On learning pseudo-sensors to improve egomotion estimation for mobile autonomy

by

Valentin Peretroukhin

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Institute for Aerospace Studies
University of Toronto

Abstract

On learning pseudo-sensors to improve egomotion estimation for mobile autonomy

Valentin Peretroukhin
Doctor of Philosophy
Graduate Department of Institute for Aerospace Studies
University of Toronto
2019

The ability to estimate *egomotion*, that is, to track one's own pose through an unknown environment, is at the heart of safe and reliable mobile autonomy. By inferring pose changes from sequential sensor measurements, egomotion estimation forms the basis of mapping and navigation pipelines, and permits mobile robots to self-localize within environments where external localization sources are intermittent or unavailable. Visual and inertial egomotion estimation, in particular, have become ubiquitous in mobile robotics due to the availability of high-quality, compact, and inexpensive sensors that capture rich representations of the world. To remain computationally tractable, 'classical' visual-inertial pipelines (like visual odometry and visual SLAM) make simplifying assumptions that, while permitting reliable operation in ideal conditions, often lead to systematic error. In this thesis, we present several data-driven learned *pseudo-sensors* that serve to complement conventional pipelines by inferring latent information from the same data stream. Our approach retains much of the benefits of traditional pipelines, while leveraging high-capacity hyper-parametric models to extract complementary information that can be used to improve uncertainty quantification, correct for systematic bias, and improve robustness to difficult-to-model deleterious effects. We validate our pseudo-sensors on several kilometres of sensor data collected in sundry settings such as urban roads, indoor labs, and planetary analogue sites in the Canadian high arctic.

Epigraph

A little learning is a dangerous thing;
drink deep, or taste not the Pierian
spring: there shallow draughts
intoxicate the brain, and drinking
largely sobers us again.

Alexander Pope

The universe is no narrow thing and the order within it is not constrained by any latitude in its conception to repeat what exists in one part in any other part. Even in this world more things exist without our knowledge than with it and the order in creation which you see is that which you have put there, like a string in a maze, so that you shall not lose your way. For existence has its own order and that no man's mind can compass, that mind itself being but a fact among others.

Cormac McCarthy

Elephants don't play chess.

Rodney Brooks

To all those who encouraged (or, at least, *never discouraged*) my intellectual wanderlust.

Acknowledgements

This document would not have been possible without the generous support and guidance of my supervisor[1], the perennial love of my family and friends[2], and the limitless patience of my lab mates[3]. Thank you all.

[1] as well as all of my collaborators and academic mentors
[2] especially the support and encouragement of Elyse
[3] in humouring my insatiable need for debate and banter

Contents

1 Mathematical Foundations **2**

1.1 Coordinate Frames . . . 2
1.2 Rotations . . . 3
1.2.1 Unit Quaternions . . . 4
1.3 Spatial Transforms . . . 5
1.3.1 Applying Transforms . . . 6
1.4 Perturbations . . . 6
1.5 Uncertainty . . . 8

2 Classical Visual Odometry **9**

2.1 A taxonomy of VO . . . 10
2.2 A classical VO pipeline . . . 10
2.2.1 Preprocessing . . . 11
2.2.2 Data association . . . 11
2.2.3 Maximum Likelihood Motion solution . . . 13
2.3 Robust Estimation . . . 15
2.4 Outstanding Issues . . . 16

3 Predictive Robust Estimation **17**

3.1 Introduction . . . 17
3.2 Motivation . . . 18
3.3 Related Work . . . 18
3.4 PROBE: Scalar k-Nearest Neighbours . . . 19
3.4.1 Mathematical Formulation . . . 20
3.4.2 Training . . . 21
3.4.3 Evaluation . . . 21
3.4.4 Prediction Space . . . 23
3.5 Generalized Kernels . . . 28

3.5.1 Predictive noise models for visual odometry 28
3.5.2 Inference without ground truth 31
3.5.3 Experiments 33
3.5.4 KITTI 34
3.6 Summary 40

Appendices **41**

Bibliography **42**

Notation

a	:	Symbols in this font are real scalars.
$\mathbf{a}$	:	Symbols in this font are real column vectors.
$\mathbf{A}$	:	Symbols in this font are real matrices.
$\mathcal{N}(\boldsymbol{\mu}, \mathbf{R})$	:	Normally distributed with mean $\boldsymbol{\mu}$ and covariance $\mathbf{R}$.
$E[\cdot]$	:	The expectation operator.
$\underrightarrow{\boldsymbol{\mathcal{F}}}_a$	:	A reference frame in three dimensions.
$(\cdot)^\wedge$	:	An operator associated with the Lie algebra for rotations and poses. It produces a matrix from a column vector.
$(\cdot)^\vee$	:	The inverse operation of $(\cdot)^\wedge$
$\mathbf{1}$	:	The identity matrix.
$\mathbf{0}$	:	The zero matrix.
$\mathbf{p}_a^{c,b}$	:	A vector from point b to point c (denoted by the superscript) and expressed in $\underrightarrow{\boldsymbol{\mathcal{F}}}_a$ (denoted by the subscript). This vector can be in homogenous coordinates depending on context.
$\mathbf{C}_{a,b}$	:	The 3×3 rotation matrix that transforms vectors from $\underrightarrow{\boldsymbol{\mathcal{F}}}_b$ to $\underrightarrow{\boldsymbol{\mathcal{F}}}_a$: $\mathbf{p}_a^{c,b} = \mathbf{C}_{a,b}\mathbf{p}_b^{c,b}$.
$\mathbf{T}_{a,b}$	:	The 4×4 transformation matrix that transforms homogeneous points from $\underrightarrow{\boldsymbol{\mathcal{F}}}_b$ to $\underrightarrow{\boldsymbol{\mathcal{F}}}_a$: $\mathbf{p}_a^{c,a} = \mathbf{T}_{a,b}\mathbf{p}_b^{c,b}$.

Chapter 1

Mathematical Foundations

> By relieving the brain of all unnecessary work, a good notation sets it free to concentrate on more advanced problems, and, in effect, increases the mental power of the race.
>
> — Alfred North Whitehead

1.1 Coordinate Frames

Before we can present the main contributions of this thesis, it will be useful to first outline the notation and mathematical foundations that underly the work. Throughout this thesis, we largely follow the notation of Barfoot (2017) when dealing with three-dimensional rigid-body kinematics.

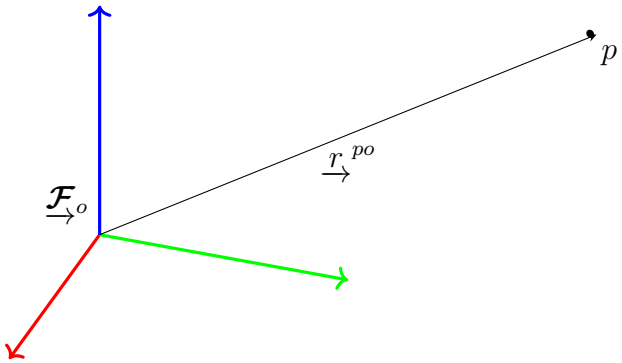


Figure 1.1: A position vector expressed in a coordinate frame.

We refer to a three-dimensional position vector, $\underrightarrow{r}^{po}$, as one that originates at the origin of a coordinate reference frame, $\underrightarrow{\mathcal{F}}_o$, and terminates at the point p. This geometric quantity has

the numerical coordinates $\mathbf{r}_o^{po}$ when expressed in $\underrightarrow{\mathcal{F}}_o$. Often, we will refer to two reference frames such as a world or *inertial* frame, $\underrightarrow{\mathcal{F}}_i$, and a vehicle frame, $\underrightarrow{\mathcal{F}}_v$. Rotation matrices or rigid-body transformations that convert coordinates from $\underrightarrow{\mathcal{F}}_i$ to $\underrightarrow{\mathcal{F}}_v$ will be represented as $\mathbf{T}_{vi}$, and $\mathbf{C}_{vi}$[1], respectively.

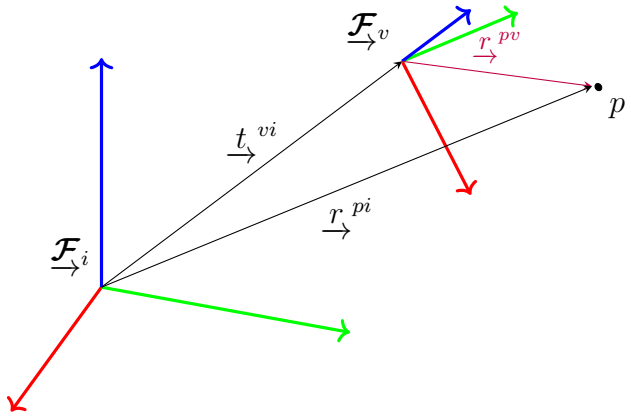


Figure 1.2: Two common references frames used throughout this thesis.

1.2 Rotations

The rotation matrix $\mathbf{C}$ is a member of the matrix Lie group $\mathrm{SO}(3)$ (the Special Orthogonal group) and can be defined as a matrix as follows:

$$\mathrm{SO}(3) = \{\mathbf{C} \in \mathbb{R}^{3\times 3} | \ \mathbf{C}^T\mathbf{C} = \mathbf{1}, \det \mathbf{C} = 1\}. \tag{1.1}$$

Active vs. Passive

An active (or *alibi*) rotation changes the coordinates of a position directly while implicitly assuming that the reference frame is fixed. A passive (or *alias*) rotation rotates the reference frame. Following Barfoot (2017), all rotation matrices in this thesis are passive unless otherwise noted.

Exponential and Logarithmic Maps

Since rotations form a matrix Lie group (we refer the reader to Solà et al. (2018) and Barfoot (2017) for a thorough treatment of Lie groups for state estimation), we can define a surjective

[1]We use $\mathbf{C}$ and not $\mathbf{R}$ for rotation matrices to avoid confusion with common notation for measurement model covariance.

exponential map[2] from three axis-angle parameters, $\boldsymbol{\phi} = \phi\mathbf{a}, \quad \phi \in \mathbb{R}, \mathbf{a} \in S^2$, to a rotation matrix, $\mathbf{C}$:

$$\mathbf{C} = \mathrm{Exp}\left(\boldsymbol{\phi}\right) = \exp\left(\boldsymbol{\phi}^{\wedge}\right) = \sum_{n=0}^{\infty} \frac{1}{n!}(\boldsymbol{\phi}^{\wedge})^n \tag{1.2}$$

$$= \cos\phi\mathbf{1} + (1 - \cos\phi)\mathbf{a}\mathbf{a}^T + \sin\phi\mathbf{a}^{\wedge}, \tag{1.3}$$

where the wedge operator $(\cdot)^{\wedge}$[3] is defined as

$$\mathbf{a}^{\wedge} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}^{\wedge} = \begin{bmatrix} 0 & -a_2 & a_1 \\ a_2 & 0 & -a_0 \\ -a_1 & a_0 & 0 \end{bmatrix}. \tag{1.4}$$

Equation (1.3) is known as the Euler-Rodriguez formula and it can also be derived geometrically, starting from Euler's theorem that any rotation can be expressed as an axis of rotation and an angle of rotation about that axis. Although the map in Equation (1.2) is surjective, we can define an inverse map if we restrict its domain to $0 \leq \phi < \pi$:

$$\boldsymbol{\phi} = \mathrm{Log}\left(\mathbf{C}\right) = \log\left(\mathbf{C}\right)^{\vee} = \frac{\phi(\mathbf{C} - \mathbf{C}^T)^{\vee}}{2\sin\phi}, \tag{1.5}$$

where $\phi = \arccos\frac{\mathrm{tr}(\mathbf{C})-1}{2}$ and the *vee* operator, $(\cdot)^{\vee} : \mathbb{R}^{3\times3} \to \mathbb{R}^3$, is defined as the unique inverse of the wedge operator $(\cdot)^{\wedge}$. Note Equation (1.5) is undefined at both $\phi = 0$ and at $\phi = \pi$. In the former case, we can use a small-angle approximation and define

$$\mathrm{Log}\left(\mathbf{C}\right) \approx (\mathbf{C} - \mathbf{1})^{\vee} \quad \text{when} \quad \phi \approx 0. \tag{1.6}$$

The latter case, (when $\phi = \pi$), defines the *cut locus* of the space where $\mathrm{Exp}\left(\cdot\right)$ is not a covering map and both $+\boldsymbol{\phi}$ and $-\boldsymbol{\phi}$ map to the same $\mathbf{C}$. This *cut locus* is related to the idea that any three parameterization of $\mathrm{SO}(3)$ will have singularities associated with it.

1.2.1 Unit Quaternions

Another way (and historically, the original way) to represent a general rotation is to use a unit quaternion, $\mathbf{q}$. A unit quaternion has four parameters, a scalar q_ω and a three-dimensional vector component, $\mathbf{q}_v$:

[2] We follow Solà et al. (2018) and also define *capitalized* map for notational clarity.

[3] This operator is often expressed as $(\cdot)^{\times}$ and is known as the skew-symmetric operator.

$$\mathbf{q} = \begin{bmatrix} q_\omega \\ \mathbf{q}_v \end{bmatrix} \in S^3, \quad (\|\mathbf{q}\| = 1). \tag{1.7}$$

Unit quaternions also form a Lie group (Solà et al., 2018) and and lie on a three-dimensional unit sphere within $\mathbb{R}^4$. This manifold represents a double cover of $\mathrm{SO}(3)$ (since both $\mathbf{q}$ and $-\mathbf{q}$ represent the same rotation). As with rotation matrices, we can define a surjective map from three parameters to the group itself,

$$\mathbf{q} = \mathrm{Exp}\left(\boldsymbol{\phi}\right) = \begin{bmatrix} \cos\phi/2 \\ \mathbf{a}\sin\phi/2 \end{bmatrix}. \tag{1.8}$$

Similarly, we can also define a logarithmic map,

$$\boldsymbol{\phi} = \mathrm{Log}\left(\mathbf{q}\right) = 2\mathbf{q}_v \frac{\arctan\left(\|\mathbf{q}_v\|, q_\omega\right)}{\|\mathbf{q}_v\|}. \tag{1.9}$$

To avoid issues with the double cover, we replace $\mathbf{q}$ with $-\mathbf{q}$ if q_ω is negative before evaluating Equation (1.9). Also note again that Equation (1.9) is undefined when $\phi = 0$, but, importantly, we do not face any issues when $\phi = \pi$ due to the half angle. As with rotation matrices, we can use small angle approximations to define:

$$\mathrm{Log}\left(\mathbf{q}\right) \approx \frac{\mathbf{q}_v}{q_\omega}\left(1 - \frac{\|\mathbf{q}_v\|^2}{3q_\omega^2}\right) \quad \text{when} \quad \phi \approx 0. \tag{1.10}$$

A fantastic summary of the history of rotation parameterizations, unit quaternions and the story of Hamilton and Rodriguez can be found in Altmann (1989).

