Calibrated and Partially Calibrated Semi-Generalized Homographies

Snehal Bhayani[1] Torsten Sattler[2] Daniel Barath[3] Patrik Beliansky[4]
Janne Heikkilä [1] Zuzana Kukelova[5]
[1]Center for Machine Vision and Signal Analysis, University of Oulu, Finland
[2]Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague
[3]Computer Vision and Geometry Group, Department of Computer Science, ETH Zürich
[4]Faculty of Mathematics and Physics, Charles University, Prague
[5]Visual Recognition Group, Faculty of Electrical Engineering, Czech Technical University in Prague

Abstract

In this paper, we propose the first minimal solutions for estimating the semi-generalized homography given a perspective and a generalized camera. The proposed solvers use five 2D-2D image point correspondences induced by a scene plane. One group of solvers assumes the perspective camera to be fully calibrated, while the other estimates the unknown focal length together with the absolute pose parameters. This setup is particularly important in structure-from-motion and visual localization pipelines, where a new camera is localized in each step with respect to a set of known cameras and 2D-3D correspondences might not be available. Thanks to a clever parametrization and the elimination ideal method, our solvers only need to solve a univariate polynomial of degree five or three, respectively a system of polynomial equations in two variables. All proposed solvers are stable and efficient as demonstrated by a number of synthetic and real-world experiments.

1. Introduction

Estimating the homography between two cameras observing a planar scene is a crucial problem in computer vision with applications, *e.g.*, in structure-from-motion (SfM) [44, 46, 51, 56], localization [7, 39, 41], visual odometry [33, 34], camera calibration [45, 57], and image retrieval [37, 55]. It is one of the oldest camera geometry problems with many solutions including the well-known normalized direct linear transform (DLT) method [20] for estimating the homography from a minimum of four point correspondences; the minimal solutions based on affine [3, 24] or SIFT correspondences [2, 4]; solutions assuming known gravity direction [15, 43] or cameras with radial distortion [8, 11, 17, 20, 22, 27]. All above-mentioned algorithms assume that both cameras satisfy the central perspective projection model (potentially, with radial distortion), *i.e.*, they assume that both cameras have a single center of projection. Surprisingly, the problem of estimating a homography has not been studied for generalized cameras.

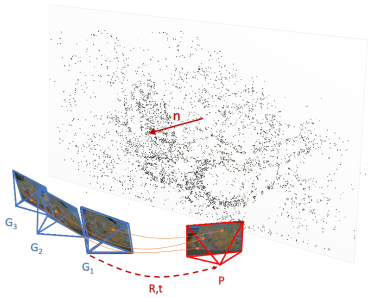

Figure 1. An illustration of the problem configuration.

A generalized camera [38] is a camera that captures some arbitrary set of rays and does not adhere to the central perspective projection model. Such a camera model is practical and appears, *e.g.*, in applications that exploit multi-camera configurations, like stereo-pairs, SfM [58], or in localization pipelines [47, 54]. Such pipelines often are based on sequences of images, where there might be a set of cameras with known poses and we are given a new image which is to be registered to a generalized camera composed of the known perspective ones. Estimating the camera pose, w.r.t. the generalized camera, in such situations often leads to results superior, in terms of accuracy, to considering only pairwise epipolar geometries, especially thanks to a larger field-of-view of the generalized camera [50]. Also, it has the advantage of recovering the absolute pose [58], *i.e.*, the scale of the translation, which is a severe deficiency of epipolar geometry-based relative pose estimation.

While the problem of estimating the absolute pose of a generalized camera can be solved very efficiently [28], *i.e.*, there exists a solution that solves only 3 quadratic equations in 3 unknowns and runs in a few μs, the problem of estimating the relative pose of two generalized cameras is

significantly more complex [49]. This problem results in a system of 15 polynomial equations, each of degree 6, with 64 solutions. The final solver based on the Gröbner basis method [49] is infeasible for real-time applications.

In [58], the authors considered a semi-generalized epipolar geometry problem, *i.e.*, the problem of estimating the relative pose together with the scale of the translation between one perspective and one generalized camera from 2D-2D correspondences. In this paper, four minimal solvers were presented, *i.e.*, $\texttt{E}_{\mathbf{5+1}}$ and $\texttt{E}_{\mathbf{4+2}}$ for calibrated pinhole cameras, and $\texttt{Ef}_{\mathbf{6+1}}$ and $\texttt{Ef}_{\mathbf{5+2}}$ for pinhole cameras with unknown focal length. Here, 4+2 denotes a configuration where four point correspondences come from one camera G_i of the generalized camera G and the remaining two from one or two other cameras. The authors showed the applicability of the proposed $\texttt{E}_{\mathbf{4+2}}$ and $\texttt{Ef}_{\mathbf{5+2}}$ solvers for incremental SfM in the absence of 2D-3D point correspondences. However, the $\texttt{E}_{\mathbf{4+2}}$ and $\texttt{Ef}_{\mathbf{5+2}}$ solvers perform operations on large matrices and, thus, are impractical for real-time applications, with running times of $1.2ms$ and $13.6ms$, respectively. The $\texttt{E}_{\mathbf{5+1}}$ and $\texttt{Ef}_{\mathbf{6+1}}$ solvers are based on the existing efficient five-point $\texttt{E}\mathbf{5}$ [35] and the six point $\texttt{E}\mathbf{6f}$ [10] relative pose methods. These solvers actually do not benefit from the generalized camera setup, except that one additional point correspondence is used to estimate the scale of the translation. Moreover, the $\texttt{E}_{\mathbf{5+1}}$ and $\texttt{Ef}_{\mathbf{6+1}}$ solvers require 5 (6 for unknown focal length) point correspondences to be detected by the same camera. This criterion may be problematic in the absence of enough inlier point matches. Note that [58] does not solve all possible configurations of point correspondences that can appear in the semi-generalized setup due to the complicated systems of polynomial equations. Further, [58] cannot handle generalized cameras with more than three cameras and having fewer than four correspondences with all cameras.

In this paper, we study a similar setup as [58], *i.e.*, one perspective and one generalized camera. However, we assume that these cameras observe a planar scene, see Fig. 1. We present the first minimal solutions for estimating the pose between a perspective and a generalized camera from 2D-2D correspondences induced by a plane, *i.e.*, the first minimal solutions for the so-called *semi-generalized homography*. The proposed solvers use five 2D-2D image point correspondences and assume either a calibrated or a perspective camera with unknown focal length. This setup is particularly important in SfM and localization, where a new camera is localized with respect to a set of known ones and 2D-3D correspondences might not be available, *e.g.*, due to memory restrictions or to avoid matching features between individual cameras in the generalized camera.

The main **contributions** of the paper are as follows: **1)** A theoretical analysis of the **new semi-generalized homography** problem for calibrated and partially calibrated cameras and a formulation of the problem as a system of linear equations in twelve unknowns.

2) Derivation of **new constraints** for the semi-generalized homography using the elimination ideal theory [29].

3) A class of efficient minimal solvers for calibrated cameras, $\texttt{sH5}_{\mathbf{2}}$, $\texttt{sH5}_{\mathbf{3}}$, $\texttt{sH5}_{\mathbf{4}}$, $\texttt{sH4.5}_{\mathbf{2}}$ and $\texttt{sH4.5}_{\mathbf{3}}$ that only need to solve a 5^{th} (3^{rd}) degree univariate polynomial, a linear system, or a system of equations in two unknowns.

4) Two new efficient minimal solvers for partially calibrated cameras $\texttt{sH5f}_{\mathbf{2}}$ and $\texttt{sH5f}_{\mathbf{3}}$ that need to solve a univariate polynomial of degree five (three).

5) Our solvers do not need 3D points or 2D-2D matches between individual cameras from the generalized camera and cover all scenarios where it is possible to estimate the scale.

6) Compared to [58], our solvers cover **all possible minimal configurations of point correspondences** as well as numbers of cameras in the generalized camera. Code is available at github.com/snehalbhayani/SemiGeneralizedHomography.

