Comparative Analysis of Two-View and Three-View Pose Estimation Algorithms for Image-Based 3D Reconstruction: Fundamental Matrix vs Trifocal Tensor IACV 2023-2024 Project

Giovanni Versiglioni[1], Luca Magri[2], Federica Arrigoni[2], and Vincenzo Caglioti[2]

[1]**[Student]** giovanni.versiglioni@mail.polimi.it
[2]**[Supervisors]** luca.magri@polimi.it, federica.arrigoni@polimi.it, vincenzo.caglioti@polimi.it

Abstract

Image-based 3D reconstruction plays a pivotal role in various fields, from computer vision to augmented reality. Traditionally, algorithms for such reconstructions have predominantly relied on two-view reconstructions. However, the availability of three-view setups offers a unique opportunity to impose more stringent geometric constraints, potentially enhancing reconstruction accuracy. This research project aims to conduct a critical comparison between existing solvers for two-view and three-view pose estimation. By evaluating their performance in practical applications, this study seeks to ascertain whether it is more advantageous to exploit pairs or triplets of images for accurate scene reconstruction. Through comprehensive experimentation and analysis, this research aims to provide valuable insights into the effectiveness of different approaches, thereby aiding practitioners in selecting the most suitable algorithms for image-based 3D reconstruction tasks.

Keywords *Multiple-View Geometry, Pose Estimation, Fundamental Matrix, Trifocal Tensor*

Contents

1. Introduction

Since computer vision's inception, the study of cameras and images has been a central focus of the field's efforts. At its heart are complex processes like determining positions and reconstructing 3D shapes, both heavily dependent on understanding the complex relationship between points in space and how they appear in images, based on the principles of perspective projection in pinhole cameras. This understanding paves the way for triangulating spatial points from their corresponding image projections.

Embedded within this framework, the fundamental matrix emerges as a pivotal algebraic entity, encoding the essential correspondence between matching image points. It serves as a pathway for understanding the relative positions and orientations of pairs of camera viewpoints, essential for various applications in computer vision. Broadening this framework to include three perspectives introduces the concept of the trifocal tensor, a mathematical construct that encapsulates the algebraic constraints governing the relationships among three corresponding image points, often referred to as trilinearities. Although theoretically plausible to derive a multi-view matrix accommodating an arbitrary number of views, practical constraints predominantly stem from pairs or triplets of views. Consequently, most multi-view structure-from-motion pipelines pivot around initial view pairs or triplets for practical implementation.

Traditionally, the trifocal tensor has been preferred over the fundamental matrix when dealing with a triplet of views. We undertake the task of challenging this established preference by conducting an in-depth investigation into the comparative performance of the trifocal tensor versus that of the fundamental matrix.

In Section (3), we meticulously define and parameterize the trifocal tensor, delving into its intricacies. Subsequently, Section (4) explores the techniques employed for its estimation and subsequent pose determination. Our journey culminates in Section (5), where we present empirical findings quantifying the performance of both methodologies. These results, analyzed in Section (6), lead us to the conclusion that while the trifocal tensor does offer certain advantages, they are not substantial enough to unequivocally declare it superior to the fundamental matrix.

1.1. Notation

In this paper, we adopt specific notation conventions: vectors are denoted by lowercase (v), matrices by uppercase (M), tensors by calligraphic bold uppercase ($\boldsymbol{\mathcal{T}}$), and tensors' correlation slices (*i.e.*, matrices) by bold uppercase ($\boldsymbol{T}_i$).

The 3×3 matrix representation of the cross product with a 3-vector v is indicated by $[v]_\times w$, *i.e.*, $[v]_\times w = v\times w$, where w represents any given vector.

The L^2 norm of a vector v is denoted as $\|v\|$, while for matrices or tensors, it represents the L^2 norm of the vector constructed from their coefficients. The Frobenius norm of a matrix M is denoted as $\|M\|$, while for a tensor $\boldsymbol{\mathcal{T}}$, it signifies the square root of the sum of squares of all its elements, denoted as $\|\boldsymbol{\mathcal{T}}\| := \sqrt{\sum_{i,j,k}(\boldsymbol{T}_i^{jk})^2}$. Additionally, $|M|$ refers to the determinant of matrix M.

2. The Fundamental Matrix

In this section, we first introduce the definition of the fundamental matrix. We then proceed to describe numerical methods for estimating the fundamental matrix given a set of point correspondences between two images. Initially, we leverage linear equations derived from epipolar constraints to establish a foundational framework.

Subsequently, we explore Gauss-Helmert optimization, aiming to enhance precision and robustness in our analysis.

2.1. Definition

The **Fundamental Matrix (FM)** is defined by the equation

$$x'^{\top} F x = 0 \tag{2.1}$$

for any pair of matching points $x \leftrightarrow x'$ in two images.

2.2. Linear Computation

Given sufficiently many point matches (*i.e.*, at least 7), Equation (2.1) can be used to compute the unknown matrix F. In particular, each point match gives rise to one linear equation in the unknown entries of F. Specifically, the equation corresponding to a pair of points $(x, y, 1)$ and $(x', y', 1)$ is

$$x'x f_{11} + x'y f_{12} + x' f_{13} + y'x f_{21} + y'y f_{22} + y' f_{23} + x f_{31} + y f_{32} + f_{33} = 0. \tag{2.2}$$

From a set of n point matches, we derive the set of linear equations

$$Af = \begin{bmatrix} x'_1x_1 & x'_1y_1 & x'_1 & y'_1x_1 & y'_1y_1 & y'_1 & x_1 & y_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x'_nx_n & x'_ny_n & x'_n & y'_nx_n & y'_ny_n & y'_n & x_n & y_n & 1 \end{bmatrix} f = 0, \tag{2.3}$$

where f is the 9-vector made up of the entries of F in row-major order.

The 8-point algorithm stands as the most straightforward approach for computing the fundamental matrix. It entails constructing and solving a set of linear equations, typically through the method of least squares. The original algorithm is due to [5].

Algorithm 1: Normalized Eight Point Algorithm (Linear FM Computation)

Objective: Given $n \geq 8$ image point correspondences $\{x_i \leftrightarrow x'_i\}$, determine the fundamental matrix F such that ${x'}_i^{\top} F x_i = 0$.

Algorithm:

(i) **Normalization:** Transform the image coordinates according to $\hat{x}_i = Tx_i$ and $\hat{x}'_i = Tx'_i$, where T and T' are normalizing transformations consisting of a translation and a scaling.

(ii) Find the Fundamental Matrix $\hat{F}'$ corresponding to the matches $\{x_i \leftrightarrow x'_i\}$ by

  (a) **Linear solution:** Determine $\hat{F}$ from the singular vector corresponding to the smallest singular value of $\hat{A}$, where $\hat{A}$ is composed from the matches $\{x_i \leftrightarrow x'_i\}$ as defined in Equation (2.3).

  (b) **Constraint enforcement:** Replace $\hat{F}$ by $\hat{F}'$ such that $|\hat{F}'| = 0$ using the SVD.

(iii) **Denormalization:** Set $F = T'^{\top} \hat{F}' T$. Matrix F is the fundamental matrix corresponding to the original data $\{x_i \leftrightarrow x'_i\}$.

2.3. Optimization

3. The Trifocal Tensor

3.1. Definition

In this section, we explore the trifocal tensor's perspective through the incidence relationship among three corresponding lines. When a 3D line appears in three different views, it imposes constraints on the resulting image lines. These constraints are grounded in geometry: the back-projected planes from each view's lines must converge at a single line in 3D space, corresponding to the 3D line projected onto the matched lines in the images. This geometric condition imposes genuine constraints on sets of corresponding lines, which we then translate into algebraic form.

We examine a set of corresponding lines denoted as $l \leftrightarrow l' \leftrightarrow l''$, alongside canonical camera matrices for the three views: $P = [I|0]$, $P' = [A|a_4]$, and $P'' = [B|b_4]$, where A and B are 3×3 matrices, and a_i and b_i represent the columns of their respective camera matrices. The epipoles a_4 and b_4 in views two and three, derived from the first camera, are denoted as e' and e'', respectively, with $e' = P'C$ and $e'' = P''C$, where C is the firs camera center.

