Relative Pose Estimation From Image Correspondences Under a Remote Center of Motion Constraint

Francisco Vasconcelos, Evangelos Mazomentos, John Kelly, Sebastien Ourselin, and Danail Stoyanov

***Abstract*—This letter proposes an algorithm to estimate the relative pose between two image view-points assuming that a camera is moving under a remote center of motion constraint. This is useful in minimally invasive robotic surgery, where the motion of a laparoscopic camera is constrained by the keyhole insertion point. Our method uses point correspondences between the two images and does not require any knowledge about the position of the remote center of motion. The pipeline consists of a 4-point minimal closed-form solver, used within a robust RANSAC framework to filter outlier correspondences, followed by a Levenberg–Marquardt refinement step. Our method compares favorably against the classic relative pose solution for unconstrained motion (5-point algorithm) both with synthetic data and a real footage of endoscopic robotic surgery.**

***Index Terms*—Surgical robotics: laparoscopy, visual-based navigation.**

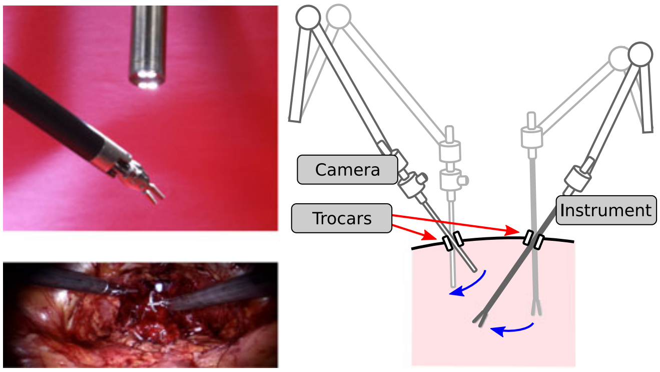


Fig. 1. Image guided minimally invasive procedure performed with a surgical robot. The motion of both the camera and the tools is constrained by the trocar placement.

I. Introduction

RELATIVE pose estimation between two image views is a common way to obtain visual odometry of a moving camera sensor [1]. It is a basic component of more complex navigation and 3D reconstruction systems such as Simultaneous Localisation and Mapping (SLAM) [2] or Structure-From-Motion (SfM) [3]. The classic relative pose problem, considering a six degree-of-freedom unconstrained motion in the 3D space has been widely validated in computer vision and robotics applications [4]. When prior knowledge about the camera motion is available, other formulations have been proposed that add new constraints to reduce the number of pose parameters to be estimated [5]–[8]. These algorithms generally outperform the general relative pose solution in their respective domains.

In this letter we address the relative pose estimation problem in the context of image guided minimally invasive surgery. In this type of procedures, the surgical tools are manipulated through trocars that are placed on small incisions on the patient (Fig. 1), and are guided by an endoscopic camera that is also inserted through a trocar. Due to this set-up the camera motion is bounded by the trocar placement in a way that is usually modelled by a remote center of motion constraint [9], i.e., the endoscope must always intersect the 3D point where the trocar is located. This means that the endoscope motion has only 4 degrees of freedom: three rotation parameters and a single translation component.

Some minimally invasive procedures are currently performed with a surgical robot (e.g., prostatectomy [10]) that enforces the trocar motion constraints by assuming a static remote center of motion[11], [12]. Given that in practice a trocar is not strictly static due to patient motion or breathing, some approaches propose a more flexible kinematic control by incorporating force feedback [13].

There have been previous works on localisation problems related to a center of motion constraint, including trocar localisation and detection from the robot kinematics [14], [15], or tool pose tracking under remote center of motion constraints [16]. However, to the best of our knowledge, the relative pose problem between two camera views under a remote center of motion constraint has not been previously addressed. A solution to this problem can be useful for tackling multiple problems

Manuscript received September 10, 2017; accepted February 5, 2018. Date of publication February 27, 2018; date of current version May 8, 2018. This letter was recommended for publication by Associate Editor I. I. Iordachita and Editor K. Masamune upon evaluation of the reviewers comments. This work was supported in part by the Wellcome Trust [WT101957] through an Innovative Engineering for Health award and in part by the Engineering and Physical Sciences Research Council (EPSRC) [NS/A000027/1]. *(Corresponding Author: Francisco Vasconcelos.)*

F. Vasconcelos, E. Mazomentos, S. Ourselin, and D. Stoyanov are with the Centre for Medical Image Computing (CMIC), University College London, London W1W 7TS, U.K. (e-mail: v.vasconcelos@ucl.ac.uk; e.mazomenos@ucl.ac.uk; s.ourselin@ucl.ac.uk; danail.stoyanov@ucl.ac.uk).

J. Kelly is with the Division of Surgery and Interventional Science, University of College London, London NW3 2PS, U.K. (e-mail: j.d.kelly@ucl.ac.uk).

Digital Object Identifier 10.1109/LRA.2018.2809617

in image guided surgery, including real-time localisation of the endoscopic camera and the surgical tools, accurate 3D reconstructions of the anatomical site, as well as 3D registration with pre-operative imaging.

With a surgical robot, the camera motion can be estimated through the kinematic chain of the manipulator holding the camera. However, visual localisation across multiple views is still necessary for representing both camera and human anatomy in the same reference frame, or whenever robot hand-eye calibration is challenging in the surgical setting.

