2020 IEEE International Conference on Robotics and Automation (ICRA)
31 May - 31 August, 2020. Paris, France

Reliable frame-to-frame motion estimation for vehicle-mounted surround-view camera systems

Yifu Wang*, Kun Huang$^{\oplus}$, Xin Peng$^{\oplus}$, Hongdong Li*, Laurent Kneip$^{\oplus}$

***Abstract*— Modern vehicles are often equipped with a surround-view multi-camera system. The current interest in autonomous driving invites the investigation of how to use such systems for a reliable estimation of relative vehicle displacement. Existing camera pose algorithms either work for a single camera, make overly simplified assumptions, are computationally expensive, or simply become degenerate under non-holonomic vehicle motion. In this paper, we introduce a new, reliable solution able to handle all kinds of relative displacements in the plane despite the possibly non-holonomic characteristics. We furthermore introduce a novel two-view optimization scheme which minimizes a geometrically relevant error without relying on 3D point related optimization variables. Our method leads to highly reliable and accurate frame-to-frame visual odometry with a full-size, vehicle-mounted surround-view camera system.**

I. Introduction

Autonomous driving represents a significant technological development with potentially very strong impact in several domains including traffic efficiency and road safety. Regular video cameras are cost-effective sensors which have already become part of the standard sensing equipment of modern cars. Such sensors may unlock lower-level autonomy in more controlled and less critical scenarios, even in the absence of other exteroceptive sensors. The present paper relies on a vehicle-mounted surround-view camera system, and notably aims at a solution to the problem of finding the planar motion of the vehicle. Such systems often include four fish-eye cameras pointing into the forward, backward, and side-ways directions. Moreover, as the overlap between the cameras' view-points is often very limited, our solution furthermore employs only temporal image correspondences measured by the same camera. Non-overlapping surround-view camera systems are described by the generalized camera model. Solutions to the generalized relative pose problem are therefore applicable to our case.

While a number of successful works have already been presented, the conclusion is that no method is all-powerful and able to handle the possibly non-holonomic planar vehicle motion[1]. Generalized relative pose solvers solving for all 6 degrees-of-freedom have substantial computational complexity and solution multiplicity in the minimal case [18], or require too many samples and linearisations in the non-minimal case [14], thus leading to unstable results under noise. Some generalized solvers even degenerate in the case of non-holonomic motion [10]. Several relative pose solvers [16], [8] specialise in the aforementioned non-holonomic planar vehicle motion estimation, including solvers designed for multi-camera systems [13]. While these methods are very robust, they rely on the ideal assumption of a fixed steering angle; the centre of rotation as introduced by the Ackermann motion model is however a dynamic point for the majority of time during which a vehicle is taking a turn. Although a single-view solver relying on the more correct planar motion model has been presented [1], a generalized planar motion solver that simultaneously exploits the information from multiple cameras appears as a gap in the literature.

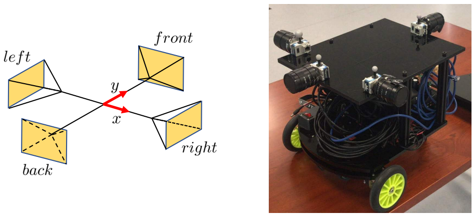


Fig. 1. Non-overlapping, surround-view four-camera system as analyzed in this paper.

Our contributions are as follows:

- We present the first generalized relative pose solver for general planar vehicle motion and non-overlapping multi-camera systems.
- We present a new univariate objective function that relies on a parallel evaluation of the epipolar geometry for each individual camera, rather than simply replacing the motion parametrization in existing formulations that rely on the generalized essential matrix (which degenerates in the case of non-overlapping multi-camera arrays).
- We introduce a modified, iteratively reweighted optimization of the planar motion that minimizes the geometrically relevant object space error. Most notably, the objective function remains a uni-variate expression, and therefore outperforms two-view bundle adjustment in terms of computational efficiency.

The paper is organized as follows. More detailed related work is discussed in Section II. Section III provides a brief review of epipolar geometry and its formulation as an eigenvalue problem. Section IV introduces our objective functions and solution strategy. Section V finally demonstrates our results on both simulated and real data. Our simulations compare robustness, accuracy, and computational efficiency

*Australian National University; $^{\oplus}$ShanghaiTech University; The authors would like to thank the fundings sponsored by Natural Science Foundation of Shanghai (grant number: 19ZR1434000)and Natural Science Foundation of China (grant number: 61950410612).

[1] Non-holonomic motion exists if several degrees of freedom are coupled. In the case of a vehicle, the non-holonomicity is described by the Ackermann motion model, which predicts that cars move along circular arcs.

