mathbf{T}_1\mathbf{H}''$ and $\mathbf{B}_2 = \mathbf{H}'^T\mathbf{T}_2\mathbf{H}''$, eq. (10) becomes:

$$u_1\mathbf{u}^T\mathbf{B}_1\mathbf{u} + u_2\mathbf{u}^T\mathbf{B}_2\mathbf{u} = 0. \tag{11}$$

Equation (11) must be certain for any point $\mathbf{u} = [u_1 u_2]^T$, so it must be the zero polynomial in u_1 and u_2 , and therefore the coefficients should be zero. Thus, four new constraints for the elements of the tensor can be stated:

$$B_2(1,1) + B_1(2,1) + B_1(1,2) = 0; \quad B_1(1,1) = 0; \\ B_1(2,2) + B_2(2,1) + B_2(1,2) = 0; \quad B_2(2,2) = 0. \tag{12}$$

It is known that 3 matched features are enough to compute a 1D homography from visual data in two 1D projections [21]. Thus, the coplanarity constraint reduces by one the minimum number of features required to compute the 1D trifocal tensor, which gives a constraint for all landmarks in a general scene. It can be computed from four observed features, three of them coplanar (collinear in the 2D scene) to estimate the homography, and a fourth correspondence out of that vertical plane. Let us name this method TT4 from now on.

This reduction of the minimum number of matched features is advantageous because of the robust technique used. In this case, instead of doing a random search in a 5 d.o.f. space of solutions, we perform a search in a 3 d.o.f. space to robustly estimate the homography; plus a second search in a 1 d.o.f. space to find the fourth required match. For example, let us suppose a RANSAC estimation with 40% of outliers, and a probability $P = 99\%$ of not failing in the random search. Now, the RANSAC algorithm would need 19+6 iterations, less than the iterations needed for TT5 (57 iterations). Notice that this big difference is realistic only if the vertical plane is dominant in the images. Another important advantage of the estimation of the homography is that it can give us some hint about degenerate situations (e.g. when a sole homography explains all the scene) and a model selection would be useful [22].

V. Experimental results

The first part of this section evaluates the performance of the different algorithms with simulated data. Afterwards, several real data experiments test their behavior in robotics.

A. Simulated data experiments

We implemented a simulator of 2D scenes which are projected into 1D virtual cameras with $53°$ of field of view and 1024 pixels. All tests were done for two different movements of the robot (Fig 2). The first one (MovA) could fit a common multi-robot configuration. The second one (MovB) represents a typical situation with a mobile robot going forward.

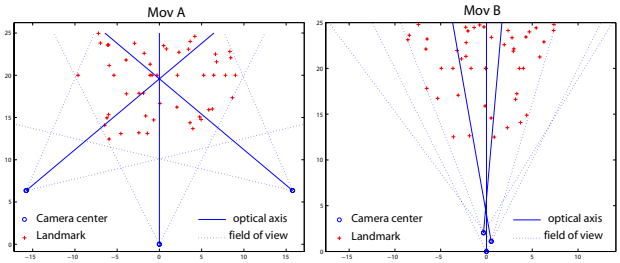


Fig. 2. *MovA & MovB.* Scheme of the two simulated scenarios.

In all simulation tests, measuring errors were represented as Gaussian random noise (with zero mean and standard deviation varying from 0 to 1 pixel) added to feature image coordinates. The evaluation parameters shown for each method are the RMS (root-mean-square) error in the rotation angles (θ' and θ'') and in the translation directions (t' and t'') of the camera location.

1) Experiment 1Sim. General scenario: This experiment compares the performance of TT7 and TT5 in general scenarios. To estimate the trifocal tensor we chose 30 random matches all over the scene. Fig. 3 shows the localization errors achieved in these experiments. It is clear that TT5 method performs better than the classical TT7: it provided more accurate and stable results for both evaluated parameters (rotation and translation direction), as TT5 always gave smaller errors and standard deviations than TT7. The same experiments were also done using 20 matches instead of 30, and we obtained similar behavior, indeed with slightly higher errors (around 10% more).

