The generalized relative pose and scale problem: View-graph fusion via 2D-2D registration

Laurent Kneip[1)], Chris Sweeney[2)], Richard Hartley[1,3)]
1) Australian National University
2) University of California Santa Barbara
3) Data61/CSIRO

Abstract

It is well-known that the relative pose problem can be generalized to non-central cameras. We present a further generalization, denoted the generalized relative pose and scale problem. It has surprising importance for classical problems such as solving similarity transformations for view-graph concatenation in hierarchical structure from motion and loop-closure in visual SLAM, both posed as a 2D-2D registration problem. The relative pose problem and all its generalizations constitute a family of similar symmetric eigenvalue problems, which allow us to compress data and find a geometrically meaningful solution by an efficient search in the space of rotations. While the derivation of a completely general closed-form solver appears intractable, we make use of a simple heuristic global energy minimization scheme based on local minimum suppression, returning outstanding performance in practically relevant scenarios. Efficiency and reliability of our algorithm are demonstrated on both simulated and real data, supporting our claim of superior performance with respect to both generalized 2D-3D and 3D-3D registration approaches. By directly employing image information, we avoid the common noise in point clouds occuring especially along the depth direction.

1. Introduction

Computing the *relative pose* between two cameras certainly ranks among the most prominent topics in geometric computer vision. Its eminent role resides in bootstrapping structure from motion, when no information about either 3D structure or motion is yet available. The present paper focuses on the calibrated case. Traditional solutions are given by the linear n-point solver presented in [6], and the non-linear minimal solver presented in [14]. The body of related work is vaste, and the interested reader is kindly referred to [7] for further reading on so-called 2D-2D registration with central cameras, as this work aims at generalizations of the problem applicable beyond a pair of central projections.

As already observed in more recent work, the relative pose problem can also be solved for *generalized cameras* [5, 12, 11, 8, 15]. A generalized camera is formed by abstracting a set of landmark observations into spatial rays that are no longer required to originate from a common point (i.e. the focal point). This formalism allows us to describe the measurements of a number of interesting camera systems, such as a camera looking at an arbitrarily shaped mirror, or multiple cameras mounted on a static rig. It only requires the knowledge of the camera calibration parameters in order to subtract the influence of intrinsic and extrinsic parameters, and transform the measurements into a bundle of non-centric rays. Again, both non-linear minimal [15] and linear non-minimal [11] solutions have been presented in the literature, which derive the relative pose between two generalized camera frames from Plücker line coordinates.

We aim at yet a further generalization, the *generalized relative pose and scale* problem. It introduces a further unknown: a relative scale factor between the ray origins in both generalized camera frames. While this may at first appear as an artificial problem, it actually has surprising importance in classical structure from motion with central cameras. A view-graph is a set of central projections (i.e. camera frames) that are registered with respect to each other in terms of their relative pose. Each landmark observation in each frame can therefore easily be expressed by a spatial ray defined inside a common frame for the entire view-graph, thus forming a virtual generalized camera. Now consider the fact that many problems in structure from motion involve the registration of partial view-graphs. By treating each view-graph as a generalized camera, this registration could be done directly based on 2D-2D correspondences by solving the generalized relative pose problem. However, the scale in each view-graph is arbitrary and registering two view-graphs requires reconciling the relative scale between them. This means that registering two view-graphs requires finding a 7 degree-of-freedom similarity transformation, a problem that has typically been solved based on 2D-3D or 3D-3D correspondences. The 2D-3D registration method is

given by the generalized absolute pose and scale algorithm presented in [20, 17], and in the 3D-3D registration case, it is the orthogonal procrustes approach presented in [13]. We show that this is equivalent to solving the generalized relative pose and scale problem from 2D-2D correspondences.

Important applications of partial view-graph registration are given by hierarchical structure from motion, loop closure, and map fusion. The first is a highly parallelizable and scalable approach doing hierarchical pairwise concatenation of partial view-graphs. On the lower levels and thus early stages of the computation it is especially the case that small view-graphs would lead to noisy 3D information, and a solution based on the original measurements leads to a provable improvement of the final registration result. In monocular visual SLAM, the local scale potentially suffers from drift along a loop. Accurate loop closure uses information from multiple frames of either ends of the loop, and—for the aforementioned reasons—benefits from a 2D-2D registration procedure as well. The latter does no longer depend on 3D points that often are the noisy result of triangulated image correspondences, and therefore intuitively outperform 2D-3D or even 3D-3D registration methods that rely on intermediate and necessarily inaccurate computations. The problem of increased triangulation noise appears especially along the depth direction, and deteriorates as the ratio between baseline and scene depth decreases.

The generalized relative pose and scale problem was first introduced and solved for known vertical direction in [18]. In this work, we drop the assumption of known vertical direction, and solve for a full Euclidean transformation. Our approach is based on a recently proposed formalism that casts the relative pose and generalized relative pose problems as a symmetric eigenvalue problem [9, 8]. This method allows for linear complexity in the number of points, and finds the solution by an efficient direct search in the space of rotations. We use heuristic global optimization based on minimum suppression (i.e. multi-start clustering), providing outstanding performance in practically relevant situations. The global optimum is approximated under an explicit geometrically meaningful metric. We test our algorithm against more traditional 2D-3D and 3D-3D registration approches on both simulated and real data, showcasing the benefits of *fusing view-graphs via 2D-2D registration*.

2. Theory

This section introduces a hierarchy of relative pose problems. It leads to a family of symmetric eigenvalue problems, thus permitting a unified solution strategy based on various rank minimization approaches.