An alternative is to estimate the camera motion directly from the change of perspective of different frames using a Structure-from-Motion or SLAM approach [17]. This is a very challenging task, as a reliable motion estimation requires that a sufficiently descriptive part of the scene remains static across different frames. This contrasts with the highly dynamic nature of most surgical scenes that include deformable tissue and moving tools. In this letter we address this problem by using the remote center of motion constraints in order to reduce the strict requirements on both the quantity and the quality of static image features required to estimate an accurate relative pose between two views. The contributions of this letter are summarised as follows:

- A simplified model (*aligned axis assumption*) for expressing the remote center of motion constraints as a single linear equation in terms of one essential matrix parameter, or by a quadratic equation in terms of translation and rotation parameters.
- Formulation of the relative camera pose problem with remote center of motion constraints, leading to a minimal solution that requires only 4 point correspondences between two images.
- Comparison between our algorithm and the classic 5-point relative pose solution for unconstrained motion [4]. Our solution outperforms the 5-point algorithm with both synthetic data and real video footage from a radical prostatectomy procedure performed with the da Vinci surgical robot [18].
- Robustness evaluation of our algorithm when the *aligned axis assumption* is not strictly verified. In simulation, our solution shows no signals of degradation for moderate deviations to the *aligned axis assumption*, considering motions where there is a sufficient field of view overlap between two views. With real data, our model is a sufficiently good approximation to make our 4-point solution work, even though a stereo camera that does not conform to the *aligned axis assumption* is used.

II. Notation

Scalars are represented by plain letters, e.g., λ, vectors are indicated by bold symbols, e.g., $\mathbf{t}$, and matrices are denoted by letters in sans serif font, e.g., T. 2D points and lines are expressed in homogeneous coordinates as 3×1 vectors. The operator $[\mathbf{v}]_\times$ designates the 3×3 skew symmetric matrix of a 3×1 vector $\mathbf{v}$, such that $[\mathbf{v}]_\times \mathbf{x} = \mathbf{v} \times \mathbf{x}$.

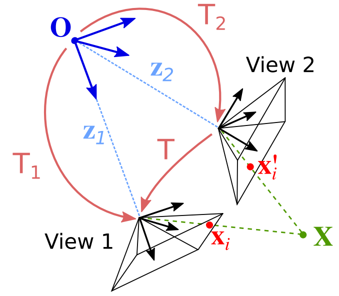


Fig. 2. Remote center of motion formulation under the aligned axis assumption.

III. Problem Formulation

Consider a rigid endoscopic camera with known intrinsic parameters being inserted into a patient through a keyhole incision point $\mathbf{O}$ (Fig. 2). We aim at estimating the relative pose with rotation R and translation $\mathbf{t}$

$$\mathsf{T} = \mathsf{T}_1 \mathsf{T}_2^{-1} = \begin{pmatrix} \mathsf{R} & \mathbf{t} \\ 0 & 1 \end{pmatrix} \tag{1}$$

when the endoscope moves between the world-to-camera transformations T_1 and T_2, given a set of pairwise point correspondences $(\mathbf{x}_i, \mathbf{x}'_i)$ between the two views. We start by briefly reviewing the classic relative pose estimation with unconstrained motion, and then we introduce the remote center of motion constraint defined by point $\mathbf{O}$.

A. Unconstrained Relative Pose

Two image point correspondences $\mathbf{x}_i$, $\mathbf{x}'_i$ that represent the same 3D point $\mathbf{X}$ under two different calibrated views are related by the epipolar constraint

$$\mathbf{x}'^{\mathsf{T}}_i \mathsf{E} \mathbf{x}_i = 0 \tag{2}$$

where the essential matrix [19]

$$\mathsf{E} = [\mathbf{t}]_\times \mathsf{R} \tag{3}$$

must verify the following cubic relations

$$\mathsf{E}\mathsf{E}^{\mathsf{T}}\mathsf{E} - \frac{1}{2}\mathrm{trace}(\mathsf{E}\mathsf{E}^{\mathsf{T}})\mathsf{E} = 0, \qquad \det \mathsf{E} = 0 \tag{4}$$

The essential matrix E has 5 degrees of freedom, and can be estimated from a minimum of 5 point correspondences [4]. Although multiple 5-point algorithm implementations exist, they typically proceed as follows: first, a 4-dimensional linear solution subspace for E is generated from 5 or more instances of (2); then, the 4 remaining unknown parameters are determined by solving a cubic system of ten equations (4). This procedure generates up to 10 algebraic solutions for the matrix E, which can only be disambiguated by verifying the epipolar consistency (2) of at least 6 correspondences. Finally a rotation R and an up-to-scale translation $\mathbf{t}$ can be uniquely factorised from E [19].

Additionally, the performance of this 5-point algorithm is greatly enhanced by using it within a RANSAC framework

[20] for outlier filtering, followed by an iterative Levenberg-Marquardt refinement step that minimises the re-projection error of inlier correspondences.

B. Remote Center of Motion Constraint

Consider now that the endoscope motion is constrained such that it must go through the remote center of motion $\mathbf{O}$. To model this constraint we work under the following assumption: for any possible camera pose, the optical axis of the endoscopic camera intersects the remote center of motion $\mathbf{O}$ (Fig. 2). In the context of this letter we designate this as the *aligned axis assumption*. Note that this assumption might not be strictly verified in practice with a real endoscope, however, we leave the discussion of its validity for later sections.

The remote center of motion constraint under the *aligned axis assumption* is a generalisation of the spherical camera motion as modelled in [5] for the case of a varying sphere radius. Therefore, we follow an analogous strategy to this work in order to derive our formulation.