2. Problem Formulation

First, we set up notations and conventions that we will follow for the rest of the paper. Let $\mathcal{P}$ denote the perspective camera, while the generalized camera is denoted as $\mathcal{G}$. We assume that the generalized camera $\mathcal{G}$ is fully calibrated, and it consists of a set of perspective cameras $\{\mathcal{G}_1, \mathcal{G}_2, \dots \mathcal{G}_k\}$. For the pinhole camera $\mathcal{P}$, we consider two different cases, *i.e.*, the case where $\mathcal{P}$ is fully calibrated, and the case when its calibration matrix is of the form $\mathtt{K} = \text{diag}(f, f, 1)$ with unknown focal length f.

In the following text, we will consider several different coordinate systems, *i.e.*, the global coordinate system, the local coordinate system of the perspective camera $\mathcal{P}$, and local coordinate systems of perspective cameras $\mathcal{G}_i$. Let $\mathtt{R}_{\mathtt{G}_i}$, $\mathbf{t}_{\mathtt{G}_i}$ and $\mathtt{R}_{\mathtt{P}}$, $\mathbf{t}_{\mathtt{P}}$ denote the rotations and translations required to align local the coordinate systems of $\mathcal{G}_i$ respectively $\mathcal{P}$, to the global coordinate system. Without loss of generality, we can assume that the global coordinate system coincides with the local coordinate system of $\mathcal{G}_1$, *i.e.*, $\mathtt{R}_{\mathtt{G}_1} = \mathtt{I}$ and $\mathbf{t}_{\mathtt{G}_1} = [0, 0, 0]^\top$. Sometimes we will call this system the local coordinate system of the generalized camera $\mathcal{G}$. Therefore, the global coordinate system, the local coordinate system of $\mathcal{G}_1$, and the local coordinate system of $\mathcal{G}$ are interchangeable. We will use the upper index to denote the coordinate system. For example, $\mathbf{X}^{\mathtt{P}} \in \mathbb{R}^3$ and $\mathbf{X}^{\mathtt{G}} \in \mathbb{R}^3$ are the coordinates of the point $\mathbf{X}$ in the local coordinate system of $\mathcal{P}$ and the local coordinate system of $\mathcal{G}$, respectively, and it holds that $\mathbf{X}^{\mathtt{G}} = \mathtt{R}_{\mathtt{P}}\mathbf{X}^{\mathtt{P}} + \mathbf{t}_{\mathtt{P}}$.

Our objective is to estimate the rotation $\mathtt{R} \in \mathbf{SO(3)}$ and translation $\mathbf{t} \in \mathbb{R}^3$ between the perspective camera $\mathcal{P}$ and the generalized camera $\mathcal{G}$, *i.e.*, the rotation and translation that align the local coordinate system of $\mathcal{P}$ to that of $\mathcal{G}$. Note

that, since the local coordinate system of $\mathcal{G}$ coincides with the global coordinate system, $\mathtt{R} = \mathtt{R}_{\mathtt{P}}$ and $\mathbf{t} = \mathbf{t}_{\mathtt{P}}$.

For the estimation of $\mathtt{R}$ and $\mathbf{t}$, we will use 2D-2D correspondences detected between $\mathcal{P}$ and the cameras in $\mathcal{G}$. We assume that these point correspondences are projections of co-planar 3D points $\mathbf{X}_j$ satisfying: $\mathbf{n}^\top \mathbf{X}_j + d = 0$, where $\mathbf{n} \in \mathbb{R}^3$ denotes the normal of the scene plane π and d denotes the plane intercept. Note that the same plane can be defined by the normal $\widetilde{\mathbf{n}} = \mathbf{n}/d$ and the equation $\widetilde{\mathbf{n}}^\top \mathbf{X}_j + 1 = 0$.

2.1. Semi-Generalized Homography

Consider a 3D point $\mathbf{X}_j$ observed by the perspective camera $\mathcal{P}$ and the camera $\mathcal{G}_i$, *i.e.*, the i-th constituent perspective camera from the generalized camera $\mathcal{G}$. Let us denote the image points detected in $\mathcal{P}$ and $\mathcal{G}_i$ as $\mathbf{p}_j = [x_j, y_j, 1]^\top$ and $\mathbf{g}_{ij} = [x_j^{G_i}, y_j^{G_i}, 1]^\top$, respectively. With this notation, the coordinates of the 3D point $\mathbf{X}_j$ in the local coordinate system of $\mathcal{P}$ can be expressed as

$$\mathbf{X}_j^{\mathtt{P}} = \alpha_j \mathtt{K}^{-1} \mathbf{p}_j \ , \tag{1}$$

where $\mathtt{K}$ is the calibration matrix of the camera $\mathcal{P}$ and α_j represents the depth of the point $\mathbf{X}_j$ in $\mathcal{P}$. A similar relationship holds for the coordinates of the 3D point $\mathbf{X}_j$ in the local coordinate system of $\mathcal{G}_i$,

$$\mathbf{X}_j^{\mathtt{G}_i} = \beta_{ij} \mathtt{K}_{\mathtt{G}_i}^{-1} \mathbf{g}_{ij} \ , \tag{2}$$

where $\mathtt{K}_{\mathtt{G}_i}$ is the calibration matrix of the camera $\mathcal{G}_i$ and β_{ij} represents the depth of the point $\mathbf{X}_j$ in $\mathcal{G}_i$.

To obtain the relationship between $\mathbf{X}_j^{\mathtt{P}}$ and $\mathbf{X}_j^{\mathtt{G}_i}$, we have to transform them into the same coordinate system, *i.e.*, in this case the global coordinate system. This gives us the following constraint

$$\alpha_j \mathtt{R}\mathtt{K}^{-1} \mathbf{p}_j + \mathbf{t} = \beta_{ij} \mathtt{R}_{\mathtt{G}_i} \mathtt{K}_{\mathtt{G}_i}^{-1} \mathbf{g}_{ij} + \mathbf{t}_{\mathtt{G}_i} \ . \tag{3}$$

Note that here we use the fact that $\mathtt{R} = \mathtt{R}_{\mathtt{P}}$ and $\mathbf{t} = \mathbf{t}_{\mathtt{P}}$. Since in our case $\mathtt{R}_{\mathtt{G}_i}, \mathbf{t}_{\mathtt{G}_i}$ and $\mathtt{K}_{\mathtt{G}_i}^{-1}$ are known, we will, for better readability, substitute $\mathbf{q}_{ij} = \mathtt{R}_{\mathtt{G}_i} \mathtt{K}_{\mathtt{G}_i}^{-1} \mathbf{g}_{ij}$ and obtain

$$\alpha_j \mathtt{R}\mathtt{K}^{-1} \mathbf{p}_j + \mathbf{t} = \beta_{ij} \mathbf{q}_{ij} + \mathbf{t}_{\mathtt{G}_i} \ . \tag{4}$$

The 3D point $\mathbf{X}_j$ is lying on the plane π, *i.e.*, $\mathbf{X}_j^{\mathtt{P}}$ should satisfy $(\widetilde{\mathbf{n}}^{\mathtt{P}})^\top \mathbf{X}_j^{\mathtt{P}} + 1 = 0$, where $\widetilde{\mathbf{n}}^{\mathtt{P}} \in \mathbb{R}^3$ is the normal of the plane π in the local coordinate system of $\mathcal{P}$. For simplicity, in the rest of the text, we will omit the upper index $\mathtt{P}$ in $\widetilde{\mathbf{n}}^{\mathtt{P}}$. The depth α_j in (1) can be then expressed using the normal $\widetilde{\mathbf{n}}$ as

$$\alpha_j = \frac{-1}{\widetilde{\mathbf{n}}^\top \mathtt{K}^{-1} \vec{\mathbf{p}_j}} \ . \tag{5}$$

Let us consider a 3×3 homography matrix $\mathtt{H}$ of the form $\mathtt{H} = \mathtt{R} - \mathbf{t}\widetilde{\mathbf{n}}^\top$. By substituting (5) into (4) we obtain

$$\alpha_j \mathtt{H}\mathtt{K}^{-1} \mathbf{p}_j = \beta_{ij} \mathbf{q}_{ij} + \mathbf{t}_{\mathtt{G}_i} \ . \tag{6}$$