2.1. Central case

The measurements in the calibrated central case are given by a bundle of n centric rays in each camera, described by unit direction vectors $\mathbf{f}_i$ and $\mathbf{f}'_i$. The goal consists of finding the translation $\mathbf{t}$ and rotation $\mathbf{R}$ that allow us to transform points from the second camera frame into the first one, following the convention $\alpha_i \mathbf{f}_i = \mathbf{t} + \mathbf{R}\alpha'_i \mathbf{f}'_i$, where α_i and α'_i denote the unknown depths inside frame 1 and 2, respectively. [8] showed that—after the application of simple scalar triple product rules—the classical epipolar constraint can take the form

$$(\mathbf{f}_i \times \mathbf{R}\mathbf{f}'_i)^T \mathbf{t} = \mathbf{n}_i^T \cdot \mathbf{t} = 0, \tag{1}$$

where $\mathbf{n}_i$ denotes the epipolar plane normal vector of a certain correspondence. All epipolar plane normal vectors need to be coplanar, and—using $\mathbf{N} = [\mathbf{n}_1 \ \dots \ \mathbf{n}_n]$—the covariance matrix of all normal vectors given by

$$\Sigma_{\mathbf{N}} = \mathbf{N}\mathbf{N}^T = \sum_{i=1}^{n} (\mathbf{f}_i \times \mathbf{R}\mathbf{f}'_i)(\mathbf{f}_i \times \mathbf{R}\mathbf{f}'_i)^T \tag{2}$$

in turn needs to have a rank deficiency of one in the general noise-free case. The only unknown in $\Sigma_{\mathbf{N}}$ is $\mathbf{R}$, and—as shown in the supplemental material—the latter may be elegantly factorized inside the expression such that our original data given by all $\mathbf{f}_i$'s and $\mathbf{f}'_i$'s is compressed into multiple summation terms. The problem of finding the solution $\hat{\mathbf{R}}$ may be formulated algebraically as

$$\hat{\mathbf{R}} = \underset{\mathbf{R} \in SO(3)}{\operatorname{argmin}} \left(\operatorname{rank}(\Sigma_{\mathbf{N}}(\mathbf{R}))\right). \tag{3}$$

$\Sigma_{\mathbf{N}}$ is a 3×3 SPD matrix with rank 2 in the ideal case. Given that we can compress the measurement data in linear time, the complexity of the actual solution turns out to be independent of the number of correspondences n.

2.2. Generalized relative pose

The measurements in each frame in the generalized camera case are expressed by non-centric rays. They may for instance originate from a rig with m central cameras in two different poses, leading to m centric bundles in each generalized camera frame. Let $\mathbf{f}_i$ and $\mathbf{f}'_i$ still denote the direction vectors of the rays of a correspondence between two generalized cameras, and the ray origins with respect to both frames be given by $\mathbf{v}_i$ and $\mathbf{v}'_i$. The origins are indeed easily recovered from extrinsic calibration parameters defining the position of each camera center inside a common frame for the entire rig. The relative transformation parameters now follow the rule $\alpha_i \mathbf{f}_i + \mathbf{v}_i = \mathbf{t} + \mathbf{R}(\alpha'_i \mathbf{f}'_i + \mathbf{v}'_i)$, where α_i and α'_i still denote the depths along the rays. [8] again showed that the generalized epipolar constraint may appear as

$$(\mathbf{f}_i \times \mathbf{R}\mathbf{f}'_i)^T \mathbf{t} + \mathbf{f}_i^T (\lfloor \mathbf{v}_i \rfloor_\times \mathbf{R} - \mathbf{R} \lfloor \mathbf{v}'_i \rfloor_\times) \mathbf{f}'_i = \mathbf{g}_i^T \tilde{\mathbf{t}} = 0, \tag{4}$$

with

$$\mathbf{g}_i = \begin{bmatrix} \mathbf{f}_i \times \mathbf{R}\mathbf{f}'_i \\ \mathbf{f}_i^T (\lfloor \mathbf{v}_i \rfloor_\times \mathbf{R} - \mathbf{R} \lfloor \mathbf{v}'_i \rfloor_\times) \mathbf{f}'_i \end{bmatrix}, \text{ and } \tilde{\mathbf{t}} = \begin{bmatrix} \mathbf{t} \\ 1 \end{bmatrix}.$$

$\mathbf{g}_i$ denotes a generalization of an epipolar plane normal vector, and $\tilde{\mathbf{t}}$ the homogeneous translation. Using $\mathbf{G} = [\mathbf{g}_1 \ \dots \ \mathbf{g}_n]$, the covariance of this distribution becomes

$$\Sigma_{\mathbf{G}} = \mathbf{G}\mathbf{G}^T = \sum_{i=1}^{n} \mathbf{g}_i \mathbf{g}_i^T, \tag{5}$$

and it again needs to have a rank deficiency of one in the ideal case. $\Sigma_{\mathbf{G}}$ is a 4×4 SPD matrix that depends only on $\mathbf{R}$ and needs to have rank 3. The supplemental material proves that it has similar properties to $\Sigma_{\mathbf{N}}$ with respect to $\mathbf{R}$, and the solution $\hat{\mathbf{R}}$ can be found by solving

$$\hat{\mathbf{R}} = \underset{\mathbf{R} \in SO(3)}{\operatorname{argmin}} \left(\operatorname{rank}(\Sigma_{\mathbf{G}}(\mathbf{R}))\right). \tag{6}$$