Consider that $\mathbf{O}$ is the origin of the world reference frame W. From the aligned axis assumption, it follows that any transformation T_i between W and the camera reference frame can be represented as

$$\mathsf{T}_i = \begin{pmatrix} \mathsf{R}_i & \mathbf{z}_i \\ 0 & 1 \end{pmatrix} \tag{5}$$

where the translation $\mathbf{z}_i = (0 \quad 0 \quad z_i)^\mathsf{T}$ has only one degree of freedom and represents the distance between $\mathbf{O}$ and the principal point of the camera.

Consider now a camera motion between transformations T_1 and T_2 (Fig. 2). Assume that, without loss of generality,

$$\mathsf{T}_1 = \begin{pmatrix} \mathtt{I} & \mathbf{z}_1 \\ 0 & 1 \end{pmatrix}, \qquad \mathsf{T}_2 = \begin{pmatrix} \mathsf{R}^\mathsf{T} & \mathbf{z}_2 \\ 0 & 1 \end{pmatrix} \tag{6}$$

with $\mathbf{z}_1 = (0 \quad 0 \quad z_1)^\mathsf{T}$, $\mathbf{z}_2 = (0 \quad 0 \quad z_2)^\mathsf{T}$, and $\mathtt{I}$ being the 3×3 identity matrix. The relative pose T between the two views becomes

$$\mathsf{T} = \mathsf{T}_1 \mathsf{T}_2^{-1} = \begin{pmatrix} \mathsf{R} & \mathbf{z}_1 - \mathsf{R}\mathbf{z}_2 \\ 0 & 1 \end{pmatrix} \tag{7}$$

By substituting this into (3), the essential matrix under the remote center of motion constraint has the following format

$$\mathsf{E} = [\mathbf{z}_1 - \mathsf{R}\mathbf{z}_2]_\times \mathsf{R} \quad = \quad [\mathbf{z}_1]_\times \mathsf{R} - \mathsf{R}[\mathbf{z}_2]_\times \tag{8}$$

$$= \begin{pmatrix} -r_{1,2}z_2 - r_{2,1}z_1 & r_{1,1}z_2 - r_{2,2}z_1 & -r_{2,3}z_1 \\ r_{1,1}z_1 - r_{2,2}z_2 & r_{1,2}z_1 + r_{2,1}z_2 & r_{1,3}z_1 \\ -r_{3,2}z_2 & r_{3,1}z_2 & 0 \end{pmatrix} \tag{9}$$

where, $r_{i,j}$ is the element from ith row and jth column of R. From this follows that the remote center of motion, under the aligned axis assumption, is constrained by

$$e_{3,3} = 0 \tag{10}$$

where $e_{3,3}$ is the element from the third row and third column of E. Additionally, this constraint can also be represented in terms of translation and rotation as

$$r_{2,3}t_1 - r_{1,2}t_2 = 0 \tag{11}$$

where t_1 and t_2 are the first and second components of the relative translation $\mathbf{t}$.

IV. 4-Point Minimal Solution

The constraint from (10) eliminates one degree of freedom for the essential matrix E, and thus it can now be estimated minimally from 4 point correspondences instead of the 5 required for unconstrained motion. Given the simplicity of the remote center of motion constraint, we propose a 4-point relative pose algorithm that is extremely similar to its 5-point counterpart. Given a set of $N >= 4$ correspondences $(\mathbf{x}_i, \mathbf{x}'_i)$, up to 10 algebraic solutions for the relative pose solutions can be obtained as follows

1) Build a linear system by stacking 4 or more instances of (2) in terms of the 8 up-to-scale unknown parameters of the essential matrix ($e_{i,j}$):

$$\begin{pmatrix} {\mathbf{x}_1}^\mathsf{T} x'_{1,1} & {\mathbf{x}_1}^\mathsf{T} x'_{1,2} & x_{1,1}x'_{1,3} & x_{1,2}x'_{1,3} \\ {\mathbf{x}_2}^\mathsf{T} x'_{2,1} & {\mathbf{x}_2}^\mathsf{T} x'_{2,2} & x_{2,1}x'_{2,3} & x_{2,2}x'_{2,3} \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} \begin{pmatrix} e_{1,1} \\ e_{2,1} \\ e_{3,1} \\ e_{1,2} \\ e_{2,2} \\ e_{3,2} \\ e_{1,3} \\ e_{2,3} \end{pmatrix} = 0 \tag{12}$$

2) Determine a 4-dimensional linear solution subspace to (12) using SVD decomposition. This defines

$$\mathsf{E} = a\mathsf{E}_1 + b\mathsf{E}_2 + c\mathsf{E}_3 + \mathsf{E}_4 \tag{13}$$

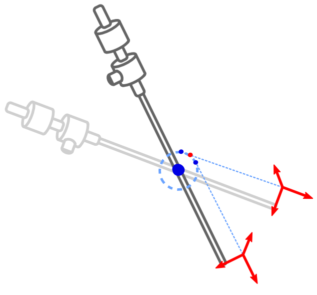
where $\{\mathsf{E}_1, \mathsf{E}_2, \mathsf{E}_3, \mathsf{E}_4\}$ is the linear base for the solution and a, b, c are unknown parameters.

3) Substitute (13) into the cubic constraints of (4), forming a polynomial system of 10 equations in 3 unknowns a, b, c.
4) Solve the polynomial system using the action matrix method [21]
5) Factorise E into rotation R and translation $\mathbf{t}$ [19]

V. Relative Pose Estimation Pipeline

Our relative pose estimation pipeline follows the same structure as the traditional pipeline for unconstrained motion that uses the 5-point algorithm [4]. The 4-point minimal solution is used within a RANSAC framework to remove outlier correspondences. Considering a camera with known intrinsics K, the result is then refined with Levenberg-Marquardt non-linear optimisation by minimising the distances $\mathbf{r}_i$, $\mathbf{r}'_i$ in pixels, between image point correspondences $(\mathbf{x}, \mathbf{x}'_i)$ and their corresponding epipolar lines ($\mathbf{l}_i = \mathsf{E}^\mathsf{T}\mathbf{x}'_i, \mathbf{l}'_i = \mathsf{E}\mathbf{x}_i$). The epipolar distances $\mathbf{r}_i$, $\mathbf{r}'_i$ are equivalent to the residue of (2) when normalised to the pixel units of each camera view.