1.3 Spatial Transforms

The rigid body transform $\mathbf{T}$ is a also a member of the matrix Lie group, the Special Euclidian group $\mathrm{SE}(3)$ and can be defined as a 4×4 matrix as follows:

$$\mathrm{SE}(3) = \{\mathbf{T} = \begin{bmatrix} \mathbf{C} & \mathbf{t} \\ \mathbf{0}^T & 1 \end{bmatrix} \in \mathbb{R}^{4\times 4} |\ \mathbf{C} \in \mathrm{SO}(3), \mathbf{t} \in \mathbb{R}^3\}. \tag{1.11}$$

As a member of a matrix Lie group, it also admits a surjective exponential map,

$$\mathbf{T} = \mathrm{Exp}\left(\boldsymbol{\xi}\right) = \exp\left(\boldsymbol{\xi}^\wedge\right) = \sum_{n=0}^{\infty} \frac{1}{n!}(\boldsymbol{\xi}^\wedge)^n \tag{1.12}$$

where $\boldsymbol{\xi} = \begin{bmatrix} \boldsymbol{\rho} \\ \boldsymbol{\phi} \end{bmatrix} \in \mathbb{R}^6$ and the wedge operator is overloaded (following Barfoot (2017)) as follows:

$$\boldsymbol{\xi}^\wedge \triangleq \begin{bmatrix} \boldsymbol{\rho} \\ \boldsymbol{\phi} \end{bmatrix}^\wedge = \begin{bmatrix} \boldsymbol{\phi}^\wedge & \boldsymbol{\rho} \\ \mathbf{0}^T & 0 \end{bmatrix}. \tag{1.13}$$

In practice, we can evaluate the exponential map through the Euler-Rodriguez formula (Equation (1.3)) and by computing the left-Jacobian of $\mathrm{SO}(3)$, $\mathbf{J}$,

$$\mathbf{T} = \mathrm{Exp}\left(\begin{bmatrix} \boldsymbol{\rho} \\ \boldsymbol{\phi} \end{bmatrix}\right) = \begin{bmatrix} \mathbf{C}(\boldsymbol{\phi}) & \mathbf{J}(\boldsymbol{\phi})\boldsymbol{\rho} \\ \mathbf{0}^T & 1 \end{bmatrix}, \tag{1.14}$$

where

$$\mathbf{J}(\boldsymbol{\phi}) = \frac{\sin\phi}{\phi}\mathbf{1} + (1 - \frac{\sin\phi}{\phi})\mathbf{a}\mathbf{a}^T + \frac{1-\cos\phi}{\phi}\mathbf{a}^\wedge. \tag{1.15}$$

1.3.1 Applying Transforms

Applying our notation for coordinate frames (and referring back to Section 1.1), a transform, $\mathbf{T}_{vi}$ can be expressed as

$$\mathbf{T}_{vi} = \begin{bmatrix} \mathbf{C}_{vi} & \mathbf{t}_v^{iv} \\ \mathbf{0}^T & 1 \end{bmatrix}. \tag{1.16}$$

This allows us to use the homogenous point representation for $\mathbf{r}_i^{pi}$ and express the following relation:

$$\begin{bmatrix} \mathbf{r}_v^{pi} \\ 1 \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{C}_{vi} & \mathbf{t}_v^{iv} \\ \mathbf{0}^T & 1 \end{bmatrix}}_{\mathbf{T}_{vi}} \begin{bmatrix} \mathbf{r}_i^{pi} \\ 1 \end{bmatrix} \tag{1.17}$$

which is numerically equivalent to

$$\mathbf{r}_v^{pi} = \mathbf{C}_{vi}\mathbf{r}_i^{pi} + \mathbf{t}_v^{iv} \tag{1.18}$$

1.4 Perturbations

It is often useful to consider a small *perturbation* about an operating point (whether that be a rotation or rigid-body transform). By leveraging a core property of Lie groups (that they are locally 'Euclidian'), we can convert difficult non-linear problems into ones that have local linear approximations.

Using rotations as an example, we can perturb an operating point, $\overline{\mathbf{C}} \triangleq \operatorname{Exp}\left(\overline{\boldsymbol{\phi}}\right)$, in three different ways:

$$\mathbf{C} = \begin{cases} \operatorname{Exp}\left(\delta\boldsymbol{\phi}^{\ell}\right)\overline{\mathbf{C}} & \text{left perturbation,} \\ \operatorname{Exp}\left(\overline{\boldsymbol{\phi}} + \delta\boldsymbol{\phi}^{m}\right) & \text{middle perturbation,} \\ \overline{\mathbf{C}}\operatorname{Exp}\left(\delta\boldsymbol{\phi}^{r}\right) & \text{right perturbation.} \end{cases} \tag{1.19}$$

We can relate all the left and middle perturbations through the left Jacobian of $\mathrm{SO}(3)$ with the following useful identity,

$$\operatorname{Exp}\left(\left(\boldsymbol{\phi} + \delta\boldsymbol{\phi}^{m}\right)\right) \approx \operatorname{Exp}\left(\mathbf{J}(\boldsymbol{\phi})\delta\boldsymbol{\phi}^{m}\right)\operatorname{Exp}\left(\boldsymbol{\phi}\right). \tag{1.20}$$

This allows us to write $\delta\boldsymbol{\phi}^{\ell} \approx \mathbf{J}(\boldsymbol{\phi})\delta\boldsymbol{\phi}^{m}$ and elucidates why $\mathbf{J}$ is called the *left* Jacobian.

In this thesis, we will use the left and middle perturbations when appropriate. Using small angle approximations, the Euler-Rodriguez formula (Equation (1.3)) yields $\operatorname{Exp}\left(\delta\boldsymbol{\phi}\right) \approx \mathbf{1} + \delta\boldsymbol{\phi}^{\wedge}$, which allows us to write the useful formula for the left perturbation:

$$\mathbf{C} = (\mathbf{1} + (\delta\boldsymbol{\phi}^{\ell})^{\wedge})\overline{\mathbf{C}}. \tag{1.21}$$

Similarly, we can write analogous expressions for a rigid body transform, $\mathbf{T} \in \mathrm{SE}(3)$, as composed of an operating point $\overline{\mathbf{T}} \triangleq \operatorname{Exp}\left(\overline{\boldsymbol{\xi}}\right)$, and a small perturbation about that operating point:

$$\mathbf{T} = \begin{cases} \operatorname{Exp}\left(\delta\boldsymbol{\xi}^{\ell}\right)\overline{\mathbf{T}} & \text{left perturbation,} \\ \operatorname{Exp}\left(\overline{\boldsymbol{\xi}} + \delta\boldsymbol{\xi}^{m}\right) & \text{middle perturbation,} \\ \overline{\mathbf{T}}\operatorname{Exp}\left(\delta\boldsymbol{\xi}^{r}\right) & \text{right perturbation.} \end{cases} \tag{1.22}$$

Now, we can also note a similar identity for $\mathrm{SE}(3)$,

$$\operatorname{Exp}\left(\left(\boldsymbol{\xi} + \delta\boldsymbol{\xi}^{m}\right)\right) \approx \operatorname{Exp}\left(\left(\boldsymbol{\mathcal{J}}(\boldsymbol{\xi})\delta\boldsymbol{\xi}^{m}\right)\right)\operatorname{Exp}\left(\boldsymbol{\xi}\right), \tag{1.23}$$

where $\boldsymbol{\mathcal{J}}$ is the left Jacobian of $\mathrm{SE}(3)$ and defined as

$$\boldsymbol{\mathcal{J}}(\boldsymbol{\xi}) \triangleq \begin{bmatrix} \mathbf{J}(\boldsymbol{\phi}) & \mathbf{Q}(\boldsymbol{\xi}) \\ \mathbf{0} & \mathbf{J}(\boldsymbol{\phi}) \end{bmatrix}, \tag{1.24}$$

where $\mathbf{Q}(\boldsymbol{\xi})$ can be evaluated analytically (see Barfoot (2017)). This again allows us to write $\delta\boldsymbol{\xi}^{\ell} \approx \boldsymbol{\mathcal{J}}(\boldsymbol{\xi})\delta\boldsymbol{\xi}^{m}$ and form a similar expression,

$$\mathbf{T} = (\mathbf{1} + (\delta\boldsymbol{\xi}^{\ell})^{\wedge})\overline{\mathbf{T}}. \tag{1.25}$$

To derive locally linear systems from sets of rigid-body transforms, or 'poses', we can apply Equation (1.25). To update an operating point, we solve for $\delta\boldsymbol{\xi}^{\ell}$ and then use the constraint-sensitive update $\mathbf{T} \leftarrow \mathrm{Exp}\left(\delta\boldsymbol{\xi}^{\ell}\right)\overline{\mathbf{T}}$.

1.5 Uncertainty

We can also use perturbation theory to implicitly define uncertainty on constrained manifolds (see Barfoot and Furgale (2014) for a thorough discussion).

Given a concentrated normal density, $\delta\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{6\times 6})$, we can *inject* this unconstrained density onto the Lie group through left perturbations about some mean:

$$\mathbf{T} = \mathrm{Exp}\left(\delta\boldsymbol{\xi}\right)\overline{\mathbf{T}} \tag{1.26}$$

This allows us to keep track of a random variable, $\mathbf{T}$, by keeping its mean in group form, $\overline{\mathbf{T}}$, while its second statistical moment is stored as a standard 6×6 covariance matrix, $\boldsymbol{\Sigma}$.

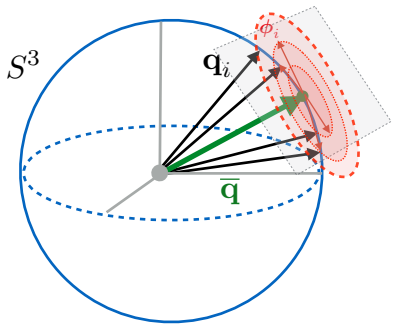


Figure 1.3: We can define uncertainty in the left tangent space of a mean element of a Lie group (here illustrated for unit quaternions).

We can define an analogous density for rotation matrices given normal densities over rotation perturbations $\delta\boldsymbol{\phi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{3\times 3})$,

$$\mathbf{C} = \mathrm{Exp}\left(\delta\boldsymbol{\phi}\right)\overline{\mathbf{C}}, \tag{1.27}$$

and also, for unit quaternions,

$$\mathbf{q} = \mathrm{Exp}\left(\delta\boldsymbol{\phi}\right) \otimes \overline{\mathbf{q}} \tag{1.28}$$

where $\otimes$ refers to the standard quaternion product operator Sola (2017).

Chapter 2

Classical Visual Odometry

> Eventually, my eyes were opened, and I really understood nature.
>
> — Claude Monet

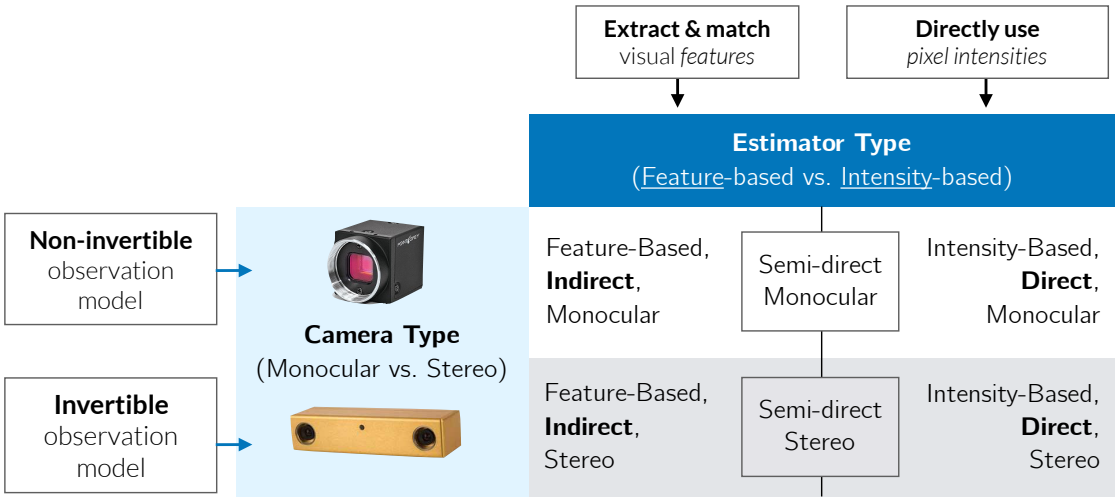


Figure 2.1: A taxonomy of different types of visual odometry.

Visual odometry (VO) has a rich history in mobile robotics and computer vision. As this dissertation largely deals with the improvement of a baseline visual odometry pipeline, we first outline the components of what we have chosen to be a canonical VO system. For a seminal tutorial on visual odometry and its more general cousin, visual SLAM, we refer the reader to two seminal papers: Scaramuzza and Fraundorfer (2011a) and Cadena et al. (2016).

2.1 A taxonomy of VO

VO can be largely divided along two dimensions (c.f. Figure 2.1): the type of camera (monocular vs. stereo) and the type of data association (indirect, or feature-based vs. direct, or pixel intensity-based).

Monocular vs. Stereo: The first distinction is based on the type of camera used by the VO pipeline. Monocular VO methods use a single camera to infer motion and can use a single compact, low-power vision sensor. They do not require any extrinsic calibration but must rely on known visual cues or external information (e.g., wheel odometry, inertial measurements) to provide metric egomotion estimates. Conversely, stereo VO methods use a stereo camera to triangulate objects with metric scale. This allows stereo VO to provide metricaly-acurate egomotion estimates. However, stereo methods rely on accurate extrinsic calibration, and their ability to resolve depth is limited by the baseline distance between the stereo pair.