Eq. (6) is the basic semi-generalized homography constraint. The depths β_{ij} can be easily eliminated from this constraint (6) by multiplying it with the skew-symmetric matrix $[\mathbf{q_j}]_\times$ from the left side, resulting in

$$[\mathbf{q}_{ij}]_\times (\alpha_j \mathtt{H}\mathtt{K}^{-1} \mathbf{p}_j - \mathbf{t}_{\mathtt{G}_i}) = \mathbf{0} \ . \tag{7}$$

Let us denote $\mathtt{G} = \mathtt{H}\mathtt{K}^{-1}$. By dividing (7) with α_j and using (5), we obtain the equations

$$[\mathbf{q}_{ij}]_\times (\mathtt{G}\mathbf{p}_j + (\mathbf{m}^\top \mathbf{p_j}) \mathbf{t}_{\mathtt{G}_i}) = \mathbf{0} \ , \tag{8}$$

where $\mathbf{m} = \mathtt{K}^{-1} \widetilde{\mathbf{n}}$. Note that we are able to eliminate the unknown depths α_j from (7), and to derive simple linear constraints (8) for the semi-generalized homography thanks to the special parameterization (5), based on the normal $\widetilde{\mathbf{n}}^{\mathtt{P}}$ expressed in the coordinate system of $\mathcal{P}$. Each 2D-2D correspondence $\mathbf{p}_j \leftrightarrow \mathbf{q}_{ij}$ gives us three equations of the form (8), from which only two are linearly independent.

2.2. Semi-Generalized Homography Constraints

Besides the constraints (8) induced by a 2D-2D correspondence $\mathbf{p}_j \leftrightarrow \mathbf{q}_{ij}$, there are other ones arising from the form of the matrix $\mathtt{G}$. The constraints (8) are linear in the 12 unknowns, *i.e.*, elements of the matrix $\mathtt{G}$ and the vector $\mathbf{m}$. However, $\mathtt{G}$ and $\mathbf{m}$ are not independent since $\mathtt{G} = \mathtt{H}\mathtt{K}^{-1} = \mathtt{R}\mathtt{K}^{-1} - \mathbf{t}\mathbf{m}^\top$. Moreover, the rotation matrix $\mathtt{R} \in \mathbf{SO(3)}$ introduces additional constraints. All these constraints, *i.e.*, constraints originating from the form

$$\mathtt{G} - \mathtt{R}\mathtt{K}^{-1} + \mathbf{t}\mathbf{m}^\top = \mathtt{0}, \quad \mathtt{R}^\top \mathtt{R} = \mathtt{R}\mathtt{R}^\top = \mathtt{I}_{3\times 3} \ , \tag{9}$$

can be used to define an ideal $I \subset \mathbb{C}[\varepsilon]$ [14], where ε contains 9 unknowns from $\mathtt{G}$, 9 from $\mathtt{R}$, 3 from $\mathbf{t}$, 3 from $\mathbf{m}$ and the inverse of the focal length $w = \frac{1}{f}$. Now we can use the elimination ideal technique [29] to eliminate 9 unknowns of $\mathtt{R}$, 3 of $\mathbf{t}$, and w from this ideal. *I.e.*, we compute an elimination ideal I_1 that will contain only polynomials in 12 unknowns from $\mathtt{G}$ and $\mathbf{m}$. Eq. (9) does not include the constraint $det(\mathtt{R}) = 1$, as it does not change I_1, whose generators are the same for $\mathtt{G}$ and $-\mathtt{G}$, *i.e.*, both $\mathtt{R}, \mathbf{t}$ and $-\mathtt{R}, -\mathbf{t}$ are valid solutions. This ambiguity is resolved at a later stage when decomposing the homography matrix $\mathtt{H}$. The elimination ideal I_1 can be computed offline using some algebraic geometry software like Macaulay 2 [19]. We found that such an elimination ideal is generated by 4 polynomials (three of degree 3 and one of degree 4) in 12 unknowns. Similarly, for the calibrated case, *i.e.*, $\mathtt{K} = \mathtt{I}$ there are 10 such generators of I_1. For more details on elimination ideals we refer to [14, 29].

These 10 new constraints (4 for the unknown focal length case) in 12 unknowns from $\mathtt{G}$ and $\mathbf{m}$ together with the linear equations (8) in these unknowns can be used to solve for $\mathtt{R}, \mathtt{t}$ (and f). After a null-space re-parameterization of $\mathtt{G}$ and $\mathbf{m}$ using the linear equations (8) for 5 point matches, we can

transform these equations to 10 (4) polynomial equations in 2 unknowns, respectively 3 unknowns for the 4.5 point matches required for the calibrated case. Such systems can be solved, *e.g.*, using the automatic generator of Gröbner basis solvers [25, 31] and they return up to 16 (calibrated), respectively 6 (unknown f), real solutions to $\mathtt{G}$ and $\mathbf{m}$. For details on solvers sizes see the supp. material (SM).

In order to make the solvers more efficient, we introduce an additional change of variables by assuming that one of the elements of $\mathtt{G}$ is non-zero, *e.g.*, $g_{33} \neq 0$. This assumption can introduce a degeneracy. However, such a degeneracy is not crucial in practical applications and can be avoided as discussed in [13]. Moreover, our change of coordinate system used for the calibrated solver directly avoids this degeneracy. The situation for the focal length case is discussed in more detail in the SM.

With this assumption, we introduce new variables $g'_{kl} = \frac{g_{kl}}{g_{33}}$, for $\forall kl \neq 33$ and $m'_k = \frac{m_k}{g_{33}}$, $k = 1, 2, 3$, where g_{kl} are elements from the k^{th} row and l^{th} column of the matrix $\mathtt{G}$ and m_k are elements of the vector $\mathbf{m}$. The variable change is also applied to the 10 (4) generators of I_1, defining a new ideal I'_1. Using again the elimination ideal technique [29], we can eliminate g_{33} from the ideal I'_1, leading to a new ideal I'_2. For the calibrated case, the ideal I'_2 is generated by five polynomials e_i, each of degree 5, in the $8+3$ unknowns, g'_{kl} for $\forall kl \neq 33$ and m'_k, $k = 1, 2, 3$. For the unknown focal length case, it is generated by only a single polynomial e of degree five. More details on the form of the generators of the elimination ideals I_1 and I'_2 for both the cases, together with the input code for Macaulay2 used to compute these generators, is provided in the SM. Note that the derivation of these 5^{th} degree constraints as well as the above mentioned generators of I_1 is crucial for the efficiency of the final solvers. Without using the elimination ideal tricks, one would need to work directly with the parameterization of $\mathtt{G}$ using the rotation matrix, leading to complex systems of polynomial equations in many unknowns and with huge solvers. Next we show how to solve these equations for a calibrated $\mathcal{P}$ and then for the case when the focal length of $\mathcal{P}$ is unknown.

2.3. Calibrated Camera Solvers

In this case, we can assume that $\mathtt{K} = \mathtt{I}_{3\times3}$, leading to $\mathtt{G} = \mathtt{H} = \mathtt{R} - \mathbf{t}\widetilde{\mathbf{n}}^\top$. Here, we have 9 DOF, 3 for each $\mathtt{R}$, $\mathbf{t}$ and $\widetilde{\mathbf{n}}$. Each point correspondence leads to 2 linearly independent homogeneous constraints of the form (8) and thus 4.5 correspondences are sufficient to solve this problem. However we still need to sample 5 point correspondences, $\mathbf{p}_j \leftrightarrow \mathbf{q}_{ij}$, $j = 1, \dots 5$, resulting in 10 constraints.

There are two ways to deal with this over-constrained formulation: 1) use all 10 constraints from the 5 point correspondences and only one constraint from the 5 generators e_i of the ideal I'_2. 2) use 9 constraints by considering only 4.5 point correspondences, and instead use all 5 constraints e_i on the $\mathtt{G}$ matrix[1]. The first approach results in solvers $\mathtt{sH5_2}$ and $\mathtt{sH5_3}$ that have to find the roots of a univariate polynomial. The second approach leads to solvers $\mathtt{sH4.5_2}$ and $\mathtt{sH4.5_3}$ that have to solve a system of 5 equations in 2 unknowns. The 4.5 point solvers are slightly slower than the 5 point solvers. However, since they use all constraints on $\mathtt{G}$, they result in a correctly decomposable homography and therefore smaller errors in the presence of noise. These solvers are described in detail in the SM. Next we describe the 5 point solvers, $\mathtt{sH5_2}$ and $\mathtt{sH5_3}$.