978-1-7281-7395-5/20/$31.00 ©2020 IEEE

of multiple algorithms, and demonstrate that our proposed method is able to outperform in all aspects. The potential of our method is finally confirmed on multiple real-world datasets, where we demonstrate highly reliable and accurate visual odometry performance.

II. Related work

The most important related solvers that either exploit the geometry of generalized cameras [18], [14], [10] or non-holonomic motion constraints [16], [8], [13] have already been introduced in the introduction. The 6-point solver presented in [18] proposes the solution to the generalized relative pose problem based on the Gröbner basis theory, and uses 6 ray-correspondences in order to come up with 64 solutions. Li et al. [14] provide a linear algorithm that requires 17 correspondences in general, and 16 or 14 correspondences in certain special situations to solve the generalized relative pose problem. A solution that factorizes the generalized relative pose problem as an iterative optimization over relative rotation is presented in [10]. The first work that exploits the non-holonomic constraints of planar vehicles to parameterize the motion with only 1 feature correspondence is given by [16] and extended to n views in [8]. [13] furthermore applies the model to multi-perspective camera systems.

Further related work is given by [15], who is the first to introduce the paradigm of *using many cameras as one*, and Lee et al. [12], who look at the generalized relative pose problem with a known reference direction. This problem is highly related to ours in that it only solves for a one-dimensional degree of freedom rotation. However, as shown in their paper, the algorithm again potentially degenerates for planar vehicle motion, most notably if the relative rotation becomes identity. Our work is also naturally related to the standard relative pose problem. A good overview of epipolar geometry is given in [7]. The most popular solvers for the relative pose problem are given by [6] and [17]. Another foundational work for ours is presented by Kneip et al. [11], who are the first to formulate epipolar geometry as an eigenvalue minimization problem in the space of rotations.

III. Foundations

This section reviews the basic idea of direct optimization of frame-to-frame rotation. We start by introducing the geometry of generalized cameras. We furthermore summarize the prior centralized and generalized methods, and conclude with a brief motivation of our new solver.

A. Notations and prior assumptions

We assume that we have an intrinsically and extrinsically calibrated multi-camera system. Without loss of generality, each 2D image point can therefore easily be expressed as a normalized 3D bearing vector. Considering two consecutive frames, let $\mathbf{f}_i^j$ and $\mathbf{f'}_i^j$ be the unit bearing vectors pointing at the same 3D world point $\mathbf{p}_i$ from the jth camera at the first and second frame, respectively. Let $\mathbf{R}_{c_j}$ furthermore be the rotation from the camera to the common body frame b, and $\mathbf{t}_{c_j}$ the position of camera j inside the body frame. Let $\mathbf{t}_b$ and $\mathbf{R}_b$ furthermore denote the relative pose of the body frame between two subsequent view-points, such that $\mathbf{p}_i = \mathbf{R}_b\mathbf{p}'_i + \mathbf{t}_b$ transforms points from the second frame b' back to the first frame b. To conclude, let $\mathbf{t}_j^c$ and $\mathbf{R}_j^c$ be the equivalent transformation parameters seen from camera j, i.e. transforming points from camera frame c'_j back to camera frame c_j.

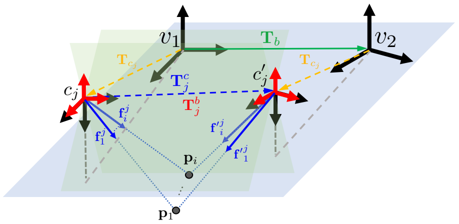


Fig. 2. Geometry of the generalized camera system.

B. Brief review of epipolar geometry and its formulation as an eigenvalue problem

The epipolar incidence relationship is given by $\mathbf{f}_i^{jT}\lfloor\mathbf{t}_j^c\rfloor_\times\mathbf{R}_j^c\mathbf{f'}_i^j = 0$, and most algebraic solvers therefore minimise the sum of squared errors

$$\underset{\mathbf{t}_j^c,\mathbf{R}_j^c}{\operatorname{argmin}} \sum_i (\mathbf{f}_i^{jT}\lfloor\mathbf{t}_j^c\rfloor_\times\mathbf{R}_j^c\mathbf{f'}_i^j)^2 \tag{1}$$