Direct vs. Indirect: The second distinction is based on the type of data association used to match sequential images. Direct methods make the assumption of brightness constancy, and attempt to *directly* maximize the similarity of pixel intensities. Indirect methods, however, rely on image features detectors to extract a set of salient landmarks, and then match these landmarks across images (typically through some sort of invariant descriptor).

2.2 A classical VO pipeline

In this thesis, we apply our learned pseudo-sensors to a baseline stereo, indirect visual odometry pipeline largely based on the work of Furgale (2011). We choose this baseline system for its computational efficiency and robustness. We briefly summarize the main components of the dpipeline here.

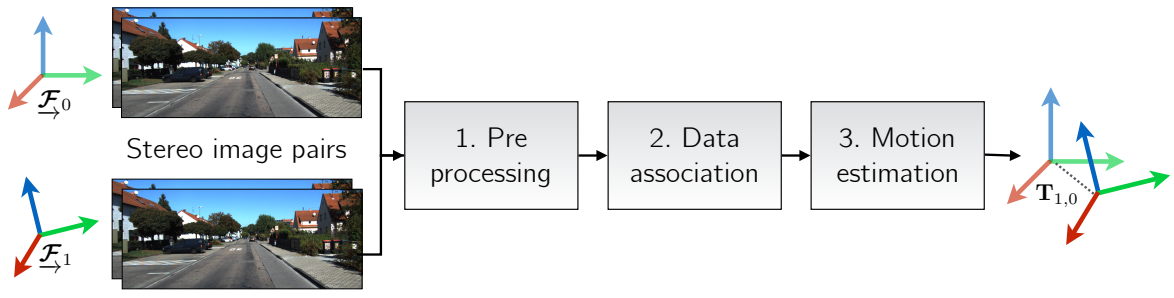
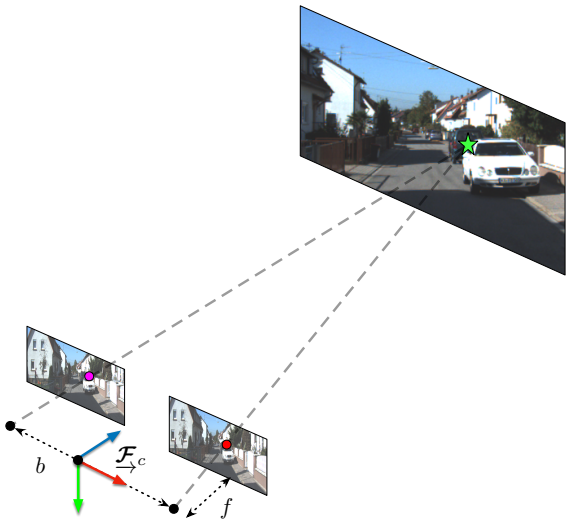
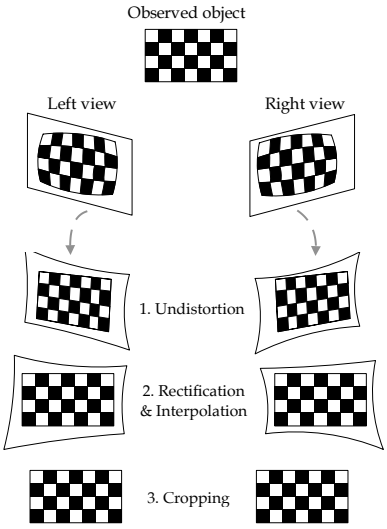


Figure 2.2: A 'classical' stereo visual odometry pipeline consists of several distinct components that have interpretable inputs and outputs.

2.2.1 Preprocessing



(a) Ideal stereo camera.



(b) Rectification and undistortion process. Figure adapted from Florez (2010).

Figure 2.3: Preprocessing components.

During preprocessing, we use a lens model (assumed to be known apriori) to undistort each stereo image. Further, using the camera extrinsic parameters (also assumed to be known), we *rectify* the stereo pair such that the images can be assumed to come from two cameras whose principal axes are parallel (Figure 2.3). Finally, we also assume that the stereo camera intrinsics are known a priori or compute them through a calibration process (Furgale et al., 2013).

2.2.2 Data association

Feature extraction and matching

In this thesis, we focus on indirect stereo visual odometry for its computational efficiency. Although a number of different types of indirect feature extraction and matching methods can be used towards this end, we choose to use the `viso2` (Geiger et al., 2011b) image feature extraction and matching algorithm. In `viso2`, features are extracted using blob and corner masks with non-minimum and non-maximum suppression. Unlike other features detectors that do not assume a particular camera motion, `viso2` assumes a smooth camera trajectory that permits fast matching through a simple sum-of-absolute-difference error metric of 11×11 windows of Sobel filter responses. Finally, features are matched across a stereo-pair and

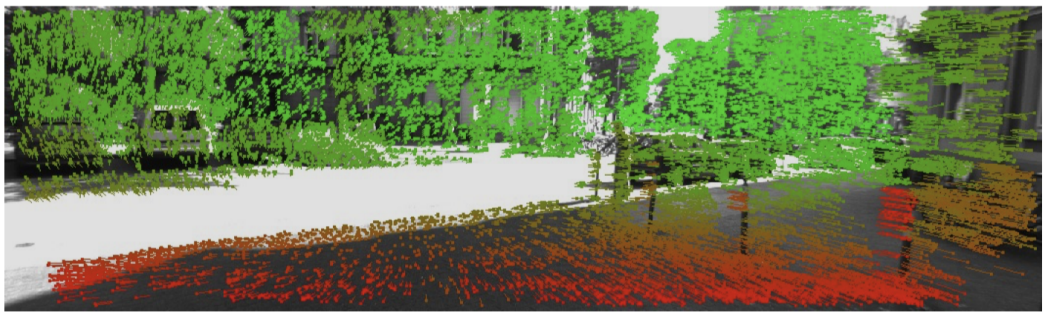


Figure 2.4: Feature tracking using `libviso2`, taken from Geiger et al. (2011a). Colours correspond to depth.

forward in time, to ensure that a single feature exists across two consecutive stereo camera poses.

Each extract feature corresponds to a point in space, expressed in homogeneous coordinates in the camera frame as $\boldsymbol{p}_{i,t} := \begin{bmatrix} p_1 & p_2 & p_3 & p_4 \end{bmatrix}^T \in \mathbb{P}^3$. Given our intrinsics and extrinsic calibration parameters, our idealized stereo-camera model, $\mathbf{f}$, projects a landmark expressed in homogeneous coordinates into image space, so that $\mathbf{y}_{i,t}$, the stereo pixel coordinates of landmark i in the first camera pose at time t, is given by

$$\mathbf{y}_{i,t} = \begin{bmatrix} u_l \\ v_l \\ u_r \\ v_r \end{bmatrix} = \mathbf{f}(\boldsymbol{p}_{i,t}) = \mathbf{M}\frac{1}{p_3}\boldsymbol{p}_{i,t}, \tag{2.1}$$

where

$$\mathbf{M} = \begin{bmatrix} f & 0 & c_u & f\frac{b}{2} \\ 0 & f & c_v & 0 \\ f & 0 & c_u & -f\frac{b}{2} \\ 0 & f & c_v & 0 \end{bmatrix}. \tag{2.2}$$

Here, $\{c_u, c_v\}$, $\{f_u, f_v\}$, and b are the principal points, focal lengths and baseline of the stereo camera respectively. Note that in this formulation, the stereo camera frame is centered between the two individual lenses.

Outlier rejection

To filter out any residual outlier matches, we use a three-point random sample consensus algorithm (RANSAC, Fischler and Bolles (1981b)) based on an analytic solution to the six

degree-of-freedom motion (Umeyama, 1991).

2.2.3 Maximum Likelihood Motion solution

We will define $\mathbf{T}_t \in \mathrm{SE}(3)$, as the rigid transform between two subsequent stereo camera poses, $\underrightarrow{\mathcal{F}}_{c_0}$ and $\underrightarrow{\mathcal{F}}_{c_1}$

$$\mathbf{T}_t = \mathbf{T}_{c_1 w} \mathbf{T}_{c_0 w}^{-1}, \tag{2.3}$$

where $\underrightarrow{\mathcal{F}}_{w}$ is a privileged world frame. After data association, we assume we have a set of N_t matches, $\{\mathbf{y}_{i,c_0}, \mathbf{y}_{i,c_1}\}_{i=1}^{N_t}$, between visual landmarks in the subsequent camera frames. For each match, we define an error function, $\mathbf{e}_i(\mathbf{T}_t)$, that relates the rigid transform to these stereo feature matches. Throughout this dissertation, we assume that these errors are corrupted by zero-mean independent Gaussian noise with the (potentially heteroscedastic) covariance, $\mathbf{\Sigma}_{i,t}$;

$$\mathbf{e}_i(\mathbf{T}_t) \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_{i,t}\right). \tag{2.4}$$

Under this noise model, the maximum likelihood transform, $\mathbf{T}_t^*$, is given by

$$\mathbf{T}_t^* = \underset{\mathbf{T}\in\mathrm{SE}(3)}{\operatorname{argmax}} \prod_{i=1}^{N_t} p(\mathbf{e}_i(\mathbf{T}_t)) = \underset{\mathbf{T}\in\mathrm{SE}(3)}{\operatorname{argmin}} \sum_{i=1}^{N_t} \mathbf{e}_i(\mathbf{T}_t)^T \mathbf{\Sigma}_{i,t}^{-1} \mathbf{e}_i(\mathbf{T}_t). \tag{2.5}$$

We will define the error function in two different ways.

Point Cloud Error

First, we can follow classical approach (Maimone et al., 2007) and define $\mathbf{e}_i(\mathbf{T}_t)$ based on a three-dimensional point cloud error. To do this, we invert our stereo camera model to triangulate pairs of points in each frame, $\boldsymbol{p}_{i,c_0} = \mathbf{f}^{-1}(\mathbf{y}_{i,c_0})$ and $\boldsymbol{p}_{i,c_1} = \mathbf{f}^{-1}(\mathbf{y}_{i,c_1})$,

$$\mathbf{e}_i(\mathbf{T}_t) = \mathbf{D}(\boldsymbol{p}_{i,c_1} - \mathbf{T}_t \boldsymbol{p}_{i,c_0}), \tag{2.6}$$

where $\mathbf{D} = \begin{bmatrix} \mathbf{1}_{3\times 3} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{3\times 4}$ converts homogenous coordinates into Euclidian coordinates.

We follow Maimone et al. (2007) and assume each stereo projection is corrupted by additive Gaussian noise,

$$\mathbf{y}_{i,c} \sim \mathcal{N}\left(\bar{\mathbf{y}}_{i,c}, \mathbf{R}_{i,c}\right), \tag{2.7}$$

then we can compute a density on the error function itself through first order noise prop-

agation as

$$\mathbf{e}_i(\mathbf{T}_t) \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_{i,t}\right), \tag{2.8}$$

where

$$\mathbf{\Sigma}_{i,t} = \mathbf{D}\mathbf{G}_{i,c_1}\mathbf{R}_{i,c_1}\mathbf{G}_{i,c_1}^T\mathbf{D}^T + \mathbf{D}\mathbf{T}_t\mathbf{G}_{i,c_0}\mathbf{R}_{i,c_0}\mathbf{G}_{i,c_0}^T\mathbf{T}_t^T\mathbf{D}^T \tag{2.9}$$

with $\mathbf{G}_{i,c} = \left.\dfrac{\partial \mathbf{f}^{-1}}{\partial \mathbf{y}}\right|_{\mathbf{y}_{i,c}}$.

Reprojection Error

Alternatively, we can represent reprojection errors in the second frame directly as

$$\mathbf{e}_i(\mathbf{T}_t) = \mathbf{y}_{i,c_1} - \mathbf{f}(\mathbf{T}_t\mathbf{f}^{-1}(\mathbf{y}_{i,c_0})), \tag{2.10}$$

and model reprojection errors directly as

$$\mathbf{e}_i(\mathbf{T}_t) \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_{i,t}\right) = \mathcal{N}\left(\mathbf{0}, \mathbf{R}_{i,t}\right), \tag{2.11}$$

where we abuse notation (slightly) and replace the index for the camera frames c_0 or c_1 with t to indicate that this covariance refers to the reprojection error that involves both sets of features.

Solution via Gauss-Newton Optimization

In either case, we have now defined a weighted nonlinear least squares problem which can be solved iteratively using standard techniques. For our purposes, we opt to use Gauss-Newton optimization and follow Barfoot (2017) to optimize constrained poses.

Namely, at a given iteration n, we linearize the error function $\mathbf{e}_i(\mathbf{T}_t)$, about an operating point $\mathbf{T}_t^{(n)} \in \mathrm{SE}(3)$, which results in a quadratic approximation to Equation (3.3). To linearize, we consider the left perturbations $\delta\boldsymbol{\xi} \in \mathbb{R}^6$ represented in exponential coordinates:

$$\mathbf{T}_t = \mathrm{Exp}\left(\delta\boldsymbol{\xi}\right)\mathbf{T}_t^{(n)} \approx (\mathbf{1} + \delta\boldsymbol{\xi}^{\wedge})\mathbf{T}_t^{(n)}. \tag{2.12}$$

This allows us to transform Equation (3.3) into a linear least squares objective in $\delta\boldsymbol{\xi}$:

$$\mathcal{L}(\delta\boldsymbol{\xi}) = \frac{1}{2}\sum_{i=1}^{N_t}\left(\mathbf{e}_i - \mathbf{J}_i\delta\boldsymbol{\xi}\right)^T\mathbf{\Sigma}_i^{-1}\left(\mathbf{e}_i - \mathbf{J}_i\delta\boldsymbol{\xi}\right) \tag{2.13}$$

where $\mathbf{J}_i = \left.\frac{\partial \mathbf{e}_i}{\partial \delta \boldsymbol{\xi}}\right|_{\mathbf{T}_t^{(n)}}$, $\mathbf{e}_i = \mathbf{e}_i(\mathbf{T}_t^{(n)})$, and $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_{i,t}(\mathbf{T}_t^{(n)})$. The minimum to this objective can be solved for analytically by solving the normal equations. This results in the optimal parameters,

$$\delta \boldsymbol{\xi}^\star = \left(\sum_{i=1}^{N_t} \mathbf{J}_i^T \, \boldsymbol{\Sigma}_i^{-1} \mathbf{J}_i \right)^{-1} \sum_{i=1}^{N_t} \mathbf{J}_i^T \, \boldsymbol{\Sigma}_i^{-1} \mathbf{e}_i. \tag{2.14}$$

We then update the operating point and proceed to the next iteration,

$$\mathbf{T}^{(n+1)} = \operatorname{Exp}\left(\delta \boldsymbol{\xi}^\star\right) \mathbf{T}^{(n)}. \tag{2.15}$$

There are many reasonable choices for both the initial transform $\mathbf{T}^{(0)}$ and for the conditions under which we terminate iteration. We initialize the estimated transform to identity, and iteratively perform the update given by eq. (2.15) until we see a relative change in the squared error of less than one percent after an update.

