Motion estimation of non-holonomic ground vehicles from a single feature correspondence measured over n views

Kun Huang
ShanghaiTech University
huangkun1@shanghaitech.edu.cn

Yifu Wang
The Australian National University
u5434194@anu.edu.au

Laurent Kneip
ShanghaiTech University
lkneip@shanghaitech.edu.cn

Abstract

The planar motion of ground vehicles is often non-holonomic, which enables a solution of the two-view relative pose problem from a single point feature correspondence. Man-made environments such as underground parking lots are however dominated by line features. Inspired by the planar tri-focal tensor and its ability to handle lines, we establish an n-linear constraint on the locally circular motion of non-holonomic vehicles able to handle an arbitrarily large and dense window of views. We prove that this stays a uni-variate problem under the assumption of locally constant vehicle speed, and it can transparently handle both point and vertical line correspondences. In particular, we prove that an application of Viète's formulas for extrapolating trigonometric functions of angle multiples and the Weierstrass substitution casts the problem as one that merely seeks the roots of a uni-variate polynomial. We present the complete theory of this novel solver, and test it on both simulated and real data. Our results prove that it successfully handles a variety of relevant scenarios, eventually outperforming the 1-point two-view solver.

1. Introduction

Autonomous vehicles promise to be a future disrupting technology on the market. The topic is currently investigated intensively across both industry and academia, and any successful outcome depends heavily on a reliable, online solution to the interdependent problems of self-localisation and environment mapping. While the most powerful solutions rely on a multitude of sensors including lidars and cameras, the community maintains a high interest in developing vision-only alternatives, too. Cameras are economic close-to-market sensors that may unlock lower-level autonomy in more controlled and less critical scenarios, even in the absence of other exteroceptive sensors [6].

While direct sparse [2], semi-dense [3], and dense methods [26] have already been presented, the more traditional way of sparse feature correspondence-based localization and mapping continues to be of high importance. Besides outstanding computational efficiency, it also produces valuable information for place recognition and enables seamless integration into global optimization objectives [25]. A particularly interesting case is given by the estimation with a single, forward-looking camera, as such sensors can already be found in today's vehicles on the market. The present paper focuses on the related fundamental problem of relative pose estimation with a single camera mounted on a car. The solution to this problem is required at the initialization stage, when no prior information about either the motion of the vehicle or the environment is available. Sequential application furthermore enables a straightforward solution of the visual odometry problem [28], thus enabling online relative self-localization of the vehicle.

The classical solution to the calibrated relative pose problem with a single camera requires at least five point-feature correspondences. However, as shown in [31], the non-holonomic motion of ground vehicles has fewer degrees of freedom, which reduces the number of required correspondences. The motion of ground vehicles can be approximated by the Ackermann model, which forces the local trajectory to be a circular arc in the plane, and the heading of the vehicle to remain tangential to the arc. The model depends on only two parameters, which are the radius of the circle and the inscribed angle of the arc. Furthermore, using only a single camera renders scale unobservable, and the latter affects only the radius of the circle. As outlined in [31], this fact permits the resolution of the relative rotation angle from only a single feature correspondence.

Draw-backs of the 1-point solver presented in [31] are a

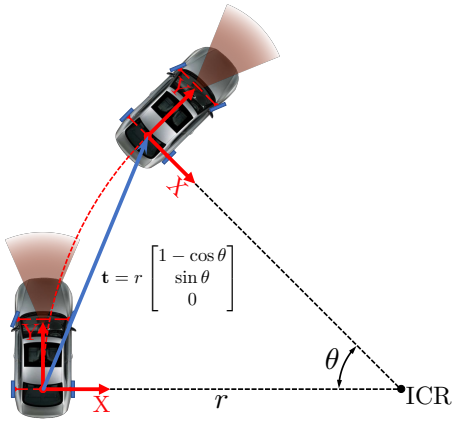

Figure 1. Our work aims at exploiting the non-holonomic motion of ground vehicles along circular trajectories to simplify visual motion estimation. We use a single, forward-looking camera.

restriction to two views and an inability to handle line correspondences. As encompassed by tri-focal tensor geometry [12], the inclusion of lines notably requires the presence of at least three views. For general motion with 6 degrees of freedom, the minimal case leads us to the three-view four-point problem, which remains unsolved to date. Though the case of planar motion has been successfully addressed by the planar trifocal tensor [11], an adaptation to the non-holonomic motion of most ground vehicles is yet to be presented.

Inspired by the planar tri-focal tensor, we present the following contributions:

- We introduce an n-linear constraint adapted to the case of planar non-holonomic motion. It can transparently handle both point and vertical line feature correspondences measured over an arbitrary number of views.
- We prove that the resolution of the motion can be regarded as a rank minimisation problem over a single degree of freedom.
- We apply Viète's formulas for extrapolating trigonometric functions of angle multiples and the Weierstrass substitution to transform the rank minimisation objective into a uni-variate polynomial.

The hereby obtained solver is able to handle either points or vertical lines over an arbitrary number of views (at least three) and increases the flexibility in how we can bootstrap and realise online visual localisation pipelines. In particular—as we will demonstrate through our experiments on both simulated and real data—using more views increases the signal-to-noise ratio, and therefore leads to improved accuracy over existing solvers. Our paper is organised as follows. After a discussion on further related work in Sections 2 and 3, Section 4 presents the derivation of our novel solver. Section 5 presents successful results on both simulated and real data, before Section 6 concludes with a brief discussion of our contribution.

