Trifocal Relative Pose From Lines at Points

To cite this version:
Ricardo Fabbri, Timothy Duff, Hongyi Fan, Margaret Regan, David da Costa de Pinho, et al.. Trifocal Relative Pose From Lines at Points. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45 (6), pp.7870-7884. 10.1109/TPAMI.2022.3226165. hal-04226115

HAL Id: hal-04226115
https://inria.hal.science/hal-04226115
Submitted on 5 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Abstract

We present a method for solving two minimal problems for relative camera pose estimation from three views, which are based on three view correspondences of (i) three points and one line and (ii) three points and two lines through two of the points. These problems are too difficult to be efficiently solved by the state of the art Gröbner basis methods. Our method is based on a new efficient homotopy continuation (HC) solver, which dramatically speeds up previous HC solving by specializing HC methods to generic cases of our problems. We characterize their number of solutions and show with simulated experiments that our solvers are numerically robust and stable under image noise. We show in real experiments that (i) SIFT feature location and orientation provide good enough point-and-line correspondences for three-view reconstruction and (ii) that we can solve difficult cases with too few or too noisy tentative matches where the state of the art structure from motion initialization fails.

1. Introduction

3D reconstruction has made an impact [4] by mostly relying on points in Structure from Motion (SfM) [1, 67, 23, 49]. Still, even production-quality SfM technology fails [4] when the images contain (i) large homogeneous areas with few or no features; (ii) repeated textures, like brick walls, giving rise to a large number of ambiguously correlated features; (iii) blurred areas, arising from moving cameras or objects; (iv) large scale changes where the overlap is not sufficiently significant; or (v) multiple and independently moving objects each lacking a sufficient number of features.

The failure of bifocal pose estimation using RANSAC on hypothesized correspondences, e.g., using 5 points [48], is highlighted in a dataset of images of mugs, Figure 1 (similar to the dataset in [51] but without a calibration board), for which the failure rate using the standard SfM pipeline COLMAP [63] is 75%. The failure of just directly apply-
ing the 5-point algorithm in this example is even higher. A similar situation exists for images containing repeated patterns where there are plenty of features, but determining correspondences is challenging. Most traditional multiview pipelines estimate the relative pose of the two best views and then register the remaining views using a P3P algorithm [68] to reduce the failure rate. The focus of this paper is to address the issue of failure of traditional bifocal algorithms in such cases.

The failure of bifocal algorithms motivates the use of (i) more complex features, i.e., having additional attributes and (ii) more diverse features. We propose that orientation (in the sense of inclination) is a key attribute to disambiguate correspondences and we show that SIFT orientation in particular is a stable feature across views for trifocal pose estimation. Orientation can also come from curve tangents [18, 17, 6], and the orientation of a straight line in multiple views also constrains pose. Observe, however, that orientation cannot be constrained in two views alone: SIFT orientation or line orientations in two views are uncorrelated, but together can identify their 3D counterparts and thus can constrain orientation in a third view. This motivates trifocal pose estimation based on point features endowed with orientation or including straight line features.

Camera estimation from trifocal tensors is long believed to augment two-view pose estimation [21], although a recent study suggests no significant improvements over bifocal pairwise estimation [31]. The calibrated trinocular relative pose estimation from four points, 3v4p, is notably difficult to solve [50, 59, 60, 17], and is not a minimal problem — it is over-constrained. The first working trinocular solver [50] effectively parametrizes the relative pose between two cameras as a curve of degree ten representing possible epipoles. A third view is then used to select the epipole that minimizes reprojection errors. In this sense, trinocular pose estimation has not truly been tackled as a minimal problem.