2.3 Robust Estimation

Since eq. (2.13) assigns cost values that grow quadratically with measurement error, it is very sensitive to outlier measurements. A common solution to this problem is to replace the L_2 cost function with one that is less sensitive to large measurement errors (MacTavish and Barfoot, 2015). These robust cost functions are collectively known as M-estimators, and many variants exist. Each uses a re-weighting function, $\rho(\cdot)$,

$$\mathbf{T}^* = \underset{\mathbf{T} \in \mathrm{SE}(3)}{\operatorname{argmin}} \sum_{i=1}^{N_t} \rho\left({\mathbf{e}_i}^T \, \boldsymbol{\Sigma}_{i,t}^{-1} \mathbf{e}_i \right) = \underset{\mathbf{T} \in \mathrm{SE}(3)}{\operatorname{argmin}} \sum_{i=1}^{N_t} \rho(\epsilon_i), \tag{2.16}$$

where. given a parameter c, some common examples include:

$$\rho(\epsilon) = \begin{cases} \frac{c^2}{2} \log\left(1 + \frac{\epsilon^2}{c^2}\right) & \text{Cauchy,} \\ \frac{1}{2} \frac{\epsilon^2}{c^2 + \epsilon^2} & \text{Geman-McClure (Geman et al., 1992),} \\ \begin{cases} \frac{\epsilon^2}{2} & \text{if } \|\epsilon\| < c \\ c \|\epsilon\| - \frac{c^2}{2} & \text{if } \|\epsilon\| \geq c \end{cases} & \text{Huber (Huber, 1964).} \end{cases} \tag{2.17}$$

2.4 Outstanding Issues

There are several outstanding limitations of classical visual odometry pipelines that we can address with learned pseudo-sensors.

Table 2.1: **Data efficiency vs. computational efficiency**

Synopsis	**Addressed by**
Classical VO pipelines face a difficult-to-optimize trade-off between using all of the information contained within image and while still remaining computationally tractable.	PROBE, DPC-Net, Sun-BCNN, HydraNet

Table 2.2: **Systematic bias**

Synopsis	**Addressed by**
Stereo visual odometry can incur systematic bias through poor extrinsic or intrinsic calibration, stereo triangulation errors, poor feature *spread* (i.e., concentration of features on one side of an image), and poor data association due self-similar textures.	DPC-Net

Table 2.3: **Homoscedastic uncertainty**

Synopsis	**Addressed by**
Stationary, homoscedastic noise in observation models can often reduce the consistency and accuracy of state estimates. This is especially true for complex, inferred measurement models. In visual data, inferred visual observations can be degraded not only due to sensor imperfections (e.g. poor intrinsic calibration, digitization effects, motion blur), but also as a result of the observed environment (e.g. self-similar scenes, specular surfaces, textureless environments).	PROBE, Sun-BCNN, HydraNet

Chapter 3

Predictive Robust Estimation

> Information is the resolution of uncertainty.
>
> Claude Shannon

3.1 Introduction

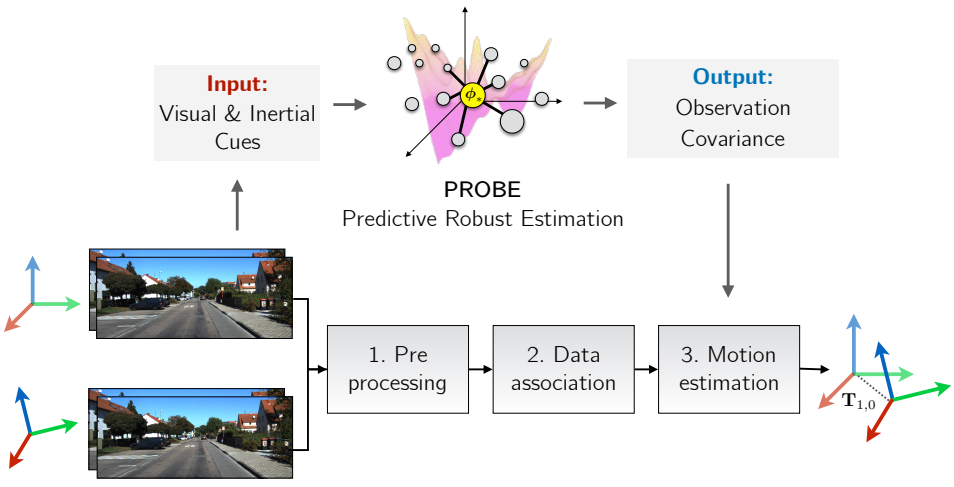


Figure 3.1: PROBE builds a predictive noise model for stereo visual odometry.

The first pseudo-sensor we present is a general technique we call PRedictive ROBust Estimation, or PROBE. This approach uses non-parametric learning to predict observation covariances for a stereo visual odometry pipeline, effectively scaling a least squares objective in a predictive fashion. We present two different methods to learn and incorporate these covariances. First we use a simple k-nearest-neighbours approach to learn isotropic covariances for

three dimensional point-cloud matching. Second, we extend this significantly by applying the method of Generalized Kernels to a Bayesian treatment of covariance learning. We show that by assuming a particular covariance prior over re-projection errors, we can derive a robust least squares loss with parameters that are predicted for each error by our approach.

There are three publications associated with this work:

1. Peretroukhin, V., Clement, L., Giamou, M., and Kelly, J. (2015a). PROBE: Predictive robust estimation for visual-inertial navigation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'15)*, pages 3668–3675, Hamburg, Germany

2. Peretroukhin, V., Clement, L., and Kelly, J. (2015c). Get to the point: Active covariance scaling for feature tracking through motion blur. In *Proceedings of the IEEE International Conference on Robotics and Automation Workshop on Scaling Up Active Perception*, Seattle, Washington, USA

3. Peretroukhin, V., Vega-Brown, W., Roy, N., and Kelly, J. (2016). PROBE-GK: Predictive robust estimation using generalized kernels. In *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, pages 817–824.

3.2 Motivation

Robot navigation relies on an accurate quantification of sensor noise or uncertainty in order to produce reliable state estimates. In practice, this uncertainty is often fixed for a given sensor and experiment, whether by automatic calibration or by manual tuning. Although a fixed measure of uncertainty may be reasonable in certain static environments, dynamic scenes frequently exhibit many effects that corrupt a portion of the available observations. For visual sensors, these effects include, for example, self-similar textures, variations in lighting, moving objects, and motion blur. We assert that there may be useful information available in these observations that would normally be rejected by a fixed-threshold outlier rejection scheme. Ideally, we would like to retain some of these observations in our estimator, while still placing more trust in observations that do not suffer from such effects.

3.3 Related Work

There is a large and growing body of work on the problem of deriving accurate, consistent state estimates from visual data. Although our approach to noise modelling is applicable in

other domains, for simplicity we focus our attention on the problem of inferring egomotion from features extracted from sequential pairs of stereo images; see Sünderhauf and Protzel (2007) for a survey of techniques. The spectrum of alternative approaches to visual state estimation include monocular techniques, which may be feature-based (Scaramuzza and Fraundorfer, 2011b), direct (Irani and Anandan, 2000), or semi-direct (Forster et al., 2014).

Apart from simply rejecting outliers, a number of recent approaches attempt to select the optimal set of features to produce an accurate localization estimate from tracked visual features. For example, Tsotsos et al. (2015) amend Random Sample Consensus (RANSAC) with statistical hypothesis testing to ensure that tracked visual features have normally distributed residuals before including them in the estimator. Unlike our predictive approach, their technique relies on the availability of feature tracks, and requires scene overlap to work continuously. In a different approach, Zhang and Vela (2015) choose an optimally observable feature subset for a monocular SLAM pipeline by selecting features with the highest *informativeness* - a measure calculated based on the observability of the SLAM subsystem. Observability, however, is governed by the 3D location of the features, and therefore cannot predict systematic feature degradation due to environmental or sensor-based effects.

3.4 PROBE: Scalar k-Nearest Neighbours

In our initial exploratory work, we explored the idea of scaling With Predictive ROBust Estimation, we aim to improve localization accuracy in the presence of such effects by building a model of the uncertainty in the affected visual observations. We learn the model in an offline training procedure and then use it online to predict the uncertainty of incoming observations as a function of their location in a predefined *prediction space*. Our model can be learned in completely unknown environments with frequent or infrequent ground truth data.

The primary contributions of this research are a flexible framework for learning the quality of visual features with respect to navigation estimates, and a straightforward way to incorporate this information into a navigation pipeline. On its own, PROBE can produce more accurate estimates than a binary outlier rejection scheme like Random Sample Consensus (RANSAC) because it can simultaneously reduce the influence of outliers while intelligently weighting inliers. PROBE reduces the need to develop finely-tuned uncertainty models for complex sensors such as cameras, and better accounts for the effects observed in complex, dynamic scenes than typical fixed-uncertainty models. While we present PROBE in the context of visual feature-based navigation, we stress that it is not limited to visual measurements and could also be applied to other sensor modalities.

The aim of PROBE is to learn a model for the quality of visual features, with the goal of

reducing the impact of deleterious visual effects such as moving objects, motion blur, and shadows on navigation estimates. Feature quality is characterized by a scalar weight, β_i, for each visual feature in an environment. To compute β_i we define a prediction space (similar to Vega-Brown et al. (2013)) that consists of a set of visual-inertial predictors computed from the local image region around the feature and the inertial state of the vehicle (Section 3.4.4 details our choice of predictors). We then scale the image covariance of each feature by β_i during the non-linear optimization.

In a similar manner to M-estimation, PROBE achieves robustness by varying the influence of certain measurements. However, in contrast to robust cost functions that weight measurements based purely on estimation error, PROBE weights measurements based on their assessed quality.

To learn the model, we require training data that consists of a traversal through a typical environment with some measure of ground truth for the path, but not for the visual features themselves. Like many machine learning techniques, we assume that the training data is representative of the test environments in which the learned model will be used.

We learn the quality of visual features *indirectly* through their effect on navigation estimates. We define high quality features as those that result in estimates that are close to ground truth. Our framework is flexible enough that we do not require ground truth at every image and we can learn the model based on even a single loop closure error.

3.4.1 Mathematical Formulation

To solve for the relative egomotion between two camera frames, $\underrightarrow{\mathcal{F}}_{c_0}$ and $\underrightarrow{\mathcal{F}}_{c_1}$, we follow technique described in Section 2.2.3 to convert stereo observations into point-clouds and then solve for the maximum likelihood $\mathrm{SE}(3)$ transformation. We associate with each match $\{\mathbf{y}_{i,c_0}, \mathbf{y}_{i,c_1}\}$ a vector of *predictors*, $\boldsymbol{\phi}_{i,c}$. Each predictor can be computed using both visual[1] and inertial cues , allowing us to model effects like motion blur and self-similar textures. We then compute the covariance as a function of these predictors, so that $\mathbf{R}_{i,c} = \mathbf{R}(\boldsymbol{\phi}_{i,c})$, and we use the same covariance function for features in both frames[2],

$$\mathbf{y}_{i,c_0} \sim \mathcal{N}\left(\bar{\mathbf{y}}_{i,c_0}, \mathbf{R}_{i,c}\right) = \mathcal{N}\left(\bar{\mathbf{y}}_{i,c_0}, \mathbf{R}(\boldsymbol{\phi}_{i,c})\right) \tag{3.1}$$

$$\mathbf{y}_{i,c_1} \sim \mathcal{N}\left(\bar{\mathbf{y}}_{i,c_1}, \mathbf{R}_{i,c}\right) = \mathcal{N}\left(\bar{\mathbf{y}}_{i,c_1}, \mathbf{R}(\boldsymbol{\phi}_{i,c})\right) \tag{3.2}$$

This then builds the following weighted least squares objective,

[1] Including potentially data from all four images in the pair of stereo images.

[2] We conjecture that this is reasonable in a VO setup, where images change minimally between consecutive frames.

$$\mathbf{T}_t^* = \underset{\mathbf{T}\in \mathrm{SE}(3)}{\operatorname{argmin}} \sum_{i=1}^{N_t} \mathbf{e}_i(\mathbf{T}_t)^T \boldsymbol{\Sigma}_{i,t}^{-1} \mathbf{e}_i(\mathbf{T}_t). \tag{3.3}$$

where $\boldsymbol{\Sigma}_{i,t}$ is now given by,

$$\boldsymbol{\Sigma}_{i,t} = \mathbf{D}\mathbf{G}_{i,c_1}\mathbf{R}(\boldsymbol{\phi}_{i,c})\mathbf{G}_{i,c_1}^T\mathbf{D}^T + \mathbf{D}\mathbf{T}_t\mathbf{G}_{i,c_0}\mathbf{R}(\boldsymbol{\phi}_{i,c})\mathbf{G}_{i,c_0}^T\mathbf{T}_t^T\mathbf{D}^T \tag{3.4}$$

We build a model for $\mathbf{R}(\boldsymbol{\phi}_{i,c})$ as,

$$\mathbf{R}(\boldsymbol{\phi}_{i,c}) = \beta(\boldsymbol{\phi}_{i,c})\bar{\mathbf{R}}, \tag{3.5}$$

with

$$\beta(\boldsymbol{\phi}_{i,c}) = \left(\frac{1}{\epsilon_{\text{avg}}K}\sum_{k=1}^{K}\epsilon_k\right)^{\gamma},\ \epsilon_k \in \text{K-NN}(\boldsymbol{\phi}_{i,c}), \tag{3.6}$$

where $\{\epsilon_k\}_{k=1}^K$ are K egomotion errors that are 'nearest' to $\boldsymbol{\phi}_{i,c}$, ϵ_{avg} is an average error, and $\gamma > 1$ is a hyper-parameter designed to exaggerate the effect of small changes in position error.