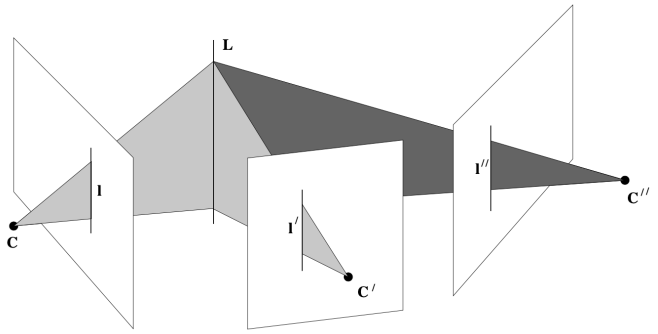

Figure 1. A line L in 3-space is imaged as the corresponding triplet $l \leftrightarrow l' \leftrightarrow l''$ in three views indicated by their centres, C, C' , C'' , and image planes. Conversely, corresponding lines back-projected from the first, second and third images all intersect in a single 3D line in space.

Considering projective transformations, we focus on properties such as image coordinates and 3D incidence relations, which remain invariant. Each image line is projected back to a plane, with these planes constrained to intersect at the common line in 3D space. This constraint is algebraically expressed by ensuring that a specific matrix $M = [\pi, \pi', \pi'']$ has a rank of 2. Here, π, π', and π'' represent the back-projected planes of the image lines in each view

$$\pi = P^\top l = \begin{pmatrix} l \\ 0 \end{pmatrix} \quad \pi' = P'^\top l' = \begin{pmatrix} A^\top l' \\ a_4^\top l' \end{pmatrix} \quad \pi'' = P''^\top l'' = \begin{pmatrix} B^\top l'' \\ b_4^\top l'' \end{pmatrix}. \tag{3.1}$$

The latter intersection constraint induces the following incidence relation amongst the image lines

$$l_i = l'^\top \boldsymbol{T}_i l'', \tag{3.2}$$

where $\boldsymbol{T}_i = a_i b_4^\top - a_4 b_i^\top$, $i = 1, 2, 3$. The set of three matrices $[\boldsymbol{T}_1, \boldsymbol{T}_2, \boldsymbol{T}_3]$ constitute the **Trifocal Tensor (TFT) in matrix notation**. Hence, the incidence relation (3.2) can be expressed as

$$l^\top = l'^\top [\boldsymbol{T}_1, \boldsymbol{T}_2, \boldsymbol{T}_3] l''. \tag{3.3}$$

3.2. Tensor Notation

Image points and lines are represented by homogeneous column and row 3-vectors, respectively. The ij-th entry of a matrix A is denoted by a_i^j, index i being the contravariant (row) index and j being the covariant (column) index. If the canonical 3×4 camera matrices are $P = [I|0]$, $P' = [a_j^i]$, and $P'' = [b_j^i]$, the definition of the **Trifocal Tensor (TFT) in tensor notation** becomes

$$\boldsymbol{\mathcal{T}}_i^{jk} = a_i^j b_4^k - a_4^j b_i^k. \tag{3.4}$$

The placement of indices in the tensor (two contravariant and one covariant) follows the arrangement of indices on the right side of the equation. Hence, the trifocal tensor is a mixed contravariant-covariant valency 3 tensor denoted by a homogeneous array of size $3 \times 3 \times 3$ (*i.e.*, 27 elements) and possesses 18 degrees of freedom.

Thus, the fundamental incidence relation (3.2) is expressed as

$$l_i = l'_j l''_k \boldsymbol{\mathcal{T}}_i^{jk}. \tag{3.5}$$

3.3. Trilinearities

Similarly to the fundamental matrix in two-view geometry, the trifocal tensor encodes relationships between points and lines across three perspectives. These relationships are denoted as trilinearities: "tri" since every monomial in the relation involves a coordinate from each of the three image elements involved, and linear because the relations are linear in each of the algebraic entities (*i.e.*, the three arguments of the tensor). The following equation portrays a point-point-point (P-P-P) correspondence

$$[x']_\times \left(\sum_i x^i \boldsymbol{T}_i \right) [x'']_\times = 0_{3\times 3} \tag{3.6}$$

with x, x', and x'' being the homogeneous coordinates of corresponding points in three images. However, other trilinear relations can be derived, such as L-L-L, P-L-L, P-L-P, P-P-L, and P-P-P, where P stands for point and L stands for line. These trilinearities are invariant under projective transformations, ensuring robustness across different camera configurations and scenes.

Among the nine scalar equations in (3.6), only four are linearly independent. They manifest linearity with respect to the parameters of the trifocal tensor and trilinearity with respect to the image coordinates. When viewed in pairs, the incidence relationships established by the fundamental matrices for the corresponding triplet x, x', and x'' consist of a group of three equations that are linear with respect to the parameters of the fundamental matrices and bilinear with respect to the image points

$$x_2^\top F_{21} x_1 = 0 \quad x_3^\top F_{31} x_1 = 0 \quad x_3^\top F_{32} x_2 = 0, \tag{3.7}$$

where the involved fundamental matrices are

$$F_{21} = [a_4]_\times A \quad F_{31} = [b_4]_\times B \quad F_{32} = [b_4 - BA^{-1}a_4]_\times BA^{-1}. \tag{3.8}$$

3.4. Linear Computation

The trifocal tensor can be derived from a linear system described by the trilinear relationships outlined in (3.3). Each triplet yields nine equations that are linear with respect to the tensor's parameters, yet only four of these equations are linearly independent. To solve this linear system, a minimum of seven correspondences is required, with the additional constraint $||\boldsymbol{\mathcal{T}}|| = 1$. If more triplets are available, a solution minimizing the algebraic error can be

obtained via Singular Value Decomposition (SVD). However, the resulting trifocal tensor may not always be valid.

To fix this, a valid trifocal tensor can be computed through an algebraic minimization algorithm that parallels the linear process employed to find the fundamental matrix.

Algorithm 2: Algebraic Minimization Algorithm (Linear TFT Computation)

Objective: Given a set of point and line correspondences in three views, compute the trifocal tensor.
Algorithm:

(i) From the set of point and line correspondences compute the set of equations of the form $At = 0$, where t is the 27-vector made up of the entries of the trifocal tensor.

(ii) Solve these equations using the least-squares solution to constrained systems, in order to find an initial estimate of the trifocal tensor $\boldsymbol{\mathcal{T}}_i^{jk}$.

(iii) Find the two epipoles e' and e'' from $\boldsymbol{\mathcal{T}}_i^{jk}$ as the common perpendicular to the left null-vectors of the three slices $\boldsymbol{T}_i$.

(iv) Construct the 27×18 matrix E such that $t = Ea$, where a is the vector representing entries of a_i^j and b_i^k, and where E espresses the linear relationship $\boldsymbol{\mathcal{T}}_i^{jk} = a_i^j e''^k - e'^j b_i^k$.

(v) Minimize $||AEa||$ subject to $||Ea|| = 1$, . Compute the error vector $\epsilon = AEa$.

(vi) **Iteration:** The mapping $(e', e'') \mapsto \epsilon$ is a mapping from $\mathbb{R}^6$ to $\mathbb{R}^2 7$. Iterate on the last two steps with varying e' and e'' using the Levenberg-Marquardt algorithm to find the optimal e', e'' pair. Hence find the optimal $t = Ea$ containing the entries $\boldsymbol{\mathcal{T}}_i^{jk}$.

3.5. Optimization

Several potential concise descriptions of the trifocal tensor have been suggested in prior literature [1, 2, 3, 6, 7, 8, 9, 11]. We've opted to concentrate on four representative ones that can be seamlessly integrated into the pose estimation procedure.

Ressl Ressl, in his thesis [9], introduced a minimal parameterization for the trifocal tensor, relying on algebraic constraints within correlation slices. This formulation consists of 20 parameters and 2 constraints. It enables the comprehensive characterization of the trifocal tensor for three views. The trifocal tensor, represented by the three matrices T_i, can be succinctly parameterized as follows

$$\boldsymbol{T}_i = [s_i, vs_i + m_i e_{31}, ws_i + n_i e_{31}]^\top, \tag{3.9}$$

where $s_i \in \mathbb{R}^3$ are such that $||(s_1 s_2 s_3)|| = 1$, $e_{31} \in \mathbb{R}$ with $||e_{31}|| = 1$, and $v, w, m_i, n_i \in \mathbb{R}$.

This parameterization directly links to the epipoles: where $e_{31} = b_4$ signifies the epipole, the projection of the first camera center onto the third image, and $e_{21} = a_4$ is proportionate to $(1, v, w)^\top$. Moreover, it's tied to an equivalent parameterization of three canonical projective matrices.