Trinocular pose estimation requires the determination of 11 degrees of freedom: six unknowns for each pair of rotation R and translation t, less one for metric ambiguity. Three types of constraints arise in matching triplets of point features endowed with orientation. First, the epipolar constraint provides an equation for each pair of correspondences in two views. Second, in a triplet of correspondences, each pair of correspondences are required to match scale, providing another constraint: a total of three equations per triplet. It is easy to see, informally, that three points are insufficient to determine trifocal pose, while four points are too many. Third, each triplet of oriented feature points provides one orientation constraint. Thus, with three points, only two points need to be endowed with orientation, giving a total of 11 constraints for the 11 unknowns. We refer to this problem of three triplets of corresponding points, with two of the points having oriented features as “Chicago.” In the second scenario, i.e., using straight lines as features, with three points, only one free (unattached to a point) straight line feature is required. We refer to the problem of three triplets of corresponding points and one triplet of corresponding free lines as “Cleveland.” This paper addresses trifocal pose estimation for the above two scenarios, shows that both are minimal problems, and develops efficient solvers for the resulting polynomial systems.

Specifically, each problem comprises eleven trifocal constraints that in principle give systems of eleven polynomials in eleven unknowns. These systems are not trivial to solve and require techniques from numerical algebraic geometry [9, 14, 41] (i) to probe whether the system is over or under constrained or otherwise minimal; (ii) to understand the range of the number of real solutions and estimate a tight upper bound; and (iii) to develop efficient and practically relevant methods for finding solutions which are real and represent camera configurations. This paper shows that the Chicago problem is minimal and has up to 312 solutions (the area code of Chicago is 312) of which typically 3-4 end up becoming relevant to camera configurations. Similarly, we show that the Cleveland problem is minimal and has up to 216 solutions. The minimality of combinations of points and lines for the general case [15] is a parallel development to the more concrete treatment presented here.

The numerical solution of polynomial systems with several hundred solutions is challenging. We devised a custom-optimized Homotopy Continuation (HC) procedure which iteratively tracks solutions with a guarantee of global convergence [14]. Our framework specializes the general HC approach to minimal problems typical of multiple view geometry, thereby dramatically speeding up the implementation. Specifically, our Chicago and Cleveland solvers are not only the first solvers for such high degree problems, but are orders of magnitude faster than solvers for such scale of problems: 660ms on average on an Intel core i7-7920HQ processor with four threads. They share the same generic core procedure with plenty of room to be further optimized for specific applications. Most significantly, since finding each solution is a completely independent integration path from the others, the solvers are suitable for implementation on a GPU, as a batch for RANSAC, which would then reduce the run time by the number of tracks, i.e., by two orders of magnitude. We hope that our developments can be a template for solving other computer vision problems involving systems of polynomials with a large number of solutions, and in fact the provided C++ framework is fully templated to include new minimal problems seamlessly.

It should be emphasized that trifocal pose estimation as a more expensive operation is not intended as a competitor of bifocal estimation algorithms. Rather, the trifocal approach can be considered as a fallback option in situations where bifocal pose estimation fails.
Experiments are initially reported on complex synthetic data to demonstrate that the system is robust and stable under spatial and orientation noise and under a significant level of outliers. Experiments on real data first demonstrates that SIFT orientation is a remarkably stable cue over a wide variation in view. We then show that our approach is successful in all cases where the traditional SfM pipeline succeeds, but of course at higher computational cost. What is critically important is that the proposed approach succeeds in many other cases where the SfM pipeline fails, e.g., on the EPFL [70] and Amsterdam Teahouse datasets [71], as shown in Figures 9 and 10. Those cases where the bifocal scheme fails – flagged by the number of inliers, for example – can consider the application of a currently more expensive but more capable trifocal scheme to allow for reconstructions that would otherwise be unsolved.

1.1. Literature Review

Trifocal Geometry Calibrated trifocal geometry estimation is a hard problem [50, 59, 60, 62]. There are no publicly available solvers we are aware of. The state of the art solver [50], based on four corresponding points (3v4p), has not yet found many practical applications [37].