3.4.2 Training

Training proceeds by traversing the training path, selecting a subset of visual features at each step, and using them to compute an incremental position estimate. By comparing the estimated position to the ground truth position, we compute the translational Root Mean Squared Error (RMSE), ϵ, and store it at each feature's position in the prediction space. The full algorithm is summarized in Algorithm 1.

3.4.3 Evaluation

To use the PROBE model in a test environment, we compute the location of each observed visual feature in our prediction space, and then compute its relative weight β_i as a function of its K nearest neighbours in the training set. For efficiency, the K nearest neighbours are found using a k-d tree. The final scaling factor β_i is a function of the mean of the α values corresponding to the K nearest neighbours, normalized by ϵ_{avg}, the mean α value of the entire training set.

The value of K can be determined through cross-validation, and in practice depends on the size of the training set and the environment. The computation of β_i is designed to map small differences in learned α values to scalar weights that span several orders of magnitude.

Algorithm 1 Train PROBE based on a dataset ($\mathcal{D}$) of pairs of input sensor data ($\mathcal{I}_s$) and ground truth egomotion ($\mathbf{T}_s$).

function BUILDPROBEMODEL($\mathcal{D}$)
  for $l \leftarrow [1, ..., N_{iter}]$ **do**
    for all $\mathcal{I}_s$, $\mathbf{T}_s$ in $\mathcal{D}$ **do**
      $\mathcal{F} \leftarrow$ EXTRACTFEATURES($\mathcal{I}_s$)
      $\{f_1, \ldots, f_J\} \leftarrow$ SAMPLE($\mathcal{F}$)
      $\hat{\mathbf{T}} \leftarrow$ COMPUTETRANSFORM($\{f_1, \ldots, f_J\}$)
      $\epsilon \leftarrow$ ERROR($\hat{\mathbf{T}}, \mathbf{T}_s$)
      $\{\boldsymbol{\phi}_{s,1}, \ldots, \boldsymbol{\phi}_{s,J}\} \leftarrow$ PREDICTOR($\{f_1, \ldots, f_J\}$)
      Insert $\{\boldsymbol{\phi}_{s,1}, \ldots, \boldsymbol{\phi}_{s,J}\}$ into $\mathcal{M}$ and store ϵ at all J locations
    end for
  end for
  return $\mathcal{M}$
end function

Algorithm 2 Compute scalar covariance factors, β_i, for a set of stereo feature tracks (and IMU data), $\mathcal{F}$, given a PROBE model $\mathcal{M}$.

function USEPROBE($\mathcal{M}$, $\mathcal{F}$, γ)
  $\epsilon_{\text{avg}} \leftarrow$ AVERAGEERROR($\mathcal{M}$)
  for all f_i in $\mathcal{F}$ **do**
    $\boldsymbol{\phi}_i \leftarrow$ PREDICTOR(f_i)
    $\epsilon_1, ..., \epsilon_K \leftarrow$ FINDKNN($\boldsymbol{\phi}_i, K, \mathcal{M}$)
    $\beta_i \leftarrow \left(\frac{1}{\epsilon_{\text{avg}} K} \sum_{k=1}^{K} \epsilon_k\right)^{\gamma}$
  end for
  return $\boldsymbol{\beta} = \{\beta_i\}$
end function

An appropriate value of γ can be found by searching through a set range of candidate values and choosing the value that minimizes the average RMSE (ARMSE) on the training set.

3.4.4 Prediction Space

A crucial component of our technique is the choice of prediction space. In practice, feature tracking quality is often degraded by a variety of effects such as motion blur, moving objects, and textureless or self-similar image regions. The challenge is in determining predictors that account for such effects without requiring excessive computation. In our implementation, we use the following predictors, but stress that the choice of predictors can be tailored to suit particular applications and environments:

- Angular velocity and linear acceleration magnitudes
- Local image entropy
- Blur (quantified by the blur metric of Crete et al. (2007))
- Optical flow variance score
- Image frequency composition

We discuss each of these predictors in turn.

Angular velocity and linear acceleration

While most of the predictors in our system are computed directly from image data, the magnitudes of the angular velocities and linear accelerations reported by the IMU are in themselves good predictors of image degradation (e.g., image blur) and hence poor feature tracking.

Local image entropy

Entropy is a statistical measure of randomness that can be used to characterize the texture in an image or patch. Since the quality of feature detection is strongly influenced by the strength of the texture in the vicinity of the feature point, we expect the entropy of a patch centered on the feature to be a good predictor of its quality. We evaluate the entropy S in an image patch by sorting pixel intensities into N bins and computing

$$S = -\sum_{i=1}^{N} c_i \log_2(c_i), \tag{3.7}$$

where c_i is the number of pixels counted in the i^{th} bin.

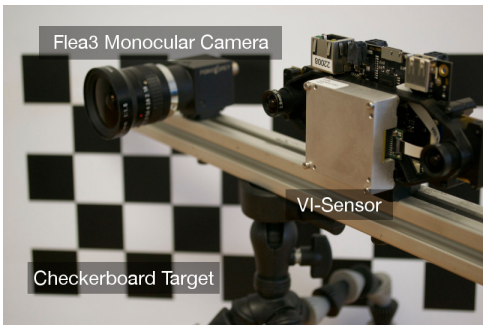


Figure 3.2: The Skybotix VI-Sensor, Point Grey Flea3, and checkerboard target used in our motion blur experiments.

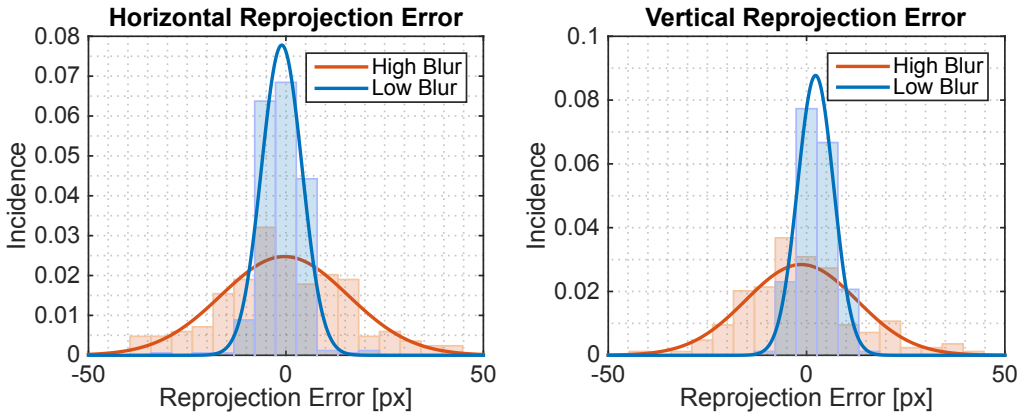


Figure 3.3: Reprojection error of checkerboard corners triangulated from the VI-Sensor and reprojected into the Flea3.

Blur

Blur can arise from a number of sources including motion, dirty lenses, and sensor defects. All of these have deleterious effects on feature tracking quality. To assess the effect of blur in detail, we performed a separate experiment. We recorded images of 32 interior corners of a standard checkerboard calibration target using a low frame-rate (20 FPS) Skybotix VI-Sensor stereo camera and a high frame-rate (125 FPS) Point Grey Flea3 monocular camera rigidly connected by a bar (Figure 3.2). Prior to the experiment, we determined the intrinsic and extrinsic calibration parameters of our rig using the KALIBR[3] package Furgale et al. (2013). The apparatus underwent both slow and fast translational and rotational motion, which induced different levels of motion blur as quantified by the blur metric proposed by Crete et al. (2007).

We detected checkerboard corners in each camera at synchronized time steps, computed their 3D coordinates in the VI-Sensor frame, then reprojected these 3D coordinates into the Flea3 frame. We then computed the reprojection error as the distance between the reprojected

[3] https://github.com/ethz-asl/kalibr

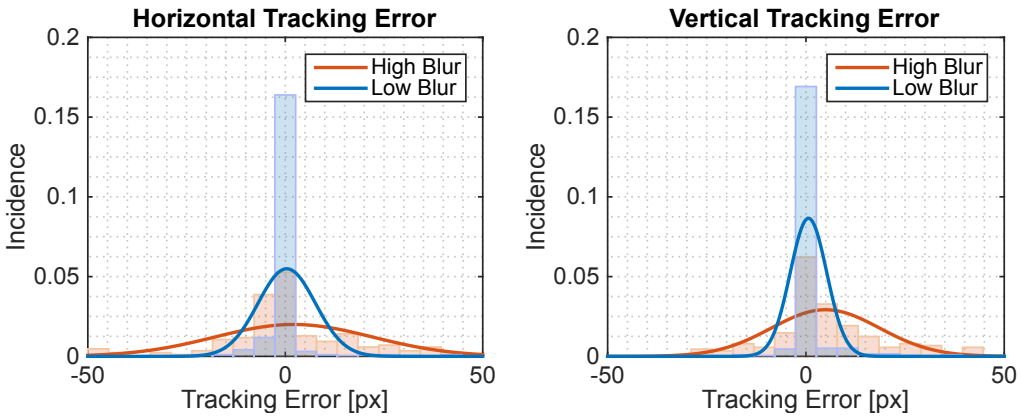


Figure 3.4: Effect of blur on reprojection and tracking error for the slow-then-fast checkerboard dataset. We distinguish between high and low blur by thresholding the blur metric Crete et al. (2007). The variance in both errors increases with blur.

image coordinates and the true image coordinates in the Flea3 frame. Since the Flea3 operated at a much higher frame rate than the VI-Sensor, it was less susceptible to motion blur and so we treated its observations as ground truth. We also computed a tracking error by comparing the image coordinates of checkerboard corners in the left camera of the VI-Sensor computed from both KLT tracking Lucas and Kanade (1981) and re-detection.

Figure 3.4 shows histograms and fitted normal distributions for both reprojection error and tracking error. From these distributions we can see that the errors remain approximately zero-mean, but that their variance increases with blur. This result is compelling evidence that the effect of blur on feature tracking quality can be accounted for by scaling the feature covariance matrix by a function of the blur metric.

Optical flow variance score

To detect moving objects, we compute a score for each feature based on the ratio of the variance in optical flow vectors in a small region around the feature to the variance in flow vectors of a larger region. Intuitively, if the flow variance in the small region differs significantly from that in the larger region, we might expect the feature in question to belong to a moving object, and we would therefore like to trust the feature less. Since we consider only the variance in optical flow vectors, we expect this predictor to be reasonably invariant to scene geometry.

We compute this optical flow variance score according to

$$\log\left(\frac{\bar{\sigma}_s^2}{\bar{\sigma}_l^2}\right), \tag{3.8}$$

where $\bar{\sigma}_s^2, \bar{\sigma}_l^2$ are the means of the variance of the vertical and horizontal optical flow vector

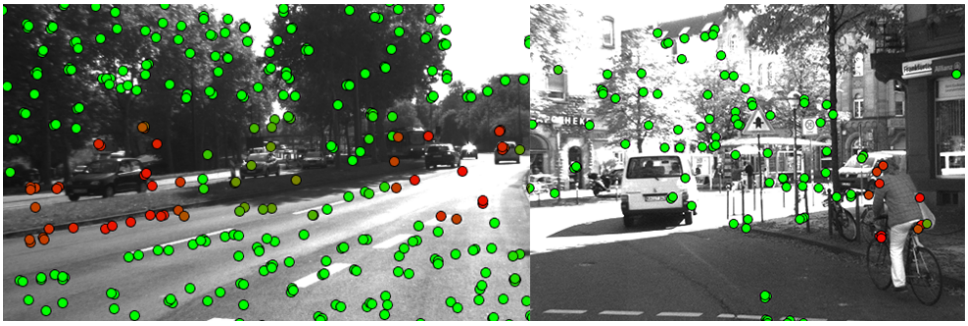


Figure 3.5: The optical flow variance predictor can help in detecting moving objects. Red circles correspond to higher values of the optical flow variance score (i.e., features more likely to belong to a moving object).

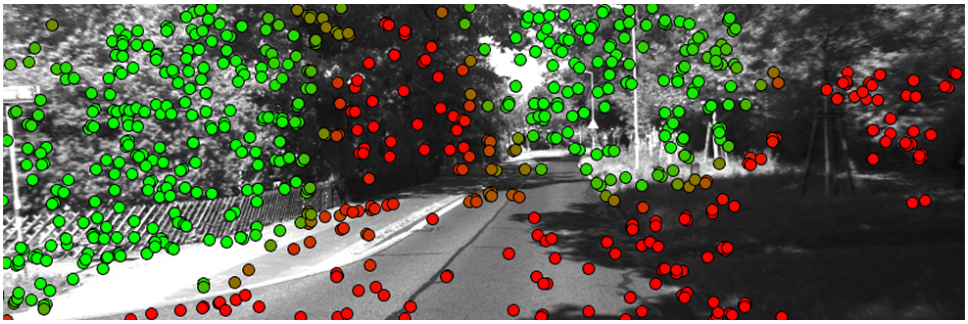


Figure 3.6: A high-frequency predictor can distinguish between regions of high and low texture such as foliage and shadows. Green indicates higher values.

components in the small and large regions respectively. Figure 3.5 shows sample results of this scoring procedure for two images in the KITTI dataset. Our optical flow variance score generally picks out moving objects such as vehicles and cyclists in diverse scenes.