Without loss of generality (w.l.o.g.), we assume that the first point correspondence is observed in camera $\mathcal{G}_1$, *i.e.*, $i = 1$ where $\mathbf{t}_{\mathcal{G}_1} = [0, 0, 0]^\top$, and pre-rotate the local coordinate systems of $\mathcal{P}$ and $\mathcal{G}_1$ such that $\mathbf{p}_1 = [0, 0, 1]^\top$ and $\mathbf{q}_{11} = [0, 0, 1]^\top$. This simplifies the equations and after substituting into (8), we have $g_{13} = 0,\ g_{23} = 0$. Moreover, we can safely assume $g_{33} \neq 0$ and divide these equations by g_{33}, transforming them into non-homogeneous equations in 9 unknowns, $\varepsilon' = \{g'_{11}, g'_{12}, g'_{21}, g'_{22}, g'_{31}, g'_{32}, m'_1, m'_2, m'_3\}$. The remaining 4 point correspondences lead to 8 linearly independent equations that can be written in matrix form

$$\mathtt{C}\mathbf{b} = 0 \ , \tag{10}$$

where $\mathtt{C}$ is a 8×10 coefficient matrix and $\mathbf{b}$ is a 10×1 vectorized form of the set of $\varepsilon' \cup \{1\}$. Next, we consider three cases based on the maximum number of correspondences coming from one camera $\mathcal{G}_i$.

$\mathtt{sH5_2}$ solver: In this case no more than 2 correspondences come from the same camera $\mathcal{G}_i$. Hence the matrix $\mathtt{C}$ in (10) has a two dimensional null-space $\{\mathbf{b_1}, \mathbf{b_2}\}$. A solution to $\mathbf{b}$ can be obtained as a linear combination $\mathbf{b} = \gamma_1\mathbf{b_1} + \gamma_2\mathbf{b_2}$. Using $b_{10} = 1$, we can express γ_2 as a linear polynomial in γ_1. Hence the variables in ε' can be parameterized as linear polynomials of γ_1. This parameterization can be substituted into the generators of the ideal I'_2 for the calibrated case. This leads to 5 univariate polynomials $e_i(\gamma_1)$, each of degree 5. We choose one of these polynomials, which we solve using Sturm sequences [21]. This results in up to five real solutions to ε'.

Next, we extract solutions to g_{33}. Writing $\mathtt{G} = \mathtt{H} = \mathtt{R} - \mathbf{t}\mathbf{m}^\top$, we obtain a set of polynomial constraints. By variable elimination and substitutions, we obtain 1 solution to g_{33}, unique up to a sign, which is fixed by constraining the solution of the plane vector $\mathbf{n}$ so that the corresponding 3D point in $\mathcal{P}$ is in the front of the camera. Solutions to $\mathtt{G}$ as well as $\mathbf{m}$ can be extracted from the solutions to ε' and g_{33}. We then decompose $\mathtt{H}$ such that $det(\mathtt{R}) = 1$ to obtain a set of relative poses $\mathtt{R}$ and $\mathbf{t}$. Note that the constraints e_i that were not used to obtain the solutions can be used to eliminate infeasible solutions.

[1] Note that the x- and y-coordinates of all 5 correspondences are used. However, one constraint from (8) originating from the matches is omitted.

	$\mathtt{E_{5+1}}$	$\mathtt{E_{4+2}}$	$\mathtt{Ef_{6+1}}$	$\mathtt{Ef_{5+2}}$	P3P+N	P5Pf+N	$\mathtt{sH5_2}$	$\mathtt{sH4.5_2}$	$\mathtt{sH5_3}$	$\mathtt{sH4.5_3}$	$\mathtt{sH5_4}$	$\mathtt{sH5f_2}$	$\mathtt{sH5f_3}$
Ref	[58]	[58]	[58]	[58]									
Focal			✓	✓		✓						✓	✓
# pt	6	6	7	7	6	8	5	4.5	5	4.5	5	5	5
# sol	10	40	10	50	4	4	5	16	3	12	1	5	3
						Complexity							
G-J/LU	10×20	73×113	11×20	378×428		5×8	8×10	11×27	8×10	23×35	8×9	8×10	8×10
Eigen		40×40	9×9	50×50	3×3			16×16		12×12			
Sturm	10						5		3			5	3
QR	5×9					5×8							

Table 1. Comparison of the proposed solvers (gray) vs. the state-of-the-art.

$\mathtt{sH5_3}$ solver: If there are 3 2D-2D point correspondences from the same camera $\mathcal{G}_i$, the situation is a bit different. Let us assume, w.l.o.g., that the points $\mathbf{q}_{i2}$ and $\mathbf{q}_{i3}$ are observed in camera $\mathcal{G}_1$, *i.e.*, $i = 1$. This is the same camera that observed the point $\mathbf{q}_{11}$. The remaining two points can be observed by one camera $\mathcal{G}_j \neq \mathcal{G}_1$ or by two different cameras $\mathcal{G}_j \neq \mathcal{G}_k \neq \mathcal{G}_1$. In this case, G-J elimination of the matrix $\mathtt{C}$ in (10) leads to a matrix of a special form

$$\begin{bmatrix} \mathtt{I}_{6\times6} & \mathtt{0}_{6\times2} & \mathbf{0}_{6\times1} & \mathbf{c}_{6\times1} \\ \mathtt{0}_{2\times6} & \mathtt{I}_{2\times2} & \mathbf{d}_{2\times1} & \mathbf{e}_{2\times1} \end{bmatrix} \mathbf{b} = \mathbf{0} \ , \tag{11}$$

where the indices of the matrices and vectors indicate their sizes. Since $\mathbf{b} = [g'_{11}, g'_{12}, g'_{21}, g'_{22}, g'_{31}, g'_{32}, m'_1, m'_2, m'_3, 1]^\top$, the first six rows of (11) directly give us a solution to g'_{kl}. The last two rows can be used to express m'_1, m'_2 as a linear function of m'_3. Substituting this parameterization into the generators of I'_2 yields five univariate polynomials, each of degree three. We can obtain up to three real solutions to ε' by solving one of these polynomials. The remaining steps are similar to the $s\mathtt{H5}_2$ solver.

$\mathtt{sH5_4}$ solver: In this case, 4 points come from the same camera $\mathcal{G}_i$. Let us assume, w.l.o.g., that the points $\mathbf{q}_{i2}$, $\mathbf{q}_{i3}$ and $\mathbf{q}_{i4}$ are observed in camera $\mathcal{G}_1$, *i.e.*, $i = 1$ and the fifth point $\mathbf{q}_{25}$ is observed by camera $\mathcal{G}_2 \neq \mathcal{G}_1$. We estimate the semi-generalized homography $\mathtt{G}$ by considering it as standard homography estimation problem from 4 point correspondences [20]. Decomposing $\mathtt{G} = \mathtt{R} - \mathbf{t}\mathbf{m}^\top$ we obtain the rotation matrix $\mathtt{R}$ and the translation vector $\mathbf{t}$ up to scale. This scale can be computed from the constraint of the form (4), induced by the fifth correspondence $\mathbf{p}_5 \leftrightarrow \mathbf{q}_{25}$, as

$$\text{scale} = \frac{\mathbf{t}_{\mathtt{G}_2}^\top [\mathtt{R}\mathbf{p}_5]_\times \mathbf{q}_{25}}{\mathbf{t}^\top [\mathtt{R}\mathbf{p}_5]_\times \mathbf{q}_{25}} \ . \tag{12}$$

2.4. Unknown Focal Length Solvers

In this case, we assume an unknown focal length f in the calibration matrix $\mathtt{K} = \text{diag}(f, f, 1)$. Therefore we have 10 DOF and we need five full 2D-2D correspondences to solve this problem. Based on the maximum number of correspondences coming from one camera $\mathcal{G}_i$, we have two solvers: one where there are up to 2 points from the same camera, $\mathtt{sH5f_2}$, and one where there are 3 points from the same camera, $\mathtt{sH5f_3}$. These solvers solve univariate polynomials of degree 5, respectively 3, and they follow similar steps as those of the calibrated solvers $\mathtt{sH5_2}$ and $\mathtt{sH5}_3$. For more details on these solvers see the SM. Note that for the case of 4 points coming from the same camera $\mathcal{G}_i$, one additional correspondence from camera $\mathcal{G}_j$ will not add sufficient constraints to recover both the unknown focal length and the unknown scale, which is proved in the SM.

