What Correspondences Reveal about Unknown Camera and Motion Models?

Thomas Probst[1], Ajad Chhatkuli[1], Danda Pani Paudel[1], Luc Van Gool[1,2]

[1]Computer Vision Laboratory, ETH Zurich, Switzerland
[2]VISICS, ESAT/PSI, KU Leuven, Belgium

Abstract

In two-view geometry, camera models and motion types are used as key knowledge along with the image point correspondences in order to solve several key problems of 3D vision. Problems such as Structure-from-Motion (SfM) and camera self-calibration are tackled under the assumptions of a specific camera projection model and motion type. However, these key assumptions may not be always justified, i.e.*, we may often know neither the camera model nor the motion type beforehand. In that context, one can extract only the point correspondences between images. From such correspondences, recovering two-view relationship – expressed by the unknown camera model and motion type– remains to be an unsolved problem. In this paper, we tackle this problem in two steps. First, we propose a method that computes the correct two-view relationship in the presence of noise and outliers. Later, we study different possibilities to disambiguate the obtained relationships into camera model and motion type. By extensive experiments on both synthetic and real data, we verify our theory and assumptions in practical settings.*

1. Introduction

Structure-from-Motion (SfM) [14, 17, 21] and a vast majority of 3D vision applications rely on feature point correspondences between images, while assuming a known camera model. A key component of SfM and other 3D vision methods is consensus maximization [10, 7, 28, 2, 13, 3, 15, 31] where the correct set of inliers is computed by searching for specific relationships between image correspondences. Such approaches have made possible the use of millions of images for reconstructing 3D, as unreliable point correspondences are weeded out through maximizing the consensus. However, this is only true when one is certain about the most suitable two-view relationship. In fact, methods developed to tackle many 3D vision problems can be used only for certain camera and motion types. For example, current state-of-the-art SfM methods assume that the camera obeys a perspective projection model and that the camera motion involves at least some translation [24, 21, 14]. Similarly, existing camera calibration methods (self-calibration or with known patterns) assume a known camera model [30, 6], and sometimes also a known motion type [12]. In such cases, experiments using images from an affine camera when the method assumes a perspective model are doomed to fail. It is therefore essential to know the camera model and the motion type beforehand, apart from the two-view relationship. In fact, one requires to know the motion (if not its type) to reason about the camera model, and vice versa. Therefore, jointly recovering the camera model and motion becomes very difficult, leading to the causality dilemma.

In that context, the problem can be broken down into two important sub-problems. The first is that of choosing the correct two-view relationship without knowing the camera type and motion beforehand. Fitting an unknown model to point correspondences is a challenging problem and is in general NP hard. On the one hand, one can only hope to know the correct two-view relationship after trying out all possible models. This is exactly what is done in the context of SfM for the Fundamental matrix and the 2D projective homography as geometric verification [27, 25, 16]. However, it is not always clear how one should select among the several models for a given problem even after trying out all of them. For example, we can always fit a relationship less constrained than the actual model in order to obtain a higher inlier count. Therefore, simply choosing the model with the largest inlier set may lead to an undesired outcome. Additionally, there exists a natural conflict between the desired camera motion baseline and the matching performance in real images [18, 1]. In such cases, one can benefit by constraining the motion to be sparse [23, 11]. However, it may not be known which motions are absent beforehand. The second problem is that of obtaining the camera model and the motion type from the correctly recovered two-view relation. Although many key results are

already known [9, 17, 14, 26], we seek to answer and summarize a different question, *i.e.*, what is the correct camera model and motion type given a two-view relationship?

In this paper, we provide contributions to the two problems discussed above: *i)* computing the two-view polynomial relationship with point correspondences when the exact type of relationship is not known and *ii)* disambiguating the camera model and motion type from the two-view relationship. In order to tackle the first problem, we present a unified approach to fitting polynomials to image point correspondences despite the unknown model type and outliers. To that end, we define the model search as that of finding the sparsest set of polynomials which agree with more than half the point correspondences up to some small fixed error. We constrain the polynomials by using a basis of monomials where the solution is known to exist. This is done using the so-called Vandermonde matrix [4, 5]. In doing so we solve for both the model parameters as well as the inlier correspondences similar to that of Random Sampling and Consensus (RANSAC) [10]. Unlike in RANSAC, we can recover multiple polynomials that describe a given set of measurements (correspondences) by iteratively searching for the orthogonal set of sparse bases of the monomial coefficients. Additionally, we also encourage the motion to be sparse, constraining the model search better for small baselines. Our approach only requires one to know beforehand the maximum degree of the polynomial and not the actual number of the polynomials. We express the problem as a Mixed Integer Program (MIP) and solve it using the Branch and Bound (BnB) approach. Our second contribution is the analysis of the camera model and the motion type using the computed two-view relationship. We consider each camera and motion type and analyze the resulting two-view relationship. We provide conditions when such camera model and motion recovery are ambiguous and why.

In order to quantify the model fitting accuracy, we evaluate the proposed method on both synthetic and real data. In the synthetic case, we simulate image point correspondences with outliers for various camera models and motion types. On average our method performs better than RANSAC for unknown model fitting as well as inlier-outlier classification. We also show that our method performs similarly to RANSAC with known camera and motion model on the same tasks. We use the real data in order to show motion disambiguation on driving sequences demonstrating the importance of such disambiguation in a practical scenario.

2. Preliminaries

Notations. We denote matrices with upper case letters and their elements with double-indexed lower case letters: $\mathsf{A} = (a_{ij})$. Similarly, we write vectors and index them as: $\mathsf{a} = (a_i)$. We use special uppercase Latin or uppercase Greek letters for sets such as $\mathcal{P}$. We use lowercase Latin letters for scalars as in a. Finally, we use $\sigma_i(.)$ for a function which gives the i-th largest singular value. We write the ring of polynomials parameterized by variables $\mathsf{x} \in \mathbb{R}^n$ as $\mathcal{R}[\mathsf{x}]$. A polynomial $p(\mathsf{x}) \in \mathcal{R}[\mathsf{x}]$ is represented using the basis of coefficients $\mathcal{B}$. We use $\|\mathsf{v}\|_p$ to denote the ℓ–p norm of any vector v.

2.1. Problem Formulation

We consider two cameras related by a motion. Let there be m point correspondences $\{\mathsf{u}_i, \mathsf{v}_i\}_{i=1}^m$ between the images of the two cameras. Then we are interested to solve the following problem.