As illustrated in [11], we can apply the scalar triple product rule to the algebraic incidence relationship, and—by defining the epipolar plane normal vector as

$$\mathbf{n}_i^j = \mathbf{f}_i^j \times \mathbf{R}_j^c\mathbf{f'}_i^j \tag{2}$$

—we easily arrive at the following modified objective for the algebraic energy minimization

$$\underset{\mathbf{t}_j^c,\mathbf{R}_j^c}{\operatorname{argmin}}\ \mathbf{t}_j^{cT}\left(\sum_i \mathbf{n}_i^j\mathbf{n}_i^{jT}\right)\mathbf{t}_j^c. \tag{3}$$

This objective is simple to solve by an eigenvalue minimisation of the matrix $\sum_i \mathbf{n}_i^j\mathbf{n}_i^{jT}$, which only depends on $\mathbf{R}_j^c$. Furthermore, as illustrated in [10], the matrix can be augmented to a 4×4 matrix to solve for the generalized case. However, as further explained in [10], this objective is not well suited for multi-camera arrays, as the latter case leads to a zero energy for identity rotation in the eigenvalue minimisation objective. As further illustrated in [14] and [12], critical motions with non-overlapping multi-camera systems even affect linear formulations using the generalized essential matrix. The following section introduces a solution able to handle all kinds of planar motion.

IV. Theory

We start by seeing a new univariate objective function, which enables the parallel evaluation of the epipolar geometry for each individual camera as a multi-eigenvalue problem. We then proceed to see both algebraic and geometric variants of the energy minimized in our paper, which is optimized over the space of rotations only, and introduce an iteratively reweighted optimization of the planar motion that minimizes the geometrically relevant object space error. The section concludes with the derivation of the relative translation.

A. Formulation as a multi-eigenvalue problem

We now proceed to the core of our contribution, which is a novel algorithm for estimating general planar motion for non-overlapping multi-camera systems. Let $\mathbf{n}_i^j$ still be an epipolar plane normal vector, given by (2). As illustrated in Figure 2, in the case of a calibrated multi-camera system, we can use the known extrinsic rotation $\mathbf{R}_{c_j}$ to rotate all the observed unit bearing vectors of each camera into a frame that is still centered at camera c_j but has similar orientation than the local body frame. We can thus obtain an alternative multi-camera system in which all cameras simply have the same orientation than the local body frame. It can be easily observed that all the new normal vectors expressed in the body frame and given by

$$\mathbf{n}_i^{b_j} = (\mathbf{R}_{c_j}\mathbf{f}_i^j) \times \mathbf{R}_b(\mathbf{R}_{c_j}\mathbf{f'}_i^j) \tag{4}$$

still span a plane that is orthogonal to the translation $\mathbf{t}_j$. Similar to [11], [10], our target remains a solution to the relative displacement that depends only on $\mathbf{R}_b$.

In order to enforce all normal vectors of each camera to obey the coplanarity condition, the basic approach consists of stacking the normal vectors from corresponding cameras into the matrix $\mathbf{N}_j = [\mathbf{n}_1^j \quad \dots \quad \mathbf{n}_i^j]^T$ such that $\mathbf{N}_j\mathbf{t}_j^c = 0$. Thus, the relative rotation $\mathbf{R}_b$ can be derived by jointly minimizing the smallest eigenvalue of the matrices $\mathbf{M}_j = \mathbf{N}_j\mathbf{N}_j^T$ from each camera. If $\lambda_{\mathbf{M}_j,min}$ denotes the smallest eigenvalue of $\mathbf{M}_j$, our final objective becomes

$$\mathbf{R}_b = \operatorname{argmin}_{\mathbf{R}_b} \sum_j (\lambda_{\mathbf{M}_j,min})^2, \text{ where} \tag{5}$$

$$\mathbf{M}_j = \sum_{i=1}^{n_j} ((\mathbf{R}_{c_j}\mathbf{f}_i^j) \times \mathbf{R}_b(\mathbf{R}_{c_j}\mathbf{f'}_i^j))((\mathbf{R}_{c_j}\mathbf{f}_i^j) \times \mathbf{R}_b(\mathbf{R}_{c_j}\mathbf{f'}_i^j))^T. \tag{6}$$

It is important to realize that this objective is different from (3) in [11]. Unless proceeding to a generalization as presented in [10], it is not possible to add all normal vectors to one co-planarity condition as each camera has a potentially different translation vector, and therefore defines a different plane for its epipolar plane normal vectors. However, the rotation is the same for each camera, and—owing to the fact that the eigenvalue formulation only depends on the relative rotation—we may still jointly minimize all objectives. In the following, we concentrate on the case of planar motion, for which the relative rotation has only a single degree of freedom. Note however that the formulation makes no assumptions about the translation, and may therefore be equally applied to any relative displacements for which at least 2 of the rotational degrees of freedom are known (e.g. zero, or measured by an alternative sensor).