3. Experiments

This section studies the performance of the proposed solvers, $\mathtt{sH5_2}$, $\mathtt{sH5_3}$, $\mathtt{sH4.5_2}$, $\mathtt{sH4.5_3}$, $\mathtt{sH5_4}$, $\mathtt{sH5f_2}$, and $\mathtt{sH5f_3}$, both on synthetic and real-world images. For comparison, we use four state-of-the-art minimal solvers for estimating the semi-generalized epipolar geometry [58], *i.e.*, the $\mathtt{E_{5+1}}$, $\mathtt{E_{4+2}}$, $\mathtt{Ef_{6+1}}$, and $\mathtt{Ef_{5+2}}$ solvers. Note, that for the experiments where we do not need the scale of the translation, $\mathtt{E_{5+1}}$ reduces to the well-known 5pt solver $\mathtt{E5}$ [35] and the $\mathtt{Ef_{6+1}}$ problem reduces to the one-sided focal length 6pt solver $\mathtt{E6f}$ [10]. In such experiments we also consider the 4pt homography solver $\mathtt{H4}$ [20]. We excluded the $\mathtt{Ef_{5+2}}$ solver from real experiments since it was too slow when used on large datasets inside RANSAC [16]. For a fair comparison, we further consider P3P/P5Pf + **N** solvers designed for our semi-generalized homography setup: we first use three correspondences between two calibrated generalized cameras $\mathcal{G}_i$ and $\mathcal{G}_j$ to estimate the normal $\mathbf{n}$ and intercept d of the observed plane π. The plane is then used to lift 3/5 2D-2D matches between the perspective camera $\mathcal{P}$ and arbitrary cameras in the generalized camera $\mathcal{G}$ to 2D-3D correspondences. Finally, the pose of the perspective camera is computed using the P3P [36] or P5Pf [26] solvers[2]. Note that these P3P/P5Pf + **N** solvers, compared to the proposed solvers, require point matches between two cameras $\mathcal{G}_i$ and $\mathcal{G}_j$ in the generalized camera. However, these correspondences are used only in the first step, when estimating the normal and plane intercept. Thus, they do not need to be visible in $\mathcal{P}$ as in the standard 2D-3D pipeline. We are

[2] In the unknown focal length case, we use the non-minimal P5Pf solver since the available implementations of the minimal P4Pf one were either much slower [9] than the P5Pf solver or did not work for planar scenes [28].

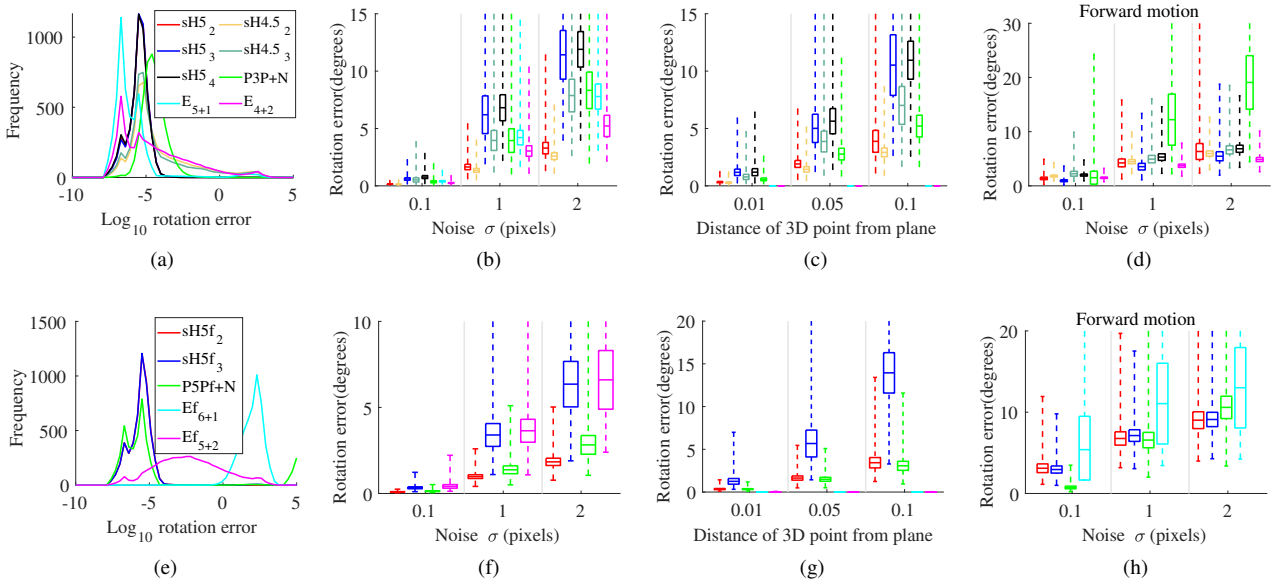

Figure 2. **Top**: solvers for calibrated cameras. **Bottom**: solvers for partially calibrated cameras: (a,e) numerical stability, (b,f) performance in the presence of image noise, (c,g) close-to-planar scenes, (d,h) forward motion in the presence of image noise.

not aware that such solvers have been used in the literature.

3.1. Synthetic Scenes

We study the performance of our proposed solvers on synthetically generated 3D scenes with known ground truth. The 3D points are randomly distributed on a plane of size 10×10. Each 3D point is projected into up to six cameras with realistic focal lengths. Five of these cameras represent the generalized camera $\mathcal{G}$ and one camera is considered as the camera $\mathcal{P}$. The orientations and positions of the cameras are selected at random such that they roughly look towards the scene from a random distance, varying from 20 to 35, from the plane. The simulated images have a resolution of 1000×1000 px. Here, we focus on the errors in the estimated rotations $\mathtt{R}$ for the calibrated and unknown focal length solvers. The rotation error is computed as the angle in the axis-angle representation of $\mathtt{R}_{GT}^{-1}\mathtt{R}$, where $\mathtt{R}_{GT}$ is the ground truth and $\mathtt{R}$ is the estimated rotation. Plots for the position and focal length errors can be found in the SM.

Numerical stability. We measure the numerical stability of the solvers by evaluating 5k camera setups for planar scenes. We compare the accuracy of the rotations estimated by the proposed solvers $\mathtt{sH5_2}$, $\mathtt{sH4.5_2}$, $\mathtt{sH5_3}$, $\mathtt{sH4.5_3}$, and $\mathtt{sH5_4}$ with that of $\mathtt{E_{5+1}}$, $\mathtt{E_{4+2}}$, and P3P + N in Fig. 2(a). Our solvers, $\mathtt{sH5_2}$, $\mathtt{sH5_3}$, and $\mathtt{sH5_4}$, achieve better stability with fewer failures (*i.e.*, no peak on the right side). $\mathtt{sH4.5_2}$ and $\mathtt{sH4.5_3}$ have comparable stability as the other solvers. Fig. 2(e) compares the numerical stability of $\mathtt{sH5f_2}$ and $\mathtt{sH5f_3}$ with that of the solvers $\mathtt{Ef_{6+1}}$, $\mathtt{Ef_{5+2}}$, and P5Pf + N. Note that a planar scene is a degenerate configuration for the $\mathtt{Ef_{6+1}}$ solver, which explains the reported performance.

Image noise. Next, we test the performance of all solvers in the presence of Gaussian noise with standard deviation σ, added to the image points in all cameras. Fig. 2(b,f) show the rotation error (in degrees) for solvers for calibrated (b) as well as partially calibrated (f) cameras. Here, we depict the results as box plots which show the 25% to 75% quantile values as boxes with a horizontal line for the median. We note that our proposed solvers $\mathtt{sH5_2}$, $\mathtt{sH4.5_2}$, and $\mathtt{sH5f_2}$ have better or comparable performance than the competing solvers in the presence of image noise. Moreover, we observe that $\mathtt{sH4.5_2}$ and $\mathtt{sH4.5_3}$ are more stable than $\mathtt{sH5_2}$ and $\mathtt{sH5_3}$, respectively, in the presence of image noise.