Nordberg Another approach to parameterize the trifocal tensor involves three 3×3 orthogonal matrices, U, V, and W, as mentioned in [6]. These matrices transform the original tensor into a sparse form, denoted as $\widetilde{\boldsymbol{\mathcal{T}}}$,

containing only 10 non-zero parameters, up to scale

$$\widetilde{\boldsymbol{\mathcal{T}}} = \boldsymbol{\mathcal{T}}(U \otimes V \otimes W), \tag{3.10}$$

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

For canonical cameras, such orthogonal matrices can be computed as

$$\begin{aligned}
U_0 &= (A^{-1}a_4, [A^{-1}a_4]_\times{}^2 B^{-1}b_4, [A^{-1}a_4]_\times B^{-1}b_4) \\
U &= U_0(U_0^\top U_0)^{-\frac{1}{2}} \\
V_0 &= (a_4, [a_4]_\times AB^{-1}b_4, [a_4]_\times{}^2 AB^{-1}b_4) \\
V &= V_0(V_0^\top V_0)^{-\frac{1}{2}} \\
W_0 &= (b_4, [b_4]_\times BA{-}1a_4, [b_4]_\times{}^2 BA^{-1}a_4) \\
W &= W_0(W_0^\top W_0)^{-\frac{1}{2}},
\end{aligned} \tag{3.11}$$

and each matrix can be represented by 3 parameters, resulting in a total of 19 parameters for the trifocal tensor $\boldsymbol{\mathcal{T}}$, along with one constraint to determine the scale of $\widetilde{\boldsymbol{\mathcal{T}}}$.

However, a notable drawback of this particular parameterization arises when the camera centers are collinear. In such cases, the matrices U_0, V_0, and W_0 become singular, rendering it impossible to compute orthogonal matrices from them. Consequently, this parameterization is only applicable when the camera centers are non-collinear.

Faugeras and Papadopoulo The work outlined in [7] introduces a set of 12 algebraic equations, serving as constraints to fully define a trifocal tensor. These include three constraints of third-degree, corresponding to the slices' determinants being zero, $|T_i| = 0$ for $i \in \{1, 2, 3\}$, and an additional nine constraints of sixth-degree. These sixth-degree constraints involve combinations of various determinants of the tensor's elements

$$\begin{aligned}
&|\boldsymbol{T}_{\cdot}^{j_1 k_1}\boldsymbol{T}_{\cdot}^{j_1 k_2}\boldsymbol{T}_{\cdot}^{j_2 k_2}||\boldsymbol{T}_{\cdot}^{j_1 k_1}\boldsymbol{T}_{\cdot}^{j_2 k_1}\boldsymbol{T}_{\cdot}^{j_2 k_2}|- \\
&|\boldsymbol{T}_{\cdot}^{j_2 k_1}\boldsymbol{T}_{\cdot}^{j_1 k_2}\boldsymbol{T}_{\cdot}^{j_2 k_2}||\boldsymbol{T}_{\cdot}^{j_1 k_1}\boldsymbol{T}_{\cdot}^{j_2 k_2}\boldsymbol{T}_{\cdot}^{j_1 k_2}| = 0,
\end{aligned} \tag{3.12}$$

where $j_1, j_2, k_1, k_2 \in \{1, 2, 3\}$ with $j_1 \neq j_2$ and $k_1 \neq k_2$, and where $\boldsymbol{T}_{\cdot}^{jk}$ represents the vector $(\boldsymbol{T}_1^{jk}, \boldsymbol{T}_2^{jk}, \boldsymbol{T}_3^{jk})^\top$.

This collection isn't minimal because it requires just 9 constraints to fully define a valid trifocal tensor. The authors suggest a method for achieving a minimal parameterization using these constraints, which involves solving a second-degree polynomial, resulting in two potential tensors. We find it preferable to pursue constraint minimization rather than minimal parameters for a simpler implementation.

Ponce and Hebert Π Matrices An alternative method of characterizing the 3-view model has been investigated in [8]. By analyzing how three lines intersect in space, a trio of matrices has been derived, each associated with principal lines. These matrices, comprising a total of 27 parameters, impose constraints on the correspondence between three image points. Analogous to the TFT, they serve a crucial role. For a configuration involving three cameras with non-collinear centers and three image points, denoted as x_1, x_2, and x_3, there exist three 4×3 matrices, denoted as $\Pi_i = (\pi_{1i}, \pi_{2i}, \pi_{3i}, \pi_{4i})^\top$, each scalable, where $\pi_{ii} = (000)^\top$, and they satisfy

$$x_1^\top(\pi_{41}\pi_{32}\top - \pi_{31}\pi_{42}\top)x_2 = 0 \tag{3.13}$$

$$x_1^\top(\pi_{41}\pi_{23}\top - \pi_{21}\pi_{43}\top)x_3 = 0 \tag{3.14}$$

$$x_3^\top(\pi_{43}\pi_{13}\top - \pi_{12}\pi_{43}\top)x_3 = 0 \tag{3.15}$$

$$(\pi_{21}^\top x_1)(\pi_{32}^\top x_2)(\pi_{13}^\top x_3) = (\pi_{31}^\top x_1)(\pi_{12}^\top x_2)(\pi_{23}^\top x_3), \tag{3.16}$$

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

$$\begin{gathered}\pi_{21}^1 = \pi_{32}^2 = \pi_{13}^3 =, 0 \\ \pi_{31}^2 = \pi_{41}^3, \quad \pi_{12}^3 = \pi_{42}^1, \quad \pi_{23}^1 = \pi_{43}^2.\end{gathered} \tag{3.17}$$

This can be accomplished through a projective transformation of the space, reducing the parameters to 21. By imposing three norm constraints on the matrices, $||\Pi_i|| = 1$, we achieve the most concise representation.

Analogous to the trilinearities (3.6) in the trifocal tensor scenario, these parameters yield four equations detailing the incidence relation for image points. Equations (3.13), (3.14), and (3.15) are bilinear regarding the points and are entirely analogous to the epipolar equations provided by the fundamental matrices. Equation (3.16), however, is trilinear concerning the image points, playing a pivotal role in characterizing the correspondence of three points that fundamental matrices falter at when one point resides on the line connecting two epipoles. This underscores the geometric significance of leveraging three views instead of individual pairs in characterizing matches.

Much like Nordberg's parameterization of the trifocal tensor, a primary limitation of the Π matrices is their applicability solely to non-collinear camera centers. In instances of collinear camera centers, Ponce and Hebert also proposed equivalent matrices incorporating an additional trilinear constraint.

4. Pose Estimation

We can derive the epipoles, which are the projections of the first camera center onto the second and third images, from a trifocal tensor $\boldsymbol{\mathcal{T}}$. The epipole e_{31} is determined as the shared point of intersection among the lines represented by the right null-vectors of $\boldsymbol{T}_1$, $\boldsymbol{T}_2$, and $\boldsymbol{T}_3$. Similarly, the epipole e_{21} is found as the common point of intersection among the lines represented by the left null-vectors of $\boldsymbol{T}_1$, $\boldsymbol{T}_2$, and $\boldsymbol{T}_3$. Subsequently, we can compute the fundamental matrices

$$\begin{aligned}F_{21} &= [e_{21}]_\times[\boldsymbol{T}_1 e_{31}, \boldsymbol{T}_2 e_{31}, \boldsymbol{T}_3 e_{31}] \\ F_{31} &= [e_{31}]_\times[\boldsymbol{T}_1^\top e_{21}, \boldsymbol{T}_2^\top e_{21}, \boldsymbol{T}_3^\top e_{21}].\end{aligned} \tag{4.1}$$

The essential matrices can be derived from the fundamental matrices and the calibration matrices K_i using the formula $[t_{ij}]_\times R_{ij} = E_{ij} = K_i^\top F_{ij} K_j$. From these essential matrices, the relative orientations (R_{21}, t_{21}) and (R_{31}, t_{31}) can be extracted through the singular value decomposition of E_{21} and E_{31}, with each translation vector's scale remaining unknown. To establish an overall scale, we set $\|t_{21} = 1\|$, while the relative scale λ of t_{31} can be determined by triangulating the space points $\{X^n\}_n$ from the first two cameras' projections and minimizing the algebraic error relative to the third image, as shown

$$\arg\min_{\lambda\in\mathbb{R}} \sum_{n=1}^{N} \left\| x_3^n \times \left(K_3 \left(R_{31} X^n + \lambda \frac{t_{31}}{\|t_{31}\|} \right) \right) \right\|. \tag{4.2}$$

The latter admits a closed form solution. So, either from the trifocal tensor or the fundamental matrices, we possess a method for computing the camera poses.