We choose the Cayley [3] parameters $\mathbf{v} = [0 \ 0 \ z]^T$ to represent the rotation $\mathbf{R}_b$, the latter being given as

$$\mathbf{R}_b = 2(\mathbf{v}\mathbf{v}^T - \lfloor \mathbf{v} \rfloor_\times) + (1 - \mathbf{v}^T\mathbf{v})\mathbf{I}. \tag{7}$$

Note that we omit the scale factor as it equally affects all terms in all energies. The result is a very efficient non-linear optimization over a single parameter only. Note that the direction of each camera's relative translation can be recovered by looking at the eigenvector that corresponds to the smallest eigenvalue. As shown in Section IV-C, they can be further used to compute the scaled relative translation between the two viewpoints once the relative rotation $\mathbf{R}_b$ has been found.

Similar to [11], the non-linear problem can be efficiently solved by implementing a Levenberg-Marquardt scheme. In our constraint (5), the only unknown parameter with respect to the sum of squares of the smallest eigenvalues is the Cayley parameter z of rotation $\mathbf{R}_b$. Owing to the fact that the angular velocities of the vehicle's motion are similar from frame to frame, a starting point for the optimization is easily given by propagating the relative rotation between the previous pair of viewpoints. An exhaustive search can be used to initialize the rotation if no prior is available.

B. Object-space error based refinement

For a perspective camera, a purely translational displacement that is parallel to the image plane can cause a very similar disparity than a pure rotation around an orthogonal axis in the image plane, and vice-versa. The hereby described *rotation-translation ambiguity* is furthermore amplified by sideways looking, fronto-parallel cameras, especially if they have a very limited field of view (FoV). The separation into 4 eigenvalue problems that are solved in parallel naturally raises the question how this affects the algorithm's ability to deal with such ambiguities. Though the algebraic solution is not geometrically or statistically meaningful, it generally leads to satisfying results at a low computational cost. However, the rotation-translation ambiguity can easily lead to local minima in the algebraic objective error in the above described case. In an aim to solve this problem, we introduce an object-space error based objective as an iterative refinement step. It is to be understood as an efficient replacement of 2-view bundle adjustment using the more traditional reprojection error. We define the object-space error as the distance between the rays defined by $\mathbf{f}_i^j$ and $\mathbf{f'}_i^j$. Starting from the definition of the distance between two skew lines [5], we derive the geometric object-space error for a single correspondence to be

$$d = \frac{((\mathbf{R}_{c_j}\mathbf{f}_i^j) \times \mathbf{R}_b(\mathbf{R}_{c_j}\mathbf{f'}_i^j)) \cdot \vec{\mathbf{t}}_j^c}{\|(\mathbf{R}_{c_j}\mathbf{f}_i^j) \times \mathbf{R}_b(\mathbf{R}_{c_j}\mathbf{f'}_i^j)\|}. \tag{8}$$

$\vec{\mathbf{t}}_j^c$ represents the direction of the relative translation. The optimization problem is finally given as

$$\{\mathbf{R}_b, \vec{\mathbf{t}}_j^c\} = \operatorname{argmin}_{\mathbf{R}_b, \vec{\mathbf{t}}_j^c} \sum_{i=0}^{n_j} (\frac{((\mathbf{R}_{c_j}\mathbf{f}_i^j) \times \mathbf{R}_b(\mathbf{R}_{c_j}\mathbf{f'}_i^j)) \cdot \vec{\mathbf{t}}_j^c}{\|(\mathbf{R}_{c_j}\mathbf{f}_i^j) \times \mathbf{R}_b(\mathbf{R}_{c_j}\mathbf{f'}_i^j)\|})^2. \tag{9}$$

It is easy to see that the same objective can again be minimized by solving the iteratively reweighted eigenvalue-minimization problem

$$\mathbf{R}_b = \operatorname{argmin}_{\mathbf{R}_b} \sum_j (\lambda_{\tilde{\mathbf{M}}_j,min})^2, \text{ where} \tag{10}$$

$$\tilde{\mathbf{M}}_j = \sum_{i=1}^{n_j} \frac{(\mathbf{n}_i^{b_j})(\mathbf{n}_i^{b_j})^T}{\|\mathbf{n}_i^{b_j}\|_2^2}. \tag{11}$$

The proposed object-space error minimization strategy still depends only on the relative rotation $\mathbf{R}_b$, meaning a one-dimensional optimization space in the case of planar motion. We confirmed through a series of simulation experiments that the minimization of the object-space error is much more stable for different FoVs than the algebraic objective. It can effectively avoid wrong minima caused by rotation-translation ambiguity. The computational complexity of the presented object-space error minimization objective is significantly lower than the one of standard two-view bundle adjustment. The latter not only optimizes over both rotation and translation parameters, but—if using the classical reprojection error—also over the 3D coordinates of each landmark. A dedicated experiment comparing the two iterative refinement alternatives is presented in Section V-D.