2. Related work

While, the majority of online localization and mapping frameworks for ground vehicles rely on either stereo [16, 17, 9] or even surround-view camera systems [6, 15], the most recent, state-of-the-art pipelines successfully solve the problem based on only a single forward-looking camera [25]. As outlined in the seminal work by Nistér et al [28], a basic monocular visual odometry solution may already be achieved by a sequential application of a calibrated relative pose solver. The latter algorithm is furthermore required for general bootstrapping of any point-based method in the absence of prior information about either motion or structure.

The calibrated relative pose problem can be solved with as few as five point-feature correspondences between a pair of views [27, 32, 22, 19]. Linear solutions to the problem have been presented earlier in [23, 13], which do however suffer from degenerate conditions in the planar case. However, of higher relevance to our work are solvers that rely on lower-dimensional, approximative models of the specific motion of ground vehicles. For example, [4] and [34] look at the special case of a known directional correspondence between both views, a case that is easily fulfilled on a ground vehicle for which the motion can be approximated to remain in a plane. The most important foundation for our work is given by [31], who exploits the fact that non-holonomic motion of ground vehicles is in fact a function of only two degrees of freedom. Considering that scale may be unobservable, this enables the solution of the relative pose problem from as few as a single point-feature correspondence across two views. The work has later been extended in [30], proving that a fixed, known, horizontal baseline between the wheel axis and the camera may even render scale observable. [20] furthermore applied the model to multi-perspective camera systems.

More recently, we have experienced a revival of hybrid point and line feature-based methods, with [29] and [35] providing monocular solutions, and [9] a state-of-the-art stereo alternative. Despite its difficulty, there have also been a few efforts on realizing purely line-based online structure-from-motion pipelines [24, 7, 21]. Their results are however not very encouraging as [24] states that *From the experimental and simulation results, there is no incentive to believe that lines constrain the motion sufficiently to be used alone*, and [7] and [21] only present small scale experiments without evaluating motion accuracy at all. As we will show in this work, exploiting non-holonomic constraints over multiple views bares the potential to perform accurate and robust, purely line-based visual motion estimation. Our work relies on the tri-focal tensor, which has

been introduced in [12] as a means to constrain relative pose over multiple views observing line feature correspondences. The theory has later been summarized in [14], eventually leading up to further important foundations for constraining motion over n views. Our work is also inspired by [11], which adapts the tri-focal tensor relationships to the case of planar motion and bearing measurements in the plane.

3. Foundations

We start with a brief review of the planar tri-focal tensor [11], which solves a problem that is tightly related to ours. We furthermore introduce a parametrization of the non-holonomic motion of ground vehicles that will be used in the continuation.

3.1. Short review of the planar tri-focal tensor

The general tri-focal tensor describes the incidence relationships between lines measured over three views [12]. In the calibrated case, these incidence relationships take the form

$$\mathbf{n}_1 \times (\mathbf{n}_2^T[\mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3]\mathbf{n}_3)^T = \mathbf{0}, \tag{1}$$

where $\mathbf{n}_1$, $\mathbf{n}_2$, and $\mathbf{n}_3$ are the normalized line measurements in views 1, 2, and 3, and $\mathbf{T}_1$, $\mathbf{T}_2$, and $\mathbf{T}_3$ are a function of the extrinsic Euclidean transformation parameters describing the pose of each view. The planar tri-focal tensor looks at what happens if the camera is mounted on a ground vehicle which remains in the horizontal plane, and for which the z-axis of the body frame remains aligned with the gravity vector $\mathbf{g} = [0 \quad 0 \quad -1]^T$. $\mathbf{n}_1$, $\mathbf{n}_2$, and $\mathbf{n}_3$ furthermore originate from bearing measurements in the plane. Such measurements can originate from 1) points measured in the horizontal plane, 2) vertical lines projected into the horizontal plane, or 3) arbitrary 3D points projected onto the ground plane. We explain the case of vertical lines, point features may be transparently adopted by the description, too.

Let $\mathbf{l}$ be one of the original lines measured in the image plane [1]. Vertical lines distinguish themselves from the fact that they intersect with the vanishing point of the vertical direction, as in

$$(\mathbf{K}\mathbf{R}_{\mathbf{cw}}\mathbf{g})^T\mathbf{l} = 0, \tag{2}$$

where $\mathbf{R}_{\mathbf{cw}}$ notably permits the rotation of points from the body frame of the vehicle into the camera frame (can be arbitrary), and $\mathbf{K}$ represents the matrix of intrinsic camera parameters. (2) can be reformulated as

$$\mathbf{g}^T({\mathbf{R}_{\mathbf{cw}}}^T{\mathbf{K}}^T\mathbf{l}) = \mathbf{g}^T\mathbf{n} = 0, \tag{3}$$

where $\mathbf{n}$ is the searched normalised bearing measurement expressed in the body frame of the vehicle. It is easy to recognise that—due to the form of $\mathbf{g}$—the third coordinate of $\mathbf{n}$ needs to remain 0, hence it represents a measurement in the horizontal plane. It is furthermore easy to prove that $\mathbf{n}$ in fact represents the normal of the ray corresponding to the bearing measurement.

$\mathbf{T}_1$, $\mathbf{T}_2$, and $\mathbf{T}_3$ for themselves take a special form in the case of planar motion. The rotation of each camera matrix notably remains a pure rotation about the z-axis, whereas the translation has a zero component along the vertical direction. It is again easy to prove that—under these conditions—only a single non-trivial equation from (1) is remaining. With 5 degrees of freedom (originating from 2 planar relative displacements and an unobservability of the overall scale), the problem can be solved from 5 bearing correspondences across the three views.

We leave the discussion about the planar trifocal tensor here, and refer the interested reader to [11], as our primary interest lies in the form of the bearing measurements. The motion model is now replaced to take into account the special properties of non-holonomic ground vehicles.