For the uncalibrated case, 6 points are needed [26], and Larsson et al. recently solved the longstanding trifocal minimal problem using 9 lines [38]. The case of mixed points and lines is less common [53], but has seen a growing interest in related problems [63, 58, 72]. The calibrated cases beyond 3v4p are largely unsolved, spurring more sophisticated theoretical work [2, 3, 33, 40, 43, 44, 52]. Kileel [33] studied many minimal problems in this setting, such as the Cleveland problem solved in the present paper, and reported studies using homotopy continuation. Kileel also stated that the full set of ideal generators, i.e., a given set of polynomial equations provably necessary and sufficient to describe calibrated trifocal geometry, is currently unknown.

Seminal works used curves and edges in three views to transfer differential geometry for matching [5, 61], and for pose and trifocal tensor estimation [13, 66], beyond straight lines for uncalibrated [24, 7] and calibrated [64, 63] SfM. Point-tangents – not to be confused with point-rays [11] – can be framed as quivers (1-quivers), or feature points with attributed directions (e.g., corners), initially proposed in the context of uncalibrated trifocal geometry but de-emphasizing the connection to tangents to general curves [30, 74]. We note that point-tangent fields may also be framed as vector fields, so related technology may apply to surface-induced correspondence data [17]. In the calibrated setting, point-tangents were first used for absolute pose estimation by Fabbri et al. [18, 19], using only two points, later relaxed for unknown focal length [36]. The trifocal problem with three point-tangents as a local version of trifocal pose for global curves was first formulated by Fabbri [17], presented here as a minimal version codenamed Chicago.

Homotopy Continuation The basic theory of polynomial homotopy continuation (HC) [9, 46, 69] was developed in 1976, and guarantees algorithms that are globally convergent with probability one from given start solutions. A number of general-purpose HC softwares have considerably evolved over the past decade [8, 12, 41, 73]. The computer vision community has used HC most notably in the nineties for 3D vision of curves and surfaces for tasks such as computing 3D line drawings from surface intersections, finding the stable singularities of a 3D line drawing under projections, computing occluding contours, stable poses, hidden line removal by continuation from singularities, aspect graphs, self-calibration, and pose estimation [10, 22, 27, 28, 29, 34, 35, 42, 45, 54, 55, 57], as well as for MRFs [10, 47], and in more recent work [16, 25, 65]. An implementation of the early continuation solver of Kiefer and Ponce [34] by Pollefeys is still widely available for low degree systems [56].

As an early example [27], HC was used to find an early bound of 600 solutions to trifocal pose with 6 lines. In the vision community HC is mostly used as an offline tool to carry out studies of a problem before crafting a symbolic solver. Kasten et al. [32] recently compared a general purpose HC solver [73] against their symbolic solver. However, their problem is one order of magnitude lower degree than the ones presented here, and the HC technique chosen for our solver [14] is more specific than their use of polyhedral homotopy, in the sense that fewer paths are tracked (c.f. the start system hierarchy in [69]).

2. Two Trifocal Minimal Problems

2.1. Basic Equations

Our notation follows [24] with explicit projective scales. A more elaborate notation [13, 18] can be used to express the equations in terms of tangents to curves.

Figure 2. Notation for the trifocal pose problems.
Let $X$ and $Y$ denote inhomogeneous coordinates of 3D points and $x_{pv}, y_{pv} \in \mathbb{P}^2$ denote homogeneous coordinates of image points. Subscript $p$ numbers the points and $v$ numbers the views. If only a single subscript is used, it indexes the views. Symbols $R_v, t_v$ denote the rotation and translation transforming coordinates from camera 1 to camera $v$; $d$ is an image line direction or curve tangent in homogeneous coordinates; and $D$ is the 3D line direction or space curve tangent in inhomogeneous world coordinates. Symbols $\alpha, \beta$ denote the depth of $X, Y$, respectively, and $\gamma$ is the displacement along $D$ corresponding to the displacement $\gamma$ along $d$.

We next formulate two minimal problems for points in three views and derive their general equations before turning to specific formulations. We first state the new minimal problem, Chicago, followed by an important similar problem, Cleveland.