2.3. Generalized relative pose and scale

Several applications of the generalized relative pose and scale problem—such as hierarchical structure from motion—involve the registration of two partial view-graphs each one consisting of several central view-points that are registered with respect to each other in terms of their relative pose. This enables a straightforward description of a view-graph as a generalized camera. The origins of the measurement rays of a correspondence $\mathbf{v}_i$ and $\mathbf{v}'_i$ are identical with the positions of the capturing view-points inside the view-graphs' reference frames. As indicated in Figure 1, the difference to the standard generalized relative pose problem is that each view-graph has its own, unknown inherent scale. We parametrize the relative scale by a scalar factor s by which the camera baselines $\mathbf{v}'_i$ in the second view-graph need to be multiplied in order to enable a proper registration. The relative transformation parameters now satisfy the subtly different constraint $\alpha_i \mathbf{f}_i + \mathbf{v}_i = \mathbf{t} + \mathbf{R}(\alpha'_i \mathbf{f}'_i + s\mathbf{v}'_i)$. By again employing factorization techniques, we may easily arrive at

$$(\mathbf{f}_i \times \mathbf{R}\mathbf{f}'_i)^T \mathbf{t} - \mathbf{f}_i^T \mathbf{R} \lfloor \mathbf{v}'_i \rfloor_\times \mathbf{f}'_i s + \mathbf{f}_i^T \lfloor \mathbf{v}_i \rfloor_\times \mathbf{R}\mathbf{f}'_i = \mathbf{q}_i^T \tilde{\mathbf{r}} = 0, \tag{7}$$

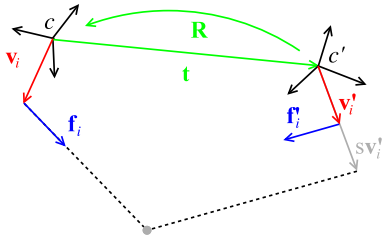

Figure 1. The geometry of the generalized relative pose and scale problem. The difference to the standard generalized case is given by the fact that the baseline of the second ray of each correspondence (i.e. in frame c') needs to be multiplied by a scale factor s in order to satisfy the geometric constraints.

with

$$\mathbf{q}_i = \begin{bmatrix} \mathbf{f}_i \times \mathbf{R}\mathbf{f}'_i \\ -\mathbf{f}_i^T \mathbf{R} \lfloor \mathbf{v}'_i \rfloor_\times \mathbf{f}'_i \\ \mathbf{f}_i^T \lfloor \mathbf{v}_i \rfloor_\times \mathbf{R}\mathbf{f}'_i \end{bmatrix}, \text{ and } \tilde{\mathbf{r}} = \begin{bmatrix} \mathbf{t} \\ s \\ 1 \end{bmatrix}.$$

Using $\mathbf{Q} = [\mathbf{q}_1 \ \dots \ \mathbf{q}_n]$, we can once more derive the covariance matrix of this distribution

$$\Sigma_{\mathbf{Q}} = \mathbf{Q}\mathbf{Q}^T = \sum_{i=1}^{n} \mathbf{q}_i \mathbf{q}_i^T, \tag{8}$$

which has a rank deficiency of one. $\Sigma_{\mathbf{Q}}$ is a 5×5 SPD matrix that depends only on $\mathbf{R}$ and needs to have rank 4. It again can be manipulated in constant time independently of n (see supplemental material), and the solution $\hat{\mathbf{R}}$ can be found by solving

$$\hat{\mathbf{R}} = \underset{\mathbf{R} \in SO(3)}{\operatorname{argmin}} \left(\operatorname{rank}(\Sigma_{\mathbf{Q}}(\mathbf{R}))\right). \tag{9}$$

2.4. Practical rank minimization

All relative pose problems have so far been formulated in terms of a generic rank-function applied to an SPD matrix Σ over the space of rotations. The eigenvalues λ_k of Σ are always positive, and a rank-deficiency of one in turn requires the smallest eigenvalue λ_{min} to be zero. The eigenvalues of Σ describe the variance of the distribution of the (generalized) epipolar plane normal vectors, and thus have a concrete geometrical meaning. This distribution must have zero dilatation in one direction, at least in the noise-free case. As demonstrated in [9, 8], the correct rotation can hence be found by an iterative minimization of λ_{min}, for which there exists a closed-form solution in the 3-dimensional and 4-dimensional cases. Furthermore, the ability to compress the data terms in Σ means that it can be manipulated in constant time independently of the number of correspondences.

Although we know from the spectral theorem and the Weyl bound [3] that an infinitesimal change of Σ leads to an infinitesimal change of λ_{min}—and hence that λ_{min} is a smooth continuously differentiable function of $\mathbf{R}$—there exists no closed-form solution in the 5-dimensional case. The problem also has increased non-linearity in the fully general case, such that none of the methods presented in [9] and [8] is able to easily retrieve the global minimum of the rank-constraint. Section 3 is notably concerned with this problematic, and presents an effective heuristic global minimization strategy of λ_{min}.

There exist alternative rank-approximations that enable us to compute both exact and approximate solutions ensuring a rank-deficiency, but are easier to compute than the smallest eigenvalue of Σ. The constraints are based on the following two important properties of SPD matrices:

- The determinant of Σ is equivalent to the product of all eigenvalues: $\det(\Sigma) = \Pi_k \lambda_k$

- The trace of Σ is equivalent to the sum of all eigenvalues: $\mathrm{trace}(\Sigma) = \sum_k \lambda_k$

Knowing that one eigenvalue has to be zero, it becomes clear that the determinant and also the trace may serve as possible (approximate) criteria to minimize the rank of Σ. The product of all eigenvalues is equivalent to the volume of a bounding box around the covariance ellipsoid, which we verified experimentally to have a considerably flat plateau around the global minimum. The trace-norm, however, is more interesting in that it is only quadratic in the rotation, and thus may serve well to quickly obtain an approximate solution as a convex upper bound.