3.2. The Ackermann motion model

As exploited in the original work of Scaramuzza et al. [31], the motion of a ground vehicle can be approximated to lie on a circular arc contained in the ground plane. As indicated in Figure 1, the heading of the vehicle furthermore remains tangential to that arc. A minimal parametrisation of the motion is hence given by the inscribed arc-angle θ, as well as the radius of this circle r. The centre of the circle is commonly called the Instantaneous Centre of Rotation (ICR), and even the front wheels of steering ground vehicles travel along circular trajectories which are centered around the ICR. We adopt the convention that the y-axis of the car is pointing in the forward-direction, while the x-axis points to the right. Let us denote the relative displacement by $\mathbf{t}$ and $\mathbf{R}$. It permits the transformation of points from the second back to the first frame using the equation $\mathbf{p}_{\mathcal{F}_0} = \mathbf{R}\mathbf{p}_{\mathcal{F}_1} + \mathbf{t}$. $\mathbf{R}$ and $\mathbf{t}$ are given by

$$\mathbf{t} = r\begin{bmatrix}1-\cos\theta\\ \sin\theta\\ 0\end{bmatrix} = \frac{d}{\sin\theta}\begin{bmatrix}1-\cos\theta\\ \sin\theta\\ 0\end{bmatrix} \tag{4}$$

$$\mathbf{R} = \begin{bmatrix}\cos\theta & \sin\theta & 0\\ -\sin\theta & \cos\theta & 0\\ 0 & 0 & 1\end{bmatrix}, \tag{5}$$

where d fixes the length of the displacement along y. Note that this form differs from the one presented in [31] in that it does not employ half angles. The equivalence of both can however be proven quite easily using half-angle formulae and by substituting $\rho = \frac{d}{\cos\frac{\theta}{2}}$ in [31]. Fixing the displacement along y helps us to ensure that the overall translational displacement does not vanish if θ approaches zero.

[1] Line vectors in the image plane are 3-vectors $\mathbf{l}$ such that any image point on the line $\mathbf{x}$ fulfils $\mathbf{x}^T\mathbf{l} = 0$.

4. 1-feature n-view solver

We now proceed to the core of our contribution, which is a novel algorithm for estimating non-holonomic motion over n views and from a single feature correspondence. We start by seeing our ideal-case assumptions and its implications on the motion model. We then see how a constraint over n views can be obtained from an n-linearity, and how this constraint permits the solution of the motion via a univariate rank minimisation objective. We proceed with a sequence of substitutions and approximations that finally result in a solver that merely requires finding the roots of a uni-variate polynomial. To the end of accurate and robust results, this solver is finally embedded into histogram voting and multi-feature optimisation procedures.

4.1. Extending the motion model

We now move to a window of n successive frames which are denoted by $\mathcal{F}_i$, where $i = \{0, \cdots, n-1\}$. Our assumption is still that the vehicle travels along a circular trajectory, and that the heading of the vehicle remains tangential to the arc. We furthermore assume that the tangential speed of the vehicle (and thus also its rotational velocity) remains constant during the capturing of the frames. Note that these assumptions are no more restrictive than the original assumptions in [31]. We merely assume that the vehicle speed can be regarded as constant over short time intervals, an assumption that is equally valid for any number of frames captured within the same time interval. More frames within the same time-span are easily captured by increasing the camera framerate.

Denoting the relative rotation angle between successive frames θ, the relative pose of subsequent frames are now simply given by employing angle-multiples:

$$\begin{aligned} \mathbf{t}_i &= \frac{d}{\sin\theta}\begin{bmatrix} 1-\cos(i\theta) \\ \sin(i\theta) \\ 0 \end{bmatrix} \\ \mathbf{R}_i &= \begin{bmatrix} \cos(i\theta) & \sin(i\theta) & 0 \\ -\sin(i\theta) & \cos(i\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}. \end{aligned} \tag{6}$$

4.2. n-linearity in the case of Ackermann motion

The derivation of our n-view solver starts with the basic incidence relationship that needs to be fulfilled in each view. Let $\mathbf{X} = \begin{bmatrix} x & y & z & 1 \end{bmatrix}^T$ be a point on the feature that is observed by the bearing measurements $\mathbf{n}_i, i \in [0 \dots n-1]$. As explained in Section 3.1, $\mathbf{n}$ corresponds to a horizontal vector that represents the normal vector of a vertical plane that contains both the camera centre and the observed feature (e.g. a vertical line or an arbitrary, single 3D point). Let $\mathbf{P}_i$ be the normalised camera projection matrix for frame

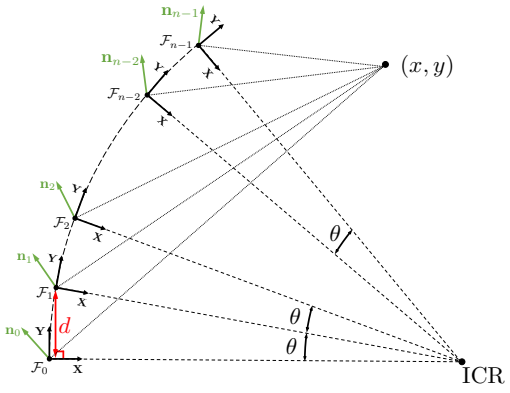

Figure 2. General Ackermann motion model used throughout this paper. Please see text for detailed explanations.