Problem 2.1 *What are the camera model $\mathcal{M}$ and motion parameters θ for the image point correspondences $\{\mathsf{u}_i, \mathsf{v}_i\}_{i=1}^m$?*

This problem is NP-hard and difficult to solve in its current form. Therefore, we make the following assumption to search the camera model and motion parameters.

Assumption 2.2 *The optimal answer to Problem 2.1,* i.e.*, $(\mathcal{M}, \theta)$ respects the point correspondences $\{\mathsf{u}_i, \mathsf{v}_i\}_{i=1}^m$ and minimizes the joint degrees of freedom of $\mathcal{M}$ and θ.*

We represent both $\mathcal{M}$ and θ using polynomials, as commonly done in the literature [9, 14, 22]. In this regard, we express Problem 2.1 under the Assumption 2.2 as an algebraic problem of finding a low dimensional variety, whose samples are the point correspondences.

Consider the ring $\mathcal{R}[\mathsf{x}] := \mathcal{R}[x_1, ..., x_n]$ of multivariate polynomials and an algebraic variety $\mathcal{V} \subseteq \mathbb{R}^n$ defined such that $\mathcal{V} := \{\mathsf{x} \in \mathbb{R}^n : p_j(\mathsf{x}) = 0, \text{for } j = 1, \ldots, r\}$. Let $\omega_i = (\mathsf{u}_i^\top, \mathsf{v}_i^\top)^\top$ be a measurement sample representing a pair of corresponding points. For a given set of samples $\Omega = \{\omega_i\}_{i=1}^m$, we wish to find the variety $\mathcal{V}$ of the lowest dimension. There is an extensive literature on computing an intrinsic dimension of the samples Ω from a variety $\mathcal{V}$ [4, 5]. However, the existing methods do not explicitly consider the noisy and outlier samples present in Ω. Therefore, they are not suitable for our task. In this work, we develop a tractable method for estimating the variety given the sample measurements which may contain noise and outliers. Primarily, we are interested in recovering polynomials $p_j(\mathsf{x})$ representing $\mathcal{V}$, from a corrupted sample set Ω. The topological space defined by $\mathcal{V}$ can be thought of as a semi-algebraic set, a differential manifold, a metric space, a Lie group, a category, a hypergraph, and many more.

Let $\mathcal{I}(\mathcal{V}) := \{\sum_j g_j(\mathsf{x}) p_j(\mathsf{x}) : g_j(\mathsf{x}) \in \mathcal{R}[\mathsf{x}]\}$ be the ideal of $\mathcal{V}$. Every polynomial in the ideal $\mathcal{I}(\mathcal{V})$ of the unknown variety $\mathcal{V}$ vanishes on samples Ω. Unfortunately, the converse is not true, *i.e.*, not all polynomials in the ideal $\mathcal{I}(\Omega)$ vanish on the variety $\mathcal{V}$. Therefore, recovering $p_j(\mathsf{x})$ to define $\mathcal{V}$ exactly is not only NP-hard, but also a non-decidable

problem. In this work, we limit the degree of the polynomials in $\mathcal{I}(\Omega)$ and assume that $\mathcal{V}$ can be recovered from $\mathcal{I}(\Omega)$ of low degree polynomials. Note that the ideal $\mathcal{I}(\Omega)$ of the finite set Ω can be computed using linear algebra, with the help of the so-called Vandermonde Matrix [4, 5].

Definition 2.3 (Vandermonde Matrix) *The Vandermonde matrix $\mathsf{M}_d(\mathsf{x})$ is a matrix with a geometric progression of monomials in each row, such that the entries m_{ij} are the monomials $\mathsf{x}^{\mathsf{e}} = x_1^{e_1} x_2^{e_2} \dots x_n^{e_n}$ of degree at most d.*

For example, if $n = 1, d = 3$, and $\Omega = \{u, v, w\}$ then $\mathsf{M}_3(\Omega)$ is the Vandermonde matrix of the form,

$$\mathsf{M}_3(\Omega) = \begin{bmatrix} u^3 & u^2 & u & 1 \\ v^3 & v^2 & v & 1 \\ w^3 & w^2 & w & 1 \end{bmatrix}.$$

When $n \geq 2$, $\mathsf{M}_d(\Omega)$ is a multivariate Vandermonde matrix with the following property:

Property 2.4 *Let $\mathcal{B}$ be a set representing a linearly independent basis of $\mathcal{R}[\mathsf{x}]$ and $\mathcal{R}_{\mathcal{B}}$ be the vector space spanned by $\mathcal{B}$. Then, the nullspace of $\mathsf{M}_d(\Omega)$ is the vector space of $\mathcal{I}(\Omega) \cap \mathcal{R}_{\mathcal{B}}$.*

We hope to learn the variety $\mathcal{V}$ by learning the ideal $\mathcal{I}(\mathcal{V})$. The ideal $\mathcal{I}(\mathcal{V})$ is learned for samples Ω in the form of $\mathcal{I}(\Omega)$, using the Vandermonde Matrix $\mathsf{M}_d(\Omega)$. In this process, we rely on the property of basis $\mathcal{B}$. The two desirable properties of $\mathcal{B}$ are:

Property 2.5 *The ideal $\mathcal{I}(\mathcal{V})$ is generated by $\mathcal{I}(\mathcal{V}) \cap \mathcal{R}_{\mathcal{B}}$.*

Property 2.6 *$\mathcal{I}(\mathcal{V}) \cap \mathcal{R}_{\mathcal{B}} \subseteq \mathcal{I}(\Omega) \cap \mathcal{R}_{\mathcal{B}}$ holds with equality.*

There is a fundamental tension between Properties 2.5 and 2.6. For small $\mathcal{B}$, Property 2.5 may not be satisfied. Similarly, Property 2.6 may be violated for large $\mathcal{B}$. Fortunately, the following theorem ensures the existence of $\mathcal{B}$.

Theorem 2.7 (Hilbert's Basis Theorem) *Every ideal in the polynomial ring $\mathcal{R}[\mathsf{x}]$ is finitely generated.*

The desired property 2.6 sets a lower bound on the sample size m. In fact, by construction m is the upper bound on the rank of $\mathsf{M}_d(\Omega)$. This implies the following lemma.

Lemma 2.8 *If Property 2.6 holds true, the inequality $m \geq dim(\mathcal{B}) - dim(\mathcal{I}(\mathcal{V}) \cap \mathcal{R}_{\mathcal{B}})$ must also be true.*

For given sample set Ω, the upper bound m on the rank is fixed. Therefore, one of the issues is choosing a suitable set $\mathcal{B}$. It is known that a suitable choice of $\mathcal{B}$ can dramatically improve the numerical accuracy. We will discuss our choice of $\mathcal{B}$, for our applications, later in this paper.