**Definition 1** (Chicago trifocal problem). Given three points $x_{1v}, x_{2v}, x_{3v}$ and two lines $\ell_{1v}, \ell_{2v}$, in views $v = 1, 2, 3$, such that $\ell_{pv}$ meet $x_{pv}$ for $p = 1, 2, v = 1, 2, 3$, compute $R_2, R_3, t_2, t_3$.

**Definition 2** (Cleveland trifocal problem). Given three points $x_{1v}, x_{2v}, x_{3v}$ in views $v = 1, 2, 3$, and given one line $\ell_{1v}$ in each image, compute $R_2, R_3, t_2, t_3$.

To setup equations, we start with image projections of points $\alpha_1 x_1 = X, \alpha_2 x_2 = R_2 X + t_2, \alpha_3 x_3 = R_3 X + t_3$ and eliminate $X$ to get

$$\alpha_v x_v = R_v \alpha_1 x_1 + t_v, \quad v = 2, 3. \tag{1}$$

Lines in space through $X$ are modeled by their points $Y = X + \eta D$ in direction $D$ from $X$. Points $Y$ are projected to images as $\beta_1 y_1 = X + \eta D, \beta_2 y_2 = R_2 (X + \eta D) + t_2, \beta_3 y_3 = R_3 (X + \eta D) + t_3$. Eliminating $X$ gives

$$\beta_1 y_1 = \alpha_1 x_1 + \eta D$$
$$\beta_2 y_2 = \alpha_2 x_2 + \eta R_2 D$$
$$\beta_3 y_3 = \alpha_3 x_3 + \eta R_3 D. \tag{2}$$

The directions $d_v$ of lines in images, which are obtained as the projection of $Y$ minus that of $X$, i.e.,

$$\beta_v \gamma_v d_v = y_v - x_v = \alpha_v x_v + \eta D - x_v, \tag{3}$$

are substituted to (2). After eliminating $D$ we get

$$(\beta_v - \alpha_v) x_v + \beta_v \gamma_v d_v = \ell_v ((\beta_1 - \alpha_1) x_1 + \beta_1 \gamma_1 d_1), \tag{4}$$

for $v = 2, 3$. To simplify notation further, we change variables as $\epsilon_v = \beta_v - \alpha_v, \mu_v = \beta_v \gamma_v$ and get

$$\epsilon_v x_v + \mu_v d_v = \ell_v (\epsilon_1 x_1 + \mu_1 d_1), \quad v = 2, 3. \tag{5}$$

For Chicago, we have three times the point equations (1) and two times the tangent equations (5). There are 12 unknowns $R_2, t_2, R_3, t_3$, and 24 unknowns $\alpha_{pv}, \epsilon_{pv}, \mu_{pv}$.

For Cleveland we need to represent a free 3D line $L$ in space. We write a general point of $L$ as $P + \lambda V$, with a point $P$ on $L$, the direction $V$ of $L$ and real $\lambda$. Considering a triplet of corresponding lines represented by their homogeneous coordinates $\ell_v$, the homogeneous coordinates of the back-projected planes are obtained as $\pi_v = [R_v | t_v]^T \ell_v$. Now, all $\pi_v$ have to contain $P$ and $V$ and thus

$$\text{rank } [(I | 0)^T \ell_1 | [R_2 | t_2]^T \ell_2 | [R_3 | t_3]^T \ell_3] < 3. \tag{6}$$

Equations 1 and 6 are the basic equations for Cleveland.

There are many ways to use elimination from these basic equations to obtain alternate formulations for these problems. A particular formulation based on vanishing minors for both Chicago and Cleveland, which produced our first working solver for Chicago, is described in 3.1.