Image frequency composition

Reliable feature tracking is often difficult in textureless or self-similar environments due to low feature counts and false matches. We detect textureless and self-similar image regions by computing the Fast Fourier Transform (FFT) of each image and analyzing its frequency composition. For each feature, we compute a coefficient for the low- and high-frequency regimes of the FFT. Figure 3.6 shows the result of the high-frequency version of this predictor on a sample image from the KITTI dataset. Our high-frequency predictor effectively distinguishes between textureless regions (e.g., shadows and roads) and texture-rich regions (e.g., foliage).

Table 3.1: Comparison of translational Average Root Mean Square Error (ARMSE) and Final Translational Error on the KITTI dataset.

			Nominal RANSAC (99% outlier rejection)		Aggressive RANSAC (99.99% outlier rejection)		PROBE	
Trial	Type	Path Length	ARMSE	Final Error	ARMSE	Final Error	ARMSE	Final Error
26_drive_0051	City [1]	251.1 m	4.84 m	12.6 m	3.30 m	8.62 m	3.48 m	8.07 m
26_drive_0104	City [1]	245.1 m	0.977 m	4.43 m	0.850 m	3.46 m	1.19 m	3.61 m
29_drive_0071	City [1]	234.0 m	5.44 m	30.3 m	5.44 m	30.4 m	3.03 m	12.8 m
26_drive_0117	City [1]	322.5 m	2.29 m	9.07 m	2.29 m	9.07 m	2.76 m	9.08 m
30_drive_0027	Residential [1, †]	667.8 m	4.22 m	12.2 m	4.30 m	10.6 m	3.64 m	4.57 m
26_drive_0022	Residential [2]	515.3 m	2.21 m	3.99 m	2.66 m	6.09 m	3.06 m	4.99 m
26_drive_0023	Residential [2]	410.8 m	1.64 m	8.20 m	1.77 m	8.27 m	1.71 m	8.13 m
26_drive_0027	Road [3]	339.9 m	1.63 m	8.75 m	1.63 m	8.65 m	1.40 m	7.57 m
26_drive_0028	Road [3]	777.5 m	4.31 m	16.9 m	3.72 m	13.1 m	3.92 m	13.2 m
30_drive_0016	Road [3]	405.0 m	4.56 m	19.5 m	3.33 m	14.6 m	2.76 m	13.9 m
UTIAS Outdoor	Snowy parking lot	302.0 m	7.24 m	10.1 m	7.02 m	10.6 m	6.85 m	6.09 m
UTIAS Indoor	Lab interior	32.83 m	—	0.854 m	—	0.738 m	—	0.617 m

[1] Trained using sequence 09_26_drive_0005. [2] Trained using sequence 09_26_drive_0046. [3] Trained using sequence 09_26_drive_0015.
[†] This residential trial was evaluated with a model trained on a sequence from the city category because of several moving vehicles that were better represented in that training dataset.

Figure 3.7: Three types of environments in the KITTI dataset, as well as 2 types of environments at the University of Toronto. We use one trial from each category to train and then evaluate separate trials in the same category.

Figure 3.8: Our four-wheeled skid-steered Clearpath Husky rover equipped with Skybotix VI-Sensor and Ashtech DGPS antenna used to collect the outdoor UTIAS dataset.

3.5 Generalized Kernels

However, not all features are created equal; most feature-based methods rely on random sample consensus algorithms (Fischler and Bolles, 1981a) to partition the extracted features into inliers and outliers, and perform estimation based only on inliers. It is common to guard against misclassifying an outlier as an inlier by using robust estimation techniques, such as the Cauchy costs employed in Kerl et al. (2013) or the dynamic covariance scaling devised by Agarwal et al. (2013). These approaches, often grouped under the title of M-estimation, aim to maintain a quadratic influence of small errors, while reducing the contribution of larger errors. The robustness and accuracy of feature-based visual odometry often hinges on the tuning of the parameters of inlier selection and robust estimation. Performance can vary significantly from one environment to the next, and most algorithms require careful tuning to work in a given environment.

In this work, we describe a principled, data-driven way to build a noise model for visual odometry. We combine our previous work (Peretroukhin et al., 2015b) on predictive robust estimation (PROBE) with our work on covariance estimation (Vega-Brown and Roy, 2013) to formulate a predictive robust estimator for a stereo visual odometry pipeline. We frame the traditional non-linear least squares optimization problem as a problem of maximum likelihood estimation with a Gaussian noise model, and infer a distribution over the covariance matrix of the Gaussian noise from a predictive model learned from training data. This results in a Student's t distribution over the noise, and naturally yields a robust nonlinear least-squares optimization problem. In this way, we can predict, in a principled manner, how informative each visual feature is with respect to the final state estimate, which allows our approach to intelligently weight observations to produce more accurate odometry estimates. Our pipeline is outlined in Figure **??**.

3.5.1 Predictive noise models for visual odometry

In order to exploit conjugacy to a Gaussian noise model, we formulate our prior knowledge about this function using an inverse Wishart (IW) distribution over positive definite $d \times d$ matrices (the IW distribution has been used as a prior on covariance matrices in other robotics and computer vision contexts, see for example, (Fitzgibbon et al., 2007)). This distribution is defined by a scale matrix $\mathbf{\Psi} \in \mathbb{R}^{d \times d} \succ 0$ and a scalar quantity called the degrees of freedom

$\nu \in \mathbb{R} > d-1$:

$$p(\mathbf{R}) = \text{IW}(\mathbf{R}; \mathbf{\Psi}, \nu) \tag{3.9}$$
$$= \frac{|\mathbf{\Psi}|^{\nu/2}}{2^{\frac{\nu d}{2}} \Gamma_d(\frac{\nu}{2})} |\mathbf{R}|^{-\frac{\nu+d+1}{2}} \exp\left(-\frac{1}{2}\text{tr}\left(\mathbf{\Psi}\mathbf{R}^{-1}\right)\right).$$

We use the scale matrix to encode our prior estimate of the covariance, and the degrees of freedom to encode our confidence in that estimate. Specifically, if we estimate the covariance $\mathbf{R}$ associated with predictor $\boldsymbol{\phi}$ to be $\hat{\mathbf{R}}$ with a confidence equivalent to seeing n independent samples of the error from $\mathcal{N}(\mathbf{0}, \hat{\mathbf{R}})$, we would choose $\nu(\boldsymbol{\phi}) = n$ and $\mathbf{\Psi}(\boldsymbol{\phi}) = n\hat{\mathbf{R}}$.

Given a sequence of observations and ground truth transformations,

$$\mathcal{D} = \{\mathcal{I}_t, \mathbf{T}_t\}, \quad t \in [1, N] \tag{3.10}$$

where

$$\mathcal{I}_t = \{\mathbf{y}_{i,t}, \mathbf{y}'_{i,t}, \boldsymbol{\phi}_{i,t}\} \quad i \in [1, N_t], \tag{3.11}$$

we can use the procedure of generalized kernel estimation (Vega-Brown et al., 2014) to infer a posterior distribution over the covariance matrix $\mathbf{R}_*$ associated with some query predictor vector $\boldsymbol{\phi}_*$:

$$p(\mathbf{R}_*|\mathcal{D}, \boldsymbol{\phi}_*) \propto \prod_{i,t} \mathcal{N}(\mathbf{e}_{i,t}|\mathbf{0}, \mathbf{R}_*)^{k(\boldsymbol{\phi}_*, \boldsymbol{\phi}_{i,t})}$$
$$\times \text{IW}(\mathbf{R}_*; \mathbf{\Psi}(\boldsymbol{\phi}_*), \nu(\boldsymbol{\phi}_*)) \tag{3.12}$$
$$= \text{IW}(\mathbf{R}_*; \mathbf{\Psi}_*, \nu_*). \tag{3.13}$$

Here, $\mathbf{e}_{i,t} = \mathbf{y}'_{i,t} - \mathbf{f}(\mathbf{T}_t \mathbf{f}^{-1}(\mathbf{y}_{i,t}))$ as before. The function $k : \mathbb{R}^M \times \mathbb{R}^M \to [0, 1]$ is a kernel function which measures the similarity of two points in predictor space. Note also that the posterior parameters $\mathbf{\Psi}_*$ and ν_* can be computed in closed form as

$$\mathbf{\Psi}_* = \mathbf{\Psi}(\boldsymbol{\phi}_*) + \sum_{i,t} k(\boldsymbol{\phi}_*, \boldsymbol{\phi}_{i,t}) \mathbf{e}_{i,t}\, \mathbf{e}_{i,t}{}^T, \tag{3.14}$$

$$\nu_* = \nu(\boldsymbol{\phi}_*) + \sum_{i,t} k(\boldsymbol{\phi}_*, \boldsymbol{\phi}_{i,t}). \tag{3.15}$$

If we marginalize over the covariance matrix, we find that the posterior predictive distri-

bution is a multivariate Student's t distribution:

$$p(\mathbf{y}'_{i,t}|\mathbf{T}_t, \mathbf{y}_{i,t}, \mathcal{D}, \boldsymbol{\phi}_{i,t}) \tag{3.16}$$

$$= \int \mathrm{d}\mathbf{R}_{i,t} \mathcal{N}\left(\mathbf{e}_{i,t}; \mathbf{0}, \mathbf{R}_{i,t}\right) \mathrm{IW}(\mathbf{R}_{i,t}; \boldsymbol{\Psi}_*, \nu_*) \tag{3.17}$$

$$= \mathrm{t}_{\nu_*-d+1}\left(\mathbf{e}_{i,t}; \mathbf{0}, \frac{1}{\nu_* - d + 1}\boldsymbol{\Psi}_*\right) \tag{3.18}$$

$$= \frac{\Gamma\left(\frac{\nu_*+1}{2}\right)}{\Gamma\left(\frac{\nu_*-d+1}{2}\right)} |\boldsymbol{\Psi}_*|^{-\frac{1}{2}} \pi^{-\frac{d}{2}} \left(1 + \mathbf{e}_{i,t}{}^T \boldsymbol{\Psi}_*{}^{-1} \mathbf{e}_{i,t}\right)^{-\frac{\nu_*+1}{2}}. \tag{3.19}$$

Given a new landmark and predictor vector, we can infer a noise model by evaluating eqs. (3.14) and (3.15). In order to accelerate this computation, it is helpful to choose a kernel function with finite support: that is, $k(\boldsymbol{\phi}, \boldsymbol{\phi}') = 0$ if $\|\boldsymbol{\phi} - \boldsymbol{\phi}'\|_2 > \rho$. Then, by indexing our training data in a spatial index such as a k-d tree, we can identify the subset of samples relevant to evaluating the sums in eqs. (3.14) and (3.15) in $\mathcal{O}(\log N + \log N_t)$ time. Algorithm 3 describes the procedure for building this model.

Algorithm 3 Build the covariance model given a sequence of observations, $\mathcal{D}$.

```
function BuildCovarianceModel(D)
    Initialize an empty spatial index M
    for all I_t, T_t in D do
        for all {y_{i,t}, y'_{i,t}, φ_{i,t}} in I_t do
            e_{i,t} = y'_{i,t} − f(T_t f^{-1}(y_{i,t}))
            Insert φ_{i,t} into M and store e_{i,t} at its location
        end for
    end for
    return M
end function
```

Once we have inferred a noise model for each landmark in a new image pair, the maximum likelihood optimization problem is given by

$$\mathbf{T}_t^* = \underset{\mathbf{T}_t \in \mathrm{SE}(3)}{\operatorname{argmin}} \sum_{i=1}^{N_t} (\nu_{i,t} + 1) \log\left(1 + \mathbf{e}_{i,t}{}^T \boldsymbol{\Psi}_{i,t}^{-1} \mathbf{e}_{i,t}\right). \tag{3.20}$$

The final optimization problem thus emerges as a nonlinear least squares problem with a rescaled Cauchy-like loss function, with error term $\mathbf{e}_{i,t}{}^T \left(\frac{1}{\nu_{i,t}+1}\boldsymbol{\Psi}_{i,t}\right)^{-1} \mathbf{e}_{i,t}$ and outlier scale $\nu_{i,t} + 1$. This is a common robust loss function which is approximately quadratic in the reprojection error for $\mathbf{e}_{i,t}{}^T \boldsymbol{\Psi}_{i,t}^{-1} \mathbf{e}_{i,t} \ll \nu_{i,t} + 1$, but grows only logarithmically for $\mathbf{e}_{i,t}{}^T \boldsymbol{\Psi}_{i,t}^{-1} \mathbf{e}_{i,t} \gg \nu_{i,t} + 1$. It follows that in the limit of large $\nu_{i,t}$—in regions of predictor space where there are

many relevant samples—our optimization problem becomes the original least-squares optimization problem.

Solving nonlinear optimization problems with the form of Equation (3.20) is a well-studied and well-understood task, and software packages to perform this computation are readily available. Algorithm 4 describes the procedure for computing the transform between a new image pair, treating the optimization of Equation (3.20) as a subroutine.

Algorithm 4 Compute the transform between two images, given a set, $\mathcal{I}_t$, of landmarks and predictors extracted from an image pair and a covariance model $\mathcal{M}$.

function COMPUTETRANSFORM($\mathcal{I}_t$, $\mathcal{M}$)
  for all $\{\mathbf{y}_{i,t}, \mathbf{y}'_{i,t}, \boldsymbol{\phi}_{i,t}\}$ in $\mathcal{I}_t$ **do**
    $\boldsymbol{\Psi}, \nu \leftarrow$ INFERNOISEMODEL($\mathcal{M}$, $\boldsymbol{\phi}_{i,t}$)
    $g(\mathbf{T}) = \mathbf{y}_{i,t} - \mathbf{f}(\mathbf{T}\mathbf{f}^{-1}(\mathbf{y}'_{i,t}))$
    $\mathcal{L} \leftarrow \mathcal{L} + (\nu + 1) \log \left(1 + g(\mathbf{T})^T \boldsymbol{\Psi}^{-1} g(\mathbf{T})\right)$
  end for
  return $\operatorname{argmin}_{\mathbf{T} \in \mathrm{SE}(3)} \mathcal{L}(\mathbf{T})$
end function
function INFERNOISEMODEL($\mathcal{M}$, $\boldsymbol{\phi}_*$)
  NEIGHBORS $\leftarrow$ GETNEIGHBORS($\mathcal{M}$, $\boldsymbol{\phi}_*$, ρ)
    ▷ ρ is the radius of the support of the kernel k
  $\boldsymbol{\Psi}_* \leftarrow \boldsymbol{\Psi}(\boldsymbol{\phi}_*)$
  $\nu_* \leftarrow \nu(\boldsymbol{\phi}_*)$
  for $(\boldsymbol{\phi}_{i,t}, \mathbf{e}_{i,t})$ in NEIGHBORS **do**
    $\boldsymbol{\Psi}_* \leftarrow \boldsymbol{\Psi}_* + k(\boldsymbol{\phi}_*, \boldsymbol{\phi}_{i,t}) \mathbf{e}_{i,t}\, \mathbf{e}_{i,t}{}^T$
    $\nu_* \leftarrow \nu_* + k(\boldsymbol{\phi}_*, \boldsymbol{\phi}_{i,t})$
  end for
  return $\boldsymbol{\Psi}_*, \nu_*$
end function

We observe that Algorithm 4 is predictively robust, in the sense that it uses past experiences not just to predict the reliability of a given image landmark, but also to introspect and estimate its own knowledge of that reliability. Landmarks which are not known to be reliable are trusted less than landmarks which look like those which have been observed previously, where "looks like" is defined by our prediction space and choice of kernel.