Another pending issue is the representation of $\mathcal{I}(\Omega) \cap \mathcal{R}_{\mathcal{B}}$. It is obvious from Property 2.4 that $\mathcal{I}(\Omega) \cap \mathcal{R}_{\mathcal{B}}$ is represented by the nullspace of $\mathsf{M}_d(\Omega)$. However, it is still unclear how to choose and compute the basis representing the null space. For example, we can obtain the orthonormal basis of $\mathcal{I}(\Omega) \cap \mathcal{R}_{\mathcal{B}}$ in the least-square sense by Singular Value Decomposition (SVD) of $\mathsf{M}_d(\Omega)$. Unfortunately, such basis are not favored as under noise this may result in a less sparse basis. Furthermore the samples Ω may contain outliers, in which case the orthonormal basis given by the SVD of $\mathsf{M}_d(\Omega)$ will be entirely wrong. In many applications, it is desirable to compute ideals $\mathcal{I}(\mathcal{V})$ with sparse generators. In particular, our Assumption 2.2 implicitly demands the basis to be sparse. Therefore, we wish to solve the following problem for the sparse basis $\mathcal{Y}$ of $\mathcal{I}(\Omega) \cap \mathcal{R}_{\mathcal{B}}$.

Problem 2.9 *Given the noise tolerance ϵ, find the orthonormal sparse basis $\mathcal{Y}$ from the nullspace of $\mathsf{M}_d(\Omega)$ by solving,*

$$\begin{aligned} \underset{\mathcal{Y}}{\textit{argmin}} \quad & \sum_{\mathsf{y}_i \in \mathcal{Y}} \|\mathsf{y}_i\|_0, \\ \textit{subject to} \quad & \|\mathsf{M}_d(\Omega)\mathsf{y}_i\|_\infty \leq \epsilon, \\ & \|\mathsf{y}_i\| \neq 0, \mathsf{y}_i^\top \mathsf{y}_j = 0, \quad \forall i, j, i \neq j. \end{aligned} \tag{1}$$

(1) involves ℓ_0 minimization, which is the holy grail of sparse approximation. Unfortunately, ℓ_0 minimization is NP-hard. Additionally, the non-linear objective for orthogonality and the search for non-trivial solutions make the problem even more difficult.

3. Sparse Basis Estimation

In this section, we develop a method to search for the polynomial constraint $\mathsf{y} \in \mathcal{Y}$ as a solution to the Problem 2.9. Our approach iteratively estimates the individual sparse orthonormal basis by solving the following Problem.

Problem 3.1 *For a given sparse basis $\mathsf{w} \in \mathcal{W} \subseteq \mathcal{Y}$, estimate a new sparse basis by solving,*

$$\begin{aligned} \underset{\mathsf{y}}{\textit{argmin}} \quad & \|\mathsf{y}\|_0, \\ \textit{subject to} \quad & \|\mathsf{M}_d(\Omega)\mathsf{y}\|_\infty \leq \epsilon, \\ & \|\mathsf{y}\|_\infty = 1, \mathsf{y}^\top \mathsf{w} = 0, \quad \forall \mathsf{w} \in \mathcal{W}. \end{aligned} \tag{2}$$

A common approximation of (2) is to replace ℓ_0 by a convex ℓ_1 objective. Here, we are rather interested to solve the exact problem of (2) using MIP.

3.1. Mixed-Integer Programming (MIP)

Proposition 3.2 *If $\mathsf{z} \in \{0, 1\}^n$ represents the sparsity of the basis vector y, Problem 3.1 is then equivalent to solving the*

following MIP.

$$\begin{aligned} \underset{\mathsf{y}\in\mathbb{R}^n,\mathsf{z}\in\{0,1\}^n}{\min} \quad & \sum_i z_i, \\ \textit{subject to} \quad & \|\mathsf{M}_d(\Omega)\mathsf{y}\|_\infty \le \epsilon, \\ & |y_i| \le z_i, \mathsf{y}^\top \mathsf{w} = 0, \quad \forall i, \forall \mathsf{w} \in \mathcal{W}, \\ & \|\mathsf{y}\|_\infty = 1, \\ & \sum_i z_i \ge 1. \end{aligned} \tag{3}$$

Proof Proof is provided in the supplementary document.

Although equivalent to (2), (3) is tractable and can be solved by BnB. Here we avoid the trivial zero solution by constraining the sum over the components of z to be greater than 1. However, (3) only optimizes for the sparse polynomial basis and fails if the measurements contain outliers. We therefore propose the following to handle outliers.

3.2. Sparse Basis in the Presence of Outliers

Proposition 3.3 *For $\mathsf{M} = \mathsf{M}_d(\Omega)$, $\mathsf{z} \in \{0,1\}^n$, $\mathsf{s} \in \{0,1\}^m$, and $\mathsf{y} \in \mathbb{R}^n$, the following MIP ensures that at least half of the correspondences respect the sparse basis obtained by solving,*

$$\begin{aligned} \underset{\mathsf{y},\mathsf{z},\mathsf{s}}{\min} \quad & \sum_{i=1}^{n} z_i, \\ \textit{subject to} \quad & \mathsf{m}_j^\top \mathsf{y} \le \epsilon + s_j m, && \forall j = 1,\dots,m, \\ & |y_i| \le z_i, \mathsf{y}^\top \mathsf{w} = 0, && \forall i, \forall \mathsf{w} \in \mathcal{W}, \\ & \|\mathsf{y}\|_\infty = 1, \\ & \sum_i z_i \ge 1, \sum_j s_j \le m/2. \end{aligned} \tag{4}$$

Proof Proof is provided in the supplementary document.

In (4), we introduce the binary variable $\mathsf{s} \in \{0,1\}^m$ to classify the polynomial from a correspondence pair $(\mathsf{u}_j, \mathsf{v}_j)$ as an outlier if $s_j = 1$ or inlier if $s_j = 0$. We express the binary constraint for every measurement row m_j of the Vandermonde matrix M using the big-M formulation [20]. In order to make the problem tractable, we assume that at least half of the points are inliers. However, in practice, the constraint may be adjusted to suit the outlier statistics. In the following section, we discuss our basis selection method in the context of the two-view camera geometry problem.

