A Critical Review of the Trifocal Tensor Estimation

Laura Julià, Pascal Monasse

A Critical Review of the Trifocal Tensor Estimation

Laura F. Julià (✉) and Pascal Monasse

LIGM (UMR 8049), École des Ponts, UPE, Champs-sur-Marne, France
laura.fernandez-julia@enpc.fr

Abstract. We explore the advantages offered by the trifocal tensor in the pose estimation of a triplet of cameras as opposed to computing the relative poses pair by pair with the fundamental matrix. Theoretically, the trinilearities characterize uniquely three corresponding image points in a tighter way than the three epipolar equations and this translates in an increasing accuracy. However, we show that this initial improvement is not enough to have a remarkable impact on the pose estimation after bundle adjustment, and the use of the fundamental matrix with image triplets remains relevant.

Keywords: Trifocal tensor, Fundamental matrix, Pose estimation

1 Introduction

The study of cameras and images has been a prominent subject since the beginning of computer vision, one of the main focus being the pose estimation and 3D reconstruction. Based on the perspective projection induced by pinhole cameras, there are constraints between the space points and their projections onto the images. Taking two images, the triangulation of the space points is possible from their projections when the poses are known. Eliminating 3D points from this model, the fundamental matrix is an algebraic operator encoding the relation between corresponding image points, which gives a way to infer the relative orientations and positions of a pair of camera viewpoints.

The natural extension is to consider three views and analyze the constraints between points to find a similar operator. The solution is the trifocal tensor; the algebraic constraints relating three corresponding image points are known as trilinearities. It was shown that a general multi-view matrix can be found for n views, but that the relations given by these n views depend only on the constraints involving two or three views at a time [6]. Theoretically, no extra geometric information about three views comes from considering additional views at once. Therefore, multi-view structure from motion pipelines always rely on initial view pairs [7,14,15] or triplets [4,8].

The conventional wisdom advocates the use of the trifocal tensor with a triplet of views rather than taking pairs and the fundamental matrix. We question this assumption with a study of the trifocal tensor and its performance

against the fundamental matrix. In Sect. 2 we present its definition and parameterizations and in Sect. 3 its estimation and pose estimation. The experiments to quantitatively measure its performance are in Sect. 4. We finally conclude in Sect. 5 that the advantages of the trifocal tensor are marginal and not sufficient to consider it superior to the fundamental matrix.

2 The Trifocal Tensor

Throughout the paper, the following notation is used: vectors are represented by lowercase (v), matrices by uppercase (M) and tensors by uppercase bold ($\mathbf{T}$). The 3×3 matrix form of the cross product on the left by a 3-vector v is denoted by $[v]_\times$, i.e., $[v]_\times w = v \times w$. For a vector v, we note $\|v\|$ its L^2 norm, and for a matrix or tensor the L^2 norm of the vector built from its coefficients. For a matrix M, $\|M\|$ is knwon as the Frobenius norm, and for a tensor $\mathbf{T}$ it represents, analogously, the square root of the sum of the squares of all its elements, $\|\mathbf{T}\| := \sqrt{\sum_{i,j,k}(T_i^{jk})^2}$. Finally, we note $|M|$ the determinant of a matrix M.

2.1 Definition

The **Trifocal Tensor** (TFT) associated to three views is a $3 \times 3 \times 3$ tensor $\mathbf{T} = [T_1, T_2, T_3]$ usually defined for three canonical projective cameras $P_1 = (Id_3|0)$, $P_2 = (A|a_4)$, $P_3 = (B|b_4)$ with each slice T_i the 3×3 matrix

$$T_i = a_i b_4^\top - a_4 b_i^\top, \tag{1}$$

where a_i and b_i are the columns of A and B. A more general definition for non canonical cameras can be found in [3].

The TFT has 27 parameters, is unique up-to-scale for any 3-view configuration and invariant by projectivity. Still, the degrees of freedom of a set of three projective cameras up-to-projectivity is 18 [3]. Hence, the parameters of the trifocal tensor must satisfy some constraints reducing the 8 remaining degrees of freedom of the trifocal tensor. However, the missing constraints are not obvious nor easily derivable. Section 2.3 presents several minimal parameterizations and constraints developed over the years.

2.2 Trilinearities

At its origin, the TFT is derived from the relation between the projections of the same 3D line in the three images. Other incidence relations can be found for this tensor, in particular, the following equation for triplets of corresponding image points x_1, x_2, x_3 (in homogeneous coordinates) is satisfied:

$$[x_2]_\times \Big(\sum_i (x_1)_i T_i \Big) [x_3]_\times = 0_{3\times 3} \ . \tag{2}$$

Among the 9 scalar equations in (2), only 4 are linearly independent. They are linear on the trifocal tensor parameters and trilinear on the image coordinates.