The rotation is parametrised as a quaternion $\mathbf{q}$ while the translation by only two of its components t_2, t_3. The missing translation component t_1 is implicitly defined by the remote center



(a) (b)

Fig. 3. Remote center of motion configurations that do not verify the *aligned axis assumption*, due to translation and rotation offsets. Within a broad range of motions, two views are still close to having intersecting optical axes at a given point (red dots). Offsets are exaggerated for visualisation purposes.

of motion constraint (11).

$$\min_{\mathbf{q},t_2,t_3} \sum_{i=1}^{N} ||\mathbf{r}_i||^2 + ||\mathbf{r}_i'||^2 \tag{14}$$

with

$$\mathbf{r}_i = \mathsf{K}\mathbf{d}_i \frac{|\mathbf{x}_i^{\mathsf{T}}\mathbf{l}_i|}{||\mathtt{I}_{2\times3}\mathbf{l}_i||}, \qquad \mathbf{r}_i' = \mathsf{K}\mathbf{d}_i' \frac{|\mathbf{x}_i'^{\mathsf{T}}\mathbf{l}_i'|}{||\mathtt{I}_{2\times3}\mathbf{l}_i'||} \tag{15}$$

where $\mathbf{d}_i$ and $\mathbf{d}_i'$ are unit 2D homogenous vectors orthogonal to the epipolar lines $\mathbf{l}_i$ and $\mathbf{l}_i'$ respectively. To ensure valid rotations, the rotation quaternion is scaled to a unit norm each time the epipolar distances are computed.

VI. Comments on the Aligned Axis Assumption

In this section we discuss the limits of the *aligned axis assumption* and its impact on the applicability of our 4-point algorithm.

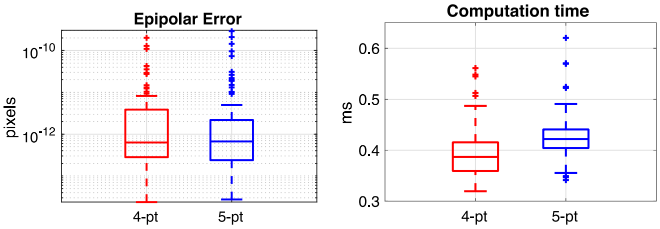
First we should note that since we do not make any assumption on the location of the remote center of motion and consider only two frames. Therefore our problem is equivalent to estimating the relative pose between any two cameras whose optical axes intersect. Note that a pure translation between two camera views (parallel optical axes) can also be estimated with our algorithm, since they intersect at infinity in the projective space and the correspondent essential matrix has the format

$$\mathsf{E} = [\mathbf{t}]_\times = \begin{pmatrix} 0 & -t_3 & t_2 \\ t_3 & 0 & -t_1 \\ -t_2 & t_1 & 0 \end{pmatrix} \tag{16}$$

which verifies (10).

On the other hand, pure rotation motions are degenerate configurations due to the elements of the essential matrix being all close to zero.

We now consider the remote center of motion constraint in cases where the *aligned axis assumption* is not verified. When there is a translation offset (Fig. 3(a)), the camera axis is always tangent to a spherical surface with radius equal to the distance between the optical axis and the remote centre of motion. When there is a rotation offset (Fig. 3(b)), the camera axis goes through a sphere whose radius is defined by the maximum or minimum depth allowed by the surgical setup. Finally we can also observe that two camera views that share a significant field of view overlap (and thus are broadly facing the same direction) generally have optical axis that are very close to intersect at a certain point (displayed in red in Fig. 3). In the experimental section we validate our algorithm using a stereo camera pair with a baseline of approximately 5 mm, and therefore both cameras correspond to a configuration similar to Fig. 3(a).



(a) (b)

Fig. 4. Minimal solver simulation results for 100 trials, using minimal data and no noise. Both 4-point and 5-point algorithms pure matlab implementations. The computation times were obtained on a Macbook Pro (Mid 2015) with a 2.5 GHz Intel Core i7.

VII. Experimental Results

We compare our 4-point algorithm against the 5-point algorithm for unconstrained motion. Although there are publicly available versions, we implemented the 5-point algorithm using the action matrix method [21] in order to use the same methodology as our 4-point implementation. Both algorithms are tested on synthetic data and real video footage from a radical prostatectomy. We also validate the robustness of our method when the *aligned axis assumption* is not verified.