4. Two-view Geometry Applications

The sparse basis computation discussed in section 3 allows one to compute the polynomials that relate the image point correspondences accurately in the presence of outliers. The basis computation directly gives the correct two-view relation, whether they are the Essential matrix, the Fundamental matrix or the 2D projective homography. On the other hand, such two-view relations may or may not say anything about the camera and motion types, and finally the actual camera motion knowing the camera type. We now discuss problem 2.1 of obtaining the camera model $\mathcal{M}$ and the motion parameters θ from the point correspondences. For that purpose, we consider various camera projection and camera motion types and analyze each condition further.

4.1. Projection and Motion Types

Considering camera projections, we analyze five different camera models: *i)* calibrated perspective, *ii)* uncalibrated perspective, *iii)* orthographic, *iv)* weak-perspective and *v)* affine camera. For each camera model we divide the motion into seven types: *i)* full motion with rotation and translation, *ii)* rotation, *iii)* translation, *iv)* rotation about x or y, *v)* rotation about z, *vi)* translation about x or y and *vii)* translation about z. Given only images, the orthographic camera and the weak-perspective camera projections differ only by a single scale factor. Therefore, we treat them as equivalent cameras in this analysis and use the term orthographic camera to discuss both types. Below we first define the transformations and the relevant two-view relationships before presenting the analysis.

Transformations and model. We consider the camera being transformed by a rotation $\mathsf{R} \in \mathrm{SO}_3$ and translation $\mathsf{t} \in \mathbb{R}^3$. Let the rotation in Euler angles be $\mathsf{r} \in \mathbb{R}^3$. We consider the camera translation t, also represented as a transformation matrix $\mathsf{T} \in \mathcal{T}$ where $\mathcal{T}$ is the space of all translations. The Essential matrix [17] for the calibrated perspective camera is $\mathsf{E} \in \mathcal{P} \subset \mathbb{R}^{3\times3}$. Let $\mathcal{P}$ represent the space of matrices that satisfy the property of having two equal non-zero singular values $\sigma_1(\mathsf{E}) = \sigma_2(\mathsf{E})$ and $\sigma_3(\mathsf{E}) = 0$. In the orthographic camera, the Essential matrix is $\mathsf{E}_O \in \mathcal{O} \subset \mathbb{R}^{3\times3}$. It has the following properties [14].

$$\mathsf{E}_O = \begin{bmatrix} 0 & 0 & c \\ 0 & 0 & d \\ a & b & e \end{bmatrix}, \quad a^2 + b^2 - c^2 - d^2 = 0, \quad a,b,c,d,e \in \mathbb{R}. \tag{5}$$

The next two-view relation model is the Fundamental matrix [9], $\mathsf{F} \in \mathcal{U} \subset \mathbb{R}^{3\times3}$. We use $\mathcal{U}$ for the space of the Fundamental matrices obtained from uncalibrated perspective cameras. The perspective fundamental matrix has two non-zero singular values and the third singular value 0. Unlike the Essential matrix from calibrated perspective cameras, the two non-zero singular values of the Fundamental matrix are in general not equal. In case of affine cameras, the Fundamental matrix is $\mathsf{F}_A \in \mathcal{A} \subset \mathbb{R}^{3\times3}$. F_A has the following property:

$$\mathsf{F}_A = \begin{bmatrix} 0 & 0 & c \\ 0 & 0 & d \\ a & b & e \end{bmatrix}, \quad a,b,c,d,e \in \mathbb{R}. \tag{6}$$

Table 1. **Summary of the sparse basis for various camera and motion types.** The sparse basis is summarized as follows. The two-view relationship is G, number/dimension of basis r, actual degrees of freedom under known camera and motion type d and the number of non-zero two-view model parameters p.

	Cal. Perspective $\mathsf{G}(r,d,p)$	Uncal. Perspective $\mathsf{G}(r,d,p)$	Orthographic $\mathsf{G}(r,d,p)$	Affine $\mathsf{G}(r,d,p)$
Full motion	$\mathsf{E}(1,5,9)$	$\mathsf{F}(1,7,9)$	$\mathsf{E}_O(1,4,5)$	$\mathsf{F}_A(1,5,5)$
Rotation	$\mathsf{H}(3,3,9)$	$\mathsf{H}(3,8,9)$	$\mathsf{E}_O(1,3,4)$	$\mathsf{F}_A(1,4,4)$
Translation	$\mathsf{E}(1,3,6)$	$\mathsf{F}(1,3,6)$	$\mathsf{H}(1,2,5)$	$\mathsf{H}(1,2,5)$
Rotation x	$\mathsf{H}(3,1,5)$	$\mathsf{H}(3,6,7)$	$\mathsf{E}_O(1,1,2)$	$\mathsf{F}_A(1,4,4)$
Rotation y	$\mathsf{H}(3,1,5)$	$\mathsf{H}(3,6,9)$	$\mathsf{E}_O(1,1,2)$	$\mathsf{F}_A(1,4,4)$
Rotation z	$\mathsf{H}(3,1,5)$	$\mathsf{H}(3,5,7)$	$\mathsf{H}(3,1,5)$	$\mathsf{H}(3,5,7)$
Translation x/y	$\mathsf{E}(1,1,2)$	$\mathsf{F}(1,3,6)$	$\mathsf{E}_O(3,1,4)$	$\mathsf{H}(3,2,5)$
Translation z	$\mathsf{E}(1,1,2)$	$\mathsf{F}(1,3,6)$	$\mathsf{E}_O(3,0,3)$	$\mathsf{H}(3,0,3)$

From the relations, we have, $\mathcal{P} \subset \mathcal{U}$ and $\mathcal{O} \subset \mathcal{A}$, while none of the sets $\mathcal{P}, \mathcal{U}$, $\mathcal{O}$ and $\mathcal{A}$ are disjoint.

The projective 2d homography $\mathsf{H} \in \mathrm{PGL}(2,\mathbb{R})$ is a full rank transformation unlike the Essential and Fundamental matrices. The homography is a point to point relation and the space of homographies $\mathrm{PGL}(2,\mathbb{R})$ is disjoint to the four spaces of the Essential and Fundamental matrices. In special cases, the homography can be represented by an affine transform $\mathsf{A} \in \mathrm{Aff}(2,\mathbb{R})$, a rotation $\mathsf{R} \in \mathrm{SO}_3$, a translation $\mathsf{T} \in \mathcal{T}$ or an identity I. All of the two-view relationships are described by models which are homogeneous quantities and therefore they are equivalent at different scales.