Considering the views pairwise, the incidence relations given by the fundamental matrices for the same corresponding triplet x_1, x_2, x_3 are a set of 3 equations linear on the fundamental matrices parameters and bilinear on the image points

$$x_2^\top F_{21} x_1 = 0, \qquad x_3^\top F_{31} x_1 = 0, \qquad x_3^\top F_{32} x_2 = 0 \ . \tag{3}$$

The involved fundamental matrices are

$$F_{21} = [a_4]_\times A, \qquad F_{31} = [b_4]_\times B, \qquad F_{32} = [b_4 - BA^{-1}a_4]_\times BA^{-1} \ . \tag{4}$$

2.3 Minimal Parameterizations and Constraints

Many possible minimal characterizations for the trifocal tensor have been proposed in the literature [1,2,10,11,12,13,17]. We chose to focus on four representative ones that can be efficiently implemented in the pose estimation process.

Ressl The minimal parameterization of the trifocal tensor proposed by Ressl in his thesis [13] is based on algebraic constraints of the correlation slices. It involves 20 parameters and 2 constraints. With this parameterization it is possible to completely characterize the trifocal tensor for three views. The three matrices of the trifocal tensor T_i can be parameterized in the following minimal form:

$$T_i = \left[s_i,\, v s_i + m_i e_{31},\, w s_i + n_i e_{31} \right]^\top \qquad i = 1, 2, 3 \tag{5}$$

where $s_i \in \mathbb{R}^3$ are such that $\left\| \begin{pmatrix} s_1 & s_2 & s_3 \end{pmatrix} \right\| = 1$, $e_{31} \in \mathbb{R}^3$ with $\|e_{31}\| = 1$, and $v, w, m_i, n_i \in \mathbb{R}$.

This parameterization is directly related to the epipoles since $e_{31} = b_4$ corresponds to the epipole, projection of the first camera center in the third image, and the epipole in the second image $e_{21} = a_4$ is proportional to $(1, v, w)^\top$. It is also related to an equivalent parameterization of three canonical projective matrices.

Nordberg The trifocal tensor can also be parameterized by three 3×3 orthogonal matrices U, V and W that transform the original tensor into a sparse one, $\tilde{\mathbf{T}}$, with only 10 non-zero parameters up-to-scale [10]:

$$\tilde{\mathbf{T}} = \mathbf{T}(U \otimes V \otimes W) \quad \Rightarrow \quad \mathbf{T} = \tilde{\mathbf{T}}(U^\top \otimes V^\top \otimes W^\top) \tag{6}$$

where the tensor operation corresponds to the matrix operation on the slices $\tilde{T}_i = V^\top (\sum_m U_{m,i} T_m) W$. The scale can be fixed by imposing $\|\tilde{\mathbf{T}}\| = 1$. For

canonical cameras, such orthogonal matrices can be computed as:

$$U_0 = (A^{-1}a_4,\ [A^{-1}a_4]_\times{}^2 B^{-1}b_4,\ [A^{-1}a_4]_\times B^{-1}b_4), \quad U = U_0(U_0^\top U_0)^{-\frac{1}{2}} \tag{7}$$

$$V_0 = (a_4,\ [a_4]_\times AB^{-1}b_4,\ [a_4]_\times{}^2 AB^{-1}b_4), \quad V = V_0(V_0^\top V_0)^{-\frac{1}{2}} \tag{8}$$

$$W_0 = (b_4,\ [b_4]_\times BA^{-1}a_4,\ [b_4]_\times{}^2 BA^{-1}a_4), \quad W = W_0(W_0^\top W_0)^{-\frac{1}{2}} \tag{9}$$

and each one can be parameterized by 3 parameters. Therefore, the trifocal tensor $\mathbf{T}$ is parameterized in this case by a total of 19 parameters and one constraint fixing the scale of $\tilde{\mathbf{T}}$.

A main disadvantage of this specific parameterization is that the matrices U_0, V_0 and W_0 become singular when the three camera centers are collinear and, therefore, no orthogonal matrix can be computed from them. It is then a parameterization only valid for non-collinear centers.