Close-to-planar scenes. We also consider the case where the scene is close to being entirely planar by placing the scene plane at $z = 0$ and sampling 3D points with varying plane-to-point distances. Fig. 2(c,g) show the rotation errors for calibrated (c) and partially calibrated (g) cameras. Our $\mathtt{sH5_2}$ and $\mathtt{sH4.5_2}$ solvers are more accurate than P3P + N while $\mathtt{sH5f_2}$ has comparable stability to P5Pf + N for close-to-planar scenes. As expected, the accuracy of the proposed solvers deteriorates with the increasing non-planarity of the scene. However, the errors, even for larger non-planarity, are comparable to the errors obtained by all general solvers in the presence of 2 px image noise.

Forward motion with image noise. Figures 2(d,h) show the rotation error for calibrated (d) and partially calibrated (h) cameras. Our solvers for calibrated cameras have similar or better stability than the competing solvers. For the unknown focal length case, our proposed solvers lead to similar rotation estimates than the competing solvers. We note that in case of a pure forward motion, the solver $\mathtt{Ef_{5+2}}$

either failed or led to very unstable results. As a result of this, we have not considered the solver $\texttt{Ef}_{\mathbf{5+2}}$ in the graphs.

3.2. Computational Complexity

Tab. 1 reports the computational complexity of the studied solvers. Since we do not have equally efficient C++ implementations of all solvers (some solvers are highly optimized, *e.g.*, $\texttt{E}\mathbf{5}$ and P3P from the PoseLib [30] library, while some do not contain any special optimization, *e.g.*, $\texttt{E}_{\mathbf{4+2}}$, $\texttt{Ef}_{\mathbf{5+2}}$ [58]), we compare only the most time consuming operations performed by these solvers. We thus focus on the matrix size for each critical matrix operation.

3.3. Real-World Experiments

Our solvers built on the fact that man-made environments frequently contain planes and planar structures both indoors and outdoors. To show the usefulness of our solvers in real applications like visual odometry and visual localization, we test them on general real-world data. Such general scenarios will of course give an advantage to our competitors, *e.g.*, [35], that work for planar as well as non-planar scenes. Yet, we show that our new proposed solvers return comparable pose and focal length estimates and sometimes even outperform the state-of-the-art general solvers. As such, we believe that our solvers can be combined with existing ones in a hybrid RANSAC [12], where the most suitable solver is selected for each scene in a data-dependent manner.

Localization experiment. We evaluate all variants of our $\texttt{sH}\mathbf{5}$ and $\texttt{sH}\mathbf{4.5}$ solvers for calibrated cameras in the context of visual localization. We use the subset of scenes from the Cambridge Landmarks dataset [23] commonly used in the literature [42]. Note that while these scenes contain one or more dominant planes, none of them is perfectly planar.

Our $\texttt{sH}\mathbf{5}$ and $\texttt{sH}\mathbf{4.5}$ and the $\texttt{E}_{\mathbf{4+2}}$, and $\texttt{E}_{\mathbf{5+1}}$ solvers enable a particularly light-weight type of structure-less localization pipelines that do not need to store a 3D model. Such representations can be easily maintained [53]. In contrast to P3P + **N** and SfM-on-the-fly [53], our solvers only need matches between the pinhole image and the generalized camera images but not within images in the generalized camera. This keeps feature matching to a minimum. We implement such a pipeline by using DenseVLAD-based image retrieval [52] to identify the 10 reference images most similar to a given query. The generalized camera is then defined using the known poses of the retrieved images.

We integrate our solvers into RANSAC with local optimization (LO-RANSAC) [32, 40]. In each iteration, we simply randomly sample 5 matches from all matches found with the retrieved images. We then select the most suitable solver for this sample, *e.g.*, $\texttt{sH}\mathbf{5_4}$ if four matches come from the same reference image. This approach is possible thanks to the fact that our solvers cover all possible combinations of 5 point correspondences. However, this approach is not suitable for the $\texttt{E}_{\mathbf{4+2}}$ and $\texttt{E}_{\mathbf{5+1}}$ solvers as the chance of randomly sampling 4 or more matches from the same reference image is very small. Instead, we first randomly select two ($\texttt{E}_{\mathbf{5+1}}$) or three ($\texttt{E}_{\mathbf{4+2}}$)[3] retrieved reference images. We then randomly select the required matches from these images. This sampling scheme is incompatible with RANSAC's standard stopping criterion. For a fair comparison, we thus run LO-RANSAC for each solver for a fixed number of iterations. The best model found by LO-RANSAC is refined over all inliers (see SM for details).

As shown in Tab. 2, our solvers outperform the $\texttt{E}_{\mathbf{4+2}}$ solver. They are consistently among the top-2 approaches based on mean / median position and orientation errors and lead to the fastest RANSAC times. Averaged over all datasets, our $\texttt{sH}\mathbf{4.5}$ solvers lead to the same median results as the $\texttt{E}_{\mathbf{5+1}}$ solver at faster run-times. The results clearly show the usefulness of our solvers. In particular, our results point towards an interesting research direction: our faster solvers can be used to quickly estimate the inlier ratios for each reference image. This can then be used for guided sampling of image pairs for the $\texttt{E}_{\mathbf{5+1}}$ solver, *e.g.*, inside a hybrid RANSAC scheme[4]. This approach should deliver the best from both types of solvers.

Tab. 2 also includes the Sift+5pt approach [59,60], which estimates the relative pose between the query and retrieved images based on SIFT feature matches and essential matrix estimation. The relative poses and the known absolute poses of the retrieved images are then used to estimate the query pose. Our approach consistently outperforms [59, 60].

Relative pose experiments. We use the 11 sequences of the KITTI benchmark [18] that are provided with the ground truth trajectories (23, 190 image pairs). In KITTI, the scenes are captured by two front-facing cameras mounted to a moving vehicle. We consider the camera pair as the generalized one and estimate the relative pose between this camera and the left image of the next frame.

As robust estimator, we use GC-RANSAC [5] that applies two different solvers: (a) one for estimating the pose from a minimal sample and (b) one for fitting to a larger-than-minimal sample when polishing the model parameters on a set of inliers. We included the compared solvers in step (a). For step (b), we applied numerical optimization [1], minimizing the Sampson distance when estimating the essential matrix, and a re-projection error when estimating the homography. Moreover, we test recovering the pose by combining the essential matrix from $\texttt{E}_{4+2}$ and multiple homographies [6] either from $\texttt{H}4$ or the proposed $\texttt{sH}\mathbf{5_3}$. The resulting pose is found by decomposing the essential matrix and homographies, and selecting the pose that has the

[3] Note that two of the three images can be identical.

[4] The hybrid RANSAC formulation from [12] deals with two sources of matches and cannot be easily extended to more sources (each retrieved reference images represents a source with its own inlier ratio).

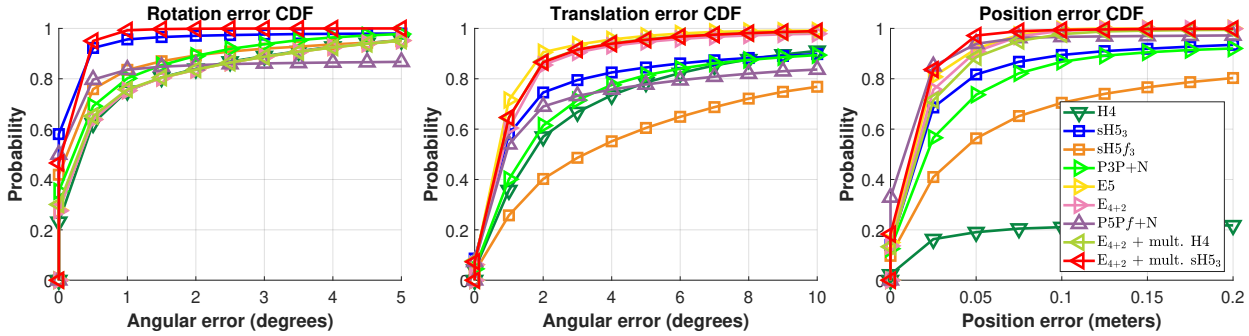

Figure 3. The CDFs of the rotation, translation (degrees) and position (meters) errors on 23,190 image pairs from the KITTI dataset. More accurate methods are closer to the top-left corner. Since most tested methods do not return the translation scale due to estimating the relative pose, we used the scale from the ground truth path to calculate the position error. Tab. 3 shows the corresponding error values.