Sparsity of two-view relationship for different conditions. We summarize the sparsity of various two-view relationships for different camera models and motion types in Table 1. In each case, the point correspondences $\{\mathsf{u}_i, \mathsf{v}_i\}_{i=1}^m$ are related either by the Fundamental matrix, the Essential matrix or the 2d projective homography. The properties of these two-view relations vary according to both the camera types and the motion types. Despite that and the varying number of model parameters, all these relations can be expressed as polynomials of degree 2 in the image point correspondences. Therefore the corresponding Vandermonde matrix used in (4) is $\mathsf{M}_2(\Omega)$ of degree 2.

One interesting problem exists in the combinations of camera and motion, where the approach of using RANSAC with 8 points results in an incorrect model. This happens when the image point correspondences are related by a homography rather than the Fundamental matrix or the Essential matrix. Solving problem (4) with the Vandermonde matrix $\mathsf{M}_2(\Omega)$ results in the correct relationship between the image point correspondences in either case. When the correspondences are related by a homography, the iterative application of problem (4) will find three independent bases corresponding to either of the following system of equations.

$$[\mathsf{u}_i^\top \; 1] \times \mathsf{H}[\mathsf{v}_i^\top \; 1]^\top = 0 \quad \text{or} \quad [\mathsf{v}_i^\top \; 1]^\top \times \mathsf{H}^{-1}[\mathsf{u}_i^\top \; 1]^\top = 0 \quad (7)$$

We recover the homography H by using a change of basis for the resulting system of equations so that the sparsity of the final system corresponds to the left equation of eq. (7).

4.2. Camera and Motion Type Recovery

Recovering the camera model $\mathcal{M}$ and the motion parameters θ from the computed two-view model is not trivial and in fact, as we show in this paper, in most cases it is not possible with only point correspondences. Knowing the camera and motion type is crucial in many 3D vision problems [17, 14, 6]. We theoretically analyze the types of the two-view models and justify when the camera type and the motion type can be disambiguated and when the ambiguities lie otherwise. We provide the summary of the properties of each two-view relationship for various camera model and motion types in Table 2. We discuss the ambiguities for each camera model below.

Calibrated perspective. The calibrated perspective camera images are either related by the Essential matrix $\mathsf{E} \in \mathcal{P}$, when there exists non-zero translation, or the homography $\mathsf{H} \in \mathrm{PGL}(2,\mathbb{R})$, when there is no translation between two cameras. For a purely rotating camera, the induced relationship is the homography [12], which is in fact the corresponding relative rotation $\mathsf{R} \in \mathrm{SO}(3)$. There are two important cases when the camera model cannot be disambiguated from the model. The first is that of pure translation, in which case, the essential matrix $\mathsf{E} = \mathsf{R}[\mathsf{t}]_\times$ is a skew-symmetric matrix, similar as in the case of an uncalibrated camera. A congruent transformation $\mathsf{K}^{-\top}\mathsf{E}\mathsf{K}^{-1}$ of a skew-symmetric matrix E results in a skew symmetric matrix with exactly two equal non-zero singular values. Therefore, the calibration cannot be verified from the image point correspondences of purely translating cameras. The second case when the camera model cannot be ascertained is when there is a pure rotation around the Z-axis. It is straight-forward to verify that an orthographic camera gives the same rotational homography in such a case. The same is true for an affine camera without the skew component. The disambiguation of motion is also not possible for pure translation without assuming a calibrated perspective camera beforehand. For example, the essential matrix for an orthographic camera rotating about X or Y axis is the same as the essential matrix for the calibrated perspective camera with pure translation on X or Y axis, respectively. The motion computation, assuming a calibrated perspective camera is always possible even when the relationship is a homography, where we do not have the usual 4-fold ambiguity of the planar homography decomposition due to camera translation.

Table 2. **Properties of the two view relationship for each camera and motion type.** Each two-view relationship is followed by three boxes showing unique disambiguation of the camera type, the motion type and the metric motion given known camera type, resp. A check mark indicates uniqueness and a cross mark indicates ambiguity.

	Cal. Perspective		Uncal. Perspective		Orthographic		Affine	
Full motion	$\sigma_1(\mathsf{E}) = \sigma_2(\mathsf{E})$	✓✓✓	$\mathsf{F} \in \mathcal{U}$	✓✓✗	$\mathsf{E}_O \in \mathcal{O}$	✓✓✗	$\mathsf{F}_A \in \mathcal{A}$	✓✓✗
Rotation	$\mathsf{H} = \mathsf{R}$	✓✓✓	$\mathsf{H}\mathsf{H}^\top \neq \mathsf{I},\ \mathsf{H} = \mathsf{K}\mathsf{R}\mathsf{K}^{-1}$	✓✓✓	$\mathsf{E}_O \in \mathcal{O},\ \mathsf{E}_{O3,3} = 0$	✗✗✗	$\mathsf{F}_A \in \mathcal{A},\ \mathsf{F}_{A3,3} = 0$	✓✓✗
Translation	$\mathsf{E} = [\mathsf{t}]_\times = -\mathsf{E}^\top$	✗✗✓	$\mathsf{F} = \mathsf{K}[\mathsf{t}]_\times\mathsf{K}^\top = -\mathsf{F}^\top$	✗✗✗	$\mathsf{H} \in \mathcal{T}$	✗✓✓	$\mathsf{H} \in \mathcal{T}$	✗✓✗
Rotation x	$\mathsf{H} = \mathsf{R}_x$	✓✓✓	$\mathsf{H}\mathsf{H}^\top \neq \mathsf{I},\ \mathsf{H} = \mathsf{K}\mathsf{R}_x\mathsf{K}^{-1}$ $\mathsf{H}_{2,1} = \mathsf{H}_{3,1} = 0$	✓✓✗	$\mathsf{E}_O = [\mathsf{r}_x]_\times = -\mathsf{E}_O^\top$	✗✗✗	$\mathsf{F}_A \in \mathcal{O}, \mathsf{F}_{A3,3} = 0$	✗✗✗
Rotation y	$\mathsf{H} = \mathsf{R}_y$	✓✓✓	$\mathsf{H}\mathsf{H}^\top \neq \mathsf{I},\ \mathsf{H} = \mathsf{K}\mathsf{R}_y\mathsf{K}^{-1}$	✓✗✗	$\mathsf{E}_O = [\mathsf{r}_y]_\times = -\mathsf{E}_O^\top$	✗✗✗	$\mathsf{F}_A = [\mathsf{r}_y]_\times = -\mathsf{F}_A^\top$	✗✗✗
Rotation z	$\mathsf{H} = \mathsf{R}_z$	✗✓✓	$\mathsf{H} \in \mathrm{Aff}(2, \mathbb{R})$	✗✓✗	$\mathsf{H} = \mathsf{R}_z$	✗✓✓	$\mathsf{H} \in \mathrm{Aff}(2, \mathbb{R})$	✗✓✗
Translation x/y	$\mathsf{E} = [\mathsf{t}_{x/y}]_\times = -\mathsf{E}^\top$	✗✗✓	$\mathsf{F} = \mathsf{K}[\mathsf{t}_{x/y}]_\times\mathsf{K}^\top = -\mathsf{F}^\top$	✗✗✗	$\mathsf{H} \in \mathcal{T}_{x/y}$	✓✓✓	$\mathsf{H} \in \mathcal{T}$	✗✗✗
Translation z	$\mathsf{E} = [\mathsf{t}_z]_\times = -\mathsf{E}^\top$	✓✓✓	$\mathsf{F} = \mathsf{K}[\mathsf{t}_z]_\times\mathsf{K}^\top = -\mathsf{F}^\top$	✗✗✗	$\mathsf{H} = \mathsf{I}$	✗✗✗	$\mathsf{H} = \mathsf{I}$	✗✗✗