Faugeras and Papadopoulo In [2] a set of 12 algebraic equations are presented as sufficient constraints to characterize a trifocal tensor. It consists of 3 constraints of degree 3 corresponding to the determinant of the slices being zero, $|T_i| = 0$ for $i \in \{1, 2, 3\}$, and 9 more constraints of degree 6 combining several determinants of the elements of $\mathbf{T}$, for $j_1, j_2, k_1, k_2 \in \{1, 2, 3\}$ with $j_1 \neq j_2,\ k_1 \neq k_2$

$$\begin{aligned} &|t_{\cdot}^{j_1k_1}\ t_{\cdot}^{j_1k_2}\ t_{\cdot}^{j_2k_2}|\,|t_{\cdot}^{j_1k_1}\ t_{\cdot}^{j_2k_1}\ t_{\cdot}^{j_2k_2}| - \\ &|t_{\cdot}^{j_2k_1}\ t_{\cdot}^{j_1k_2}\ t_{\cdot}^{j_2k_2}|\,|t_{\cdot}^{j_1k_1}\ t_{\cdot}^{j_2k_2}\ t_{\cdot}^{j_1k_2}| = 0 \end{aligned} \tag{10}$$

where $t_{\cdot}^{jk}$ represents the vector $\left(T_1^{jk},\ T_2^{jk},\ T_3^{jk}\right)^\top$.

This set is not minimal since only 9 constraints should be enough for the characterization of a valid trifocal tensor. The authors give an outline of how to obtain a minimal parameterization using the constraints that requires to solve a polynomial of degree 2, thus giving two possible tensors. We considered best to use the minimization of the constraints instead of the minimal parameters for a more straightforward implementation.

Ponce and Hebert Π matrices A completely different approach to characterize the 3-view model has been explored in [12]. Through the study on the incidence of three lines on space, a set of three matrices (related to the principal lines) that give constraints on the correspondence of three image points can be defined. These matrices have a total of 27 parameters and play a role similar to the TFT. Given three cameras with non-collinear centers and three image points $x_1,\ x_2,\ x_3$ there exist three 4×3 matrices up-to-scale $\Pi_i = (\pi_{1i}, \pi_{2i}, \pi_{3i}, \pi_{4i})^\top$ with $\pi_{ii} = (0\ 0\ 0)^\top$ and verifying:

$$x_1^\top(\pi_{41}\pi_{32}^\top - \pi_{31}\pi_{42}^\top)x_2 = 0 \tag{11}$$

$$x_1^\top(\pi_{41}\pi_{23}^\top - \pi_{21}\pi_{43}^\top)x_3 = 0 \tag{12}$$

$$x_2^\top(\pi_{42}\pi_{13}^\top - \pi_{12}\pi_{43}^\top)x_3 = 0 \tag{13}$$

$$(\pi_{21}^\top x_1)(\pi_{32}^\top x_2)(\pi_{13}^\top x_3) = (\pi_{31}^\top x_1)(\pi_{12}^\top x_2)(\pi_{23}^\top x_3) \tag{14}$$

if, and only if, the x_i form a triplet of corresponding points. Ponce and Hebert propose the 6 homogeneous constraints:

$$\pi^1_{21} \;=\; \pi^2_{32} \;=\; \pi^3_{13} = 0, \quad \pi^2_{31} = \pi^3_{41}, \quad \pi^3_{12} = \pi^1_{42}, \quad \pi^1_{23} = \pi^2_{43} \tag{15}$$

that can be achieved by a projective transformation of the space. This reduces the parameters to 21 and with 3 norm constraints on the matrices, $\|\Pi_i\| = 1$, the minimal representation is attained.

Just like with the trilinearities (2) in the trifocal tensor case, these parameters give 4 equations describing the incidence relation for image points. Here, (11) to (13) are bilinear on the points and completely equivalent to the epipolar equations given by the fundamental matrices. Equation (14) is trilinear on the image points and it is key to the characterization of the correspondence of three points, which the fundamental matrices fail to achieve when one of the points lies on the line joining two epipoles. This is precisely the geometric contribution of taking three views instead of individual pairs to the characterization of matches.

Similarly to the parameterization of the trifocal tensor by Nordberg, the main drawback of the Π matrices is that they are only valid for non-collinear camera centers. For collinear camera centers, Ponce and Hebert[12] also proposed equivalent matrices with one extra trilinear constraint.