$\mathcal{F}_i$. The transformed world point $\mathbf{X}$ needs to intersect with the measured vertical plane, which leads us to the constraint

$$\mathbf{n}_i^T \mathbf{P}_i \mathbf{X} = 0. \tag{7}$$

Each camera matrix $\mathbf{P}_i$ is a function of the motion parameters

$$\mathbf{P}_i = \begin{bmatrix} \mathbf{R}_i^T & | & -\mathbf{R}_i^T \mathbf{t}_i \end{bmatrix}. \tag{8}$$

After substituting the camera matrices in (7), replacing the pose parameters with their expressions given in (6), removing the trivial line $0 \cdot z = 0$, and factoring out d, we obtain

$$\begin{bmatrix} x_{\mathbf{n}_i} & y_{\mathbf{n}_i} \end{bmatrix} \begin{bmatrix} \cos(i\theta) & -\sin(i\theta) & \frac{1}{\sin(\theta)}(1-\cos(i\theta)) \\ \sin(i\theta) & \cos(i\theta) & -\frac{1}{\sin(\theta)}\sin(i\theta) \end{bmatrix} \begin{bmatrix} x \\ y \\ d \end{bmatrix} = 0. \tag{9}$$

The constraints from each view can be stacked into an n-linear problem, which finally leads to

$$\begin{bmatrix} \begin{pmatrix} x_{\mathbf{n}_0} & y_{\mathbf{n}_0} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \\ \vdots \\ \begin{pmatrix} x_{\mathbf{n}_i} & y_{\mathbf{n}_i} \end{pmatrix} \begin{pmatrix} \cos(i\theta) & -\sin(i\theta) & \frac{1}{\sin(\theta)}(1-\cos(i\theta)) \\ \sin(i\theta) & \cos(i\theta) & -\frac{1}{\sin(\theta)}\sin(i\theta) \end{pmatrix} \\ \vdots \\ \begin{pmatrix} x_{\mathbf{n}_{n-1}} & y_{\mathbf{n}_{n-1}} \end{pmatrix} \begin{pmatrix} \cos((n-1)\theta) & -\sin((n-1)\theta) & \frac{1}{\sin(\theta)}(1-\cos((n-1)\theta)) \\ \sin((n-1)\theta) & \cos((n-1)\theta) & -\frac{1}{\sin(\theta)}\sin((n-1)\theta) \end{pmatrix} \end{bmatrix} \begin{bmatrix} x \\ y \\ d \end{bmatrix} = 0. \tag{10}$$

Note that this form is similar to the n-linearity presented in [14] (Chapter 17), the difference being a specialization to the non-holonomic, planar case.

4.3. From rank minimisation to a univariate polynomial objective

Our objective (10) is of the form $\mathbf{A}\mathbf{x} = 0$, where $\mathbf{A}$ is a matrix function of the angle interval θ. In order to have a non-trivial solution, $\mathbf{A}$ needs to be rank deficient. The solution of θ can hence be enforced by solving the rank minimisation problem

$$\theta_{\text{opt}} = \underset{\theta}{\operatorname{argmin}} \operatorname{rank}\left(\mathbf{A}\left(\theta\right)\right). \tag{11}$$

What follows is a sequence of substitutions and simplifications that will transform this optimisation objective into finding the roots of a simple univariate polynomial. The procedure starts with ensuring that $\mathbf{A}$ becomes a polynomial matrix function.

The first substitution that we apply is given by replacing trigonometric functions of angle multiples by functions that only involve $\cos\theta$ and $\sin\theta$. This is achieved by using Viète's formulae

$$\begin{cases} \sin(i\theta) = \sum_{k=0}^{2k+1\leq i} (-1)^k \binom{i}{2k+1} \cos^{i-2k-1}(\theta)\sin^{2k+1}(\theta) \\ \cos(i\theta) = \sum_{k=0}^{2k\leq i} (-1)^k \binom{i}{2k} \cos^{i-2k}(\theta)\sin^{2k}(\theta) \end{cases}. \tag{12}$$

We then apply the Weierstrass substitution (also known as the tangent half-angle substitution) given by

$$\begin{cases} \cos\theta = \frac{1-z^2}{1+z^2} \\ \sin\theta = \frac{2z}{1+z^2} \end{cases}. \tag{13}$$

We finally substitute (13) in (12), the result into (10), and multiply the equation with $(1+z^2)^{(n-1)}$ to largely reduce the degree of the denominator. We obtain

$$\begin{bmatrix} (x_{\mathbf{n}_0} \quad y_{\mathbf{n}_0}) \begin{pmatrix} C_0 & -S_0 & \frac{1+z^2}{2z}((1+z^2)^{(n-1)} - C_0) \\ S_0 & C_0 & -\frac{1+z^2}{2z}S_0 \end{pmatrix} \\ \dots \\ (x_{\mathbf{n}_{n-1}} \quad y_{\mathbf{n}_{n-1}}) \begin{pmatrix} C_{n-1} & -S_{n-1} & \frac{1+z^2}{2z}((1+z^2)^{(n-1)} - C_{n-1}) \\ S_{n-1} & C_{n-1} & -\frac{1+z^2}{2z}S_{n-1} \end{pmatrix} \end{bmatrix} \begin{bmatrix} x \\ y \\ d \end{bmatrix} = \mathbf{0}, \tag{14}$$

where

$$\begin{cases} C_i = (1+z^2)^{n-1-i} \sum_{k=0}^{2k\leq i} (-1)^k \binom{i}{2k} (1-z^2)^{i-2k}(2z)^{2k} \\ S_i = (1+z^2)^{n-1-i} \sum_{k=0}^{2k+1\leq i} (-1)^k \binom{i}{2k+1} (1-z^2)^{i-2k-1}(2z)^{2k+1} \end{cases}. \tag{15}$$

It is interesting to observe that the denominator in this expression cancels out so the problem takes the form

$$\begin{bmatrix} p[2n-1](z) & p[2n-1](z) & p[2n-1](z) \\ \dots & & \\ p[2n-1](z) & p[2n-1](z) & p[2n-1](z) \end{bmatrix}_{n\times 3} \begin{bmatrix} x \\ y \\ d \end{bmatrix}$$

$$= \mathbf{B}[2n-1](z) \begin{bmatrix} x \\ y \\ d \end{bmatrix} = \mathbf{0}, \tag{16}$$

where $p[k](z)$ and $\mathbf{B}[k](z)$ denote a degree-k polynomial and matrix in z, respectively.