Uncalibrated perspective. An uncalibrated perspective camera in general has more ambiguities from projections. For full motion and rotation, the camera model can be disambiguated by the singular values of F or the singular values of H. As described above, pure translational motion results in $\mathsf{F} = \mathsf{K}^{-\top}[\mathsf{t}]_\times\mathsf{K}^{-1} = -\mathsf{F}^\top$, with $\mathsf{F} \in \mathcal{P}$, meaning that the uncalibrated camera cannot be disambiguated from the calibrated camera. Particularly interesting is the asymmetry of rotation about X and Y axis. The rotation about the X-axis results in the homography H with two zeros on the first column but nothing can be said for the case with pure rotation about the Y-axis. This apparent asymmetry is simply due to the choice of axes for the skew that results in the convention of the upper triangular intrinsics. The uncalibrated camera type can be disambiguated only for full motion, rotation and rotation about the X or Y-axis. For rotation of the uncalibrated camera we have the following proposition.

Proposition 4.1 *For the rotation-induced 2d homography* H*, with intrinsics* $\mathsf{K} \neq \mathsf{I}$ *and translation* $\mathsf{t} = 0$*, we have,*

$$\mathsf{H}\mathsf{H}^\top \neq \mathsf{I}, \tag{8}$$

where, $\mathsf{r} \neq \pm[0\ 0\ \pi/2]^\top$.

Proof Proof is provided in the supplementary document.

Proposition 4.1 claims that the homography in an uncalibrated perspective camera is *never* orthogonal unless the rotation is the identity or a rotation of $\pi/2$ about the Z-axis. It can be shown that when the rotation is $\pm\pi/2$ and the two focals of K are equal, H becomes orthogonal using the spectral decomposition of $\mathsf{H}\mathsf{H}^\top$. The motion in case of an uncalibrated perspective camera can be disambiguated into full motion, translation, rotation and rotation around the X-axis or Z-axis. However, even after assuming an uncalibrated perspective camera, the metric motion is always ambiguous without knowing the camera intrinsics [9]. One notable exception is pure rotation discussed in [12], where one can reason about the metric rotation.

Orthographic camera. The orthographic camera images are related by either the Essential matrix $\mathsf{E}_O \in \mathcal{O}$ or the affine Homography $\mathsf{A} \in \mathrm{Aff}(2, \mathbb{R})$. The orthographic camera can be identified from the images if the motion is full. A pure rotation will result in the Essential matrix E_O with the last diagonal element 0, the same as in the affine camera with rotation around the X axis. In the case of pure rotation around only the X or Y axis, E_O is the same as E for the pure translation of a calibrated perspective camera. Similarly, a pure rotation around the Z-axis results in a rotational homography as for the calibrated perspective camera. Motions with pure translation results in an affine transform similar to that in affine camera and thus the camera model cannot be disambiguated completely. As to the motion type disambiguation, the full motion, rotation and translation can be identified as such for the orthographic camera. The rest of the motion shows ambiguity either with the affine camera or the calibrated perspective camera as evident from the relations detailed in table 2. Finally, even with a known camera type, it is not possible to exactly decompose camera motion from two images of an orthographic camera due to the bas-relief ambiguity [14, 26]. This means that one of the rotational components can never be disambiguated, but other motions can be decomposed.

Affine camera. The affine camera images are related either by the Fundamental matrix F_A or the affine Homography $\mathsf{H} \in \mathrm{Aff}(2, \mathbb{R})$. The affine camera model can be disambiguated for full motion and pure rotation. In full motion, the affine Fundamental matrix satisfies eq. (6) but in general not the constraint of the orthographic Essential matrix eq. (5). Since the affine camera is a generalization of the orthographic camera, it has camera model ambiguity with the orthographic camera in all other motion types. The only exception is in the case of pure rotation around the Y-axis, where the ambiguity is with both the orthographic camera rotating around the Y-axis and the uncalibrated perspective

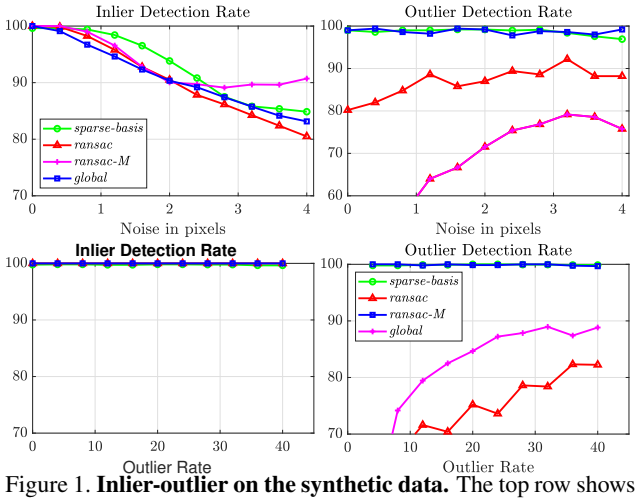

Figure 1. **Inlier-outlier on the synthetic data.** The top row shows the inlier and outlier % detection rates when varying noise magnitude while the bottom row shows the same for varying % outlier rate. Our method gives expected results in all camera type and motion type, thus producing identical results to *ransac-M*.

camera translating along the Y-axis. One more ambiguity with the uncalibrated perspective camera is for pure rotation around the Z-axis where the two-view relation is an affine transform for both camera types. The motion in an affine camera can be identified as full motion, rotation and translation. Further disambiguation is also possible in case of pure rotation around the Z-axis. Given the affine camera type, one cannot compute the metric motion due to the affine ambiguity and furthermore certain motion components [14] cannot be resolved due to the bas-relief ambiguity.