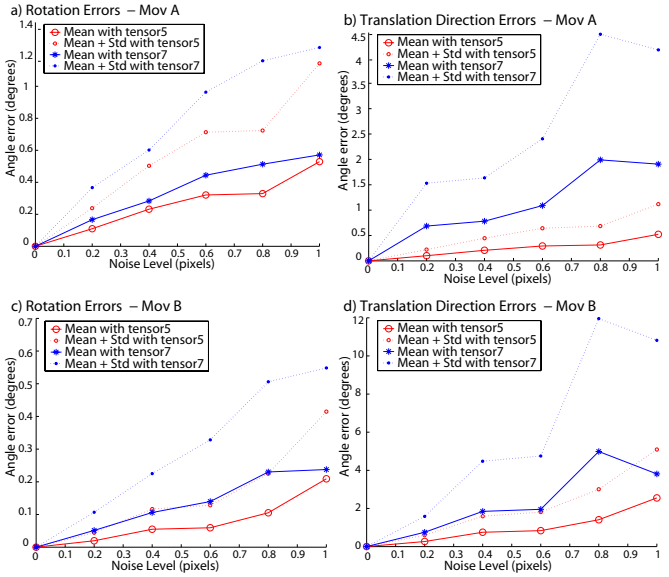


Fig. 3. *Experiment 1Sim. General scenario.* RMS error (mean and mean+std from 100 executions) in rotation and translation direction estimation using TT5 and TT7.

2) Experiment 2Sim. Scene with planes: This experiment evaluates the performance taking or not into account the coplanarity constraints; TT4 against TT5.

Two cases were considered: first a general case, where the vertical plane is not dominant (10 random matches in the plane and 20 random matches out of it). Second, the plane is dominant in the images (20 random matches in the plane and 10 out of it). The vertical plane was placed parallel to the first image, 20 units ahead the origin, in both scenarios shown in Fig. 2. For the first case, TT4 gave similar results to TT5 (both very similar to results shown in previous experiment Sim1). However, when the plane is dominant in the images, some performance improvements were detected using TT4 method. Fig. 4 shows the results of this case. Notice that TT4 behaves better than TT5, particularly when noise increases.

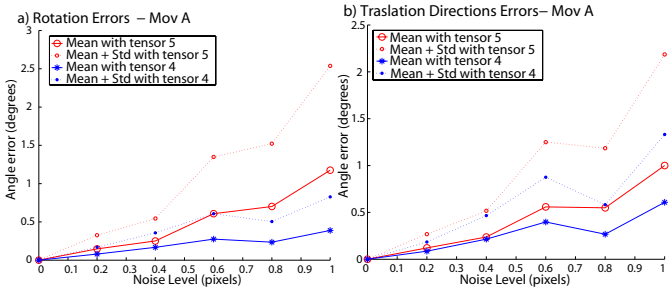


Fig. 4. *Experiment 2Sim. Dominant plane in the scene.* RMS error (mean and mean+std from 100 executions) in rotation and translation direction estimation using TT5 and TT4.

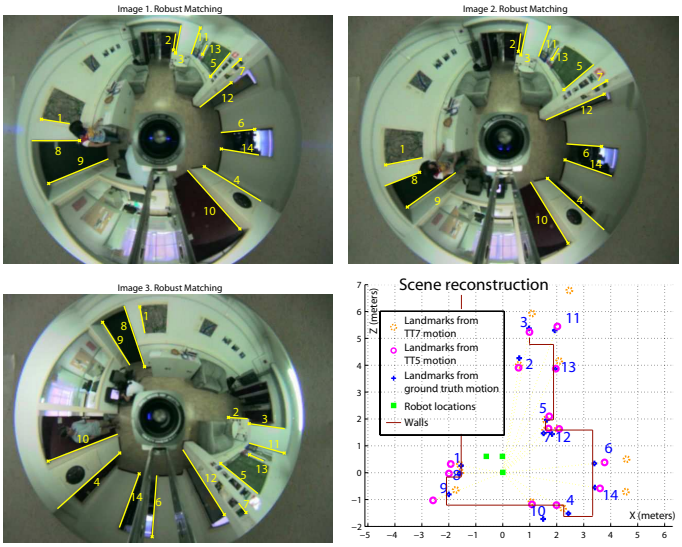
Therefore, from all the simulations we can extract two conclusions:

- TT5 is the best general method to estimate the 1D trifocal tensor, it gives better performance and is simpler than the commonly used TT7 method.