4.1. Bundle Adjustment

In pose estimation, a frequent final stage involves refining the orientations through Bundle Adjustment. This process aims to minimize the square reprojection error across potential camera orientations and spatial points. For N correspondences and M = 3 cameras

$$\min_{\{R_j,t_j\}_j,\{X^i\}_i} \epsilon^2, \quad \epsilon^2 = \sum_{i=1}^{N}\sum_{j=1}^{M} d\left(x_j^i, K_j(R_jX^i + t_j)\right)^2, \tag{4.3}$$

where x_j^i is for the homogeneous coordinates of the observed image point, and the distance d is the Euclidean distance in homogeneous coordinates

$$d\left((x,y,z)^\top,(t,u,v)^\top\right)^2 = \left(\frac{x}{z}-\frac{t}{v}\right)^2 + \left(\frac{y}{z}-\frac{u}{v}\right)^2 \tag{4.4}$$

The optimization procedure can be executed using the Levenberg-Marquardt algorithm [4].

5. Experiments

We put into action and assessed the outcomes of pose estimation for both synthetic and real data employing both the fundamental matrix and the trifocal tensor. [1] In the initial scenario, we employ linear computation for the tensor (**Linear TFT**) and apply Gauss-Helmert optimization using minimal parameterizations proposed by Ressl (**Ressl TFT**), Nordberg (**Nordberg TFT**), Faugeras and Papadopoulo (**Faugeras-Papadopoulo TFT**), and Ponce and Hebert (**Ponce-Hebert TFT**). As for the fundamental matrix, we compute it both linearly (**Linear FM**) and through Gauss-Helmert optimization (**Optimized FM**). Additionally, we present the result obtained through Bundle Adjustment (BA), which is initialized using any of the aforementioned methods. Remarkably, our experiments reveal that all initializations yield nearly identical final poses post-minimization in the majority of cases, an observation we delve into later in our discussions.

5.1. Synthetic Data

We conducted trials on synthetic data to assess pose estimation using both the fundamental matrix and the trifocal tensor across various configurations. The standard experimental setup consists of a collection of spatial points situated within a 400mm-sided cube centered at the world's origin (Figure X). Points are projected onto three views, and Gaussian noise with a standard deviation of 1 pixel is applied to the image points, unless specified otherwise. A set of 10 points is utilized for computations across various models. The image dimensions are 1800×1200 pixels, representing a sensor size of $36mm \times 24mm$, with a fixed focal length of 50mm. All cameras are aligned to focus on the origin. Results are averaged over 30 simulations of data.

Overall, the experiments consistently demonstrate that pose estimation derived from the trifocal tensor outperforms that based on the fundamental matrix. Regardless of the method used to optimize the trifocal tensor with minimal parameterization, each effectively enhances the initial linear solution, converging to the same minimum. Similarly, optimizing the fundamental matrix reduces the error of the linear solution. Despite these enhancements being evident, they do not affect the minimum attained through Bundle Adjustment, which is achieved even when initialized by the most basic method (*i.e.*, linear fundamental matrix estimation).

[1] The MATLAB code to reproduce these experiments is available at the GitHub repository: https://github.com/versi379/Two-View-Three-View-Pose-Estimation.git.

Metrics before and after Bundle Adjustment, against Gaussian noise level added to the data points, are shown respectively in Figure (2) and Figure (3).

Metrics before and after Bundle Adjustment, against the varying focal length, are shown respectively in Figure (4) and Figure (5).

Metrics before and after Bundle Adjustment, against the number of points considered in the synthetic scene, are shown respectively in Figure (6) and Figure (7).

Metrics before and after Bundle Adjustment, against the varying angle among the three camera centers, are shown respectively in Figure (8) and Figure (9).

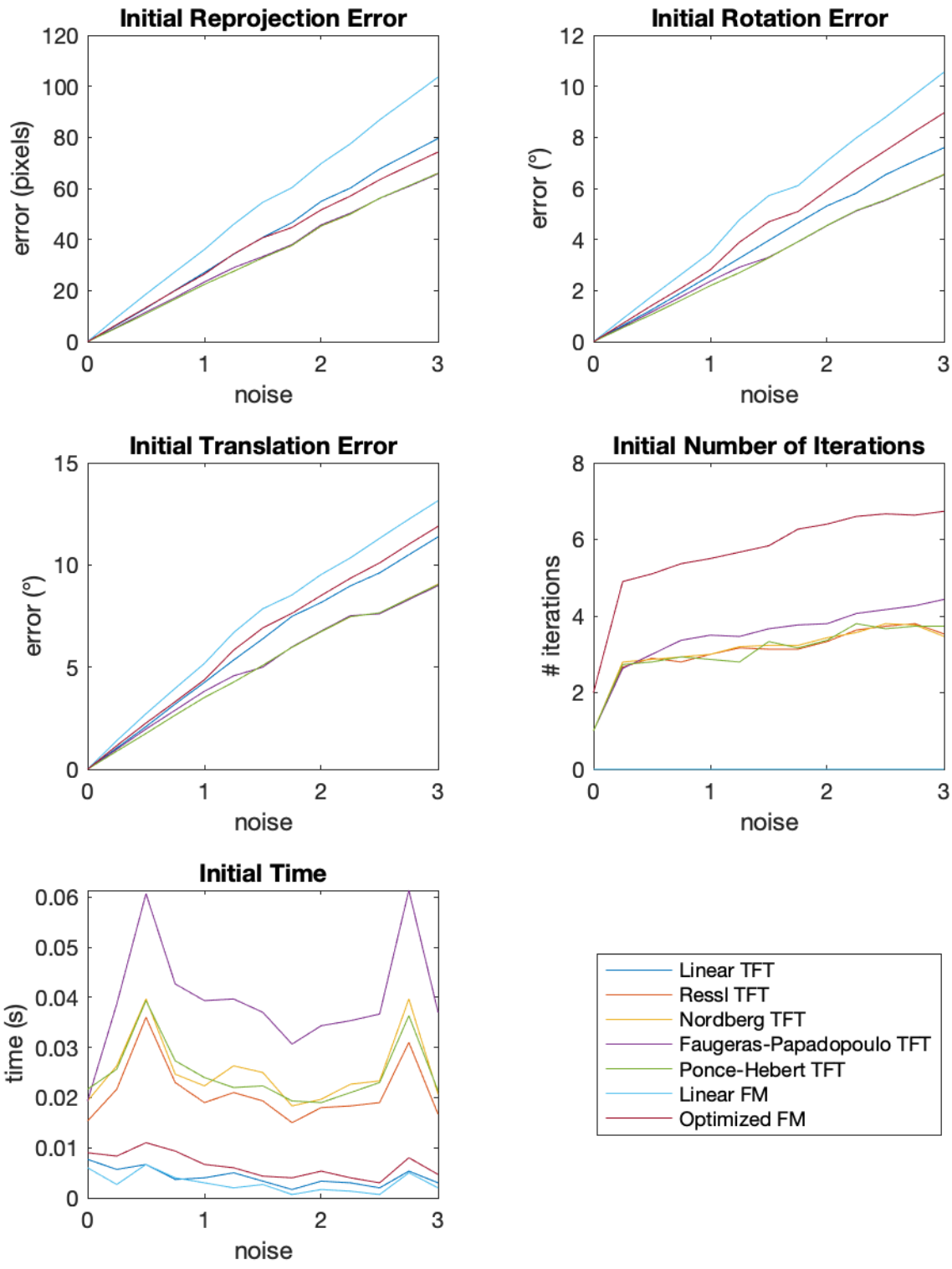

Figure 2. Initial reprojection error (top-left), rotation error (top-right), translation error (mid-left), number of iterations (mid-right), computational time (bottom-left); when varying the Gaussian noise added to the image points.