3 Pose Estimation

From a trifocal tensor $\mathbf{T}$ we can extract the epipoles, projections of the first camera center in the second and third images. The epipole e_{31} can be computed as the common intersection of the lines represented by the right null-vectors of T_1, T_2 and T_3. Analogously, the epipole e_{21} can be computed as the common intersection of the lines represented by the left null-vectors of T_1, T_2 and T_3. Then the fundamental matrices can be computed:

$$\begin{aligned} F_{21} &= [e_{21}]_\times [T_1 e_{31}, T_2 e_{31}, T_3 e_{31}] \ , \\ F_{31} &= [e_{31}]_\times [T_1^\top e_{21}, T_2^\top e_{21}, T_3^\top e_{21}] \ . \end{aligned} \tag{16}$$

From the fundamental matrices and the calibration matrices K_i, the essential matrices can be obtained as $[t_{ij}]_\times R_{ij} = E_{ij} = K_i^\top F_{ij} K_j$, from which the relative orientations (R_{21}, t_{21}) and (R_{31}, t_{31}) can be retrieved by the singular value decomposition of E_{21} and E_{31}, each translation vector being up to unknown scale. The overall scale is fixed by setting $\|t_{21}\| = 1$ and the relative scale λ of t_{31} can be computed by using a triangulation of the space points $\{X^n\}_n$ from the projections in the first two cameras and minimizing the algebraic error with respect to the third image:

$$\arg\min_{\lambda \in \mathbb{R}} \sum_{n=1}^{N} \left\| x_3^n \times \left(K_3 (R_{31} X^n + \lambda \frac{t_{31}}{\|t_{31}\|}) \right) \right\|^2 , \tag{17}$$

which admits a closed form solution.

So either from the trifocal tensor or the fundamental matrices, we have a way to compute the camera poses.

3.1 Linear Estimation of the Trifocal Tensor

The TFT can be estimated from a linear system given by the trilinearities of (2). From each triplet we get 9 equations linear on the parameters of the tensor, from which only 4 are linearly independent. At least 7 correspondences are needed to solve the linear system if we also impose $\|\mathbf{T}\| = 1$. If more triplets are available, a solution minimizing the algebraic error can be found by SVD. The resulting trifocal tensor will not necessarily be a valid tensor. To fix it, we can compute a valid trifocal tensor in the following way: extract the epipoles e_{21} and e_{31}, find matrices A and B that minimize (1) (resulting in linear systems), and finally compute a valid tensor.

Analogously, following the classical "normalized 8-point algorithm" specified in [3], the fundamental matrices can be computed linearly from the epipolar equations (3) and valid matrices can be found by imposing rank deficiency using SVD.

3.2 Optimization with Minimal Parameterization

Section 2.3 detailed four ways to parameterize minimally the 3-view model. All parameterizations involve non-linear constraints, so to be able to estimate the parameters an initialization is necessary. The linear solution from Sec. 3.1 can be used as an initial guess to estimate the different initial minimal parameterizations. Once the correct parameters of the initial model have been found they can be optimized by reinforcing the constraints and minimizing the Gold standard error (maximum likelihood estimator) with the Gauss-Helmert algorithm [9]. This algorithm finds a local optimum of the constrained least-square problem

$$\arg\min_{x,p} \|x - x_0\|^2 \qquad \text{s.t.} \qquad f(x,p) = 0, \qquad g(p) = 0 \tag{18}$$

by linearizing at each iteration the constraints f and g. The variables in vector x_0 are the observations, in p the parameters to optimize and in x the variables fitting the model, i.e., verifying f.

In the 3-view or 2-view models, the observations x_0 correspond to the matching image points and the main constraints f are the trilinearities and epipolar equations. In Table 1 the parameters and constraints to use for each minimal parameterization are summarized, as well as the ones to use to optimize a fundamental matrix.

3.3 Optimization with Bundle Adjustment

A common last step in pose estimation is a refinement of the orientations by Bundle Adjustment. It minimizes the square reprojection error over the possible cameras orientations and space points: For N correspondences and $M = 3$ cameras,

$$\min_{\{R_j,t_j\}_j,\{X^i\}_i} \epsilon^2 \qquad \epsilon^2 = \sum_{i=1}^{N}\sum_{j=1}^{M} d(x_j^i, K_j(R_j X^i + t_j))^2, \tag{19}$$

Table 1: Parameters and constraints to use in the Gauss-Helmert algorithm for the different minimal parameterizations of the 3-view model and the 2-view model.