C. Recovery of relative translation

In order to recover the translation in absolute scale, we start by formulating the hand-eye calibration constraint for camera c_j inside the multi-camera system [7]:

$$\begin{cases} \mathbf{t}_b = \mathbf{t}_{c_j} + \mathbf{R}_{c_j}\mathbf{t}^c_j - \mathbf{R}_{c_j}\mathbf{R}^c_j\mathbf{R}^T_{c_j}\mathbf{t}_{c_j} \\ \mathbf{R}_b = \mathbf{R}_{c_j}\cdot\mathbf{R}^c_j\cdot\mathbf{R}^T_{c_j} \end{cases} \tag{12}$$

As mentioned in Section IV-A, we can compensate for each camera's extrinsic rotation $\mathbf{R}_{c_j}$, and thus obtain the simpler constraint

$$\mathbf{t}_b = \mathbf{t}_{c_j} + \mathbf{t}^c_j - \mathbf{R}_b\mathbf{t}_{c_j}. \tag{13}$$

For each relative translation $\mathbf{t}^c_j = \lambda_j \cdot \vec{\mathbf{t}}^c_j$, directions $\vec{\mathbf{t}}^c_j$ can be easily computed by composing $\mathbf{M}_j$ and deriving the eigenvector corresponding to the optimized $\mathbf{R}_b$. All pair-wise constraints in the form of (13) can now be grouped into a linear problem $\mathbf{Ax} = \mathbf{b}$, where

$$\mathbf{A} = \begin{bmatrix} \vec{\mathbf{t}}^c_1 & & -\mathbf{I} \\ & \cdots & \cdots \\ & \vec{\mathbf{t}}^c_4 & -\mathbf{I} \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} (\mathbf{R}_b - \mathbf{I})\mathbf{t}_{c_1} \\ \cdots \\ (\mathbf{R}_b - \mathbf{I})\mathbf{t}_{c_4} \end{bmatrix}, \tag{14}$$

and $\mathbf{x} = \left[\lambda_1 \ \ldots \lambda_4 \ \mathbf{t}^T_b\right]^T$. $\mathbf{A}$ and $\mathbf{b}$ can be computed from the known extrinsics and the relative rotation $\mathbf{R}_b$, whereas $\mathbf{x}$ contains all unknowns. The non-homogeneous linear problem $\mathbf{Ax} = \mathbf{b}$ can be solved by a standard technique such as singular value decomposition (SVD). Note that the system shows an obvious characteristic of non-overlapping multi-camera arrays, namely that metric scale remains unobservable if $\mathbf{R}_b = \mathbf{I}$. The rotation however remains computable.

V. Experimental evaluation

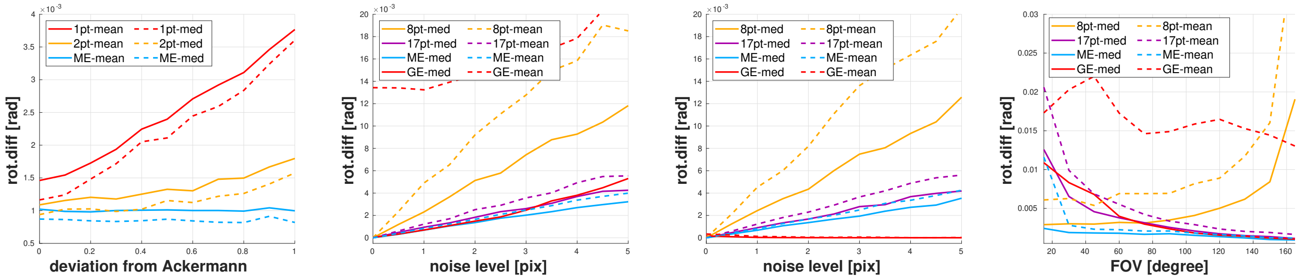
We test our algorithm on both synthetic and real data. Our solver depends on planar motion and is designed for non-overlapping multi-camera systems. Our experiments therefore focus on a comparison against previous relative pose solvers for generalized cameras, which are a 2-point Ransac algorithm relying on a non-holonomic motion assumption, the linearized 17-point algorithm, and a generalized eigenvalue minimization algorithm. In order to demonstrate the benefit of using multiple cameras pointing into different directions, we also include a comparison against centralized, single-camera algorithms such as the traditional 8-point algorithm and 1-point Ransac, the latter again relying on a non-holonomic motion assumption. We execute different comparative simulation experiments to evaluate accuracy and noise resilience, the performance of the proposed object-space error based non-linear refinement, and the performance when embedded into a Ransac scheme [4]. We conclude with continuous frame-to-frame motion estimation demonstrations on both small-scale indoor and large-scale outdoor datasets captured by a 4-camera system mounted on either a turtlebot or a full-size car (cf. video in supplementary material). Ground truth for the indoor sequence is delivered by a highly accurate external motion capture system.