**2.2. Problem Analysis**

A general camera pose problem is defined by a list of labeled features in each image, which are in correspondence. The image coordinates of each feature are given, and we aim to determine the relative poses of the cameras. The concatenated list of all the feature coordinates from all cameras is a point in the image space $Y$, while the concatenated list of the features’ locations in the world frame or camera 1 is a point in the world feature space $W$. Unless the scale of some feature is given, the scale of the relative translations is indeterminate, so relative translations are treated as in projective space. For $N$ cameras, the combined poses of cameras $2, \ldots, N$ relative to camera 1 are points in $SE(3)^{N-1}$. Let the pose space be $X$, the projectivized version of $SE(3)^{N-1}$, and so $X = 6N - 7$. Given the 3D features and the camera poses, we can compute the image coordinates of the features by considering a viewing map $V : W \times X \rightarrow Y$. A camera pose problem is: given $y \in Y$, find $(w, x) \in W \times X$ such that $V(w, x) = y$. The projection $\pi : (w, x) \mapsto x$ is the set of relative poses we seek.

**Definition 3.** A camera pose problem is minimal if $V : W \times X \rightarrow Y$ is invertible and nonsingular at a generic $y \in Y$.

A necessary condition for a map to be invertible and nonsingular is that the dimensions of its domain and range must be equal. Let us consider three kinds of features: a point, a point on a line (equivalently a point with tangent direction), and a free line (a line with no distinguished point on it). For each feature, say $F$, let $C_F$ be the number of cameras that see it. The contributions to the dimension $W$ and $X$ of each kind of feature are in the table below, where a point with a tangent counts as one point and one tangent. Thus, a point feature has several tangents if several lines intersect at it.
Accordingly, summing the contributions to $\dim Y - \dim W$ for all the features, we have the following result.

**Theorem 2.1.** Let $\langle x \rangle = \max(0, x)$. A necessary condition for a $N$-camera pose problem to be minimal is $6N - 7 = \sum_P \langle 2C_P - 3 \rangle + \sum_T \langle C_T - 2 \rangle + \sum_L \langle 2C_L - 4 \rangle$.

For trifocal problems where all cameras see all features, *i.e.*, $C_P = C_T = C_L = 3$, a pose problem with 3 feature points and 2 tangents meets the condition. A pose problem with 3 feature points and 1 free line also meets the condition. Adding any new features to these problems will make them overconstrained, having $\dim Y > \dim W \times X$.

To demonstrate sufficiency, it’s enough to find $(w, x) \in W \times X$ where the Jacobian of $V(w, x)$ is full rank. Such a rank test for a random point $(w, x)$ serves to establish non-singularity with probability one. Using floating point arithmetic this is highly indicative but not rigorous unless one bounds floating-point error, which can be done using interval or exact arithmetic. A singular value decomposition of the Jacobian using floating point showing that the Jacobian has a smallest singular value far from zero can be taken as a numerical demonstration that the problem is minimal. Similarly, a careful calculation using techniques from numerical algebraic geometry can compute a full solution list in $\mathbb{C}$ for a randomly selected example and thereby produce a numerical demonstration of the algebraic degree of the problem. Using such techniques, we make the following claims with the caveat that they have been demonstrated numerically.

**Theorem 2.2 (Numerical).** The Chicago trifocal problem is minimal with algebraic degree $312$, and the Cleveland problem is minimal with algebraic degree 216.

Proof. The previous paragraphs explain the numerical arguments, but the definite proof by computer involves symbolically computing the Gröbner basis over $\mathbb{Q}$, with special provisions, as discussed in the supplementary material. □

While this result is in agreement with degree counts for Cleveland in [33], the analysis of Chicago is novel as this problem is presented in this paper for the first time.

### 3. Homotopy Continuation Solver

In this section we describe our homotopy continuation solvers. In subsection 3.1 we reformulate the trifocal pose estimation problems as parametric polynomial systems in unknowns $R_2, R_3, t_2, t_3$ using the equations based on minors described in 3.1, while other formulations are discussed in supplementary material. We attribute relatively good run times to two factors. First, we use coefficient-parameter homotopy, outlined in 3.2, which naturally exploits the algebraic degree of the problem. Already with general-purpose software [8, 41], parameter homotopies are observed to solve the problems in a relatively efficient manner. Secondly, we optimize various aspects of the homotopy continuation routine, such as polynomial evaluation and numerical linear algebra. In subsection 3.3, we describe our optimized implementation in C++ which was used for the experiments described in section 4.