5. Experimental Results

We use MATLAB to implement our method written as *sparse-basis*. We use RANSAC [10] for the Fundamental matrix as the compared method for unknown model. We write it as *ransac*. We write RANSAC with known model as *ransac-M*, where a specific implementation of RANSAC is used according to the ground-truth two-view relationship type. We verify the theoretical discussions and its applications using both the synthetic and real data. We choose the threshold $\epsilon = 1.5$ px for our method and equivalently for *ransac*. We further test a globally optimal method of outlier rejection [8] as *global*.

5.1. Synthetic Data Experiments

We use synthetic data to validate the theoretical results in various conditions of noise and outliers for the camera and motion types discussed in section 4.1. We use a camera resolution of 512 pixels and generate matches with outlier ratio varying from 0 to 40%. We also test the compared methods with noise, by varying uniform noise from 0 to 4 pixels. We add 0 to 0.5 pixel noise of uniform distribution and also

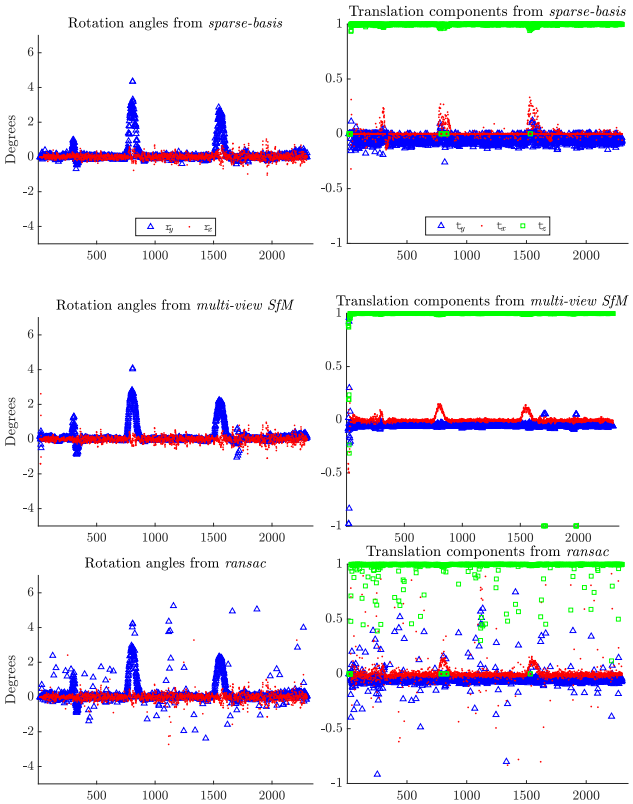

Figure 2. **Two-view poses for a sequence in the Oxford Robot Car dataset [19].** In the sequence, the car does not stop and hence we do not get any homographies. We observe that the relative camera poses obtained with *sparse-basis* is very close to the results obtained from *multi-view SfM*.

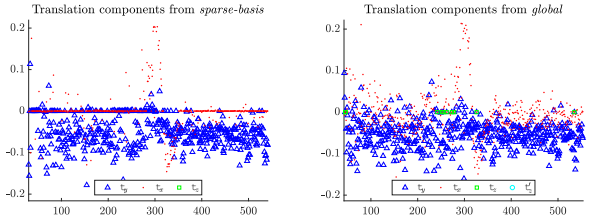

Figure 3. Results zoomed to show differences from *global* method for the first part of Oxford sequence [19] on translation estimation.

add 5% outliers to all the projections. For each condition of motion type, camera type and noise/outlier, we use 20 simulations each with 50 points to generate the experimental results. We then average out the detection results for different motion and camera types. Figure 1 shows the results of the inlier/outlier classification obtained during the model computation with our method and the compared methods. More importantly, we consistently meet the two-view relationship property of table 2 in the experiment.

While we expect the same performance of *ransac-M* and *sparse-basis* in all cases, small gap can be particularly noted in the inlier detection rate with varying noise. Our method

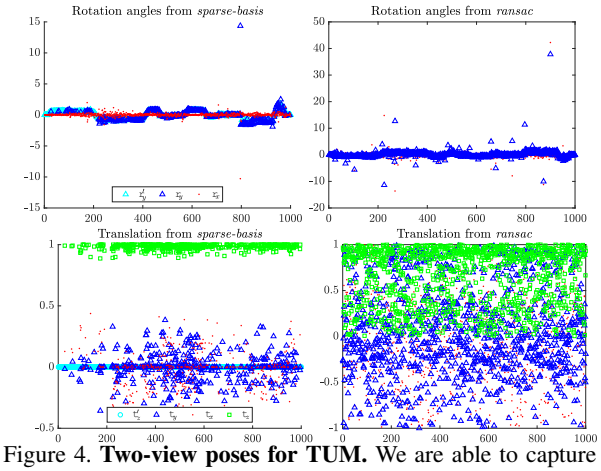

Figure 4. **Two-view poses for TUM.** We are able to capture the turns and the degenerate motions better despite the short baseline. We introduce r'_y and t'_z to show the respective motions estimated from the homography instead of the Essential matrix.

sparse-basis performs slightly better on average simply due to the threshold and the way we normalize each polynomial. Another issue is the behavior of *ransac* and *global* for the increase in outlier detection rate with the outlier rate. When the correct two-view relationship between image correspondences is a homography, a model can still find outliers by fitting the fundamental matrix. However, the fundamental matrix is a weaker constraint and particularly for the first few outliers, *ransac* and *global* can find fit them as inliers in the fundamental matrix model. However, as the outlier rate increases more outliers are correctly rejected. In contrast, *sparse-basis* always fits the correct model.

5.2. Real Data Experiments

We conduct experiments on real datasets to show how our method performs in practical settings. In the first experiment we evaluate motion disambiguation on the first 2400 frames of a sequence from the Oxford Robot car dataset [19] and a sequence in the TUM RGBD [29] dataset in figure 2 and 4 respectively, using all consecutive frames. The Oxford Robot car dataset consists of high quality images where a good set of the feature matches can be expected. In order to compute poses, we match each video frame with the one after the next frame. We keep the number of points $m = 100$ for both *sparse-basis* and *ransac* by randomly choosing the matches. Same set of matches are used for both the methods. The odometry ground-truth provided in the dataset contains drift and inaccuracies. Therefore, we reconstruct the sequence of about 2400 frames using multi-view SfM COLMAP [24]. The relative ground-truth poses are then obtained from the SfM reconstruction of COLMAP. We refer to the ground-truth poses as *multi-view SfM*. We plot the comparison between *sparse-basis* and *ransac* in figures 2 and 4 for the two datasets. We further provide a comparison between *sparse-basis* and *global* on a subsequence of figure 2.