A. Outline of the simulation experiments

The surround-view camera system we investigate in simulation highly resembles the multi-camera system on real experiments, and has four cameras pointing into all directions (cf. Figure 1). The cameras all lie in the same horizontal plane and have a distance between 0.6 and 1m away from the body origin. In each experiment iteration, we fix the frame of the first viewpoint to coincide with the world frame. We then add 6 further views by adopting a linearly changing rotational velocity, therefore generating realistic, non-circular motion trajectories. We finally calculate the relative displacement between the first and the last viewpoint. The final relative rotation angle lies between 3 and 6 degrees, and the displacement between frames is set to 0.6m. We generate random correspondences for each camera by defining random 3D landmarks located within the field of view of each camera in the first viewpoint, assign random depths between 1 and 8m, and finally reproject the obtained 3D landmarks into each camera of the second viewpoint. Noise is added to the measurements by extracting the orthogonal plane of each bearing vector, and adding noise based on a virtual spherical camera with focal length 800 pixels. Outliers are added by replacing bearing vectors such that they point towards new, randomly generated landmarks. We analyze the performance of our method under different conditions, which are a dynamic rotational velocity, purely translational motion, a varying field-of-view, and a changing outlier fraction. We execute 1000 random constellations for each experiment, and—due to scale observability issues—report primarily on the accuracy of the relative rotation estimation.

B. Comparison against minimal solvers

We compare our method (**ME**) against state-of-the-art minimal solvers for planar motion, which is the 2-point Ransac algorithm by [13] (**2-pt**) and the 1-point Ransac algorithm proposed by [16] (**1-pt**). These algorithms adopt the Ackermann motion model, and thus make the assumption that the motion is non-holonomic with a fixed centre of rotation (ICR) for the entire relative displacement. As a result, they have a reduced number of required correspondences. We generate 20 points in each camera for all algorithms, and use the minimal number of points for each method to solve the



(a) (b) (c) (d)

Fig. 3. Comparison between our proposed method **ME** and the **1pt**, **2pt**, **8pt**, **17pt**, **GE** method for different perturbation factors. Each value is averaged over 1000 random experiments. Details are provided in the text.

problem hypotheses (1 point for **1-pt**, 2 points for **2-pt** and 3 points per camera in our algorithm). No outliers are added to the data. We repeat the experiment for changing deviations from pure Ackermann motion defined by the linearly changing per-frame rotation change $\omega = 0.04k \cdot i + \omega_0$, where $i = 1, \ldots, 6$, $\omega_0 = 0.2°$, and k is varied from 0 to 10. The ICR is extrapolated by assuming the constant forward velocity $v = 0.1$m per second. The results are indicated in Figure 3(a). As expected, our model outperforms as the deviation from non-holonomic motion is increasing.

C. Comparison against non-minimal solvers

We compare our method against alternative non-minimal, generalized solvers (**17-pt** [14] and **GE** [10]) as well as a central method (**8-pt** [6]) commonly applied in vehicle motion estimation with a forward-facing camera. We use 5 points in each camera for all generalized algorithms and 8 points in the forward camera for **8-pt**. We run only the solvers and do not add nonlinear refinement. We conduct three types of experiments:

- *Significant rotation*: The field-of-view of each camera is fixed to 120°, and the noise level is varied between 0 and 5 pixels. The results are indicated in Figure 3(b). As can be observed, all generalized solvers out-perform the centralized method, and **ME** performs better than **17-pt** in terms of both the mean and median error. As stated in [10], **GE** has been designed for omnidirectional cameras and occasionally converges into wrong local minima, thus leading to an increased mean error with repect to **ME**.
- *Pure translations*: We repeat the same experiment but simply force the motion to be purely translational. The result is illustrated in Figure 3(c). As expected, **17-pt** and **ME** maintain a higher level of accuracy than **8-pt**. As furthermore explained in [10], **GE** is affected by a constant zero energy for identity rotation. While this leads to perfect performance in this experiment, it is to be interpreted as a weakness. **GE** is unable to distinguish small from zero rotation angles.
- *Variation of the field of view*: We vary the field-of-view from 15° to 165°. As shown in Figure 3(d), the centralized method **8-pt** applied in the forward facing camera performs better for very small fields-of-view. As explained in Section IV, side-ways looking cameras are affected by the rotation-translation ambiguity. The effect worsens for a decreasing field-of-view, and potentially affects all generalized camera solvers. However, as soon as the field-of-view is sufficiently large (75° for our settings), generalized solvers start to outperform. **ME** furthermore clearly outperforms other methods and beats **8-pt** for any FoV larger than 30°.

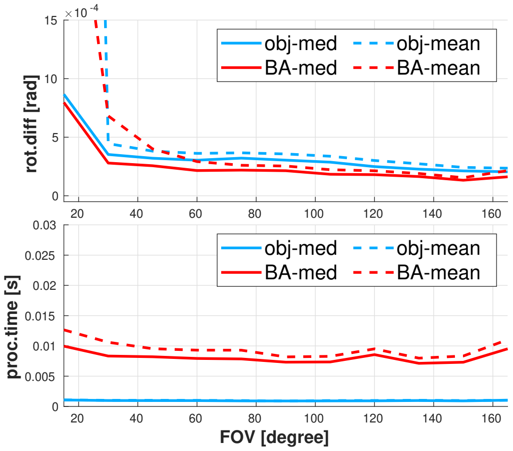


Fig. 4. Accuracy of the different geometric optimization method and average execution time.

D. Behavior of object-space error based refinement

Figure 4 shows the comparison between our proposed object-space error minimizer and standard two-view bundle adjustment. Both depend on a sufficiently good initialization. However, as stated in Section IV-B, standard two-view bundle adjustment reduces reprojection errors over rotation, translation, and structure parameters, while the proposed joint eigenvalue minimization based object-space error reduction involves only the rotational degrees of freedom (which—in the case of planar motion—is only a single degree of freedom). As indicated in Figure 4, object-space error minimization shows comparable performance than 2-view bundle adjustment for varying fields-of-view. However, owing to its univariate nature, the proposed objective is minimized 8 times faster.

E. Overall performance within Ransac

Before moving on to real data experiments, we add a final experiment with outliers to also compare the behavior with

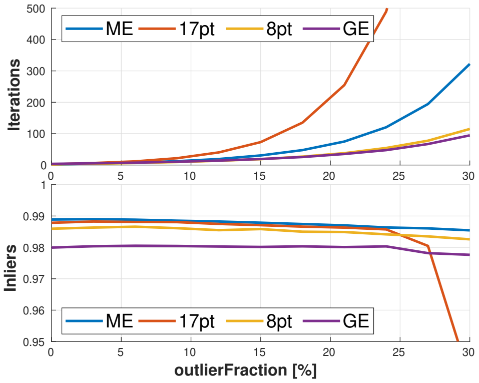


Fig. 5. Average number of iterations and found inlier rates.

respect to full generalized relative pose solvers if embedded into a Ransac scheme. We add up to 30% outliers, and use the same outlier threshold and inlier verification criterium for each algorithm. We test three elements including execution time of each algorithm, the required number of Ransac iterations, and the percentage of all true inliers found by each method. Results are depicted in Figure 5. Our method's processing time is 0.23ms, **17-pt** uses 0.11ms, **GE** uses 0.3ms, and **8-pt** is the fastest at 0.05ms. As already indicated in [14], the linear **17-pt** method requires too many samples and even 1000 iterations may be insufficient to perform successful inlier identification. Our method also shows high computational efficiency. It consumes 2 times the time of the linear solver **17-pt** and is 1.5 times faster than **GE**. To conclude, our solution finds the largest percentage of all true inliers, which demonstrates that—given a sufficiently large number of iterations—our method has a low probability to miss the global minimum of the objective function.