- When there are planes in the scene, using TT4 reduces the complexity and gives similar results to TT5. In the special case that the plane is dominant in the images, TT4 behaves better, particularly with higher values of noise. Computing the tensor through the homography has also an additional advantage, because the intermediate estimation of the homography can help to detect a singular situation. For example, if the whole image matches can be explained with a sole homography, then the scene is planar or the baseline is nearly zero. In this case, the localization results from the trifocal tensor are not good, as we have tested in [23].

B. Real data experiments

This section shows several experiments with automatic data association in different kinds of real images. In these experiments, the scale and the double solution ambiguity were solved using data from ground truth. In practice, they can be determined from additional information like the odometry, or from the relative location of two cameras if working with reference images from a database [24].



Basic matches: 20 (7 mismatches).

Robust matches with TT5: 14 (1 mismatch, '2')- with TT7: 17 (4 mismatches).

	Robot localization error				*Landmarks reconstruction error*	
	rotation		*transl. dir.*		*mean* (std)	*mean* (std)
	θ'	θ''	t'	t''	*coord-x*	*coord-z*
TT5	0.8^o	1.9^o	1.1^o	7^o	0.2m. (0.18)	0.16m. (0.15)
TT7	1.3^o	2.2^o	17^o	6^o	0.3m. (0.4)	0.28m. (0.33)

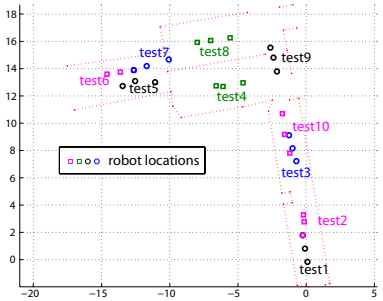
Fig. 5. *Experiment 1R-* TT7*vs.*TT5. Top: Omnidirectional images with robust line matches and scene reconstruction scheme, with landmark and camera locations obtained through the tensor estimations and ground truth. Bottom: Matching results and table with robot and landmark localization errors.

The following experiments show the behavior of the methods when the vertical placement of the camera is not perfect (the optical axis is not strictly vertical for omnidirectional cameras, or the image plane is not strictly vertical for conventional ones). There are not explicit experiments to measure this, but the images were taken with neither special care nor calibration, which probably explains part of the small performance reduction in real data experiments with regard to simulation tests.

1) Experiment 1R: This experiment compares the results obtained with TT7 and TT5 using omnidirectional images. The vertical lines (projected as radial ones in the image) are quite suitable for this kind of image. A very stable cue of these lines in omnidirectional images is their orientation, that is used as the 1D bearing-only data to estimate a 1D trifocal tensor.

Fig. 5 shows on the top an example of the three views with the line matching results obtained with TT5 and TT7. The initial *basic matching* was performed following the appearance-based algorithm explained in [24]. The *robust matching* results were better using TT5 than using TT7, which directly influenced the localization estimation, also better and more stable using TT5. For example in 30 executions, TT5 had standard deviations of 1.1^o and 3.5^o in rotation and translation direction errors respectively, while for TT7 these values were of 4^o and 12^o. The same Fig. 5 includes a scene reconstruction and a table with the localization parameters obtained in an execution where both TT7 and TT5 achieved an acceptable robust set of matches. Still in this case, where both methods performed properly the robust estimation, the localization accuracy obtained from TT5 was higher. The most important advantage of TT5 is that it gives a good robust set of matches, even in these experiments with few three-view correspondences.