3. Solving the problem

This section is concerned with the practical solution of the symmetric eigenvalue problems (3), (6), and (9), thus leading to a general framework for solving relative pose problems under an efficient geometric error criterion. After a brief overview of the entire procedure, we will move on to the core of our method: heuristic global energy minimization based on local minimum suppression. The section concludes with a comparison to traditional non-linear minimal and linear non-minimal solution approaches, which underlines the complexity of the problem, and the validity of our chosen approach.

3.1. Overview of the method

Our approach is based on minimizing the smallest eigenvalue λ_{min} of Σ, which is always positive, and may be regarded as an energy term. Since the relative rotation in practical cases is not arbitrarily large, we furthermore chose the minimal Cayley parameterization [2] of a rotation as our optimization variables. The topology of our energy minimization problem may hence be very well visualized by plotting λ_{min} in the 3D space of the Cayley parameters $\mathbf{x} = (x, y, z)$, which we call *Cayley space*. Typical examples of this cost function can be found in [9, 8].

Our method works in three steps:

- We first minimize the trace of Σ in Cayley space, which—as explained in Section 2.4—serves well to find an approximative solution for the global minimum of the rank-function. Let us denote this first solution by $\mathbf{x}_0$. It is found by simple gradient descent from the origin coupled to a line search, though a closed-form solution is possible as well. As shown in Section 4.2, we furthermore determined an upper bound s_{max} for the maximum error of $\mathbf{x}_0$ over countless realistic experiments, s.t. $\|\mathbf{x}_0 - \mathbf{x}_{\text{true}}\| < s_{\text{max}}$. s_{max} serves as a search radius around $\mathbf{x}_0$ for the final global optimization step[1].

- Our method furthermore requires an approximate value for the maximum energy within the search radius, denoted E_{max}. We set this value by finding the smallest eigenvalue for the $3 \times 3 \times 3 = 27$ nodes of a cuboid grid with edge-length $2s_{\text{max}}$ and centered around $\mathbf{x}_0$. Despite its simplicity, this sampling pattern already returns a sufficiently accurate upper bound.

- We then move on to the final heuristic global optimization step, which is explained in the following section.

3.2. Multi-start clustering: Energy minimization with approximate minima suppression

As soon as more than the minimum number of points is used—meaning at least 8 for the generalized relative pose and scale problem—, the topology of the cost volume typically shows a clearly defined global minimum. Furthermore, the number of local minima within the vicinity of the global minimum typically remains relatively low, thus leading to a problem with a manageable degree of non-linearity. Our minimization strategy is adapted to such cases, and belongs to the family of heuristic multi-start clustering global optimization methods [19]. We explain this method at the hand of the one-dimensional example illustrated in Figure 2(a), but it is applicable in arbitrary dimensions.

Multi-start clustering essentially consists of a repetitive application of gradient descent with a randomized starting point. Only exception is given by the first iteration, where we start at $\mathbf{x}_0$. Let us assume that the first run of gradient descent leads us to a local minimum $\mathbf{x}_i$ at which we have an energy E_{min} (cf. Figure 2(b)). A very simple method that would necessarily lead to an improvement consists of sampling the solution space until we have found a new starting point for gradient descent that has lower energy than E_{min}. Unless $\mathbf{x}_i$ is the global minimum, there remains a certain probability of finding a better point with lower energy, meaning that—if we sample homogeneously and long enough—the global minimum will be attained.

The important question is how long "long enough" is? The probability of improving the solution can be characterized by considering the fraction of the solution space for which the energy is smaller than E_{min}. This would for instance be the blue interval in Figure 2(b). While this question can not be answered without complete knowledge about our problem, we may still use this insight to answer a different, related question: How can we increase the probability of finding a starting point inside the convergence basin of the global minium?

The trick is easy: In order to consider a larger convergence basin we simply need to take into account points at higher energy levels than E_{min} as well. We introduce a pa-

[1] Compensating by this initial orientation ensures that we only have to find a small residual rotation in subsequent steps, and thus avoid the degeneracy of the Cayley parametrization.

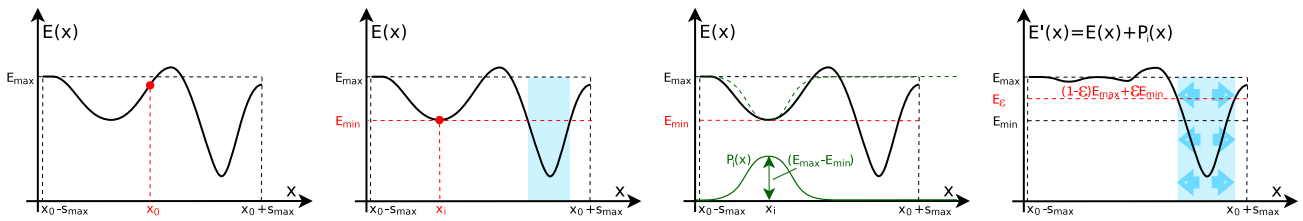

Figure 2. 1D-illustration of multi-start clustering. Figure (a): We seek the global minimum of a function $E(x)$ within a finite search space (e.g. $x \in [x_0 - s_{\max}, x_0 + s_{\max}]$) given that we know a coarse value for the maximum energy within the search interval ($E_{\max}$). Figure (b): After one round of gradient descent from the center of the search space, we arrive at a local minimum x_i with energy $E_{\min}$. By finding an even lower value through random sampling of the search space, we may ultimately converge to the global minimum (i.e. by chosing a point within the blue interval). Figure (c): Aiming at a faster convergence rate, we define a peak $P_i(x)$ at the local minimum x_i that potentially smoothes out the local minimum. Figure (d): Within the smoothed energy function $E'(x) = E(x) + P_i(x)$, we may accept values below a higher energy level $E_\epsilon > E_{min}$, and still obtain only points within the convergence basin of the global minimum. However, we may obtain this point within fewer sampling iterations, as E_ϵ leads to an enlargement of the blue interval of acceptable points.