F. Results on a real multi-camera system

In order to demonstrate the performance of our algorithm on real images, we apply it to two sequences representing both an indoor and an outdoor example. All datasets are captured by a fully calibrated and synchronized surround-view multi-camera system mounted on either a turtlebot (cf. illustrated in Figure 1) or a full-size car. The indoor dataset allows us to compare our method against highly accurate ground truth captured by a motion tracking system. It provides a mix of characteristics with both straight forward motion and significant rotation parts. All methods are embedded into a Ransac scheme and applied on a frame-to-frame basis. We use object-space error minimization for **ME** and standard 2-view bundle adjustment for all remaining algorithms as a two-view refinement procedure. We do not add a multi-frame back-end optimization module (i.e. sliding-window bundle-adjustment) as this permits the observation of the original performance of each method. Implementations are made in C++, and use OpenCV [2] and OpenGV [9] for image processing and geometry problems, respectively. All experiments are conducted on an Intel Core i7 2.4 GHz CPU with 8GB RAM. Figure 6 shows our results obtained on the indoor dataset and compares them against all alternative algorithms. The following is worth noting:

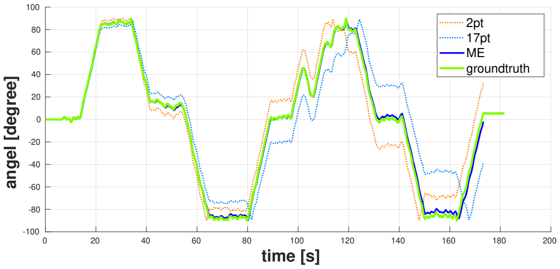


Fig. 6. Evaluation of the dead-reckoned absolute orientation of a real multi-camera rig moving in an indoor environment. Ground-truth is provided by an Optitrack motion tracking system.

- The trajectories all suffer from slow error accumulation, which means that all algorithms successfully process the entire 2000 frames without any gross errors. Our algorithm **ME** clearly outperforms both **17-pt** and **2pt**. Note that **8-pt** and **GE** are not included in the results, as they are both unable to provide competitive results.
- The observations concerning rotation-translation ambiguity are consistent with our prior analysis. We therefore implement a firewall strategy to prevent occasional convergence to wrong local minima. We check the solution obtained from only the front and back cameras, and compare it against the solution obtained from the entire system. If the two solutions have obvious differences, we down-weight the energy contribution of the sideways facing cameras. As shown in Figure 6, this strategy leads to the best result.
- Note furthermore that **2pt** is confined to the front and back cameras, as we observed that it performs much better than a full 4-camera alternative. Nonetheless, it suffers from the strict Ackermann-motion assumption, and leads to a similar error accumulation like **17-pt**.

The real-time execution of the algorithm and further qualitative results on a full-size vehicle moving outdoors can be found in the supplemental video file.

VI. CONCLUSIONS

Our work stands in contrast with many prior closed-form solutions presented in the literature, as it relies on an iterative optimization scheme. However, by exploiting simple linear algebra relationships and the planarity of the motion, the dimensionality of the energy minimization problem is reduced to one, and can hence be solved very effectively. Furthermore, the fact that the minimized energy depends only on the rotation parameters and the insight that these parameters are shared between all cameras permits us to minimize multiple single camera objectives in parallel rather than a single generalized objective. As demonstrated through our results, the formulation is free of singularities and amenable to highly accurate and reliable, continuous motion estimation for surround-view camera systems. It is our belief that this contribution must be of interest to the intelligent vehicles community, and direct our future work towards an extension of the approach over multiple temporal frames.

REFERENCES

[1] O. Booij and Z. Zivkovic. The planar two point algorithm. Technical Report IAS-UVA-09-05, University of Amsterdam, 2009.

[2] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[3] 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.

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

[5] W. Gellert, S. Gottwald, M. Hellwich, H. Kästner, and H. Künstner. *The VNR Concise Encyclopedia of Mathematics*. Van Nostrand Reinhold, New York, NY, USA, second edition, 1989.

[6] R. Hartley. In Defense of the Eight-Point Algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 19(6):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] K. Huang, Y. Wang, and L. Kneip. Motion estimation of non-holonomic ground vehicles from a single feature correspondence measured over n views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, California, USA, 2019.

[9] L. Kneip and P. Furgale. OpenGV: A Unified and Generalized Approach to Real-Time Calibrated Geometric Vision. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Hongkong, 2014.

[10] 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.

[11] 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.

[12] G. Lee, M. Pollefeys, and F. Fraundorfer. Relative pose estimation for a multi-camera system with known vertical. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[13] G. H. Lee, F. Faundorfer, and M. Pollefeys. Motion Estimation for Self-Driving Cars With a Generalized Camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2746–2753, 2013.

[14] 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.

[15] 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.

[16] D. Scaramuzza, F. Fraundorfer, and R. Siegwart. Real-time monocular visual odometry for on-road vehicles with 1-point RANSAC. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2009.

[17] 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.

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