parameterization	p	#	f	g	#
Ressl	s_i, m_i, n_i e_{31}, v, w	20	(2)	$\|(s_1, s_2, s_3)\| = 1$, $\|e_{31}\| = 1$	2
Nordberg	$\tilde{\mathbf{T}}, U, V, W$	19	(2)	$\|\tilde{\mathbf{T}}\| = 1$	1
Faug.-Papad.	$\mathbf{T}$	27	(2)	$\|T_i\| = 0$, (10)	12
Ponce-Hebert	Π_i	21	(11)–(14)	$\|\Pi_i\| = 1$	3
Fundamental	F_{21}	9	(3)	$\|F_{21}\| = 1$, $\|F_{21}\| = 0$	2

with x_j^i the homogeneous coordinates of the observed image point. The distance d is the Euclidean distance of points expressed in homogeneous coordinates:

$$d\left((x,y,z)^\top,(t,u,v)^\top\right)^2 = \left(\frac{x}{z}-\frac{t}{v}\right)^2 + \left(\frac{y}{z}-\frac{u}{v}\right)^2 \quad . \tag{20}$$

The optimization can be carried out by the Levenberg-Marquardt algorithm [5].

4 Experiments and Discussion

We implemented and evaluated the results of the pose estimation for synthetic and real data using the trifocal tensor and also using the fundamental matrix. [1] In the first case, we compute the tensor linearly (TFT-L) and applying a Gauss-Helmert optimization with the minimal parameterizations of Ressl (TFT-R), Nordberg (TFT-N), Faugeras and Papadopoulo (TFT-FP) and Ponce and Hebert (TFT-PH). For the fundamental matrix we compute it linearly (F-L) and with a Gauss-Helmert optimization (F-O). One last result is represented for the minimum found by the bundle adjustment (BA) initialized by any of the other methods. Indeed, we found that all the initializations gave the same final pose after the minimization in almost all our experiments, an important observation of our tests that we discuss later.

4.1 Synthetic Data

We tested the trifocal tensor and the fundamental matrix pose estimation on synthetic data for different configurations The standard scene for our experiments is composed of a set of space points contained in a cube of side 400mm centered at the world's origin (see Fig. 1). Points are projected onto three views and Gaussian noise is added to the image points with $\sigma = 1$ pixel, if not stated

[1] The MATLAB code to reproduce these experiments is available at the GitHub repository `https://github.com/LauraFJulia/TFT_vs_Fund.git`.

otherwise. A sample of 12 points is used for the computations of the different models. The image size is 1800×1200 pixels, corresponding to a 36mm×24mm sensor and the focal length is set to 50mm. The cameras all point at the origin. Results are averaged over 20 simulations of data.

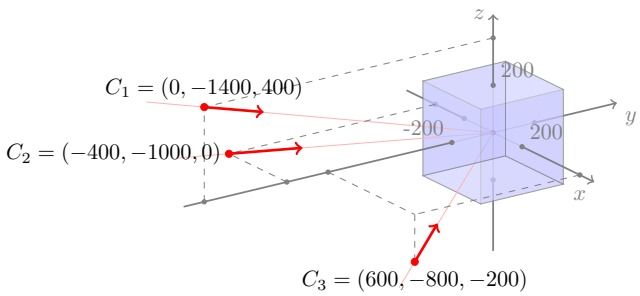


Fig. 1: Synthetic data.

The angular error in the estimated rotations and translation directions against Gaussian noise level added to the data points is shown in Fig.2. The experiments reveal that the pose estimation based on the trifocal tensor is consistently more accurate than the fundamental matrix pose estimation. All different methods optimizing the trifocal tensor with a minimal parameterization manage to improve the initial linear solution and end up in the same minimum. In the same way, the optimization of the fundamental matrix decreases the error of the linear solution. All these improvements, while clear, have no consequence on the minimum found by the bundle adjustment, which is reached even when initialized by the simplest method (F-L). Also in Fig. 2, a plot of the computational time spent on

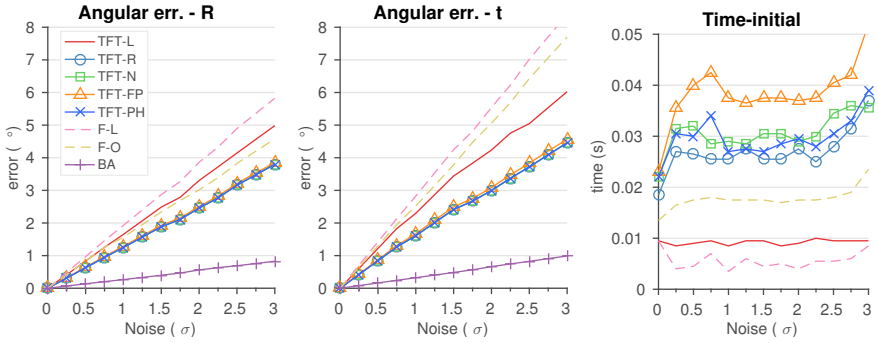


Fig. 2: Average errors for rotations (-R) on the left, for translation directions (-t) on the center, and computational time on the right, when varying the Gaussian noise added to the image points.