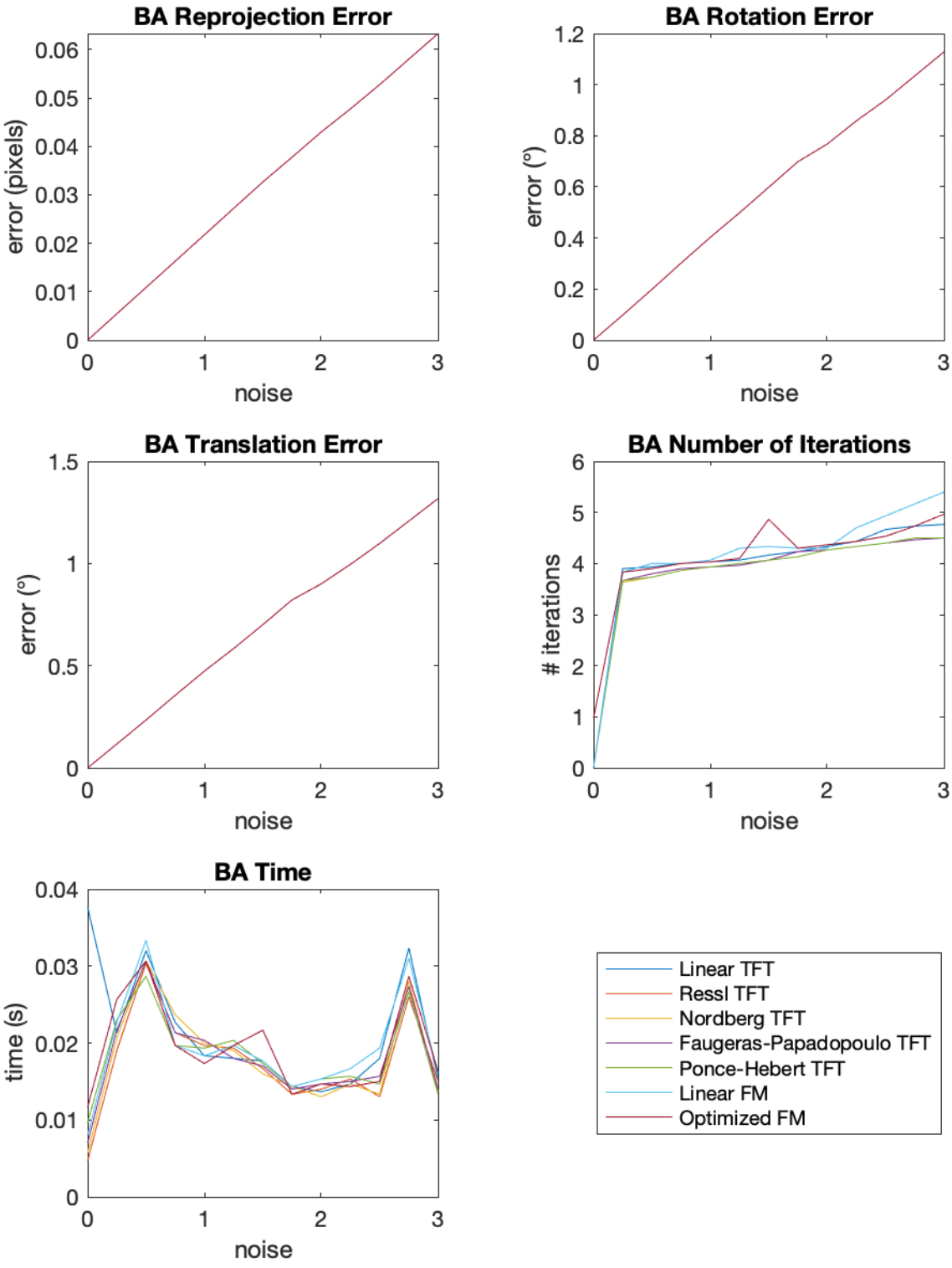

Figure 3. Reprojection error (top-left), rotation error (top-right), translation error (mid-left), number of iterations (mid-right), computational time (bottom-left) after Bundle Adjustment; when varying the focal length.

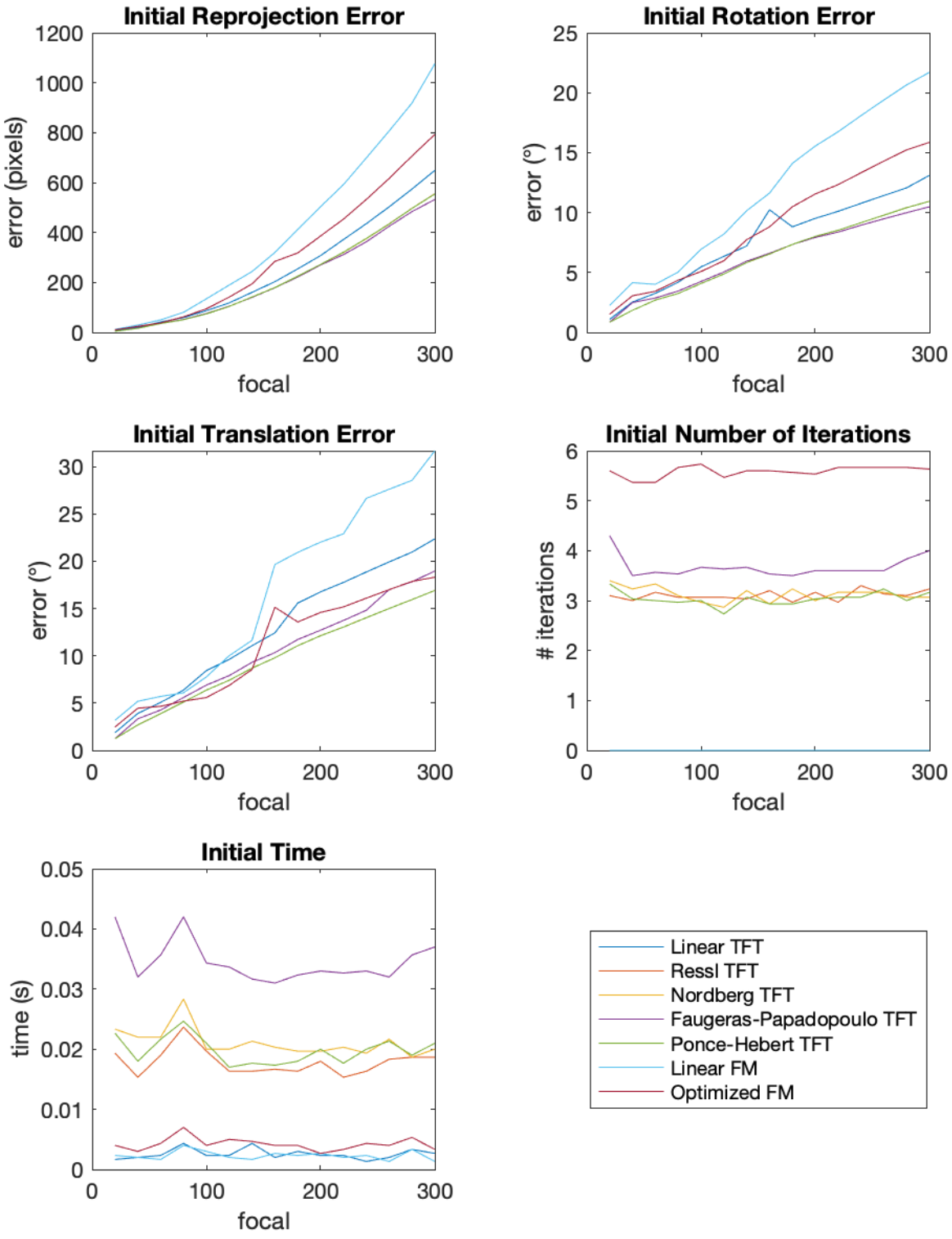

Figure 4. Initial reprojection error (top-left), rotation error (top-right), translation error (mid-left), number of iterations (mid-right), computational time (bottom-left); when varying the Gaussian noise added to the image points.