rameter $\epsilon \in [0, 1]$ used to define a new threshold energy level

$$E_\epsilon = (1 - \epsilon)E_{\max} + \epsilon E_{\min}, \tag{10}$$

where $E_{\max}$ is the approximate value for the maximum energy within our search radius. ϵ is a design variable, which we set to 0.5 throughout this paper.

Since E_ϵ obviously reintroduces a certain probability of ending up in the same local minimum $\mathbf{x}_i$, multi-start clustering requires the addition of approximate minimum suppression. We use the curvature in $\mathbf{x}_i$ as well as knowledge about the approximate maximum energy $E_{\max}$ in our search space to define a peak $P_i(\mathbf{x})$ that—when added to our original energy function $E(\mathbf{x})$—results in a new energy function $E'(\mathbf{x})$ that approximately smoothes out the local minimum at $\mathbf{x}_i$. $P_i(\mathbf{x})$ and $E'(\mathbf{x})$ are illustrated in Figures 2(c) and 2(d), respectively. As expected, this procedure changes the topology of our energy function such that—by using E_ϵ and sampling in $E'(\mathbf{x})$—it ideally increases the probability of finding a good starting point for gradient descent only in the neighbourhood of minima other than $\mathbf{x}_i$.

The technique of multi-start clustering consists of iteratively finding good starting points with energy below E_ϵ inside

$$E'(\mathbf{x}) = E(\mathbf{x}) + \sum_i P_i(\mathbf{x}), \tag{11}$$

and executing gradient descent from there. For every local minimum $\mathbf{x}_i$ found in this way, we define a new peak $P_i(\mathbf{x})$ that in turn changes the topology of our energy function by adding it to $E'(\mathbf{x})$. We furthermore refine every $\mathbf{x}_i$ by executing another round of gradient descent within $E(\mathbf{x})$. We keep track of the current best point—denoted $\mathbf{x}_{\min}$—and the corresponding energy level $E_{\min} = E(\mathbf{x}_{\min})$. By doing a final refinement over $E(\mathbf{x})$, we ensure that $\mathbf{x}_{\min}$ and $E_{\min}$ are not under the influence of any $P_i(\mathbf{x})$.

Even if not finding any more points below E_ϵ, we still execute gradient descent over $E'(\mathbf{x})$ from the lowest starting point we can find within k iterations, followed by a refinement step over $E(\mathbf{x})$. The only real pitfall occurs when the influence of minimum suppression shifts the global minimum into the convergence basin of another local minimum inside the original energy function E. It is intuitively clear that this is a very unlikely thing to happen given the relatively simple topology of our energy volume. We will also demonstrate this claim through our exhaustive experimental validation in Section 4.3. Furthermore, even if the compensation of a local minimum fails in the sense of missing out on some points below E_ϵ, the algorithm eventually still succeeds, as this remaining local minima will simply get suppressed during a following iteration.

3.2.1 Choice of the peak-approximation function

We need a basis-function to approximate peaks for flattening local minima $\mathbf{x}_i$. It should take into account the energy level with respect to $E_{\max}$. It also needs to be a smooth continuously differentiable function, as $E'(\mathbf{x})$ should remain such as well. It furthermore needs to converge to zero the further we move away from $\mathbf{x}_i$, as we aim at restricting the influence of $P_i(\mathbf{x})$ to the convergence basin of $\mathbf{x}_i$. The last requirement is given by an easy way to impose a given curvature in $\mathbf{x}_i$. A good function that fulfills these requirements is given by a gaussian bell curve, defined as

$$P_i(\mathbf{x}) = (E_{\max} - E'_i)\, e^{\frac{(\mathbf{x}-\mathbf{x}_i)^T \mathbf{H}_i (\mathbf{x}-\mathbf{x}_i)}{2(E_{\max} - E'_i)}}, \tag{12}$$

with $\mathbf{H}_i$ defined as the Hessian matrix of $E'(\mathbf{x})$ in the point $\mathbf{x}_i$, and E'_i as the energy level of $E'(\mathbf{x})$ in $\mathbf{x}_i$. Note that $E'(\mathbf{x})$ in these contexts is the reshaped energy function just before $P_i(\mathbf{x})$ is added. It is easily verified that the Hessian of $P_i(\mathbf{x})$ in the point $\mathbf{x}_i$ equals to $\mathbf{H}_i$.

By defining peaks in this way, the sum of all peaks $\sum_i P_i(\mathbf{x})$ effectively becomes a Gaussian Mixture Model (GMM) that approximates the function $E_{\max} - E(\mathbf{x})$.

3.2.2 Summary of the algorithm

A detailed summary of our algorithm is provided in the supplemental material. The Hessian is estimated by a double numerical differentiation. As for the very simple version (i.e. $\epsilon = 1$ and deactivated minimum suppression), the success of our algorithm still depends on whether we sample long enough. We therefore include a minimum number of iterations as well as a coarse minimum requirement for E_{min} into our termination condition. Furthermore, local minima are only suppressed if the final residual is below E_{min}, thus constraining the suppression mechanism to sufficiently relevant local minima.