each initial estimation is shown.[2] As expected, linear methods (TFT-L, F-L) are faster than methods involving optimization, since the former are prerequisites for initialization of the latter. However, from the latter group, the fastest one is F-O, which involves two consecutive optimizations for two fundamental matrices.

Figure 3 tests the effect of changing the number of corresponding points used for the pose estimation. It shows how the fundamental matrix is much more affected by using a minimal set of correspondences than any trifocal tensor estimators but TFT-FP. The Faugeras-Papadopoulo minimal parameterization not only fails to improve the pose given by the linear estimation of the tensor for the minimal set of 7 correspondences but it returns a much worse estimation. For initial sets of more than 7 triplets, however, it performs as well as the other TFT methods. For sets with more than 15 triplets, all models start to stabilize. On the time plot in Fig. 3 we can see that linear methods maintain a constant computation time while optimization methods increase linearly with the number of initial points used.

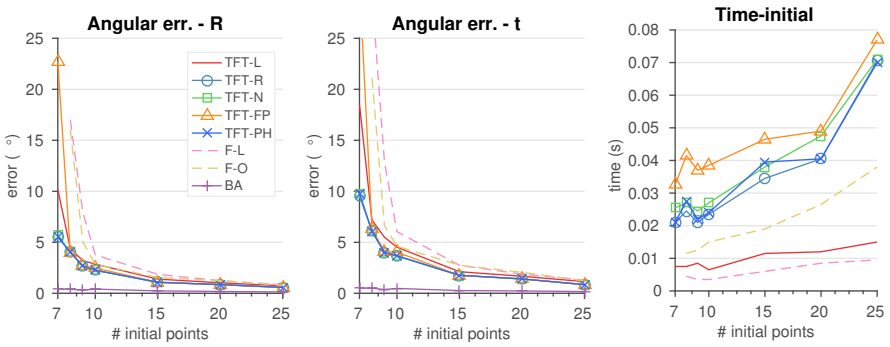
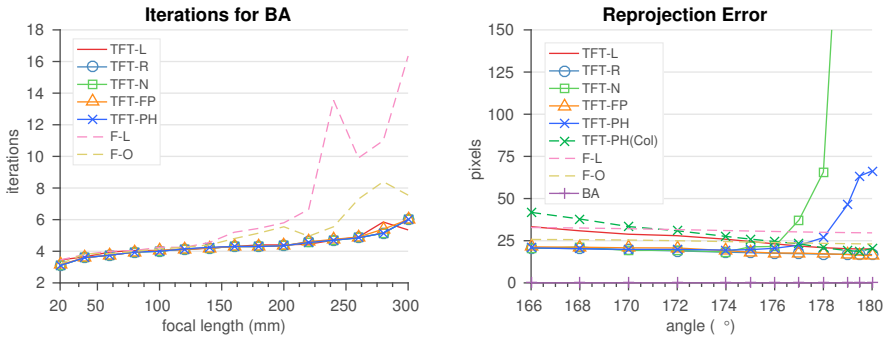


Fig. 3: Average angular errors for rotations (-R) on the left, for translation directions (-t) on the center, and computational time on the right, when the number of corresponding points is varied.

Long focal lengths are known to make difficult the camera pose estimation with the fundamental matrix. We studied the effect of increasing the focal length of our synthetic scene (while also proportionally getting the cameras farther away from the point cloud and from each other). Figure 4a shows that even if all methods get worse results in a similar way, the methods based on the fundamental matrix have an unstable higher increase of iterations for the bundle adjustment to converge after $f = 200$mm. Still, the final estimation remains the same, whatever the initialization method.

In all these experiments, all TFT-based methods generally give the exact same results, showing the equivalence of all parameterizations. However, there is

[2] based on the MATLAB code run on an Intel Xeon E5-2643 CPU at 3.3 GHz with 192 GB of RAM.