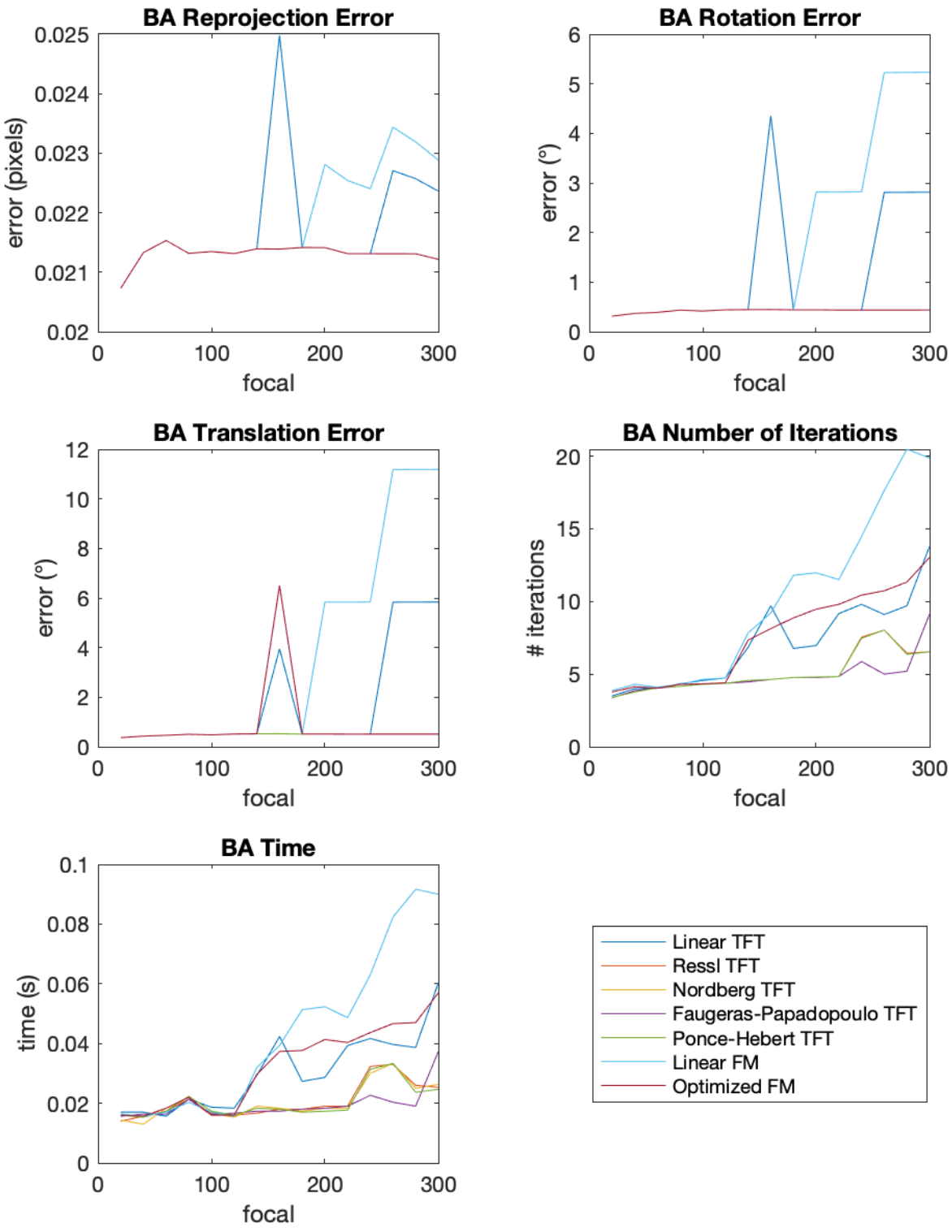

Figure 5. Reprojection error (top-left), rotation error (top-right), translation error (mid-left), number of iterations (mid-right), computational time (bottom-left) after Bundle Adjustment; when varying the focal length.

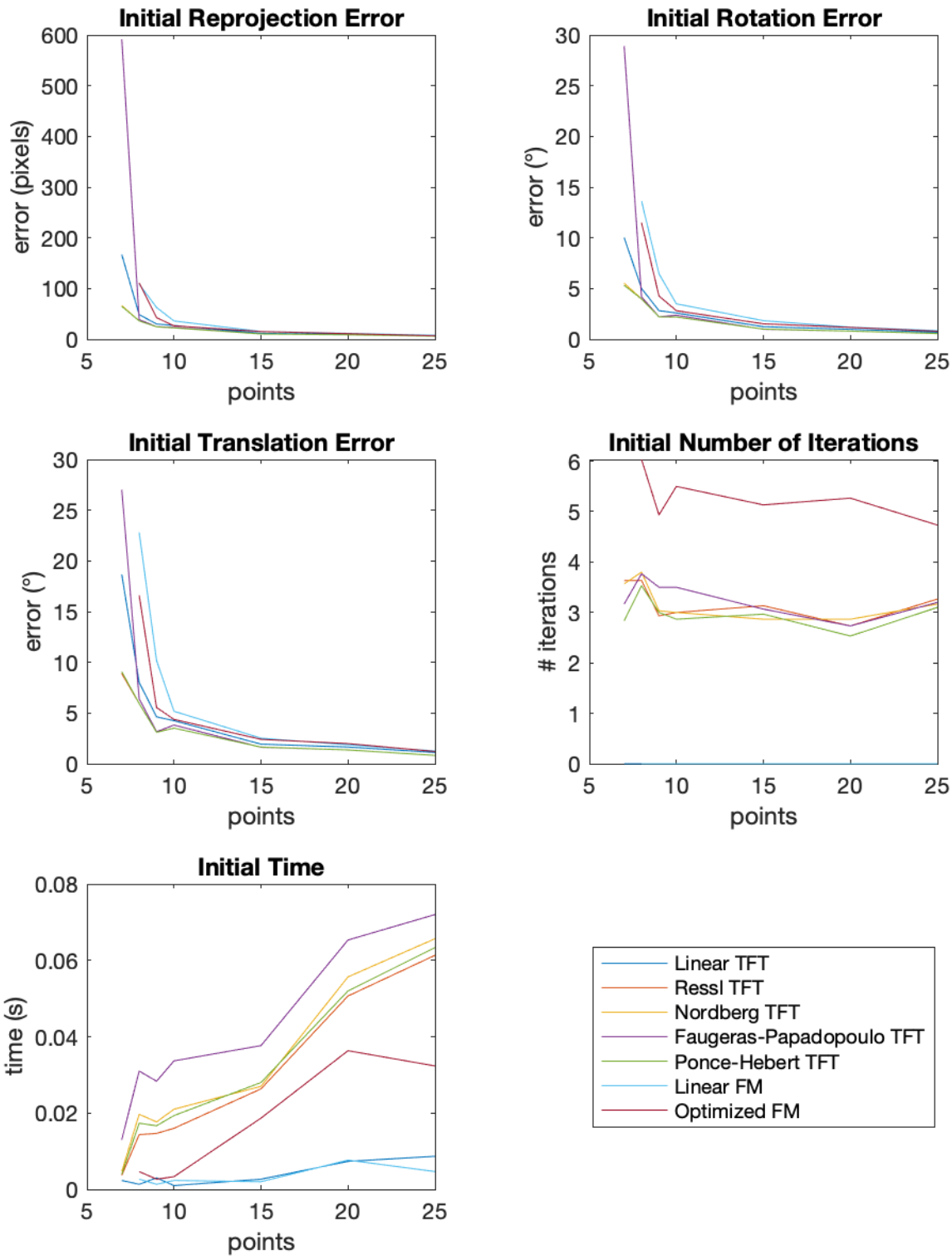

Figure 6. Initial reprojection error (top-left), rotation error (top-right), translation error (mid-left), number of iterations (mid-right), computational time (bottom-left); when varying the number of image points.