The method is applicable to problems in arbitrary dimensions, given that the following requirements are fulfilled:

- The search space of the problem is finite.
- An approximate value for the maximum energy within the search space is available.
- The problem has only "moderate" non-linearity.

3.3. Comparison to alternative approaches

Judging the significance of the presented optimization strategy naturally requires a comparison to alternative, classical solution approaches. Relative pose solvers typically rely on some algebraic error metric, which can lead to efficient closed-form solutions with linear complexity in the number of points. They show best performance when applied in the minimal case, as solutions are then computed exactly independently of the employed metric (up to numerical inaccuracies). We attempted to derive a minimal polynomial solver by extending the formulation presented in [15] by the additional relative scale factor s. We succeeded in computing a Gröbner basis for a random problem posed within a finite prime field[2]. The parametrization leads to 140 solutions, which proves the existence of a closed-form solution. However, the substantial complexity of the computation suggests that the algorithm would be highly unstable and not useful in practice. Furthermore, applying the technique presented in [10] for finding the necessary elimination template fails due to memory overload, which clearly underlines the complexity of the discussed problem.

The form of a linear solver is not attractive neither. As presented in [12], a linear solver for generalized cameras can be obtained by applying a direct linear transformation to the original generalized essential matrix constraint. Taking the additional scale factor s into account, one would need 26 correspondences for solving the problem, a number that clearly complicates the application within outlier-affected scenarios, and leads to bad noise resilience due to massive over-parametrization. In contrast, we propose a direct iterative minimization of a geometric error criterion. The complexity is linear (i.e. the iterative part does not depend on the number of points), and the method is transparently applicable to an entire class of relative pose problems.

[2]The first step of finding a minimal solver based on the Gröbner basis theory traditionally consists of applying the Buchberger algorithm [1] or F4 [4] to a problem posed in a finite field, where exact arithmetics support the identification of problem solvability and the exact form of the basis.

4. Experimental validation

The focus of this section lies on a concise evaluation of our proposed method based on exhaustive simulation experiments. We start by verifying the practical assumption of a bounded error of the trace-based initialization. We then demonstrate the performance of our algorithm in comparison to 2D-3D and 3D-3D registration approaches in the most challenging 7DoF case[3]. We furthermore illustrate the benefit of our minima suppression mechanism, namely improved algorithm convergence rate. We complete the evaluation by a succesful application to a real-world example.

4.1. Outline of the experiments

Our experiments consist of n random points shifted by a varying displacement d along the z-axis. We then define two view-graph frames, one equal to the world frame, and one shifted by a translation of maximum norm 2. Within each view-graph, we then define m central cameras plus their extrinsic parameters. All frames have identity orientation, as our initialization is able to account for the majority of the relative rotation in any case (and at the same time introduces only a small residual rotation due to the trace-based approximation). Our correspondences across cameras and view-graphs are finally obtained by projecting all points into all cameras, and adding a varying amount of noise (we assume spherical cameras with 800 pix focal length, and Gaussian noise of 1 pix std. dev.). For the 2D-3D and 3D-3D registration approaches, 3D world points are obtained by triangulating 2D points from the cameras in each view-graph, as this adds realistic noise to our 3D world points as well. We only present results for the rotation, which is our primary optimization variable. However, as mentioned in Section 2, our algorithm implicitly solves for relative translation and scale as well (the corresponding values are given by the eigenvector that corresponds to the smallest eigenvalue). We provide those results in the supplemental material, as they essentially support the same conclusions. The rotation error is easily given by the norm

[3]Note that all methods could be extended by nonlinear minimization of the reprojection error (i.e. bundle adjustment), but we restrict outselves to pure and efficient globally optimal solutions that can be employed many times within a RANSAC scheme, and over small sets of correspondences. We also omit the method presented in [18], as we only consider fully general approaches that do not require any information other than the raw correspondences. As such, the only competing methods are 2D-3D and 3D-3D alignment alternatives.

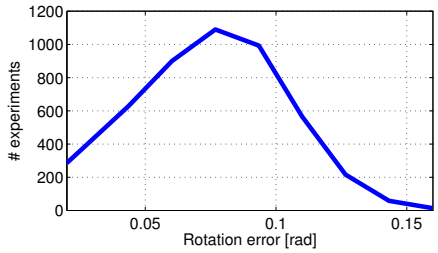

Figure 3. Histogram of initialization errors over 5000 random experiments.

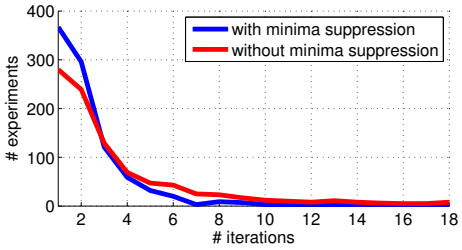

Figure 5. Histogram of the number of required iterations to obtain a solution with at most 0.005 rad error (with and without minimum suppression, each time over 1000 experiments).

of the axis-angle representation, as the true rotation always corresponds to identity.

4.2. Accuracy of the initialization

By minimizing the trace of Σ in Cayley space, we obtain initial solutions for our relative rotation. A histogram of initialization errors over 5000 random experiments is indicated in Figure 3, confirming that it is bounded and staying below about 0.2 rad. Note that—by shifing the world points along the z-axis—we create "ill-posed" and potentially ambiguous problems without omni-directional 2D measurements[4].