Since $\mathrm{rank}(\mathbf{B}) = \mathrm{rank}(\mathbf{B}^T\mathbf{B})$, our optimisation objective finally becomes

$$z_{\text{opt}} = \underset{z}{\text{argmin}}\, \mathrm{rank}\left(\mathbf{M}[4n-2](z)\right), \tag{17}$$

where $\mathbf{M}[4n-2](z) = \mathbf{B}[2n-1](z)^T\mathbf{B}[2n-1](z)$ is a 3×3 polynomial matrix function of z. $\mathbf{M}$ is a positive semi-definite matrix, and its rank can be minimized by minimising its smallest eigenvalue (which, in particular, is equivalent to minimizing the smallest singular value of the original matrix $\mathbf{A}$). The objective becomes

$$z_{\text{opt}} = \underset{z}{\text{argmin}} \min \left(\underset{\lambda}{\text{solve}} \left(\det(\mathbf{M} - \lambda\mathbf{I}) \right) \right). \tag{18}$$

While this objective looks compact, it is indeed hard to optimise as it depends on an iterative, internal resolution of the smallest eigenvalue of $\mathbf{M}$, which—though possible—is hard to compute in closed form, especially considering the elevated order of the z-polynomials contained in $\mathbf{M}$. However, it is also clear that in the ideal noise-free case, the rank deficiency is fulfilled and the smallest eigenvalue of $\mathbf{M}$ at the optimum simply becomes zero. We therefore set λ to zero, and simply solve the final objective

$$z_{\text{opt}} = \underset{z}{\text{argmin}} \left(\det\left(\mathbf{M}\right) \right). \tag{19}$$

Note that this is equivalent to seeking the real roots of a univariate polynomial in z, which we solve using Sturm's root bracketing approach. Due to our local assumption on the vehicle motion model, $\theta \in (-\frac{\pi}{2}, \frac{\pi}{2})$, thus leading to $z \in (-1, 1)$. The determinant polynomials are of order $12n-6$, and typically return only one or two real roots. It is clear that the introduction of noise will perturb this solution, but we prove in our experimental results section that this influence is marginal. Furthermore, the determinant constraint has the significant advantage of being much less affected by local minima than the original rank minimization objective, which is especially the case if only considering a single feature correspondence. The solution of the above objective therefore returns a very good starting point for a concluding refinement over all inliers.

4.4. Robustness via Histogram Voting

Since using only a single feature correspondence to obtain a hypothesis for θ, a straightforward solution to deal with outliers is given by performing histogram voting [31]. We simply take all roots identified from our determinant constraint and set up the histogram using the Freedman-Diaconis rule for automatic bin sizing [5]. We finally pick the center of the dominant bin as an initial guess, and define the inlier set as the set of correspondences that contributed to the dominant bin.

4.5. Solution refinement

The final solution is refined over all inlier correspondences. This refinement is achieved by stacking all inlier correspondences into an extended, multi-correspondence form of (10). This remains a problem of the form $\mathbf{A}(\theta)\mathbf{x} =$

$\mathbf{0}$, and can again be optimized by performing a rank minimization of $\mathbf{A}$ over θ. In our concluding refinement step, this nonlinear least-squares problem is solved exactly by minimizing the smallest singular value of $\mathbf{A}$ using a bisectioning 1D line search. We kindly refer the reader to our supplementary material for further details about the method.

5. Experimental evaluation

We test our algorithm on both synthetic and real data. It is clear that the success of our method depends on a sufficient validity of the Ackermann motion assumption. Our experiments therefore focus on a comparison against the 2-view 1-point solver proposed in [31], which operates under the same assumptions than our method. We execute different simulation experiments in order to test the accuracy and sensitivity of our method, which includes an analysis of the performance under increasing violation of the Ackermann assumption. We conclude with tests on popular, real-data benchmark sequences and compare both methods against ground truth recorded by an accurate Inertial Navigation System.

5.1. Experiments on simulated data

We start by seeing the definition of a default random experiment for our simulations. Without loss of generality, we assume that the pose of the first frame equals to $\left[\mathbf{I}|\mathbf{0}\right]$. We furthermore fix the displacement d along y to 1 for all our experiments, which again has no impact on generality due to scale unobservability. We then add 5 further views by rotating each new view by $\theta = 5^\circ$ with respect to the previous one, and apply the Ackermann motion model. The orientation of the camera on the vehicle is fixed such that the image plane is negative and the principal axis is pointing in the forward direction

$$\mathbf{R}_{cw} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}. \tag{20}$$

The camera model is assumed to be perspective with a focal length of 721.53, and a principal point in the center of the image which has a size of 1242×375. To conclude, we generate 15 random correspondences across all views by defining random points in the first image, normalizing them, defining random depths of about 8, and reprojecting the hereby obtained 3D landmarks into all views. We conclude by adding normally distributed noise with a standard deviation of 5 pixels to each image point.

The method from [31] is denoted **1-pt**, whereas our one-feature n-view method is denoted by **1FNV**. In order to be fair, **1-pt** is evaluated over each pair of subsequent views, followed by averaging of all results. **1FNV** is evaluated over the entire window. We conduct six types of experiments for which each time one of the default parameters is changed and varied over a certain range. For each experiment and each setting, we run 1000 random experiments and calculate the mean and standard deviation of the recovered θ with respect to ground truth. Our experiments are as follows:

- *Variation of θ*: In our first experiment, we change the value of the inter-frame rotation angle θ from 0 to 8 degrees. As indicated in Figure 3(a), it can be observed that a larger inter-frame rotation angle leads to a reduction of the error with respect to ground truth. The figure also demonstrates that using more views over a larger window represents an improvement over the 2-view solver **1-pt**.