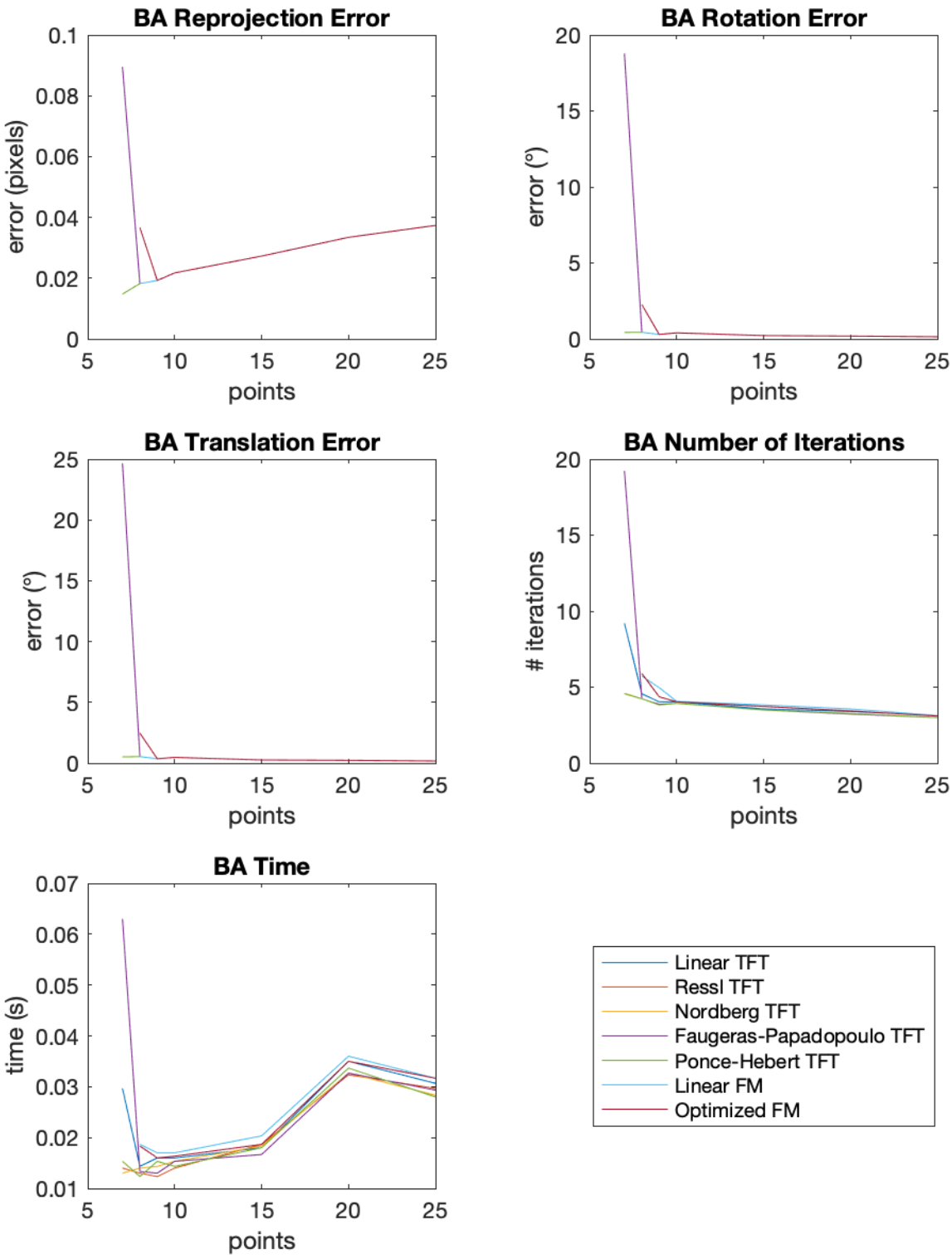

Figure 7. Reprojection error (top-left), rotation error (top-right), translation error (mid-left), number of iterations (mid-right), computational time (bottom-left) after Bundle Adjustment; when varying the number of image points.

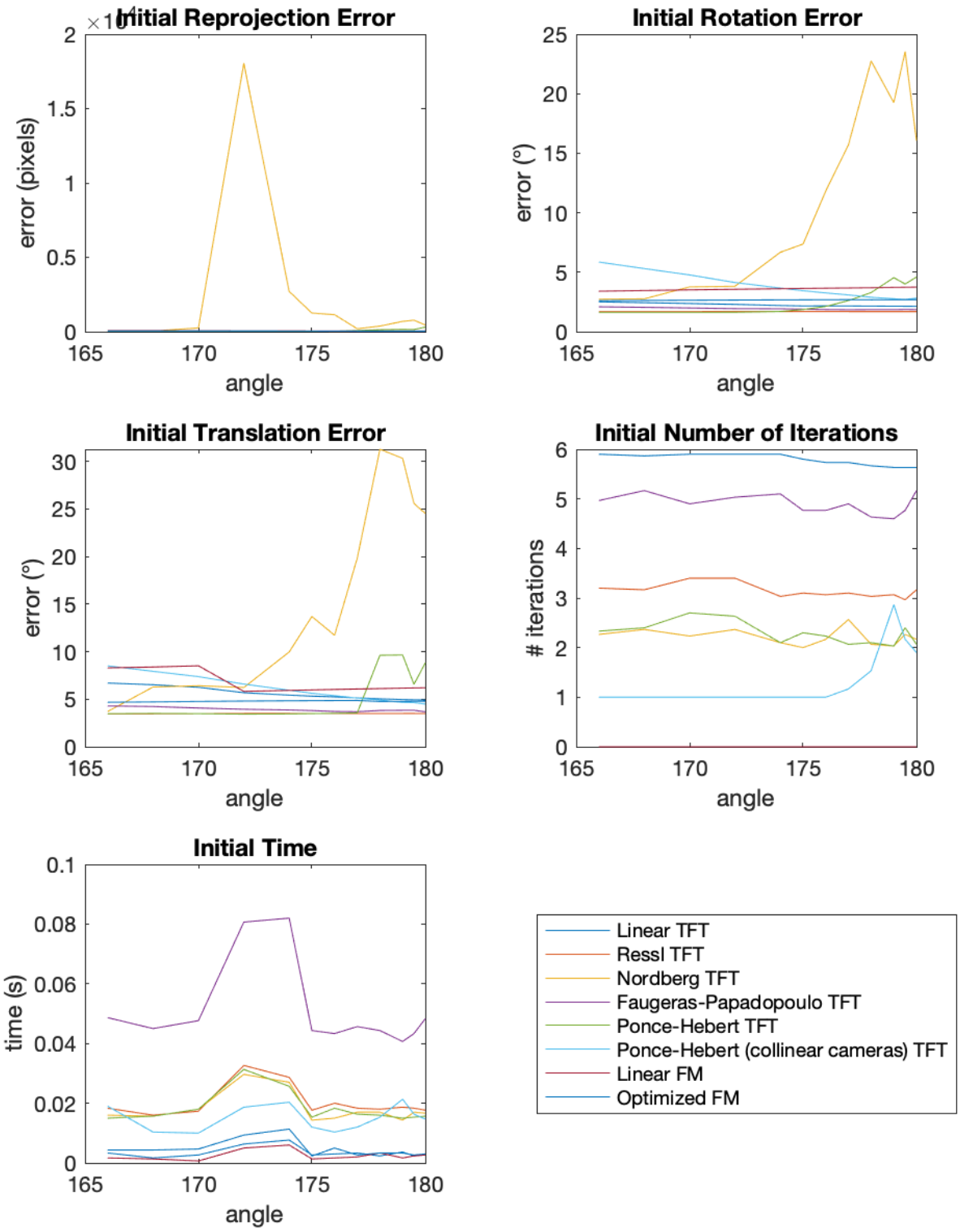

Figure 8. Initial reprojection error (top-left), rotation error (top-right), translation error (mid-left), number of iterations (mid-right), computational time (bottom-left); when varying the angle among the three camera centers.