A. Simulation

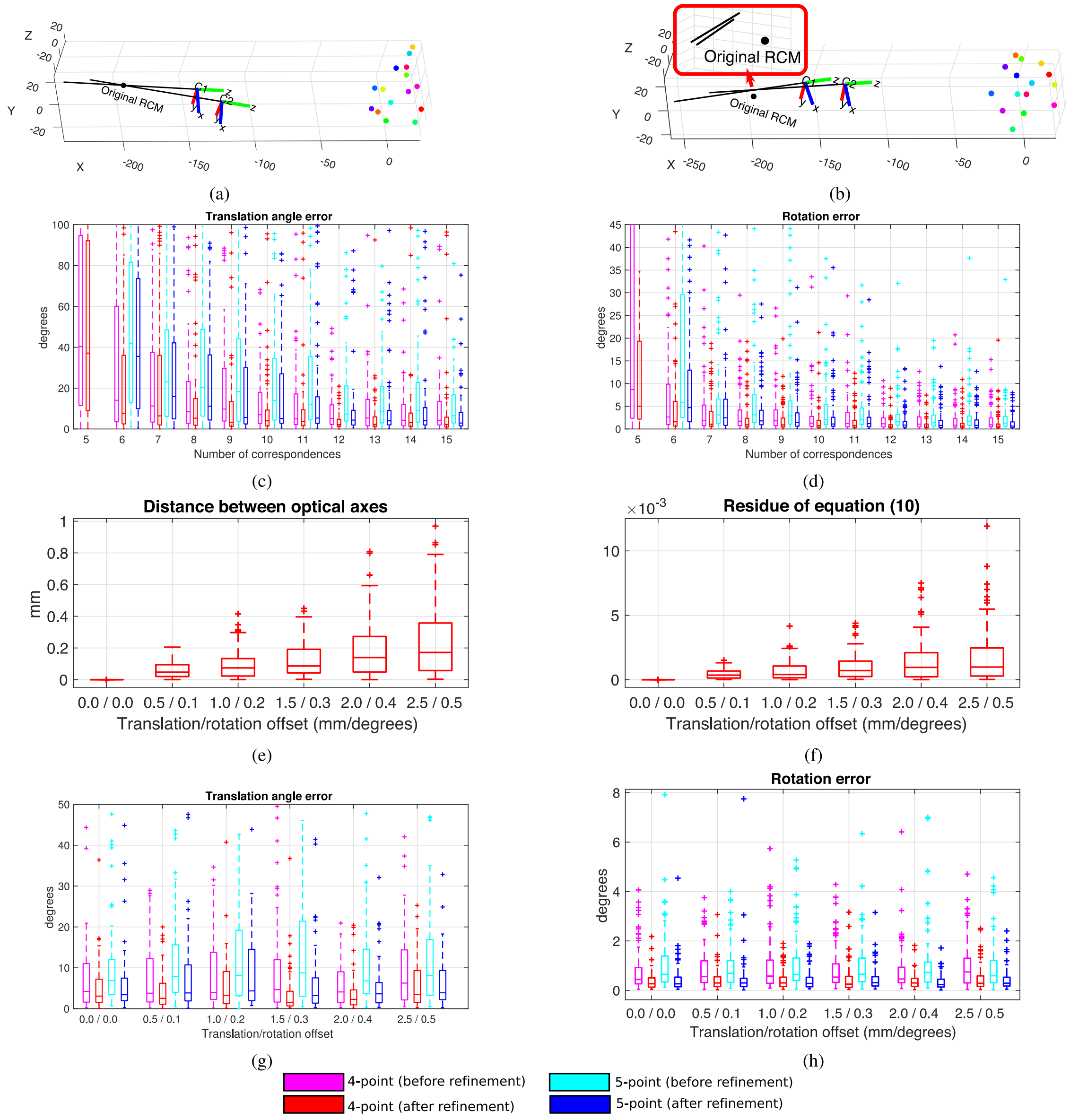
A simulator was designed to approximate the imaging conditions of a surgical robot. We consider a pinhole camera with resolution 1920 × 1080, with intrinsic parameters

$$\mathsf{K} = \begin{pmatrix} 1500 & 0.01 & 800 \\ 0 & 1400 & 600 \\ 0 & 0 & 1 \end{pmatrix} \tag{17}$$

A set of 3D points is randomly generated within a 60 mm cube. The remote center of motion is set at a distance of 200 mm from the center of mass of the scene 3D points. Camera poses are randomly generated within a distance interval between 40 and 80 mm to the remote center of motion, while the rotation is generated within the maximum range that allows all 3D points to be visible in the images.

We start by analysing the behaviour of the minimal 4-point solver with noise-free data in 100 random trials. Fig. 4(a) displays the epipolar error, as defined in (15), while Fig. 4(b) displays the computational time.

We also compare both algorithms in the presence of point correspondences with 1 pixel variance Gaussian noise and non-minimal data. In this case we use the complete relative pose pipeline including RANSAC and Levenberg-Marquardt optimisation. A threshold of 1 pixel is used to filter outliers



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 5. Simulation results: distributions are Matlab boxplots, where the center mark is the medium, the box limits are the 1st and 3rd quartiles, the whisker limits are minimum and maximum values, and cross marks are outliers. (a) Sample simulated camera motion under *aligned axis assumption*. (b) Sample simulated camera motion with an axis offset (in both rotation and translation) that makes the *aligned axis assumption* not true (zoomed in detail). (c), (d) Translation and rotation errors under the *aligned axis assumption* for a varying number of point correspondences and 1 pixel noise. Translation error is the angle between groundtruth and estimated vectors. (e) Simulated distances between optical axes when the *aligned axis assumption* is not verified, the distance between optical axes is the orthogonal euclidean distance between lines in 3D space. (f) Simulated residues for the value of $e_{3,3}$ (10) when the *aligned axis assumption* is not verified. (g), (h) Translation and rotation errors for a varying axis offset (i.e., *aligned axis assumption* not true) in both translation and rotation, using 15 point correspondences with 1 pixel variance Gaussian noise.

in RANSAC. We start by considering that the *aligned axis assumption* holds true (Fig. 5(a)). Note that due to the fact that translations are estimated up-to-scale, we use the angle between estimated and groundtruth translation vectors as the error metric. Our algorithm is consistently more accurate both in terms of rotation and translation (Fig. 5(c) and (d)). For 15 correspondences, it obtains median translation and rotation errors of 2.22 and 0.44 degrees respectively, while the 5-point solution reaches median errors of 2.91 and 0.64 degrees.