(a) (b)

Fig. 4: On the left (a), average number of iterations needed in bundle adjustment to reach a minimum for different focal lengths. On the right (b), average reprojection error of the pose estimation when making camera centers collinear.

a degenerate case specific for the Ponce-Hebert and Nordberg parameterizations: collinear camera centers. Alongside all the previously presented methods, we implemented and tested the collinear parameterization of the Π_i matrices given by Ponce and Hebert (TFT-PH(Col)). We tested all methods gradually moving the camera centers of the scene in order to make them align. The measure of collinearity is the angle $\widehat{C_2C_1C_3}$ (180° when collinear). Figure 4b shows the reprojection error, that is the error minimized in (19), with the estimated poses, for 100 points not used in the estimation. The results show an increasing accuracy on the collinear method, starting to be comparable to the others at 176°, the same point where the non-collinear parameterizations suffer a jump on the error, much greater for TFT-N than for TFT-PH. After 178° the initial poses given by the non-collinear parameterizations are no longer able to find the right minimum through bundle adjustment.

Based on these results and the instability of the Faugeras-Papadopoulo parameterization with a minimal set of initial points, the trifocal tensor parameterization of Ressl seems to be the most robust to degenerate scenes and the most recommended for pose estimation using the TFT.

4.2 Real Datasets

To evaluate the performance of these methods in real settings, we chose to use two scenes from the EPFL dense multi-view stereo test image datasets [16] that come with a reliable ground truth. These datasets consist of images of size 3072×2048 pixels taken with a 35mm equivalent focal length. The first scene is the fountain-P11 dataset which has 11 images. We tested 70 of the possible image triplets and the averages of the results are shown in Table 2. The second scene is the Herz-Jesu-P8 dataset which consists of 8 images, from which we tested 50 possible image triplets. The averages of the obtained errors are shown

in Table 3. For each triplet of images and method tested, the pose estimation is computed from a set of $N_{\text{init}} = 100$ triplets of correspondences chosen randomly from the total N inlier correspondences. The bundle adjustment optimization is carried out using a subset of $N_{\text{BA}} = 50$ correspondences from the initial set. The reprojection error, $\sqrt{\epsilon^2/(M\,N)}$ in (19), is evaluated on all N inliers.

Table 2: Average results over 70 triplets of images (one such triplet is shown) from the EPFL fountain-P11 dataset.

	repr. error (px)	R error (°)	t error (°)	init. time (s)	iter. BA
TFT-L	2.395	0.125	0.405	0.063	3.81
TFT-R	2.047	0.116	0.400	2.037	3.83
TFT-N	2.133	0.133	0.403	1.896	3.86
TFT-FP	2.365	0.119	0.403	2.063	3.84
TFT-PH	2.122	0.117	0.401	1.824	3.84
F-L	1.967	0.115	0.372	0.043	3.77
F-O	1.953	0.113	0.366	0.908	3.80
BA	0.281	0.064	0.074		

Table 3: Average results over 50 triplets of images from the EPFL Herz-Jesu-P8 dataset.

	repr. error (px)	R error (°)	t error (°)	init. time (s)	iter. BA
TFT-L	4.806	0.459	0.871	0.062	4.06
TFT-R	3.479	0.397	0.668	1.591	4.00
TFT-N	4.093	0.540	0.692	1.480	4.04
TFT-FP	4.506	0.446	0.833	1.887	4.06
TFT-PH	4.306	0.421	0.672	1.249	4.00
F-L	3.762	0.414	0.772	0.040	4.00
F-O	3.650	0.420	0.765	0.858	4.02
BA	0.372	0.063	0.068		

On the one hand, the results confirm that Ressl's parameterization is the most robust and better performing of all TFT-based methods getting the smallest error in all metrics. On the other hand, Nordberg's parameterization fails to improve the linear estimation since it gets a higher angular error in rotation. This might be due to the near-collinearity of some triplets (2 triplets in fountain-P11 and 4 in Herz-Jesu-P8 have a maximum angle between camera cen-

ters greater than 175°) which can cause great instability in the pose estimation of this methods as seen in the synthetic experiments (Fig. 4b).

We also notice how the fundamental-based methods get comparable results or even outperform the TFT-based methods in both datasets. What is more, they achieve it with less initial computation time and a similar average number of iterations to converge to the minimum in the bundle adjustment (two last columns of Tables 2 and 3).

In fact, all methods manage to reach the same minimum in the bundle adjustment optimization with around 4 iterations on average. The difference between the errors corresponding to the optimum reached and the errors from any method is much greater than the difference in the errors of the optimization-based methods and the linear methods. Therefore, one can conclude that the advantage of using an optimization to reinforce the constraints or minimal parameterization of the model before carrying out a bundle adjustment is negligible. The other lesson is that the bundle adjustment, even if performed with a small subsets of points for reduced computation time, is highly beneficial according to all error metrics.