- *Variation of the number of views*: We vary the number of views from 3 to 9. The result is illustrated in Figure 3(b). As expected, **1FNV** performs better as the number of views is increasing, and furthermore outperforms **1-pt** as soon as more than five views are taken into account.

- *Number of correspondences*: Here we change the number of correspondences from as few as one until 80. The result is indicated in Figure 3(c). It can clearly be observed that the multiple observations over an extended window of views help **1FNV** to maintain a high level of accuracy even in the minimal case of a single correspondence. This demonstrates a high ability to operate in feature-poor environments.

- *Image noise*: As shown in 3(d), taking multiple observations of features also leads to a better signal-to-noise ratio, thus again demonstrating an improvement given by **1FNV**.

- *Deviation from Ackermann*: Taking an extended window of views of course diminishes the validity of our model if the Ackermann motion assumption is violated. A limit on the size of the window is typically given by the vehicle speed and the camera framerate. We generate deviations from the Ackermann model by simulating a dynamic rotational velocity which changes linearly over time. The orientation of the vehicle is obtained by integrating this rotational velocity, and the position is extrapolated by assuming constant tangential velocity. The linear model for the rotational velocity is defined as $\omega = 0.1k \cdot \omega_0 \cdot i + \omega_0$, where ω defines a per-frame rotation change varied by k times 10% of the initial rotational velocity per frame, $\omega_0 = 5^\circ$. k is varied from 0 to 4. As expected, our model performs slightly worse than **1-pt** for deviations from the Ackermann model shown in 3(e). As we will see in Section 5.3, the validity of the model remains however sufficient for **1FNV** to outperform **1-pt** on real data.

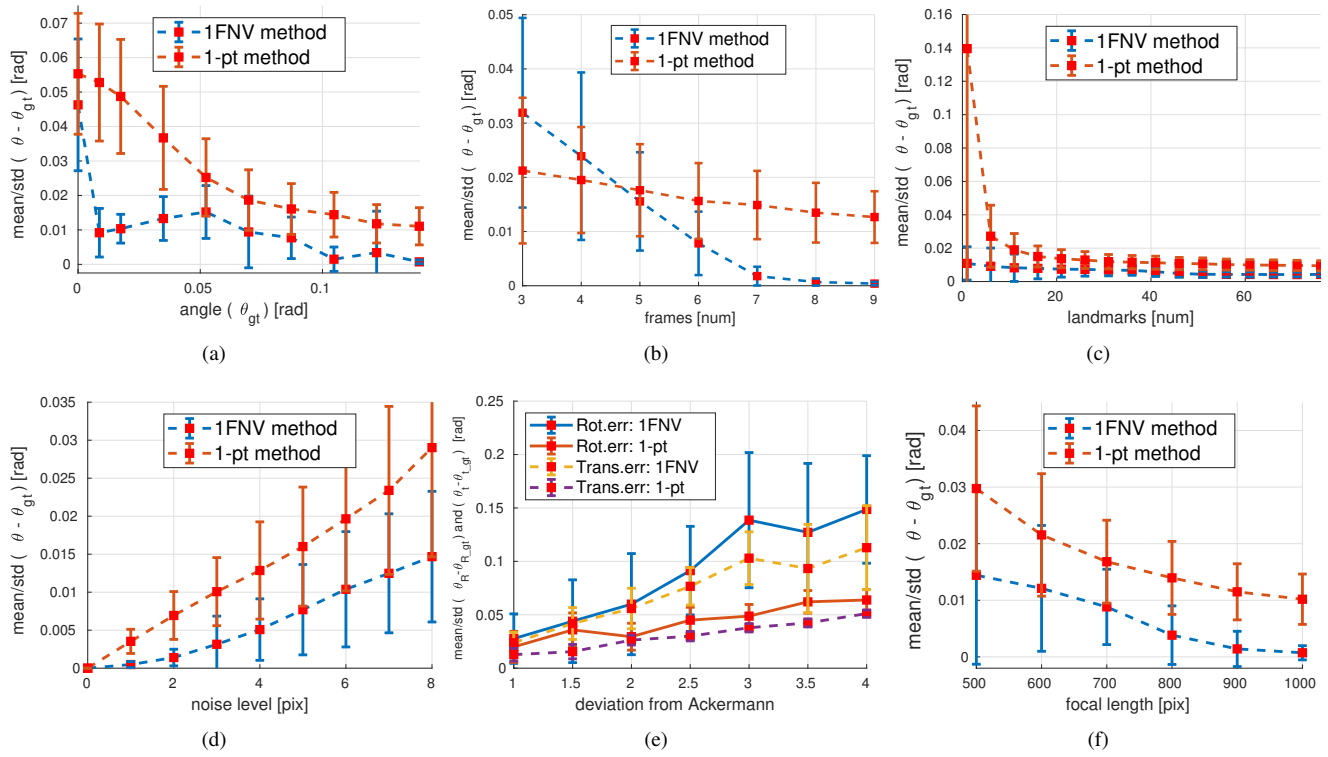

Figure 3. Comparison between our proposed method (*1FNV*) and the method from [31] (*1-pt*) for different perturbation factors. Each value is averaged over 1000 random experiments. Details are provided in the text.

- *Focal length*: The focal length impacts on the error in the bearing normals $\mathbf{n}$ after normalization. As illustrated in 3(f), both methods show only a slight increase in the error even if the focal length is reduced by 50%.