2) Experiment 2R: Once we have checked that TT5 behaves better also in real cases, this experiment performs a more exhaustive test and shows an example using another kind of feature, SIFT, from where we take only the x-coordinate to obtain



Localization Errors

Rot.	mean (std)
θ'	0.56^o (0.49^o)
θ''	0.98^o (1.71^o)
Dir.	**mean (std)**
t'	3.79^o (3.62^o)
t''	4.20^o (3.83^o)

Fig. 6. *Experiment 2R*-Left: Location of the robot in the 10 tests performed, and contour map of the building. Right: average and standard deviation errors for rotations (θ',θ'') and translation direction (t',t'') estimation.

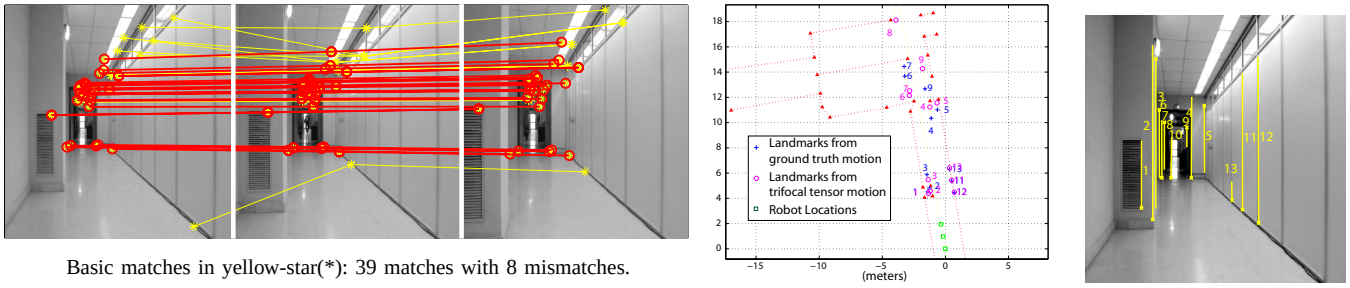
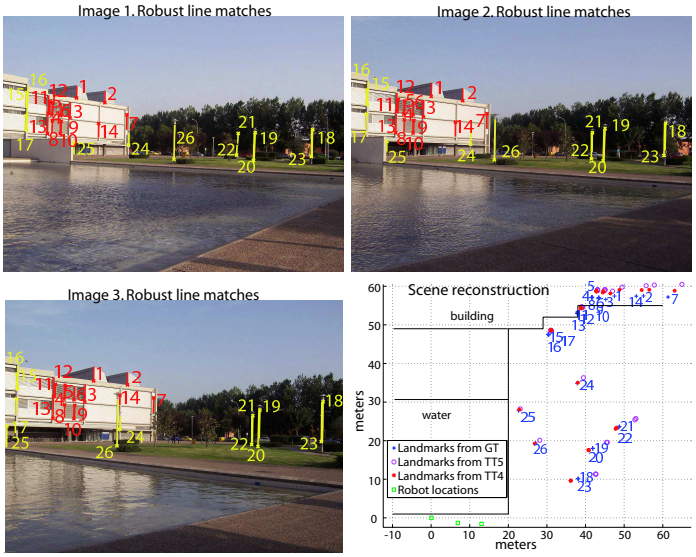


Fig. 7. *Experiment 2R* - Results for *test1* (see Fig. 6). Left: Images with basic and robust three-view SIFT matches. Right: landmarks reconstruction obtained from TT5 and from the ground truth of selected vertical features shown in image on the right.

the 1D bearing-only data. We use the original implementation of SIFT extraction and initial matching [15]. In this case the scenario was a sequential robot motion extracted from a robot indoors tour. We used the robot SLAM test-bed presented in [25]. These images were taken with a conventional camera (768x512 pixels) and a detailed ground truth (taken mostly with a theodolite) of the experiment was given together with all the necessary sensor calibration data. We chose 10 different triplets of images distributed all over the trajectory, shown in the left of Fig. 6. The table on the right of the same figure shows the average results in robot and landmark localization (for 50 different executions per case). Fig. 7 presents on the left a typical example (*test1*) of three-view SIFT matches obtained in this experiment, where all the final matches are correct. The tensor allows us to improve this initial matching by rejecting wrong SIFT correspondences, mostly due to pattern repetitions in the scene. The right part of the same Fig. 7 shows the comparison of the reconstruction obtained through the tensor (o) and with the ground truth motion (+). To measure the landmarks reconstruction with more accuracy, we chose several vertical lines. They were automatically extracted but manually selected, in order to show the most relevant scene landmarks.