Method	King's College			Old Hospital			Shop Facade			St. Mary Church			Avrg. all	
	pos.	rot.	time	pos.	rot.	time	pos.	rot.	time	pos.	rot.	time	pos.	rot.
E$_{5+1}$ [58] (100 iter.)	**0.20** / **0.44**	**0.36** / **0.61**	0.29	**0.54** / **1.30**	**1.02** / **2.12**	0.15	**0.06** / **0.10**	**0.33** / **0.46**	0.15	**0.13** / **0.20**	**0.51** / **0.73**	0.18	**0.23**	**0.56**
E$_{4+2}$ [58] (100 iter.)	0.25 / 1.58	0.42 / 1.70	0.26	1.51 / 56.2	2.82 / 6.90	0.15	0.09 / 2.77	0.44 / 3.28	0.14	0.41 / 242.4	1.42 / 5.62	0.17	0.57	1.28
ours (sH5) (100 iter.)	**0.22** / 0.71	0.39 / 1.20	**0.20**	0.88 / **2.20**	**1.68** / 3.98	**0.08**	0.09 / 0.78	0.43 / 2.23	**0.08**	0.25 / 2.52	0.95 / 6.50	**0.11**	**0.36**	**0.86**
ours (sH4.5) (100 iter.)	**0.20** / **0.32**	**0.33** / **0.49**	**0.23**	**0.51** / 48.4	**1.02** / **3.15**	**0.12**	**0.07** / **0.14**	**0.32** / **0.67**	**0.11**	**0.15** / **0.30**	**0.52** / **1.24**	**0.14**	**0.23**	**0.55**
E$_{5+1}$ [58] (1k iter.)	**0.19** / 0.33	**0.34** / **0.48**	0.82	**0.42** / **1.10**	**0.75** / **1.78**	0.39	**0.06** / **0.10**	**0.29** / **0.44**	0.37	**0.11** / **0.15**	**0.38** / **0.55**	0.49	**0.20**	**0.44**
E$_{4+2}$ [58] (1k iter.)	**0.20** / 0.42	0.35 / 0.60	0.75	0.83 / 2.51	1.55 / 3.88	0.56	**0.07** / **0.16**	**0.32** / 0.70	0.53	0.20 / 0.70	0.71 / 2.17	0.59	0.33	0.73
ours (sH5) (1k iter.)	**0.20** / **0.31**	**0.34** / **0.48**	**0.33**	0.46 / **1.03**	0.89 / 2.47	**0.16**	**0.06** / **0.10**	**0.29** / **0.45**	**0.16**	0.13 / 0.43	0.47 / 1.35	**0.20**	**0.21**	**0.50**
ours (sH4.5) (1k iter.)	**0.19** / **0.30**	**0.33** / **0.46**	**0.52**	**0.40** / 1.21	**0.74** / **1.91**	**0.27**	**0.06** / **0.10**	**0.29** / **0.44**	**0.26**	**0.12** / **0.17**	**0.40** / **0.59**	**0.33**	**0.20**	**0.44**
Sift+5Pt [59, 60]	0.48 / -	1.13 / -	-	0.88 / -	1.91 / -	-	0.17 / -	0.99 / -	-	0.35 / -	1.58 / -	-	0.47	0.88

Table 2. Localization results on Cambridge Landmarks [23]. We report the median/mean position (in meters) and rotation (in degrees) errors, and the mean RANSAC time (in seconds). We also report the average median position and rotation error over all four scenes. We show results for fixing the number of RANSAC iterations to 100 respectively 1000. Best and second best results are shown in red and blue.

Method	$\epsilon_\mathbf{R}$ (°)	$\epsilon_\mathbf{t}$ (°)	$\epsilon_\mathbf{p}$ (m)	$\epsilon_\mathbf{f}$ (px)
sH5$_3$	**0.21** / **0.45**	1.31 / 4.51	**0.03** / 0.10	–
H4 [20]	0.52 / 1.19	2.06 / 4.34	1.33 / 1.28	–
P3P + N	0.39 / 0.91	1.84 / 6.43	**0.03** / 0.10	–
E5 [48]	0.49 / 1.18	**1.21** / **1.73**	**0.02** / **0.03**	–
E$_{4+2}$ [58]	0.49 / 1.19	1.28 / 2.56	**0.02** / **0.03**	–
E$_{4+2}$ + mult. H4	0.45 / 1.15	**1.24** / 2.16	**0.03** / 0.04	–
E$_{4+2}$ + mult. sH5$_3$	0.27 / **0.32**	**1.24** / **1.84**	**0.02** / **0.02**	–
sH5f_3	**0.32** / **1.09**	**3.71** / **10.55**	**0.05** / **0.21**	**18.16** / **349.36**
P5Pf + N	**0.25** / **11.94**	**1.41** / **10.80**	**0.02** / **0.05**	**40.66** / **>10^6**

Table 3. Rotation, translation (degrees), position (meters) and focal length errors (pixels) on 23k image pairs from KITTI. Best and second best results are shown in red and blue. Since most tested methods do not return the scale of the translation due to estimating the relative pose, we used the scale from the ground truth path to calculate the position error. Fig. 3 shows the corresponding CDFs.

largest support when thresholding the re-projection error.

Tab. 3 reports the median / mean rotation, position, and focal length errors on the 23,190 image pairs. Fig. 3 shows the corresponding CDFs. Since some of the tested methods, *e.g.*, **E5**, do not return the translation scale due to estimating the relative pose, we used the scale from the ground truth path to calculate the position error. Even though the proposed **sH5$_3$** finds the most accurate rotation matrices, its translation and position errors are marginally higher than those of the essential matrix-based solvers. Using **E$_{4+2}$** and multiple homographies from **sH5$_3$**, however, leads to the most accurate poses. Amongst the solver estimating the focal length, the proposed **sH5f$_3$** solver is the most accurate one. We did not include **Ef$_{5+2}$** and **Ef$_6$** since both fail when the camera undergoes purely forward motion.

4. Conclusion

In this paper, we have considered the problem of estimating the semi-generalized homography between a pinhole and a generalized camera. We have proposed efficient solvers handling both calibrated and partially calibrated pinhole cameras with unknown focal length. Our solvers cover all possible minimal combinations of point correspondences between the pinhole and the generalized camera where it is possible to recover the scale. To the best of our knowledge, we are the first to solve this problem. Synthetic and real experiments focusing on two real-world applications show that our solvers are practically relevant. While they may not outperform more general existing solvers, which handle non-planar scenes, under all conditions, our results show that our solvers are preferable in certain conditions. Combining all these solvers into a single hybrid RANSAC approach is thus an interesting direction for future work.

Acknowledgements. This paper was funded by the OP VVV funded project CZ.02.1.01/0.0/0.0/16_019/0000765 "Research Center for Informatics", the EU Horizon 2020 project RICAIP (No 857306) and the European Regional Development Fund under project IMPACT (No. CZ.02.1.01/0.0/0.0/15_003/0000468).