Finally, we add an offset rotation and translation with increasing magnitude to the simulated camera views so that the *aligned*

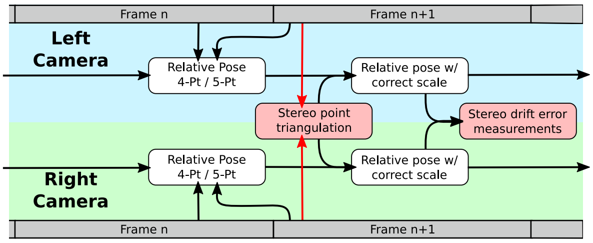

Frame n
Frame n+1
Left Camera
Relative Pose 4-Pt / 5-Pt
Relative pose w/ correct scale
Stereo point triangulation
Stereo drift error measurements
Right Camera
Relative Pose 4-Pt / 5-Pt
Relative pose w/ correct scale
Frame n
Frame n+1

Fig. 6. Stereo motion experiment. Left and right camera trajectories are estimated independently, except for the translation scale factor, which is obtained using a point cloud obtained from sparse stereo triangulation. The trajectories are compared by measuring their consistency with the fixed and known stereo extrinsic calibration.

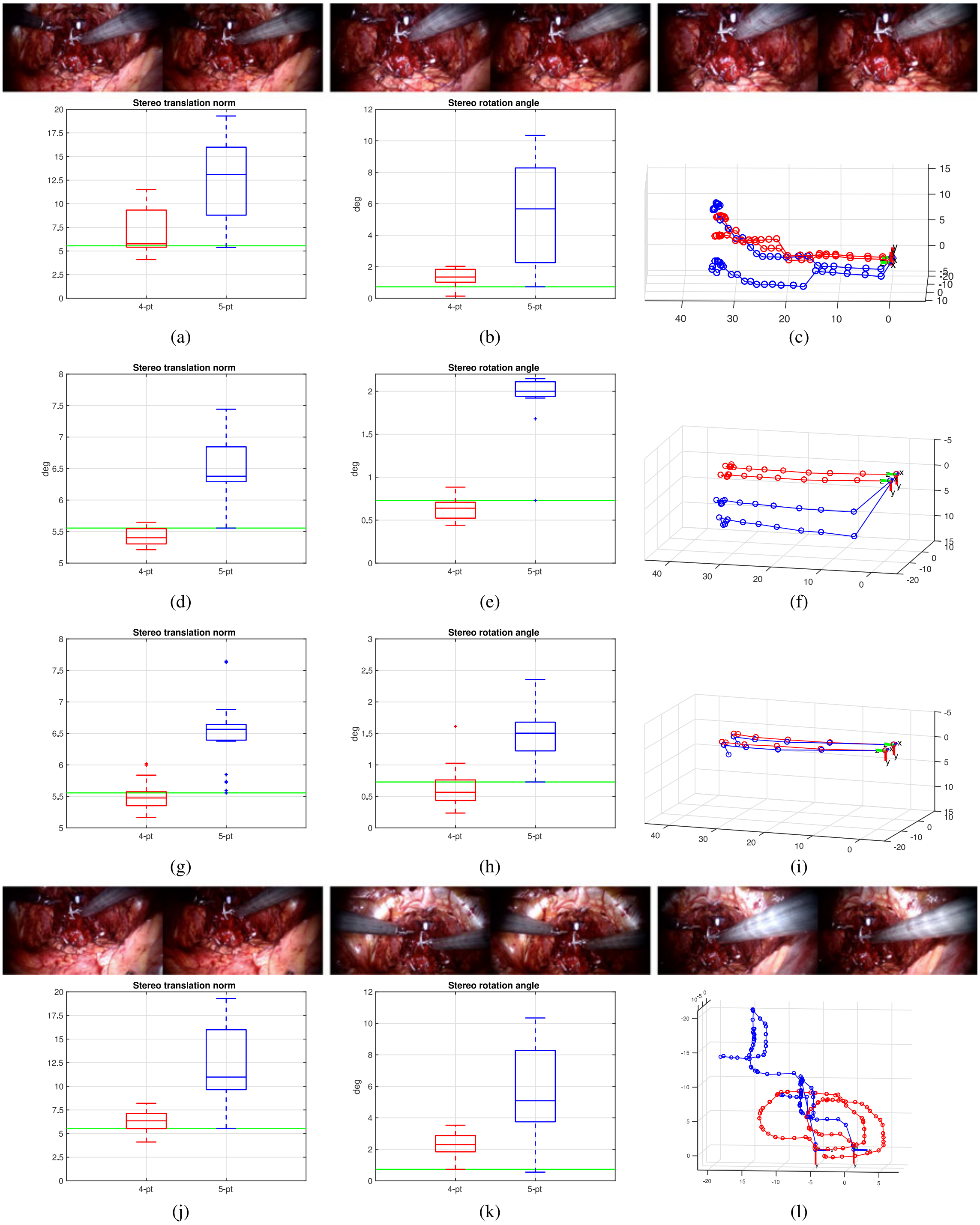


Fig. 7. Prostate stereo dataset captured with da Vinci. (a)–(i) results for the same forward motion trajectory when estimated using different frame steps: 2, 5, and 10 respectively for each row. (j)–(l) results for a circular motion trajectory. Left and right camera motions are estimated independently except for the translation scale. Only point matches between consecutive frames are considered, and no global refinement is performed. The stereo translation and rotation distributions represent the different transformations between left/right cameras obtained from the two independent trajectory estimations, while the green line represents the stereo transformation obtained from an independent offline stereo camera calibration.