4.3. Comparison to other registration methods

Figure 4 illustrates a comparison between our 2D-2D registration method, a state-of-the-art 2D-3D registration approach [17], and a 3D-3D registration method [13] for solving similarity transformations. We check the influence of the most important parameters of our experiments. Our standard experiment consists of having 4 cameras in each view-graph, and 100 points shifted by 10 units along the z-axis. In row one, we then show results for a varying number of cameras within each view-graph (from left to right: $m =$ 4, 10, 20, and 50). In row 2, we present results for a varying number of points (from left to right: $n =$ 12, 20, 100, and 1000). In row 3, we finally show the influence of a varying average depth of the points (from left to right: $d =$ 10, 20, 30, and 40).

In general, the 2D-2D registration method shows best performance, with sometimes even overlapping mean and median errors (i.e. the algorithm is very robust and consistently converges to the global minimum). The mean error of the 2D-2D registration method sometimes even stays below the median error of the alternative approaches. Perhaps the biggest weakness of our algorithm is identified in cases of low numbers of correspondences (row 2 on the left), and of increased depth with respect to the camera baselines (row 3 on the right). The residual mean error at zero noise indicates that our algorithm occasionally fails to converge in these situations. This problem could perhaps be avoided by further tuning of the algorithm parameters. However, even in those cases, our solution seems to provide the best trade-off between robustness and accuracy for realistic noise, and occasional convergence to wrong minima is tolerated within a RANSAC scheme. A further advantage compared to 3D-3D and 2D-3D registration methods is that points need to be visible in pairs of views only.

[4]In the omni-directional case, the error of the trace constraint becomes almost negligibly small.

4.4. Improvement of the convergence rate

In our final simulation experiment, we show the improvement of the convergence rate by toggling the minima suppression mechanism. Figure 5 shows a histogram of the number of required iterations to obtain a solution with at most 0.005 rad error, with and without minimum suppression. Even for our relatively simple 3-dimensional case, we can observe a clear improvement of the convergence rate.

Given the low number of required iterations and the simplicity of the compressed data handling, the algorithm easily scales up to real-time applications, and thus applicability within a RANSAC scheme. The algorithm requires approximately 1ms per iteration, and this duration is largely independent of the number of correspondences.

4.5. Validation over real data

We applied our algorithm to images of the TUM benchmark sequence *freiburg2_xyz* [16], for which accurate ground-truth camera poses as well as intrinsic camera parameters are available. We extract correspondences by taking a quadruplet of overlapping images from the sequence and finding SIFT features that match correctly across all 4 views. We furthermore put a threshold on the reprojection error of the triangulated correspondences in order to eliminate all remaining outliers, by which we finally obtain about 20 correspondences. We then split the quadruplet up into two pairs of images, thus representing two generalized cameras or view-graphs of minimal size. We simply use the ground truth camera poses to define the position and orientation within each view-graph, thus ensuring that inaccuracies are due to noise in the 2D measurements only. We apply the three algorithms as follows:

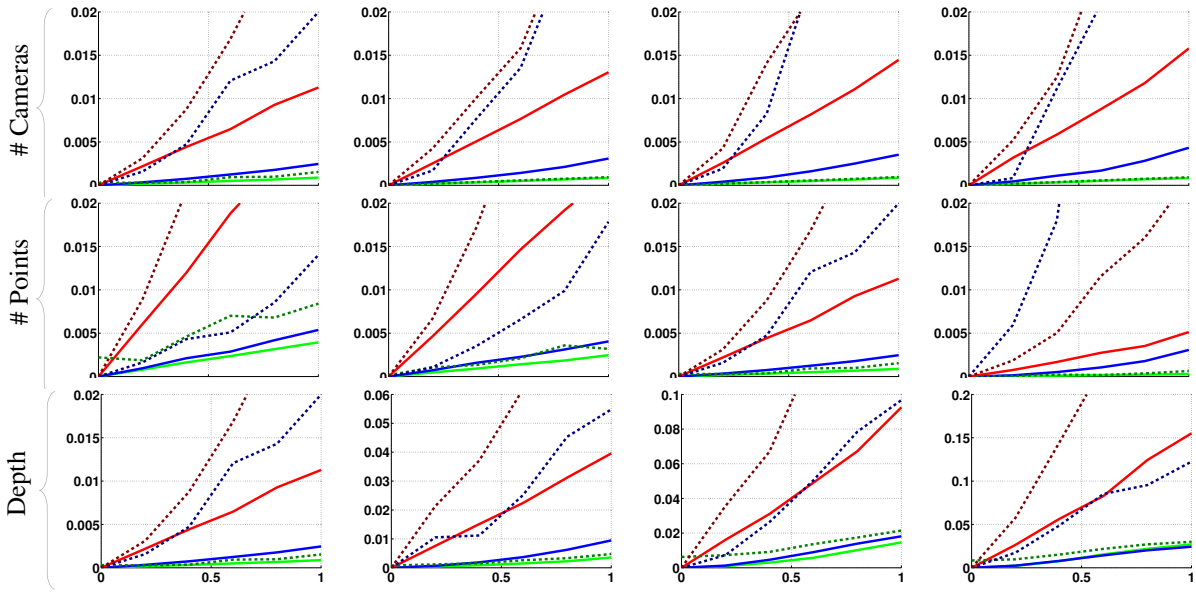

Figure 4. Comparison of errors in rotation estimation [rad] for the presented method (green), the 2D-3D method presented in [17] (blue), and the 3D-3D method from [13] (red). Median errors are indicated with bright solid lines, whereas mean errors appear with dark, dotted lines. Each plot shows results for varying noise levels [pix] along the x axis. The standard experiment consists of having 4 cameras in each view-graph, and 100 points shifted by 10 units along the z-axis. In row one, we then show results for a varying number of cameras within each view-graph (from left to right: $m = 4$, 10, 20, and 50) In row 2, we present results for a varying number of points (from left to right: $n = 12$, 20, 100, and 1000). In row 3, we show the influence of a varying depth of the points (from left to right: $d = 10$, 20, 30, and 40).