3.5.2 Inference without ground truth

Algorithm 3 requires access to the true transform between training image pairs. In practice, such ground truth data may be difficult to obtain. In these cases, we can instead formulate a likelihood model $p(\mathcal{D}'|\mathbf{T}_1, \ldots, \mathbf{T}_t)$, where $\mathcal{D}' = \{\mathcal{I}_t\}$ is a dataset consisting only of landmarks

and predictors for each training image pair. We can construct a model for future queries by inferring the most likely sequence of transforms for our training images. The likelihood has the following factorized form:

$$p(\mathcal{D}'|\mathbf{T}_{1:T}) \propto \int \prod_{i,t} \mathrm{d}\mathbf{R}_{i,t}\, p(\mathbf{y}'_{i,t}|\mathbf{y}_{i,t}, \mathbf{T}_t, \mathbf{R}_{i,t}) p(\mathbf{R}_{i,t}|\boldsymbol{\phi}_{i,t}, \mathcal{D}, \mathbf{T}_{1:T}). \tag{3.21}$$

We cannot easily maximize this likelihood, since marginalizing over the noise covariances removes the independence of the transforms between each image pair. To render the optimization tractable, we follow previous work (Vega-Brown and Roy, 2013) and formulate an iterative expectation-maximization (EM) procedure. Given an estimate $\mathbf{T}_t^{(n)}$ of the transforms, we can compute the expected log-likelihood conditioned on our current estimate:

$$Q(\mathbf{T}_{1:T}|\mathbf{T}_{1:T}^{(n)}) = \int \left(\prod_{i,t} \mathrm{d}\mathbf{R}_{i,t}\, p(\mathbf{R}_{i,t}|\mathcal{D}_{\setminus i,t}, \mathbf{T}_{1:T}^{(n)}) \right) \log \prod_{i,t} p(\mathbf{y}'_{i,t}|\mathbf{y}_{i,t}, \mathbf{T}_t, \mathbf{R}_{i,t}). \tag{3.22}$$

This has the effect of rendering the likelihood of each transform to be estimated independently. Moreover, the expected log-likelihood can be evaluated in closed form:

$$Q(\mathbf{T}_{1:T}|\mathbf{T}_{1:T}^{(n)}) \cong -\frac{1}{2} \sum_{t=1}^{T} \sum_{i=1}^{N_t} \mathbf{e}_{i,t}{}^T \left(\frac{1}{\nu_{i,t}^{(n)}} \boldsymbol{\Psi}_{i,t}^{(n)} \right)^{-1} \mathbf{e}_{i,t}. \tag{3.23}$$

The symbol $\cong$ is used to indicate equality up to an additive constant. A derivation of this observation can be found in our supplemental material.

We can iteratively refine our estimate by maximizing the expected log-likelihood

$$\mathbf{T}_{1:T}^{(n+1)} = \underset{\mathbf{T}_{1:T} \in \mathrm{SE}(3)^T}{\operatorname{argmax}} Q(\mathbf{T}_{1:T}|\mathbf{T}_{1:T}^{(n)}). \tag{3.24}$$

Due to the additive structure of $Q(\mathbf{T}_{1:T}|\mathbf{T}_{1:T}^{(n)})$, this takes the form of T separate nonlinear least-squares optimizations:

$$\mathbf{T}_t^{(n+1)} = \underset{\mathbf{T}_t \in \mathrm{SE}(3)}{\operatorname{argmin}} \sum_{i=1}^{N_t} \mathbf{e}_{i,t}{}^T \left(\frac{1}{\nu_{i,t}^{(n)}} \boldsymbol{\Psi}_{i,t}^{(n)} \right)^{-1} \mathbf{e}_{i,t}. \tag{3.25}$$

Algorithm 5 describes the process of training a model without ground truth. We refer to this process as PROBE-GK-EM, and distinguish it from PROBE-GK-GT (Ground Truth). We note that the sequence of estimated transforms, $\mathbf{T}_{1:T}^{(n)}$, is guaranteed to converge to a local maxima of the likelihood function (Dempster et al., 1977). It is also possible to use a robust

loss function (Equation (3.20)) in place of Equation (3.25) during EM training. Although not formally motivated by the derivation above, this approach often leads to lower test errors in practice. Characterizing when and why this robust learning process outperforms its non-robust alternative is part of ongoing work.

Algorithm 5 Build the covariance model without ground truth given a sequence of observations, $\mathcal{D}'$, and an initial odometry estimate $\mathbf{T}_{1:T}^{(0)}$.

function BUILDCOVARIANCEMODEL($\mathcal{D}'$, $\mathbf{T}_{1:T}^{(0)}$)
  Initialize an empty spatial index $\mathcal{M}$
  for all $\mathcal{I}_t$ in $\mathcal{D}'$ **do**
    for all $\{\mathbf{y}_{i,t}, \mathbf{y}'_{i,t}, \boldsymbol{\phi}_{i,t}\}$ in $\mathcal{I}_t$ **do**
      $\mathbf{e}_{i,t} = \mathbf{y}_{i,t} - \mathbf{f}(\mathbf{T}_t^{(0)}\mathbf{f}^{-1}(\mathbf{y}'_{i,t}))$
      Insert $\boldsymbol{\phi}_{i,t}$ into $\mathcal{M}$ and store $\mathbf{e}_{i,t}$ at its location
    end for
  end for
  repeat
    for all $\mathcal{I}_t$ in $\mathcal{D}'$ **do**
      for all $\{\mathbf{y}_{i,t}, \mathbf{y}'_{i,t}, \boldsymbol{\phi}_{i,t}\}$ in $\mathcal{I}_t$ **do**
        $\boldsymbol{\Psi}, \nu \leftarrow$ INFERNOISEMODEL($\mathcal{M}, \boldsymbol{\phi}_{i,t}$)
        $g(\mathbf{T}) = \mathbf{y}_{i,t} - \mathbf{f}(\mathbf{T}\mathbf{f}^{-1}(\mathbf{y}'_{i,t}))$
        $\mathcal{L} \leftarrow \mathcal{L} + g(\mathbf{T})^T \left(\frac{1}{\nu}\boldsymbol{\Psi}\right)^{-1} g(\mathbf{T})$
      end for
      $\mathbf{T}_t \leftarrow \operatorname{argmin}_{\mathbf{T}\in\mathrm{SE}(3)} \mathcal{L}(\mathbf{T})$
      $\mathbf{e}_{i,t} = \mathbf{y}_{i,t} - \mathbf{f}(\mathbf{T}_t^{(0)}\mathbf{f}^{-1}(\mathbf{y}'_{i,t}))$
      Update the error stored at $\boldsymbol{\phi}_{i,t}$ in $\mathcal{M}$ to $\mathbf{e}_{i,t}$
    end for
  until converged
  return $\mathcal{M}$
end function

3.5.3 Experiments

Synthetic

Next, we formulated a synthetic dataset wherein a stereo camera traverses a circular path observing 2000 randomly distributed point features. We added Gaussian noise to each of the ideal projected pixel co-ordinates for visible landmarks at every step. We varied the noise variance as a function of the vertical pixel coordinate of the feature in image space. In addition, a small subset of the landmarks received an error term drawn from a uniform distribution to simulate the presence of outliers. The prediction space was composed of the vertical and

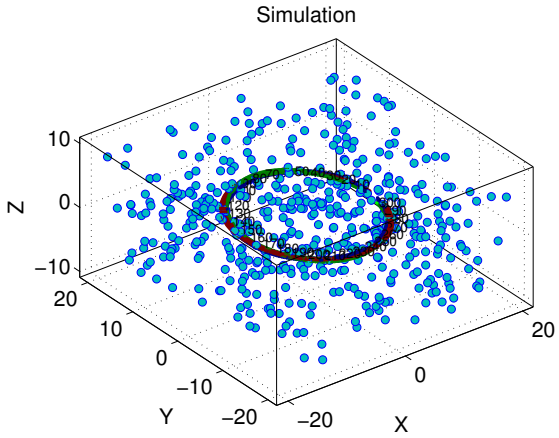


Figure 3.9: Our synthetic world. A stereo camera rig moves through a world with 2000 point features.

horizontal pixel locations in each of the stereo cameras.

We simulated independent training and test traversals, where the camera moved for 30 and 60 seconds respectively (at a forward speed of 3 metres per second for final path lengths of 90 and 180 meters). Figure 3.10 and Table 3.2 document the qualitative and quantitative comparisons of PROBE-GK (trained with and without ground-truth) against two baseline stereo odometry frameworks. Both baseline estimators were implemented based on **??**. The first utilized fixed covariances for all reprojection errors, while the second used a modified robust cost (i.e. M-estimation) based on Student's t weighting, with $\nu = 5$ (as suggested in Kerl et al. (2013)). These benchmarks served as baseline estimators (with and without robust costs) that used fixed covariance matrices and did not include a predictive component.

Using PROBE-GK with ground truth data for training, we significantly reduced both the translation and rotational Average Root Mean Squared Error (ARMSE) by approximately 50%. In our synthetic data, the Expectation Maximization approach was able to achieve nearly identical results to the ground-truth-aided model within 5 iterations.

3.5.4 KITTI

To evaluate PROBE-GK on real environments, we trained and tested several models on the KITTI Vision Benchmark suite (Geiger et al., 2012, 2013), a series of datasets collected by a car outfitted with a number of sensors driven around different parts of Karlsruhe, Germany. Within the dataset, ground truth pose information is provided by a high grade inertial navigation unit which also fuses measurements from differential GPS. Raw data is available for different types of environments through which the car was driving; for our work, we focused on the city, residential and road categories (Figure 3.11). From each category, we chose two

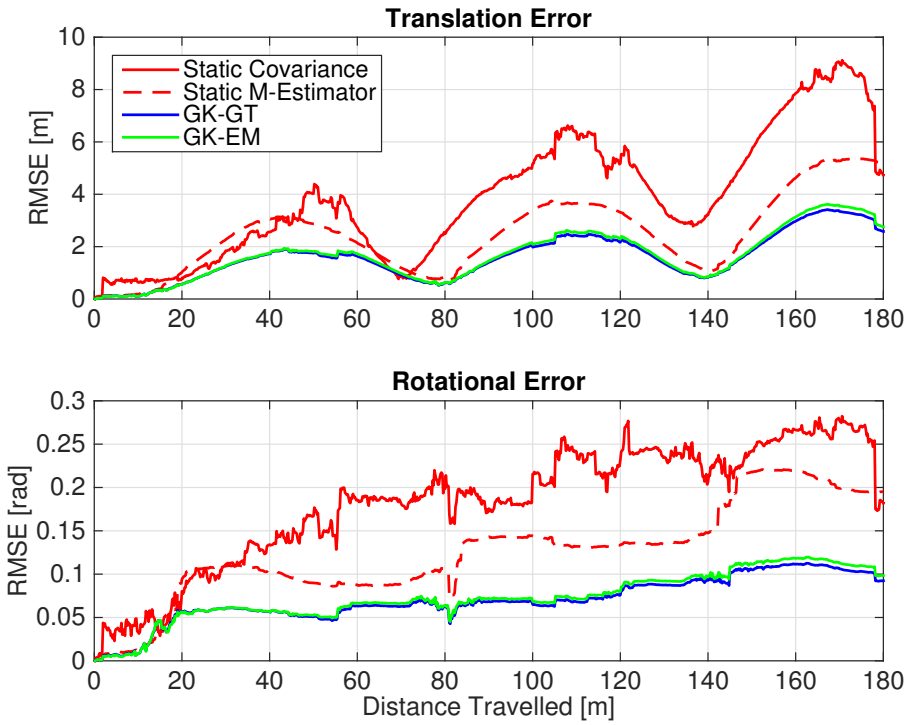


Figure 3.10: A comparison of translational and rotational Root Mean Square Error on simulated data (RMSE) for four different stereo-visual odometry pipelines: two baseline bundle adjustment procedures with and without a robust Student's t cost with a fixed and hand-tuned covariance and degrees of freedom (M-Estimation), a robust bundle adjustment with covariances learned from ground truth with algorithm 3 (GK-GT), and a robust bundle adjustment using covariances learned without ground truth using expectation maximization, with algorithm 5 (GK-EM). Note in this experiment, the RMSE curves for GK-GT and GK-EM very nearly overlap. The overall translational and rotational ARMSE values are shown in Table 3.2.

Figure 3.11: The KITTI dataset contains three different environments. We validate PROBE-GK by training on each type and testing against a baseline stereo visual odometry pipeline.