3) Experiment 3R: This experiment intends to compare in a real situation the results obtained comparing TT5 and TT4. The bearing-only data used here is the x-coordinate of image vertical lines, which were extracted from conventional images of an outdoor scene. The *basic matching* of these line features was performed similarly to [21].



	Robot localization error				*Landmarks reconstruction error*		
	rotation		*transl. dir.*		*mean image*	*mean*	*mean*
	θ'	θ''	t'	t''	*repr. err* (std)	*2D x*	*2D z*
TT5	0.14^o	0.1^o	3.8^o	3.9^o	0.3 pix.(0.3)	0.7 m.	2.7 m.
TT4	0.17^o	0.5^o	2.9^o	3.4^o	0.6 pix.(0.8)	0.5 m.	1.5 m.

Fig. 8. *Experiment 3R*- Top: Outdoor real images with line robust matches (coplanar lines [1..14] marked in red) and scene reconstruction scheme with robot and landmark locations obtained through TT5 (in pink o), through TT4 (in red *) and from the ground truth motion (in blue +). Bottom: robot and landmark localization errors.

The line matches, the scene reconstruction and the robot and landmark localization errors are shown in Fig. 8. We observe a very similar behavior for both tensor estimation methods, with a slightly higher accuracy for TT4 in reconstructing the landmarks. The ground truth was obtained making a photogrammetrical bundle adjustment, using the software *Photomodeler* for the landmarks, and from the contour map of the building and the park. Also here, the advantages for the matching provided by the tensor are proved. It allows us to correctly match several features, which initially were mismatched due to pattern

repetitions (e.g. to another window-edge or lamppost) or due to different relative position (a lamppost is on the right or on the left of another lamppost depending on the image).

VI. Conclusions

The central issue of this work is the 1D trifocal tensor and its application to robotic tasks. This tensor imposes a general constraint for triplets of bearing-only observations which helps in automatic data association. Experiments with both conventional and omnidirectional images show that the trifocal tensor improves the quality of the matching. Robot and landmark localization can be obtained from this tensor as well. This paper explains and evaluates three ways of estimating the tensor with different minimum number of matches required. From the experimental work, we can conclude that computing a constrained tensor with 5 matches provides better performance for localization than using the typical solution based on 7 matches. It is simpler, less time consuming and gives smaller localization errors. Another important contribution of this work is the constrained method to compute the 1D trifocal tensor taking advantage of planes in the scene. It is estimated from only 4 matches, 3 of them collinear in the 2D scene. Our tests also indicate that the 4 matches method provides similar results to the 5 matches one in general cases, but if the plane is dominant in the image its results are slightly better. A fourth image or simple additional information can be used to solve for the scale and the double-solution ambiguity, which are inherent issues of the structure and motion problem from 1D view triplets. The experiments performed with conventional and omnidirectional images show sufficient accuracy and stability to be successfully used in robotic tasks such as navigation or initialization of bearing-only SLAM algorithms.

References

[1] I. Shimshoni, "On mobile robot localization from landmarks bearings," *IEEE Trans. on Robotics and Automation*, vol. 18, no. 6, pp. 971–976, 2002.

[2] A. Costa, G. Kantor, and H. Choset, "Bearing-only landmark initialization with unknown data association," in *IEEE Int. Conf. on Robotics and Automation*, 2004, pp. 1164–1770.

[3] O. Faugeras, Q.-T. Luong, and T. Papadopoulo, *The Geometry of Multiple Images*. Cambridge, MA, USA: MIT Press, 2001.

[4] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge: Cambridge University Press, 2000.

[5] Z. Zhang, R. Deriche, O. Faugeras, and Q. Luong, "A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry," *Artificial Intelligence*, vol. 78, pp. 87–119, 1995.