#### 3.1. Equations based on minors

One way of building a parametric homotopy continuation solver is to formulate the problems as follows. An instance of Chicago may be described by 5 visible lines in each view. We represent each line by its defining equation in homogeneous coordinates, *i.e.*, as $\ell_1, \ldots, \ell_5 \in \mathbb{C}^{3\times 1}$ for each $v \in \{1, 2, 3\}$. With the convention that the first three lines pass through the three pairs of points in each view and that the last two pass through associated point-tangent pairs, let

$$L_j = [I | 0]^T \ell_{j1} [R_2 | t_2]^T \ell_{j2} [R_3 | t_3]^T \ell_{j3},$$

for each $j \in \{1, \ldots, 5\}$. We enforce line correspondences by setting all $3 \times 3$ minors of each $L_j$ equal to zero. Certain common point constraints must also be satisfied, *i.e.*, that the $4 \times 4$ minors of matrices $[L_1 | L_2 | L_4], [L_2 | L_3 | L_5]$, and $[L_1 | L_3]$, all vanish.

We may describe the Cleveland problem with similar equations. For this problem, we are given lines $\ell_{1v}, \ldots, \ell_{4v}$ for $v \in \{1, 2, 3\}$. We enforce line correspondences for matrices $L_1, \ldots, L_4$ defined as in (7) and common point constraints by requiring that the $4 \times 4$ minors of $[L_1 | L_2], [L_1 | L_3], [L_2 | L_3]$ all vanish. The “visible lines” representation of both problems is depicted in Figure 3.

#### 3.2. Algorithm

From the previous section, we may define a specific system of polynomials $F(R; A)$ in the unknowns $R = (R_2, R_3, t_2, t_3)$ parametrized by $A = (\ell_{11}, \ldots)$. Many representations for rotations were explored, but our main implementation employs quaternions. A fundamental technique for solving such systems, fully described in [69], is coefficient-parameter homotopy. Algorithm 1 summarizes homotopy continuation from a known set of solutions for
given parameter values to compute a set of solutions for the desired parameter values. It assumes that solutions for some starting parameters $A_0$ have already been computed via some offline, ab initio phase. For our problems of interest, the number of start solutions is precisely the algebraic degree of the problem.

Several techniques exist for the ab initio solve. For example, one can use standard homotopy continuation to solve the system $F(R; A_0) = 0$, where $A_0$ are randomly generated start parameters [9, 69]. This method may be enhanced by exploiting additional structure in the equations or using regeneration. Another technique based on monodromy, described in [14], was used to obtain a set of starting solutions and parameters for the solver described in Section 3.3.

**Algorithm 1:** Homotopy continuation solution tracker

**Input:** Polynomial system $F(R; A)$, where $R = (R_x, R_z, t_x, t_z)$, and $A$ parametrizes the data;
Start parameters $A_0$; start solutions $R_0$ where $F(R_0; A_0) = 0$; Target parameters $A^*$

**Output:** Set of target solutions $R^*$ where $F(R^*; A^*) = 0$

Setup homotopy $H(R; s) = F(R; (1 - s)A_0 + sA^*)$.

For each start solution do

for $s \leftarrow 0$

while $s < 1$ do

Select step size $\Delta s \in (0, 1 - s]$.

**Predict:** Runge-Kutta Step from $s$ to $s + \Delta s$ such that $dH/ds = 0$.

**Correct:** Newton step st. $H(R; s + \Delta s) = 0$.

$s \leftarrow s + \Delta s$

return Computed solutions $R^*$ where $H(R^*, 1) = 0$.