References

[1] Sameer Agarwal, Keir Mierle, and Others. Ceres solver. http://ceres-solver.org. 7

[2] Daniel Barath. Five-point fundamental matrix estimation for uncalibrated cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 235–243, 2018. 1

[3] Daniel Barath and Levente Hajder. A theory of point-wise homography estimation. *Pattern Recognition Letters*, 94:7–14, 2017. 1

[4] Daniel Barath and Zuzana Kukelova. Homography from two orientation-and scale-covariant features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1091–1099, 2019. 1

[5] Daniel Barath and Jiří Matas. Graph-cut RANSAC. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 7

[6] Daniel Barath and Jiri Matas. Progressive-x: Efficient, anytime, multi-model fitting algorithm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3780–3788, 2019. 7

[7] Eric Brachmann and Carsten Rother. Expert sample consensus applied to camera re-localization. In *ICCV*, 2019. 1

[8] Matthew Brown, Richard I Hartley, and David Nistér. Minimal solutions for panoramic stitching. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 1

[9] M. Bujnak, Z. Kukelova, and T. Pajdla. A general solution to the p4p problem for camera with unknown focal length. In *CVPR*, 2008. 5

[10] Martin Bujnak, Zuzana Kukelova, and Tomas Pajdla. 3d reconstruction from image collections with a single known focal length. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1803–1810. IEEE, 2009. 2, 5

[11] Martin Byröd, Matthew A Brown, and Kalle Åström. Minimal solutions for panoramic stitching with radial distortion. In *British Machine Vision Conference (BMVC)*, 2009. 1

[12] Federico Camposeco, Andrea Cohen, Marc Pollefeys, and Torsten Sattler. Hybrid camera pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 136–144, 2018. 7

[13] Ondřej Chum and Jiří Matas. Planar affine rectification from change of scale. In *Asian Conference on Computer Vision (ACCV)*, 2010. 4

[14] David A. Cox, John Little, and Donal O'shea. *Using algebraic geometry*, volume 185. Springer Science & Business Media, 2006. 3

[15] Yaqing Ding, Jian Yang, Jean Ponce, and Hui Kong. An efficient solution to the homography-based relative pose problem with a common reference direction. In *International Conference on Computer Vision (ICCV)*, 2019. 1

[16] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 5

[17] Andrew Fitzgibbon. Simultaneous linear estimation of multiple view geometry and lens distortion. In *Computer Vision and Pattern Recognition (CVPR)*, 2001. 1

[18] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *IJRR*, 32(11):1231–1237, 2013. 7

[19] Daniel R. Grayson and Michael E. Stillman. Macaulay2, a software system for research in algebraic geometry. Available at http://www.math.uiuc.edu/Macaulay2/. 3

[20] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1, 5, 8

[21] D. Hook and P. R. McAree. Using sturm sequences to bracket real roots of polynomial equations. *Graphics gems*, pages 416–422, 1990. 4

[22] Hailin Jin. A three-point minimal solution for panoramic stitching with lens distortion. In *Computer Vision and Pattern Recognition (CVPR)*, 2008. 1

[23] Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In *ICCV*, 2015. 7, 8

[24] Kevin Köser. *Geometric Estimation with Local Affine Frames and Free-form Surfaces*. Shaker, 2009. 1

[25] Zuzana Kukelova, Martin Bujnak, and Tomas Pajdla. Automatic generator of minimal problem solvers. In *European Conference on Computer Vision*, pages 302–315. Springer, 2008. 4

[26] Zuzana Kukelova, Martin Bujnak, and Tomas Pajdla. Real-time solution to the absolute pose problem with unknown radial distortion and focal length. In *2013 IEEE International Conference on Computer Vision*, pages 2816–2823, 2013. 5

[27] Zuzana Kukelova, Jan Heller, Martin Bujnak, and Tomas Pajdla. Radial distortion homography. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[28] Zuzana Kukelova, Jan Heller, and Andrew Fitzgibbon. Efficient intersection of three quadrics and applications in computer vision. In *Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 5

[29] Zuzana Kukelova, Joe Kileel, Bernd Sturmfels, and Tomas Pajdla. A clever elimination strategy for efficient minimal solvers. In *Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 4, 2017. 2, 3, 4

[30] Viktor Larsson. PoseLib - Minimal Solvers for Camera Pose Estimation. https://github.com/vlarsson/PoseLib, 2020. 7

[31] Viktor Larsson, Kalle Åström, and Magnus Oskarsson. Efficient solvers for minimal problems by syzygy-based reduction. In *CVPR*, volume 2, page 4, 2017. 4

[32] Karel Lebeda, Juan E. Sala Matas, and Ondřej Chum. Fixing the Locally Optimized RANSAC. In *British Machine Vision Conference (BMVC)*, 2012. 7

[33] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 1

[34] Raul Mur-Artal and Juan D Tardós. ORB-SLAM2: An open-source slam system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017. 1

[35] David Nistér. An efficient solution to the five-point relative pose problem. *IEEE transactions on pattern analysis and machine intelligence*, 26(6):756–770, 2004. 2, 5, 7

[36] Mikael Persson and Klas Nordberg. Lambda twist: An accurate fast robust perspective three point (p3p) solver. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 5

[37] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object Retrieval with Large Vocabularies and Fast Spatial Matching. In *CVPR*, 2007. 1

[38] Robert Pless. Camera cluster in motion: motion estimation for generalized camera designs. *IEEE Robotics & Automation Magazine*, 11(4):39–44, 2004. 1

[39] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[40] Torsten Sattler et al. RansacLib - A Template-based *SAC Implementation, 2019. 7

[41] T. Sattler, B. Leibe, and L. Kobbelt. Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. *PAMI*, 39(9):1744–1756, 2017. 1

[42] Torsten Sattler, Qunjie Zhou, Marc Pollefeys, and Laura Leal-Taixe. Understanding the limitations of cnn-based absolute camera pose regression. In *CVPR*, 2019. 7

[43] Olivier Saurer, Pascal Vasseur, Rémi Boutteau, Cédric Demonceaux, Marc Pollefeys, and Friedrich Fraundorfer. Homography based egomotion estimation with a common direction. *IEEE transactions on pattern analysis and machine intelligence*, 39(2):327–341, 2017. 1

[44] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016. 1

[45] Thomas Schöps, Viktor Larsson, Marc Pollefeys, and Torsten Sattler. Why Having 10,000 Parameters in Your Camera Model Is Better Than Twelve. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[46] Noah Snavely, Steve Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM transactions on graphics (TOG)*, volume 25, pages 835–846. ACM, 2006. 1

[47] Eric Stenborg, Torsten Sattler, and Lars Hammarstrand. Using Image Sequences for Long-Term Visual Localization. In *3DV*, 2020. 1

[48] Henrik Stewénius, David Nistér, Fredrik Kahl, and Frederik Schaffalitzky. A minimal solution for relative pose with unknown focal length. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 789–794. IEEE, 2005. 8

[49] Henrik Stewénius, D. Nistér, M. Oskarsson, and K. Åström. Solutions to minimal generalized relative pose problems. 2005. 2

[50] Chris Sweeney, Laurent Kneip, Tobias Höllerer, and Matthew Turk. Computing similarity transformations from only image correspondences. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3305–3313, 2015. 1

[51] Chris Sweeney, Torsten Sattler, Tobias Hollerer, Matthew Turk, and Marc Pollefeys. Optimizing the viewing graph for structure-from-motion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 801–809, 2015. 1

[52] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 Place Recognition by View Synthesis. In *CVPR*, 2015. 7

[53] A. Torii, H. Taira, J. Sivic, M. Pollefeys, M. Okutomi, T. Pajdla, and T. Sattler. Are Large-Scale 3D Models Really Necessary for Accurate Visual Localization? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019. 7

[54] Johanna Wald, Torsten Sattler, Stuart Golodetz, Tommaso Cavallari, and Federico Tombari. Beyond Controlled Environments: 3D Camera Re-Localization in Changing Indoor Scenes. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[55] T. Weyand and B. Leibe. Discovering Favorite Views of Popular Places with Iconoid Shift. In *ICCV*, 2011. 1

[56] Changchang Wu. Towards linear-time incremental structure from motion. In *2013 International Conference on 3D Vision-3DV 2013*, pages 127–134. IEEE, 2013. 1

[57] Zhengyou Zhang. A Flexible New Technique for Camera Calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 22:1330–1334, December 2000. 1

[58] Enliang Zheng and Changchang Wu. Structure from motion using structure-less resection. In *International Conference on Computer Vision (ICCV)*, 2015. 1, 2, 5, 7, 8

[59] Qunjie Zhou, Torsten Sattler, Marc Pollefeys, and Laura Leal-Taixe. To Learn or Not to Learn: Visual Localization from Essential Matrices. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019. 7, 8

[60] Qunjie Zhou, Torsten Sattler, Marc Pollefeys, and Laura Leal-Taixe. To Learn or Not to Learn: Visual Localization from Essential Matrices. *arXiv:1908.01293v1*, 2020. 7, 8