5.2. Validity of the determinant solver

As mentioned in Section 4.3, our final rank minimisation is approximated by the determinant function. Here we provide a separate experiment that analyses the error committed by this operation. We conduct 1500 random experiments as outlined in the previous section, and show the error after the determinant minimization, and after the minimisation of the smallest singular value. Figure 4 shows the distribution of the errors, indicating that the determinant already provides a very good approximation of the SVD solution as the mean is very close the zero, and the standard deviation insignificantly larger than the one of the singular value minimisation.

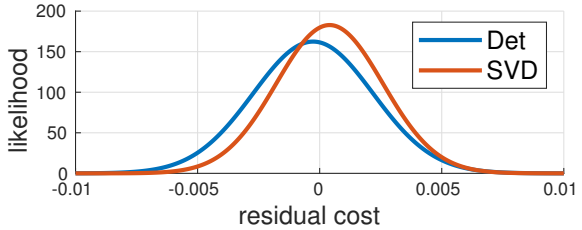

Figure 4. Distribution of errors for the determinant solver and the minimisation of the smallest singular value. Results are obtained over 1500 experiments.

5.3. Experiment on Real Data

We conclude with tests on real data, where we rely on the KITTI benchmark dataset[8]. These datasets are fully calibrated and contain images captured by a forward-looking camera mounted on a vehicle driving through a city. The dataset allows us to compare our method against ground truth which was obtained using high-standard GPS/IMU sensors. We use several different sequences that provide a mix of characteristics such as significant rotation or simple forward motion. Our camera is fixed on the windscreen and doesn't fully satisfy the requirements given by the Ackermann motion model (i.e. position on top of the back wheel axis), but—as proven in [31]—the restrictive model is still applicable if the rotation angle θ between two camera poses is sufficiently small, which is the case in our datasets. To conclude, we use the scale from ground truth to correctly rescale the translation for both algorithms. We finally evaluate the relative pose error between the ground truth and the estimated trajectory by utilising the tools from the TUM-RGBD[33] sequence, which return the Relative Pose Error (RPE) divided into rotation and translation accuracy and

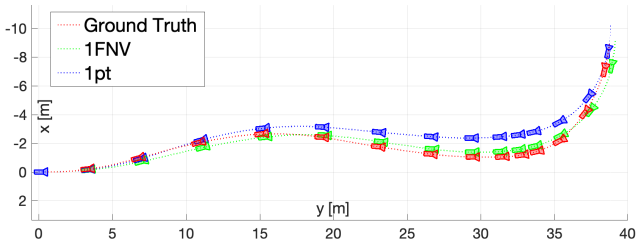

Figure 5. Comparison between **1pt** method and **1FNV** method with ground truth *_0046*.

produce both the RMS and median of errors. Our experiments are conducted on a computer with 8GB RAM and an Intel Core i7 2.4 GHz CPU, and the C++ implementation uses OpenCV [1], Eigen [10], and OpenGV [18]. An example result is indicated in Figure 5, which shows that **1FNV** performs closer to ground-truth than **1-pt**.

Results are indicated in Table 1. Our algorithm generally returns good performance. Note furthermore that the intention behind our experiments on real data is not to demonstrate outstanding performance over large scale, as the one-point solvers are only applicable if the Ackermann assumption is sufficiently fulfilled. The algorithms hence can only serve to provide a good initial guess for a final optimisation. Our experiments aim at proving that:

- Our algorithm has the ability to outperform **1-pt** particularly in the situation where the Ackermann constraint is well fulfilled. This behavior can be observed on dataset *_0046*, which does not contain sharp turns.
- We have the ability to use line features, which—if sufficiently available—can provide higher accuracy than point features. This behavior is demonstrated on dataset *_0095*, which does contain a sufficiently large amount of lines in the building facades.
- While being more susceptible to violations of the Ackermann assumption, the fact that we use correspondences over many views nonetheless returns elevated overall robustness of our method. This is demonstrated by the fact that we generally provide the lowest RMS error.

6. Conclusion

We have presented a new algorithm to estimate the planar motion of ground vehicles for which the motion is non-holonomic and can be locally approximated by a circular arc. The derivation is based on an extension of the planar tri-focal tensor to the case of an arbitrary number of views and the Ackermann motion model. We are furthermore able to include both point and vertical line features into the estimation. We prove that—starting from a sufficiently large window and a number of frames—the improved signal-to-noise ratio leads to an increase in the accuracy of the inter-frame rotation angle. This conclusion is particularly supported by our results on real data. However, a question that is not yet exhaustively addressed by our work is how the size of the window is impacting on the validity of the Ackermann model. While it seems that the conditions of our analysed real data lead to a situation in which the gain in signal-to-noise ratio prevails over a reduced validity of the Ackermann model, it is also clear that this depends on a sufficiently large camera framerate. Our future work addresses this issue by assuming a linear model for the first-order differential of the vehicle orientation. Given that the intercept of this model can be propagated over time, this formulation can still be cast as a uni-variate rank minimization problem in which only the slope of the linear model is estimated.

method	Error	Datasets		
		_0046	_0095	_0104
1pt	Rmse($\mathbf{R}$)	0.3338	0.4170	0.3633
	Median($\mathbf{R}$)	0.0037	**0.0044**	**0.0027**
	Rmse($\mathbf{t}$)	0.0465	0.0984	0.0954
	Median($\mathbf{t}$)	0.0430	0.0890	0.0731
1FNV (point features)	Rmse($\mathbf{R}$)	**0.2735**	**0.3753**	**0.3213**
	Median($\mathbf{R}$)	**0.0033**	0.0050	0.0034
	Rmse($\mathbf{t}$)	**0.0383**	0.0629	**0.0495**
	Median($\mathbf{t}$)	**0.0281**	0.0432	0.0325
1FNV (line features)	Rmse($\mathbf{R}$)	0.5537	0.4160	0.3951
	Median($\mathbf{R}$)	0.0062	0.0056	0.0053
	Rmse($\mathbf{t}$)	0.0384	**0.0628**	0.0497
	Median($\mathbf{t}$)	0.0290	**0.0427**	**0.0319**

Table 1. Performance Comparison on KITTI Datasets

Acknowledgements

We would like to acknowledge the generous startup fund 2017F0203-000-15 provided by ShanghaiTech University and the Chinese Academy of Sciences.