### 3.3. Implementation

We provide an optimized open source C++ package called MINUS – MIrnimal problem NUmerical Solver\(^1\). This is a homotopy continuation code specialized for minimal problems, templated in C++, so that efficient specialization for different problems and different formulations are possible. The most reliable and high-quality solver according to our experiments uses a $14 \times 14$ minors-based formulation. Although other formulations have demonstrated further potential for speedup by orders of magnitude, there may be reliability tradeoffs (c.f. supplementary material).

### 4. Experiments

Experiments are conducted first for synthetic data for a detailed and controlled study, followed by experiment on challenging real data. Due to space constraints, we present results for the more challenging Chicago problem, leaving Cleveland for supplementary materials.

**Synthetic data experiments:** The synthetic data from [20, 18] consists of 3D curves in a $4 \times 4 \times 4cm^3$ volume projected to 100 cameras (Figure 4), and sampled to get 5117 points enclosed with orientations (tangents of curves) that are projections of the same 3D analytic points and 3D curve tangents [20], and then degraded with noise and outliers. Camera centers are randomly sampled around an average sphere around the scene along normally distributed radii of mean 1m and $\sigma = 10mm$. Rotations are constructed via normally distributed look-at directions with mean along the sphere radius looking to the object, and $\sigma = 0.01 rad$ such that the scene does not leave the viewport, followed by uniformly distributed roll. This sampling is filtered such that no two cameras are within $15^\circ$ of each other.

Our first experiment studies the numerical stability of the solvers. The dataset provides true point correspondences, which inherit an orientation from the tangent to the analytic curve. For each sample set, three triplets of point correspondences are randomly selected with two endowed with the orientation of the tangent to the curve. The real solutions are selected from among the output, and only those that generate positive depth are retained. The unused tangent of the third triplet is used to verify the solution as it is an overconstrained problem. For each of the remaining solutions a pose is determined.

The error in pose estimation is compared with ground-truth as the angular error between normalized translation vectors and the angular error between the quaternions. The process of generating the input to pose computation is repeated 1000 times and averaged. This experiment demonstrates that: (i) pose estimation errors are negligible, Figure 5(a); (ii) the number of real solutions is small: 35 real solutions on average, pruned down to 7 on average by enforcing positive depth, and even further to about 3-4 physically realizable solutions on average employing the unused tangent of the third triplet as verification, Figure 5(b); (iii) the solver fails in about 1% of cases, which are detectable and, while not a problem for RANSAC, can be eliminated by running the solver for that solution path with higher accuracy or more parameters at a higher computational cost.

The second experiment shows that we can reliably and accurately determine camera pose with correct but noisy correspondences. Using the same dataset and a subset of the selection of three triplets of points and tangents – 200 in total – zero-mean Gaussian noise was added both to the feature locations with $\sigma \in \{0.25, 0.5, 0.75, 1.0\}$ pixels and to the orientation of the tangents with $\sigma \in \{0.05, 0.1, 0.15, 0.2\}$ radians, reflecting expected feature localization and orientation localization error. A RANSAC scheme determines the feature set that generates the highest number of inliers. Experiments indicate that the translation and rotation errors are reasonable. Figure 6 (top) shows how the extent of localization error affects pose (in terms

\(^1\)Code available at [http://github.com/rfabbri/minus](http://github.com/rfabbri/minus)
Figure 4. Sample views of our synthetic dataset. Real datasets have also been used in our experiments. (3D curves are from [18, 20]).

Figure 5. (a) Errors of computed pose are small showing that the solver is numerically stable. (b) The histogram of the numbers of real solutions in different stages.

Figure 6. Translational and rotational error distributions between cameras 1 and 2 (blue) and 1 and 3 (green) for different levels of feature localization (top) and orientation noise (bottom).

Figure 7. Distributions of reprojection error of feature location plotted against localization and orientation errors.

Figure 8. Average reprojection error on ground truth inlier points with different ratio of outliers.

determined by the other two points in a triplet, is shown in Figure 7, averaged over 100 triplets.