Both sequences pose challenging condition for computing motion due to the short baseline. Nonetheless, *sparse-basis* is able to better condition the motion computation as we look for sparse basis for the two-view relationship, whether it is the Essential matrix or the homography. In particular, we capture the smooth turning of the vehicle (rotation about Y-axis), the consistent forward motion of the car (Z-translation) and no motion (middle of the sub-sequence) which is not possible without searching for the sparse bases as shown in figure 4 for *ransac*. We also note a smooth transition between r_y and r'_y as seen in figure 4.

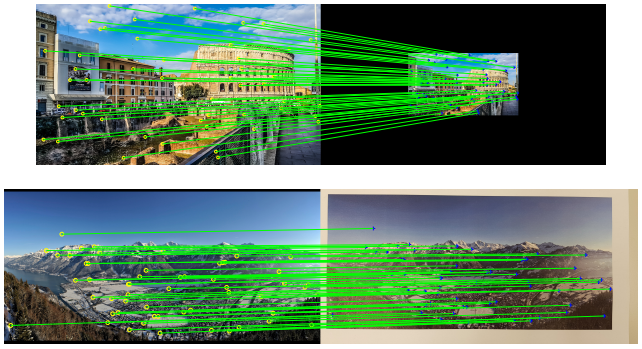

Figure 5. **Inlier detected in real images.** The top row shows the inliers for images with change in resolution and bottom row for perspective image of a printed picture. In both cases *sparse-basis* correctly fits a homography as the two-view relationship.

We further show the qualitative results of matching points using our method in two interesting examples in figure 5. We first match points using SIFT [18] descriptors in each of the three cases. In figure 5 our method *sparse-basis* is able to match identical lower and higher resolution images while rejecting outliers. At the same time, we can reason that the correspondences are from identical images as we obtain an identity homography. We provide additional experiments in the supplementary material.

6. Conclusions

We proposed a method that computes the correct two-view relationship in the presence of noise and outliers, even when the camera and the motion types are unknown by looking for sparse polynomials. In this scenario, we discussed the possibilities of disambiguating camera and motion case-by-case. The experiments verify our theory and give practical applications of our method.

Acknowledgements. This research was funded by the EU's Horizon 2020 programme under grant No. 687757 – REPLICATE and by the Swiss Commission for Technology and Innovation (CTI), Grant No. 26253.1 PFES-ES – EXASOLVED.

References

[1] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346–359, 2008. 1

[2] J.-C. Bazin, H. Li, I. S. Kweon, C. Demonceaux, P. Vasseur, and K. Ikeuchi. A branch-and-bound approach to correspondence and grouping problems. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1565–1576, 2013. 1

[3] J. C. Bazin, Y. Seo, R. I. Hartley, and M. Pollefeys. Globally optimal inlier set maximization with unknown rotation and focal length. In *ECCV*, 2014. 1

[4] . Björck and V. Pereyra. Solution of vandermonde systems of equations. *Mathematics of Computation*, 24(112):893–903, 1970. 2, 3

[5] P. Breiding, S. K. Verovsek, B. Sturmfels, and M. Weinstein. Learning algebraic varieties from samples. *arXiv preprint arXiv:1802.09436*, 2018. 2, 3

[6] M. Chandraker, S. Agarwal, F. Kahl, D. Nister, and D. Kriegman. Autocalibration via rank-constrained estimation of the absolute quadric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. 1, 5

[7] T. J. Chin, Y. H. Kee, A. Eriksson, and F. Neumann. Guaranteed outlier removal with mixed integer linear programs. In *CVPR*, 2016. 1

[8] T.-J. Chin, P. Purkait, A. Eriksson, and D. Suter. Efficient globally optimal consensus maximisation with tree search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2413–2421, 2015. 7

[9] O. D. Faugeras. What can be seen in three dimensions with an uncalibrated stereo rig. In *ECCV*, pages 563–578, 1992. 2, 4, 6

[10] 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*, 24(6):381–395, 1981. 1, 2, 7

[11] 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 *ECCV*, 2010. 1

[12] R. I. Hartley. Self-calibration from multiple views with a rotating camera. In *ECCV*, pages 471–478, 1994. 1, 5, 6

[13] R. I. Hartley and F. Kahl. Global optimization through rotation space search. *IJCV*, 82(1):64–79, 2009. 1

[14] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004. 1, 2, 4, 5, 6, 7

[15] H. Li. Consensus set maximization with guaranteed global optimality for robust geometry estimation. In *ICCV*, 2009. 1

[16] A. Locher, M. Havlena, and L. Van Gool. Progressive structure from motion. In *ECCV*, 2018. 1

[17] H. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981. 1, 2, 4, 5

[18] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 1, 8

[19] W. Maddern, G. Pascoe, C. Linegar, and P. Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research (IJRR)*, 36(1):3–15, 2017. 7, 8

[20] G. P. McCormick. Computability of global solutions to factorable nonconvex programs: Part iconvex underestimating problems. *Mathematical programming*, 10(1):147–175, 1976. 4

[21] D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(6):756–777, 2004. 1

[22] D. Nister, O. Naroditsky, and J. Bergen. Visual odometry. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 652–659, 2004. 2

[23] D. Scaramuzza. 1-point-ransac structure from motion for vehicle-mounted cameras by exploiting non-holonomic constraints. *Int. J. Comp. Vision*, 95(1):74–85, 2011. 1

[24] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 8

[25] J. L. Schonberger and J.-M. Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1

[26] L. S. Shapiro, A. Zisserman, and M. Brady. 3d motion recovery via affine epipolar geometry. *International Journal of Computer Vision*, 16(2):147–182, 1995. 2, 6

[27] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *Int. J. Comput. Vision*, 80(2):189–210, 2007. 1

[28] P. Speciale, D. P. Paudel, M. R. Oswald, T. Kroeger, L. V. Gool, and M. Pollefeys. Consensus maximization with linear matrix inequality constraints. In *CVPR*, 2017. 1

[29] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *IROS*, 2012. 8

[30] Z. Zhang and A. R. Hanson. 3d reconstruction based on homography mapping. In *ARPA Image Understanding Workshop*, 1996. 1

[31] Y. Zheng, S. Sugimoto, and M. Okutomi. Deterministically maximizing feasible subsystem for robust model fitting with unit norm constraint. In *CVPR*, 2011. 1