Although not all known parameterizations of the trifocal tensor were covered by our tests, they all involve non-linear constraints admitting no closed form solution. As a consequence, they require also an initialization phase through the linear estimation of Sect. 3.1 and the possible initial benefits in terms of reduced error are likely to be erased by the bundle adjustment; the extra computation time would not make them advantageous alternatives to the standard fundamental matrix computation.

5 Conclusion

We reviewed methods of estimation of trifocal tensor and of the pose of three views. Compared with the pose estimation obtained by the fundamental matrices from the pairs of views, our experiments show that the trifocal tensor does not offer enough improvement to be considered the preferred choice. By its simplicity and lower computation time, the recommended option is to consider only pairwise constraints through the fundamental matrix, provided some bundle adjustment is used at the end (which is also highly recommended, as it can routinely decrease the error by a significant factor). In other words, the only usage of points viewed in image triplets, in the initialization phase of that approach, is to determine the relative scales of translations. Still, it would be interesting to study whether the use of the trifocal tensor improves results when $n > 3$ views are considered. However, in such a multi-view stereo pipeline, the way the image pairs and triplets are integrated is likely to have a preponderant importance. This research brought also another issue to our attention: observing that the bundle adjustment optimization is able to reach a correct minimum, even when starting from a far initial position, motivates us to study in future work the possible extended local convexity of the minimized energy.

References

1. Canterakis, N.: A Minimal Set of Constraints for the Trifocal Tensor, pp. 84–99. Springer Berlin Heidelberg, Berlin, Heidelberg (2000). doi: 10.1007/3-540-45054-8_6
2. Faugeras, O., Papadopoulo, T.: A nonlinear method for estimating the projective geometry of 3 views. In: Sixth International Conference on Computer Vision (IEEE), pp. 477-484 (1998). doi: 10.1109/ICCV.1998.710761
3. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, second edn. Cambridge University Press, ISBN: 0521540518 (2004)
4. Havlena, M., Torii, A., Knopp, J., Pajdla, T.: Randomized structure from motion based on atomic 3d models from camera triplets. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2874–2881 (2009). doi: 10.1109/CVPR.2009.5206677
5. Levenberg, K.: A method for the solution of certain non-linear problems in least squares. Quarterly of Applied Mathematics **2**(2), 164–168 (1944). `http://www.jstor.org/stable/43633451`
6. Ma, Y., Huang, K., Vidal, R., Košecká, J., Sastry, S.: Rank conditions on the multiple-view matrix. International Journal of Computer Vision **59**(2), 115–137 (2004). doi: 10.1023/B:VISI.0000022286.53224.3d
7. Moulon, P., Monasse, P., Marlet, R.: Adaptive structure from motion with a contrario model estimation. In: Asian Conference on Computer Vision, pp. 257–270. Springer (2012)
8. Moulon, P., Monasse, P., Marlet, R.: Global fusion of relative motions for robust, accurate and scalable structure from motion. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3248–3255 (2013)
9. Neitzel, F.: Generalization of total least-squares on example of unweighted and weighted 2D similarity transformation. Journal of Geodesy **84**(12), 751–762 (2010)
10. Nordberg, K.: A minimal parameterization of the trifocal tensor. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1224–1230 (2009). doi: 10.1109/CVPR.2009.5206829
11. Papadopoulo, T., Faugeras, O.: A new characterization of the trifocal tensor, pp. 109–123. Springer Berlin Heidelberg, Berlin, Heidelberg (1998). doi: 10.1007/BFb0055662
12. Ponce, J., Hebert, M.: Trinocular geometry revisited. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 17–24 (2014). doi: 10.1109/CVPR.2014.10
13. Ressl, C.: A minimal set of constraints and a minimal parameterization for the trifocal tensor. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XXXIV, Part 3A, ISPRS-Comm. III Symposium, Graz, 9, p. 13 (2002)
14. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
15. Snavely, N.: Bundler: Structure from motion for unordered image collections (2010). `https://www.cs.cornell.edu/~snavely/bundler/`
16. Strecha, C., von Hansen, W., Gool, L.V., Fua, P., Thoennessen, U.: On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008). doi: 10.1109/CVPR.2008.4587706
17. Torr, P., Zisserman, A.: Robust parameterization and computation of the trifocal tensor. Image and Vision Computing **15**, 591–605 (1997)