The third experiment probes whether the system can reliably and accurately determine trifocal pose when correct noisy correspondences are mixed with outliers. With a fixed feature localization error of 0.25 pixels and feature orientation error of 0.1 radians, 200 triplets of features were generated, with a percentage of these replaced with samples having random location and orientation. The ratio of outliers is varied over 10%, 25% and 40%, and the experiment is repeated 100 times for each. The resulting reprojection error is small and stable across outlier ratios, Figure 8.

Computational efficiency: Each solve using our software MINUS takes 660ms on average (1.9s in the worst case) as compared to over 1 minute on average for the best prototypes using general purpose software [8, 41], both on an Intel core i7-7920HQ processor and four threads. More aggressive but potentially unsafe optimizations towards microseconds are feasible, but require assessing failure rate, as reported in the supplementary materials.

Real data experiments: Much like the standard pipeline, SIFT features are first extracted from all images. Pairwise features are found by rank-ordering measured similarities and making sure each feature’s match in another image is not ambiguous and is above accepted similarity. Pairs of features from the first and second views are then grouped with the pairs of features from the second and third views into triplets. A cycle consistency check enforces that the triplets must also support a pair from the first and third views. Three feature triplets are then selected using RANSAC and the relative pose of the three cameras is determined from two tangents with their assigned SIFT orientation and a third point without orientation.
Figure 9 shows that camera pose is reliably and accurately found using triplets of images taken from the EPFL dense multi-view stereo test image dataset [70]. Our quantitative estimates on 150 random triplets from this dataset give pose errors of $1.5 \times 10^{-3}$ radians in translation and $3.24 \times 10^{-4}$ radians in rotation. The average reprojection error is 0.31 pixels. These are comparable to or better than the trifocal relative pose estimation methods reported in [31]. Our conclusion for this dataset, whose purpose is simply to validate the solver, is that our method is at least as good and often better than the traditional ones. See supplementary data for more examples and a substantiation of this claim. Note that we do not advocate replacing the traditional method for this dataset. We simply state that our method works just as well, of course at a higher cost.

The EPFL dataset is feature-rich, typically yielding on the order of 1000 triplet features per image triplet. As such it does not portray some of the typical problems faced in challenging situations when there are few features available. The Amsterdam Teahouse Dataset [71], which also has ground-truth relative pose data, depicts scenes with fewer features. Figure 10 (top) shows a triplet of images from this dataset where there is a sufficient set of features (the soup can) to support a bifocal relative pose estimation followed by a P3P registration to a third view (using COLMAP [67]). However, when the number of features is reduced, as in Figure 10 (bottom) where the soup can is occluded, COLMAP fails to find the relative pose between pairs of these images. In contrast, our approach, which relies on three and not five features, is able to recover the camera pose for this scene. Further results are in supplementary material.

We also created another featureless dataset similar to the one in [51] but with the calibration board manually removed. This scene lacks point features, which is extremely challenging for traditional structure from motion. We built 20 triplets of images within this dataset. Within these 20 triplets, camera poses of only 5 triplets can be generated with COLMAP, but with our method, 10 out of 20 camera poses can be estimated. We reached a 100% improvement over the standard pipeline on image triplets. The sample successful cases are shown in Figure 1 and 11.

5. Conclusion

We presented a new calibrated trifocal minimal problem, an analysis demonstrating its number of solutions, and a practical solver by specializing general computation techniques from numerical algebraic geometry. Our approach is able to characterize and solve a similar difficult minimal problem with mixed points and lines. The increased ability to solve trifocal problems is key to future work on broader problems connecting the multi-view geometry of points and lines to that of points and tangents appearing when observing 3D curves, e.g., in scenes without point features, using tools of differential geometry [17, 20]. Our “100 lines of custom-made solution tracking code” will also be used to try to improve solvers of many other minimal problems which have not been solved efficiently with Gröbner bases [39].
References