References

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

[2] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 40(3):611–625, 2018.

[3] J. Engel, T. Schöps, and D. Cremers. LSD-SLAM: Large-Scale Direct Monocular SLAM. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[4] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, Heraklion, Greece, 2010.

[5] D. Freedman and P. Diaconis. On the histogram as a density estimator: L2 theory. *Probability Theory and Related Fields*, 57(4):453–476, 1981.

[6] P. Furgale, U. Schwesinger, M. Rufli, W. Derendarz, H. Grimmett, P. Muhlfellner, S. Wonneberger, B. Li, B. Schmidt, T. N. Nguyen, E. Cardarelli, S. Cattani, S. Brning, S. Horstmann, M. Stellmacher, S. Rottmann, H. Mielenz, K. Kser, J. Timpner, M. Beermann, C. Hne, L. Heng, G. H. Lee, F. Fraundorfer, R. Iser, R. Triebel, I. Posner, P. Newman, L. Wolf, M. Pollefeys, S. Brosig, J. Effertz, C. Pradalier, and R. Siegwart. Toward automated driving in cities using close-to-market sensors: an overview of the V-Charge project. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, 2013.

[7] Andrew P Gee and Walterio Mayol-Cuevas. Real-time model-based slam using line segments. In *International Symposium on Visual Computing*, pages 354–363. Springer, 2006.

[8] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[9] R. Gomez-Ojeda, F.-A. Moreno, D. Scaramuzza, and J. González Jiménez. PL-SLAM: a stereo SLAM system through the combination of points and line segments. *Arxiv Computing Research Repository*, abs/1705.09479, 2017.

[10] Gaël Guennebaud, Benoît Jacob, et al. Eigen v3. http://eigen.tuxfamily.org, 2010.

[11] J. J. Guerrero, A. C. Murillo, and C. Sagues. Localization and matching using the planar trifocal tensor with bearing-only data. *IEEE Transactions on Robotics (T-RO)*, 24(2):494–501, 2008.

[12] R.I. Hartley. Projective reconstruction from line correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 903–907, Seattle, WA, USA, 1994.

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

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

[15] L. Heng, B. Choi, Z. Cui, M. Geppert, S. Hu, B. Kuan, P. Liu, R. Nguyen, Y. C. Yeo, A. Geiger, G. H. Lee, M. Pollefeys, and T. Sattler. Project AutoVision: Localization and 3D scene perception for an autonomous vehicle with a multi-camera system. In *arXiv:1809.05477*, 2018.

[16] A. Howard. Real-time stereo visual odometry for autonomous ground vehicles. In *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, Nice, France, 2008.

[17] B. Kitt, A. Geiger, and H. Lategahn. Visual odometry based on stereo image sequences with RANSAC-based outlier rejection scheme. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, 2010.

[18] 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.

[19] Z. Kukelova, M. Bujnak, and T. Pajdla. Polynomial Eigenvalue solutions to the 5-pt and 6-pt relative pose problems. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2008.

[20] 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.

[21] Thomas Lemaire and Simon Lacroix. Monocular-vision based slam using line segments. In *Robotics and Automation, 2007 IEEE International Conference on*, pages 2791–2796. IEEE, 2007.

[22] H. Li and R. Hartley. Five-point motion estimation made easy. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, volume 1, pages 630–633, 2006.

[23] H.C. Longuet-Higgins. *Readings in computer vision: issues, problems, principles, and paradigms*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1987.

[24] C. Mei and E. Malis. Fast central catadioptric line extraction, estimation, tracking and structure from motion. In *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, Beijing, China, 2006.

[25] R. Mur-Artal, J. M. M. Montiel, and J. D. Tards. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Transactions on Robotics (T-RO)*, 31(5):1147–1163, 2015.

[26] R. Newcombe, S. Lovegrove, and A. Davison. DTAM: Dense Tracking and Mapping in Real-Time. In *Proceedings of the International Conference on Computer Vision (ICCV)*, Barcelona, Spain, 2011.

[27] D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 26(6):756–777, 2004.

[28] D. Nistér, O. Naroditsky, and J. Bergen. Visual odometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 652–659, Washington, DC, USA, 2004.

[29] A. Pumarola, A. Vakhitov, A. Agudo, A. Sanfeliu, and F. Moreno-Noguer. PL-SLAM: Real-time monocular visual SLAM with points and lines. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Singapore, 2017.

[30] D. Scaramuzza, F. Fraundorfer, M. Pollefeys, and R. Siegwart. Absolute scale in structure from motion from a single vehicle mounted camera by exploiting nonholonomic constraints. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2009.

[31] 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.

[32] 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.

[33] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012.

[34] Chris Sweeney, John Flynn, and Matthew Turk. Solving for relative pose with a partially known rotation is a quadratic eigenvalue problem. In *International Conference on 3D Vision*, 2014.

[35] X. Zuo, X. Xie, Y. Liu, and G. Huang. Robust visual SLAM with point and line features. *Arxiv Computing Research Repository*, abs/1711.08654, 2017.