axis assumption is not true any more (Fig. 5(b)). The maximum tested offset of 2.5 mm corresponds to the expected scenario in our real experiment using data from a stereo camera with 5 mm baseline between left and right cameras. Fig. 5(e) and (f) quantify how this offset affects the *aligned axis assumption*. Fig. 5(e) displays the distance between the optical axes of both views. In line with our observations in Section VI, after an offset is applied, the optical axes are still relatively close to intersecting. E.g., an offset of 2.5 mm and 0.5 degrees corresponds to a median distance between optical axes below 0.2 mm, and a maximum distance around 1 mm. Since our simulation guarantees that there is a sufficient overlap in the fields of view of both cameras, the motions are strictly restricted to a small working space, where the camera axes of both cameras are still close to intersecting. We expect this to be the case with a real scenario where point matches can be extablished between two views. Fig. 5(f) represents the groundtruth value of $e_{3,3}$ (10) after normalising the essential matrix to a unit Frobenius norm. Fig. 5(g) and (h) represent the translation and rotation errors of the 4-point and 5-point algorithms in these conditions. The 4-point algorithm does not degrade in performance for an increasing offset, 15 point correspondences and 1 pixel variance Gaussian noise.

B. Real Data

We compare the performance of 4-point and 5-point algorithms when estimating a camera trajectory on video sequences from a radical prostatectomy performed with the da Vinci Si surgical robot. The camera is a stereo laparoscope, an interesting case for two reasons: 1) it allows us to test the robustness of our algorithm in a scenario where the aligned axis assumption is not verified by design of the scope; 2) in the absence of groundtruth motion data, we can evaluate the relative pose algorithms indirectly by measuring the discrepancy between left/right trajectory estimations in terms of the left-to-right stereo transformation changes along the estimated trajectories.

We select two sequences from the procedure where there is significant camera motion. The first one is a forward motion (55 frames), and the second is a circular motion (86 frames). Both videos contain two static surgical tools and a live tissue background presenting slight deformations over time.

We use SIFT descriptors [22] to establish image point correspondences between different frames. The camera trajectories are estimated by successively applying a relative pose algorithm between frames at regular intervals. Note, however, that monocular relative pose algorithms addressed in this letter only provide an up-to-scale translation. In order to find the correct scale we compare 3D point reconstructions from a stereo pair with the correct baseline against 3D point clouds obtained from two consecutive frames. The scale estimation is the only step where stereo information is used. The consecutive up-to-scale relative poses are estimated independently for the left and right cameras. The complete pipeline for the stereo sequence experiment is summarised in Fig. 6. Although better trajectory estimations could obviously be obtained by incorporating stereo information during the relative pose estimation step, our main goal is to establish unbiased consistency metrics to compare the performance between the 5-point and 4-point algorithms.

An aspect that must be taken into account in this experiment is that relative pose estimation degenerates for very small translations [19]. This affects both the classic 5-point algorithm and our 4-point algorithm. To evaluate their breaking points we compare both algorithms on the forward motion sequence using different frame steps. We can observe that by estimating the relative pose every two frames (Fig. 7(c)), both the 5-point and the 4-point algorithms perform badly, although the 4-point is able to hold an accurate trajectory for a longer period. Estimating the relative pose every five frames (Fig. 7(f)) represents the threshold when our 4-point method starts working with greater stability, while the 5-point algorithm still fails at some frames. Finally, using every ten frames (Fig. 7(i)) both the 5-point and 4-point are able to estimate consistent trajectories, with our method taking a slight advantage.

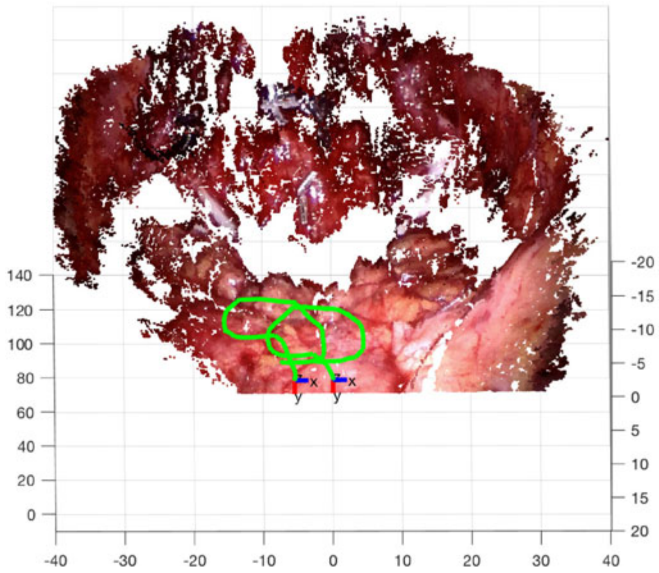


Fig. 8. Circular motion trajectory represented in the same reference frame as a 3D stereo reconstruction of the scene.

The circular motion is faster, and thus our 4-point algorithm is able to recover a consistent trajectory estimate for the whole duration of the sequence by estimating a relative pose every two frames (Fig. 7(l)). The 5-point algorithm, however, presents a significant discrepancy in terms of the stereo transformation between the two trajectories. In Fig. 8 we represent the circular trajectory with respect to the 3D scene as reconstructed from a stereo view.