[6] P. Newman, D. Cole, and K. Ho, "Outdoor slam using visual appearance and laser ranging," in *IEEE Int. Conf. on Robotics and Automation*, 2006, pp. 1180–1187.

[7] K. Åström and M. Oskarsson, "Solutions and ambiguities of the structure and motion problem for 1d retinal vision," *Journal of Mathematical Imaging and Vision*, vol. 12, no. 2, pp. 121–135, 2000.

[8] O. Faugeras, L. Quan, and P. Sturm, "Self-calibration of a 1d projective camera and its application to the self-calibration of a 2d projective camera," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1179–1185, 2000.

[9] S. Thirthala and M. Pollefeys, "The radial trifocal tensor: A tool for calibrating the radial distortion of wide-angle cameras," in *Computer Vision Pattern Recognition*, 2005, pp. 321–328.

[10] F. Dellaert and A. Stroupe, "Linear 2d localization and mapping for single and multiple robots," in *IEEE Int. Conf. on Robotics and Automation*, 2002, pp. 688–694.

[11] C. Sagüés, A. C. Murillo, J. J. Guerrero, T. Goedemé, T. Tuytelaars, and L. V. Gool, "Localization with omnidirectional images using the 1d radial trifocal tensor," in *IEEE Int. Conf. on Robotics and Automation*, 2006, pp. 551–556.

[12] K. Imai, K. Tsuji, and M. Yachida, "Iconic memory-based omnidirectional route panorama navigation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 1, pp. 78–87, 2005.

[13] P. I. Corke, D. Strelow, and S. Singh, "Omnidirectional visual odometry for a planetary rover," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2004, pp. 4007–4012.

[14] T. Goedemé, T. Tuytelaars, G. Vanacker, M. Nuttin, and L. V. Gool, "Feature based omnidirectional sparse visual path following," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2005, pp. 1003–1008.

[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[16] R. Basri, E. Rivlin, and I. Shimshoni, "Visual homing: Surfing on the epipoles," in *IEEE Int. Conf. on Computer Vision*, 1998, pp. 863–869.

[17] E. Malis, F. Chaumette, and S. Boudet, "2 1/2 d visual servoing with respect to unknown objects through a new estimation scheme of camera displacement," *Int. Journal of Computer Vision*, vol. 37, no. 1, pp. 79–97, June 2000.

[18] A. Shashua and M. Werman, "Trilinearity of three perspective views and its associate tensor," in *Int. Conf. on Computer Vision*, 1995, pp. 920–925.

[19] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Comm. of the ACM*, vol. 24, pp. 381–395, 1981.

[20] G. Cross, A. W. Fitzgibbon, and A. Zisserman, "Parallax geometry of smooth surfaces in multiple views," in *Int. Conf. on Computer Vision*, 1999, pp. 323–329.

[21] J. J. Guerrero, R. Martinez-Cantin, and C. Sagüés, "Visual map-less navigation based on homographies," *Journal of Robotic Systems*, vol. 10, no. 22, pp. 569–581, 2005.

[22] P. Torr, A. Fitzgibbon, and A. Zisserman, "The problem of degeneracy in structure and motion recovery from uncalibrated image sequences," *Int. Journal of Computer Vision*, vol. 32, no. 1, pp. 27–44, 1999.

[23] A. C. Murillo, J. J. Guerrero, and C. Sagüés, "Robot and landmark localization using scene planes and the 1d trifocal tensor," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2006, pp. 2070–2075.

[24] A. C. Murillo, C. Sagüés, J. J. Guerrero, T. Goedemé, T. Tuytelaars, and L. V. Gool, "From omnidirectional images to hierarchical localization," *Robotics and Autonomous Systems*, vol. 55, pp. 372–382, 2007.

[25] J. A. Castellanos, J. M. Martínez, J. Neira, and J. D. Tardós, "Experiments in multisensor mobile robot localization and map building," in *3rd IFAC Symp. on Intelligent Autonomous Vehicles*, 1998, pp. 173–178.