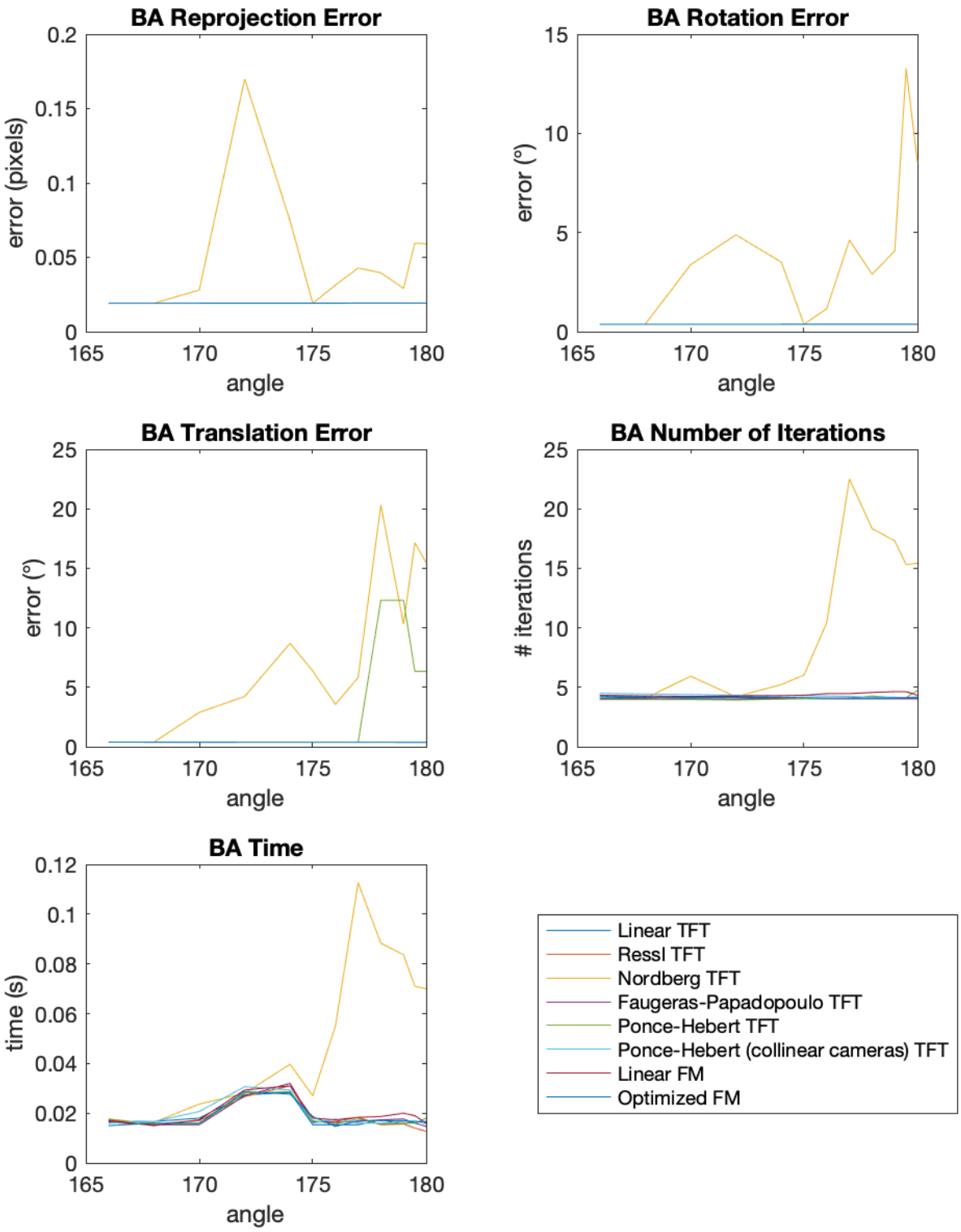

Figure 9. Reprojection error (top-left), rotation error (top-right), translation error (mid-left), number of iterations (mid-right), computational time (bottom-left) after Bundle Adjustment; when varying the angle among the three camera centers.

5.2. Real Data

In assessing the efficacy of these methods within real-world contexts, we opted to utilize scenes from the EPFL Dense Multi-View Stereo Dataset, presented in [10], provided by the CVLab at EPFL. [2]

Table (1) and (2) show metrics before and after Bundle Adjustment with respect to the *fountain-P11* set of images from the dataset.

Table 1. Initial metrics with respect to the *fountain-P11* set of images.

	repr. error (px)	R error (°)	t error (°)	# iter.	time (s)
Linear TFT	2.3953	0.1249	0.4048	0	0.0621
Ressl TFT	2.0474	0.1158	0.4003	2.8429	0.6400
Nordberg TFT	2.1322	0.1334	0.4028	2.8000	0.6280
Faugeras-Papadopoulo TFT	2.3688	0.1187	0.4055	2.7714	0.6073
Ponce-Hebert TFT	2.0871	0.1167	0.4030	2.5857	0.5554
Linear FM	1.9671	0.1149	0.3717	0	0.0273
Optimized FM	1.9530	0.1127	0.3658	4.9286	0.3209

Table 2. Metrics after Bundle Adjustment with respect to the *fountain-P11* set of images.

	repr. error (px)	R error (°)	t error (°)	# iter.	time (s)
Linear TFT	0.2814	0.0640	0.0743	3.8143	0.0743
Ressl TFT	0.2814	0.0640	0.0743	3.8286	0.0720
Nordberg TFT	0.2814	0.0640	0.0743	3.8571	0.0716
Faugeras-Papadopoulo TFT	0.2814	0.0640	0.0743	3.8429	0.0723
Ponce-Hebert TFT	0.2814	0.0640	0.0743	3.8429	0.0743
Linear FM	0.2814	0.0640	0.0743	3.7714	0.0816
Optimized FM	0.2814	0.0640	0.0743	3.8000	0.0784

Table (3) and (4) show metrics before and after Bundle Adjustment with respect to the *Herz-Jesu-P8* set of images from the dataset.

Table 3. Initial metrics with respect to the *Herz-Jesu-P8* set of images.

	repr. error (px)	R error (°)	t error (°)	# iter.	time (s)
Linear TFT	4.8062	0.4589	0.8707	0	0.0506
Ressl TFT	3.4792	0.3966	0.6677	2.7800	0.4904
Nordberg TFT	4.0656	0.5252	0.6917	2.6600	0.4816
Faugeras-Papadopoulo TFT	4.5006	0.4459	0.8324	3.4400	0.5452
Ponce-Hebert TFT	4.5293	0.4261	0.6682	2.3000	0.4116
Linear FM	3.7624	0.4142	0.7725	0	0.0224
Optimized FM	3.6503	0.4196	0.7654	5.6600	0.2906

[2]The EPFL Dense Multi-View Stereo Dataset, featuring the scenes utilized in our study, is readily accessible at the following location: https://documents.epfl.ch/groups/c/cv/cvlab-unit/www/data/multiview/denseMVS.html.

Table 4. Metrics after Bundle Adjustment with respect to the *Herz-Jesu-P8* set of images.

	repr. error (px)	R error (°)	t error (°)	# iter.	time (s)
Linear TFT	0.3719	0.0635	0.0682	4.0600	0.0792
Ressl TFT	0.3719	0.0635	0.0682	4.0000	0.0674
Nordberg TFT	0.3719	0.0635	0.0682	4.0400	0.0690
Faugeras-Papadopoulo TFT	0.3719	0.0635	0.0682	4.0600	0.0680
Ponce-Hebert TFT	0.3719	0.0635	0.0682	4.0000	0.0664
Linear FM	0.3719	0.0635	0.0682	4.0000	0.0718
Optimized FM	0.3719	0.0635	0.0682	4.0200	0.0724

6. Conclusions

In our investigation, we delved into various techniques for estimating the trifocal tensor and determining the pose of three distinct views. Upon rigorous experimentation, it became evident that while the trifocal tensor offers a methodological alternative, its advantages over pose estimation derived from fundamental matricesbased on pairs of viewsare not substantial enough to warrant its exclusive preference.

Our findings underscore the pragmatic appeal of simplicity and computational efficiency. We advocate for a strategy that emphasizes the utilization of pairwise constraints through fundamental matrices, supplemented by bundle adjustment procedures to refine results. It's worth noting that employing bundle adjustment consistently leads to a significant reduction in errors. Essentially, in this approach, the initial phase relies solely on establishing pairwise constraints to gauge the relative scales of translations, with image triplets serving primarily for this purpose.

6.1. Future Work

Nevertheless, an intriguing avenue for future exploration lies in investigating whether employing the trifocal tensor yields improved outcomes in scenarios involving more than three views (*i.e.*, $n > 3$). However, in such multi-view stereo pipelines, the manner in which image pairs and triplets are integrated is likely to wield considerable influence over the overall efficacy.

Furthermore, our research has brought to light an additional consideration: the robustness of bundle adjustment optimization. We observed that even when initialized from distant starting points, the optimization process can converge to a satisfactory minimum. This observation prompts further inquiry into the potential extended local convexity of the minimized energy landscape, which we intend to explore in future studies.

References

[1] N. Canterakis. A minimal set of constraints for the trifocal tensor. In *Computer Vision - ECCV 2000*, pages 84–99, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg. 6

[2] O. Faugeras and T. Papadopoulo. A nonlinear method for estimating the projective geometry of 3 views. pages 477 – 484, 02 1998. 6

[3] L. F. Julià and P. Monasse. A critical review of the trifocal tensor estimation. In M. Paul, C. H. Morimoto, and Q. Huang, editors, *Image and Video Technology - 8th Pacific-Rim Symposium, PSIVT 2017, Wuhan, China, November 20-24, 2017, Revised Selected Papers*, volume 10749 of *Lecture Notes in Computer Science*, pages 337–349. Springer, 2017. 6

[4] K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2(2):164–168, 1944. 9

[5] H. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. 293: 133–135, 1981. 3

[6] K. Nordberg. A minimal parameterization of the trifocal tensor. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1224–1230. IEEE, 2009. 6

[7] T. Papadopoulo and O. Faugeras. A new characterization of the trifocal tensor. In *Computer VisionECCV'98: 5th European Conference on Computer Vision Freiburg, Germany, June, 2–6, 1998 Proceedings, Volume I 5*, pages 109–123. Springer, 1998. 6, 7

[8] J. Ponce and M. Hebert. Trinocular geometry revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 17–24, 2014. 6, 7

[9] C. Ressl. A minimal set of constraints and a minimal parameterization for the trifocal tensor. *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences*, 34(3/A):277–282, 2002. 6

[10] C. Strecha, W. Von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. Ieee, 2008. 19

[11] P. H. Torr and A. Zisserman. Robust parameterization and computation of the trifocal tensor. *Image and vision Computing*, 15(8):591–605, 1997. 6

List of Figures

List of Tables