VIII. Conclusions

We propose a new algorithm for estimating the relative pose between two camera views under a remote center of motion constraint. We use the *aligned axis assumption* to greatly simplify the formulation, making our algorithm extremely simple to implement. Although the *aligned axis assumption* is not strictly verified in practice, in all our tests this did not stop our algorithm from outperforming the classic 5-point solution for unconstrained motion estimation.

Although our current formulation enforces a remote center of motion constraint between two views, it does not enforce it to be at a known position, nor to be the same point for different pairs of frames. Therefore, our method can be used for any problem where two consecutive views have intersecting camera axes. A trivial example is the planar 2D motion of a ground vehicle equipped with a non-tilted camera. Although it is possible that enforcing a fixed remote center over more than 2 frames could improve the estimation of motion sequences, it is yet unclear if the current flexibility of our formulation is able to cope better with moderate RCM motions. This trade-off requires further experiments to be properly evaluated.

A complete motion estimation pipeline for endoscopy cannot be built solely using a relative pose solution, since estimating the relative pose using the the essential matrix is not adequate for motions with very small translations. As observed in Fig. 7(c), it is very challenging for our method (as well as the 5-point algorithm) to work reliably in video sequences with high frame rates. The next step is therefore to extend the remote center of motion constraint to other components of SfM/SLAM systems, such as the ressectioning (pnp) problem [23], the relative pose between stereo pairs [24], and multi-view bundle adjustment [25].

References

[1] D. Nistér, O. Naroditsky, and J. Bergen, "Visual odometry," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, 2004, pp. I-652–I-659.

[2] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha, "Visual simultaneous localization and mapping: A survey," *Artif. Intell. Rev.*, vol. 43, no. 1, pp. 55–81, 2015.

[3] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3D," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 835–846, 2006.

[4] D. Nister, "An efficient solution to the five-point relative pose problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 6, pp. 756–770, Jun. 2004.

[5] J. Ventura, "Structure from motion on a sphere," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 53–68.

[6] B. Li, L. Heng, G. H. Lee, and M. Pollefeys, "A 4-point algorithm for relative pose estimation of a calibrated camera with a known relative rotation angle," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2013, pp. 1595–1601.

[7] D. Scaramuzza, F. Fraundorfer, and R. Siegwart, "Real-time monocular visual odometry for on-road vehicles with 1-point RANSAC," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2009, pp. 4293–4299.

[8] F. Fraundorfer, P. Tanskanen, and M. Pollefeys, "A minimal case solution to the calibrated relative pose problem for the case of two known orientation angles," in *Proc. Eur. Conf. Comput. Vis.*, pp. 269–282, 2010.

[9] R. H. Taylor, J. Funda, D. D. Grossman, J. P. Karidis, and D. A. LaRose, "Remote center-of-motion robot for surgery," US Patent 5,397,323, Mar. 14 1995.

[10] A. Tewari *et al.*, "Technique of da Vinci robot-assisted anatomic radical prostatectomy," *Urology*, vol. 60, no. 4, pp. 569–572, 2002.

[11] R. H. Taylor *et al.*, "A telerobotic assistant for laparoscopic surgery," *IEEE Eng. Med. Biol. Mag.*, vol. 14, no. 3, pp. 279–288, May/Jun. 1995.

[12] N. Aghakhani, M. Geravand, N. Shahriari, M. Vendittelli, and G. Oriolo, "Task control with remote center of motion constraint for minimally invasive robotic surgery," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2013, pp. 5807–5812.

[13] A. Krupa, C. Doignon, J. Gangloff, M. de Mathelin, L. Solert, and G. Morel, "Towards semi-autonomy in laparoscopic surgery through vision and force feedback control," in *Experimental Robotics VII*. Berlin, Germany: Springer-Verlag, 2001, pp. 189–198.

[14] L. Dong and G. Morel, "Robust trocar detection and localization during robot-assisted endoscopic surgery," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2016, pp. 4109–4114.

[15] Z. Wang *et al.*, "Vision-based calibration of dual RCM-based robot arms in human-robot collaborative minimally invasive surgery," *IEEE Robot. Autom. Lett.*, vol. 3, no. 2, pp. 672–679, Apr. 2018.

[16] C. Doignon, F. Nageotte, and M. De Mathelin , "Segmentation and guidance of multiple rigid objects for intra-operative endoscopic vision," in *Dynamical Vision*. Berlin, Germany: Springer-Verlag, 2007, pp. 314–327.

[17] P. Mountney, D. Stoyanov, A. Davison, and G.-Z. Yang, "Simultaneous stereoscope localization and soft-tissue mapping for minimal invasive surgery," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2006, pp. 347–354.

[18] G. H. Ballantyne and F. Moll, "The da Vinci telerobotic surgical system: The virtual operative field and telepresence surgery," *Surgical Clinics*, vol. 83, no. 6, pp. 1293–1304, 2003.

[19] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[20] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[21] M. Byröd, K. Josephson, and K. Åström, "Fast and stable polynomial equation solving and its application to computer vision," *Int. J. Comput. Vis.*, vol. 84, no. 3, pp. 237–256, 2009.

[22] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, Nov. 2004.

[23] F. Moreno-Noguer, V. Lepetit, and P. Fua, "Accurate non-iterative o (n) solution to the pnp problem," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[24] F. Vasconcelos and J. P. Barreto, "Towards a minimal solution for the relative pose between axial cameras," in *Proc. Brit. Mach. Vision Conf.*, 2013, p. 1241–1–1241–1.

[25] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment—A modern synthesis," in *Proc. Int. Workshop Vis. Algorithms*, 1999, pp. 298–372.