- We triangulate 3D points from each pair individually, and apply [13] to the resulting 3D-3D correspondences.
- We then consider only 3D points triangulated from the first pair of images, plus their corresponding normalized image points in the second pair of frames. We apply [17] to the resulting 2D-3D correspondences.
- We finally run our presented approach over all 2D-2D correspondences across the pairs of views.

Since the camera poses are all registered in absolute scale with respect to one and the same world frame, the algorithms should ideally come up with an identity transformation between the two view-graphs. We again look at the rotation error: [13] leads to a relatively large error of 0.1128 rad, [17] leads to 0.0184 rad, and our 2D-2D registration algorithm leads to only 0.0138 rad error in the relative rotation. We thus verify our results from simulation, and prove that solving similarity transformations based on 2D-2D registration leads to improved view-graph concatenation.

5. Discussion

The present paper contains three important contributions. First, we contribute to a relatively novel, challenging, and meaningful problem in the field of geometric computer vision, called the generalized relative pose and scale problem. It permits the solution of a common task in structure from motion as a pure 2D-2D registration problem, namely partial view-graph registration. We confirm that the independence of 3D point clouds improves the accuracy over existing 2D-3D and 3D-3D registration methods, and present the first completely general solution. Second, we show that heuristic global energy minimization with multi-start clustering effectively solves this challenging problem. Third, we provide a unified and geometrically meaningful, direct solution to all calibrated relative pose problems. Over the years, the progress around the perspective n-point problem led us from linear to non-linear, then to nonlinear and geometrically optimal, and finally to non-linear and geometrically optimal O(n)-complexity solutions. Although iterative and approximate in its nature, the class of solvers presented in this paper could be regarded as a step towards a relative pendant to the most state-of-the-art absolute pose solvers. As shown in several recent works, the employed geometric error metric indeed leads to outstanding accuracy in comparison to algebraic solutions. Future research consists of implementing the driving application behind this paper, namely hierarchical view-graph fusion.

ACKNOWLEDGMENT

This research is supported by the ARC Centre of Excellence for Robotic Vision, as well as the ARC grant DE150101365. The work is also supported by NSF Grant IIS-1219261, ONR Grant N00014-14-1- 0133 and NSF Graduate Research Fellowship Grant DGE- 1144085.

References

[1] B. Buchberger. *Multidimensional Systems Theory - Progress, Directions and Open Problems in Multidimensional Systems*. Reidel Publishing Company, Dodrecht - Boston - Lancaster, 1985.

[2] A. Cayley. About the algebraic structure of the orthogonal group and the other classical groups in a field of characteristic zero or a prime characteristic. *Reine Angewandte Mathematik*, 32, 1846.

[3] F. Dopico, J. Moro, and J. Molera. Weyl-type relative perturbation bounds for eigensystems of Hermitian matrices. *Linear Algebra and its Applications*, 309(1–3):3–18, 2000.

[4] J. Faugère. A new efficient algorithm for computing Gröbner bases (F4). *Journal of Pure and Applied Algebra*, 139(1–3):61–88, 1999.

[5] M. Grossberg and S. Nayar. A General Imaging Model and a Method for Finding its Parameters. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 108–115, Kyoto, Japan, 2001.

[6] R. Hartley. In defense of the eight-point algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 19:580–?593, 1997.

[7] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, second edition, 2004.

[8] L. Kneip and H. Li. Efficient Computation of Relative Pose for Multi-Camera Systems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, USA, 2014.

[9] L. Kneip and S. Lynen. Direct Optimization of Frame-to-Frame Rotation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, Sydney, Australia, 2013.

[10] Z. Kukelova, M. Bujnak, and T. Pajdla. Automatic generator of minimal problem solvers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 302–315, Marseille, France, 2008.

[11] H. Li, R. Hartley, and J.-H. Kim. A Linear Approach to Motion Estimation using Generalized Camera Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, Anchorage, Alaska, USA, 2008.

[12] R. Pless. Using many cameras as one. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 587–593, Madison, WI, USA, 2003.

[13] P. Schonemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31:1–10, 1966.

[14] H. Stewénius, C. Engels, and D. Nistér. Recent developments on direct relative orientation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 60(4):284–294, 2006.

[15] H. Stewénius and D. Nistér. Solutions to Minimal Generalized Relative Pose Problems. In *Workshop on Omnidirectional Vision (ICCV)*, Beijing, China, 2005.

[16] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, Vilamoura-Algarve, Portugal, 2012.

[17] C. Sweeney, V. Fragoso, T. Hollerer, and M. Turk. gDLS: A Scalable Solution to the Generalized Pose and Scale Problem. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Zurich, Switzerland, 2014.

[18] C. Sweeney, L. Kneip, T. Höllerer, and M. Turk. Computing similarity transformations from only image correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015.

[19] W. Tu and R. Mayne. Studies of multi-start clustering for global optimization. *International Journal for Numerical Methods in Engineering*, 53(9):2239–2252, 2002.

[20] J. Ventura, C. Arth, G. Reitmayr, and D. Schmalstieg. A minimal solution to the generalized pose-and-scale problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, USA, 2014.