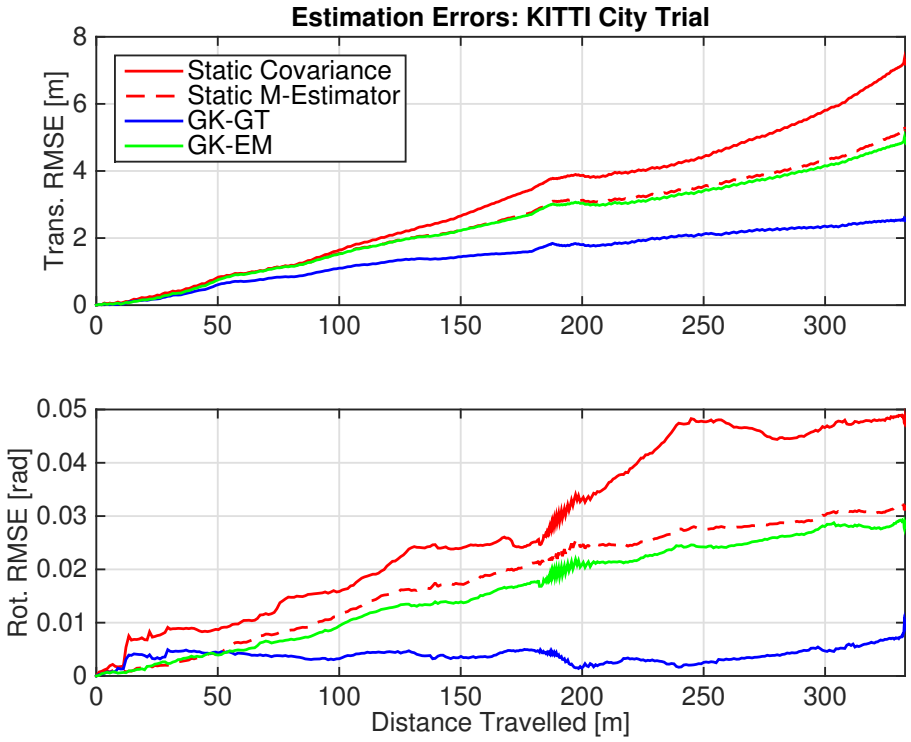


Figure 3.12: RMSE comparison of stereo odometry estimators evaluated on data from the city category in the KITTI dataset. See Table 3.2 for a quantitative summary.

separate trials for training and testing.

Our prediction space consisted of inertial magnitudes, high and low image frequency coefficients, image entropy, pixel location, and estimated transform parameters. The choice of predictors is motivated by the types of effects we wish to capture (in this case: grassy self-similar textures, as well as shadows, and motion blur). For a more detailed explanation of our choice of prediction space, see our previous work (Peretroukhin et al., 2015b).

Figures 3.12 to 3.14 show typical results; **??** presents a quantitative comparison. PROBE GK-GT produced significant reductions in ARMSE, reducing translational ARMSE by as much as 80%. In contrast, GK-EM showed more modest improvements; this is unlike our synthetic experiments, where both GK-EM and GK-GT achieved similar performance. We are still actively exploring why this is the case; we note that although our simulated data is drawn from a mixture of Gaussian distributions, the underlying noise distribution for real data may be far more complex. With no ground truth, EM has to jointly optimize the camera poses and sensor uncertainty. It is unclear whether this is feasible in the general case with no ground truth information.

Further, we observe that the performance of PROBE-GK depends on the similarity of the training data to the final test trials. A characteristic training dataset was important for consistent improvements on test trials.

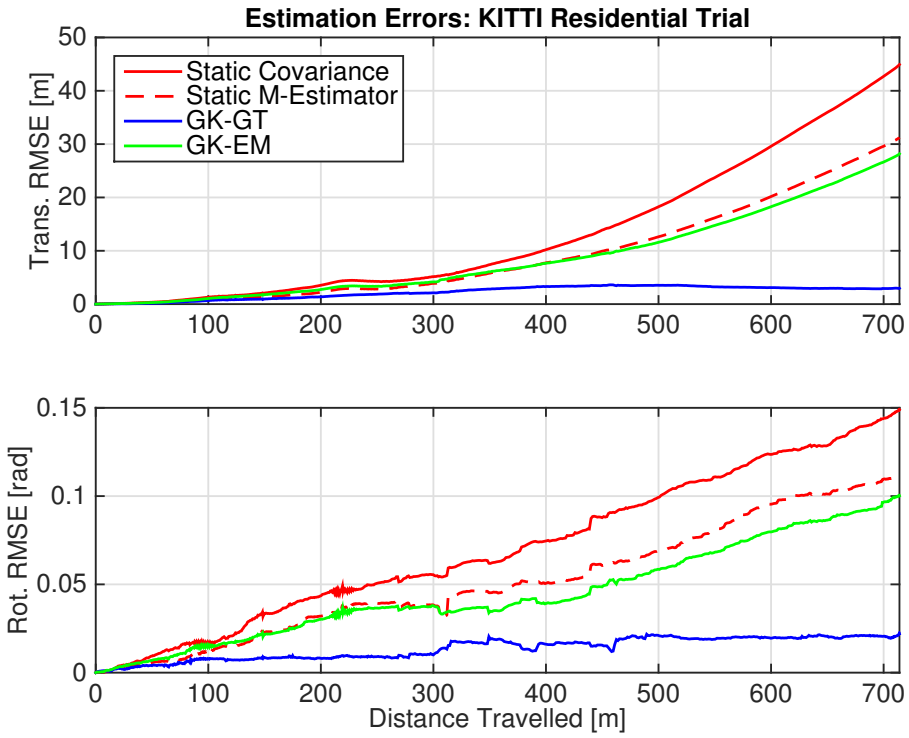


Figure 3.13: RMSE comparison of stereo odometry estimators evaluated on data from the residential category in the KITTI dataset.

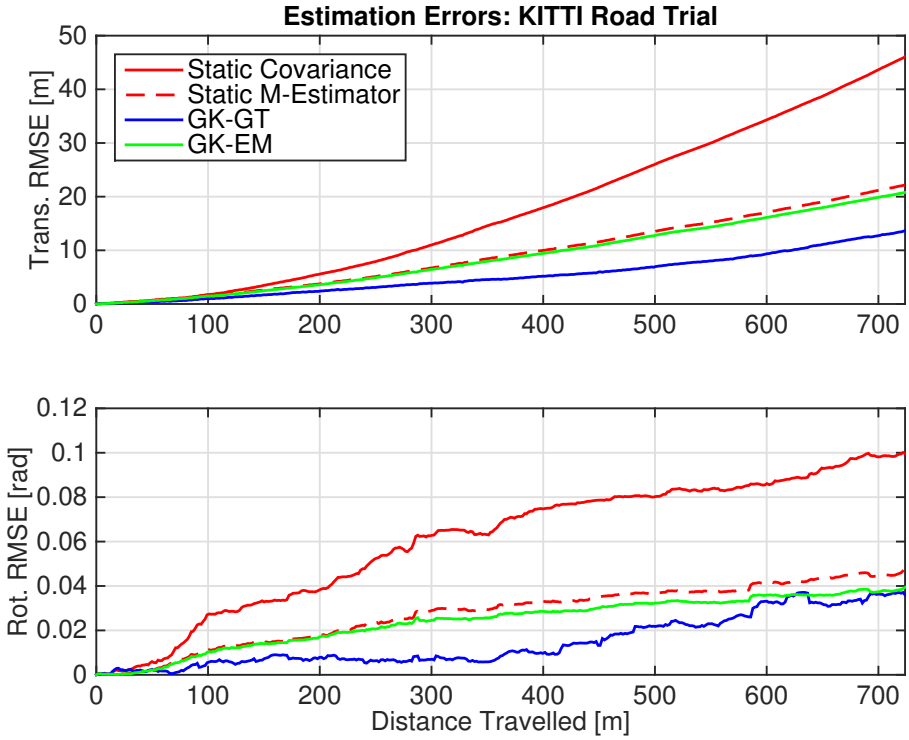


Figure 3.14: RMSE comparison of stereo odometry estimators evaluated on data from the road category in the KITTI dataset.

Table 3.2: Comparison of average root mean squared errors (ARMSE) for rotational and translational components. Each trial is trained and tested from a particular category of raw data from the synthetic and KITTI datasets.

		Trans. ARMSE [m]				Rot. ARMSE [rad]			
	Length [m]	Fixed Covar.	Static M-Estimator	GK-GT	GK-EM	Fixed Covar.	Static M-Estimator	GK-GT	GK-EM
Synthetic	180	3.87	2.49	1.59	1.66	0.18	0.13	0.070	0.073
City	332.9	3.84	2.99	1.69	2.87	0.032	0.021	0.0046	0.018
Residential	714.1	13.48	9.37	1.97	8.80	0.068	0.050	0.013	0.044
Road	723.8	17.69	9.38	5.24	8.87	0.060	0.027	0.015	0.024

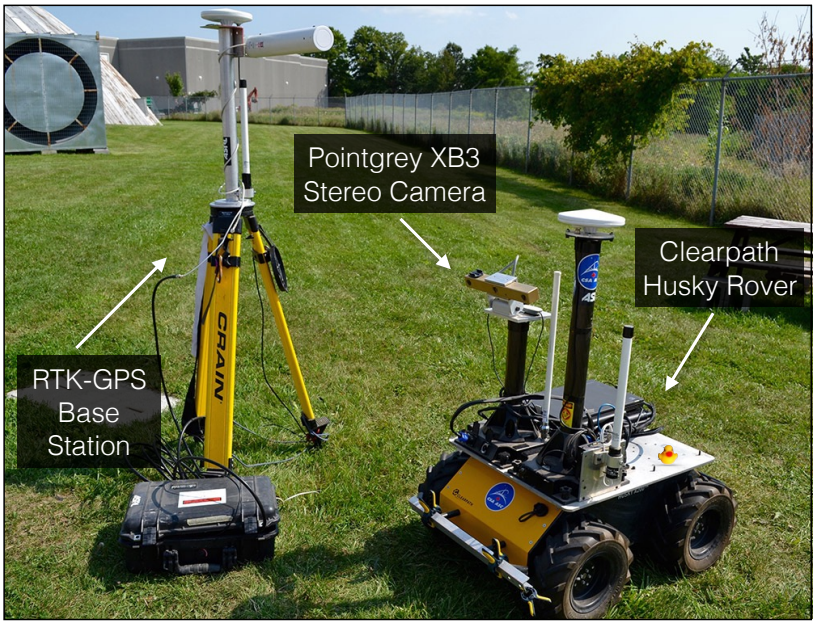


Figure 3.15: Our experimental apparatus: a Clearpath Husky rover outfitted with a PointGrey XB3 stereo camera and a differential GPS receiver and base station.

UTIAS

To further investigate the capability of our EM approach, we evaluated PROBE-GK on experimental data collected at the University of Toronto Institute for Aerospace Studies (UTIAS). For this experiment, we drove a Clearpath Husky rover outfitted with an Ashtech DG14 Differential GPS, and a PointGrey XB3 stereo camera around the MarsDome (an indoor Mars analog testing environment) at UTIAS (Figure 3.15) for five trials of a similar path. Each trial was approximately 250 m in length and we made an effort to align the start and end points of each loop. We used the wide baseline (25 cm) of the XB3 stereo camera to record the stereo images. The approximate trajectory for all 5 trials, as recorded by GPS, is shown in Figure 3.16. Note that the GPS data was not used during training, and only recorded for reference.

For the prediction space in our experiments, we mimicked the KITTI experiments, omitting inertial magnitudes as no inertial data was available. We trained PROBE-GK without ground truth, using the Expectation Maximization approach. Figure 3.17 shows the likeli-

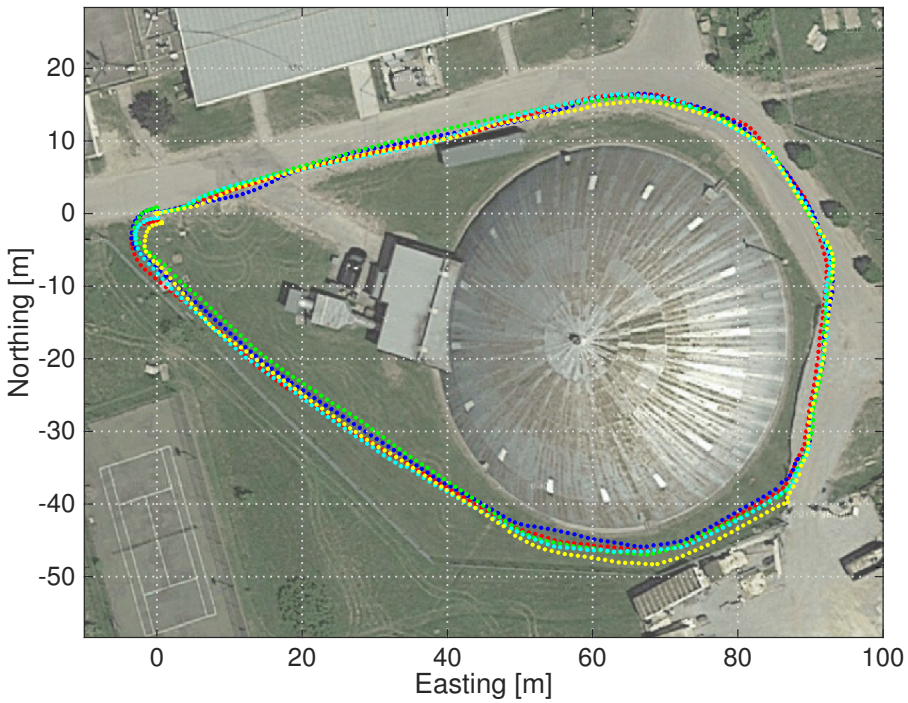


Figure 3.16: GPS ground truth for 5 experimental trials collected near the UTIAS Mars Dome. Each trial is approximately 250 m long.

Table 3.3: Comparison of loop closure errors for 4 different experimental trials with and without a learned PROBE-GK-EM model.

		Loop Closure Error [m]	
Trial	Path Length [m]	PROBE-GK-EM	Static M-Estimator
2	250.3	3.88	8.07
3	250.5	3.07	6.64
4	205.4	2.81	7.57
5	249.9	2.34	7.75

hood and loop closure error as a function of EM iteration.

The EM approach indeed produced significant error reductions on the training dataset after just a few iterations. Although it was trained with no ground truth information, our PROBE-GK model was used to produce significant reductions in the loop closure errors of the remaining 4 test trials. This reinforced our earlier hypothesis: the EM method works well when the training trajectory more closely resembles the test trials (as was the case in this experiment). Table 3.3 lists the statistics for each test.

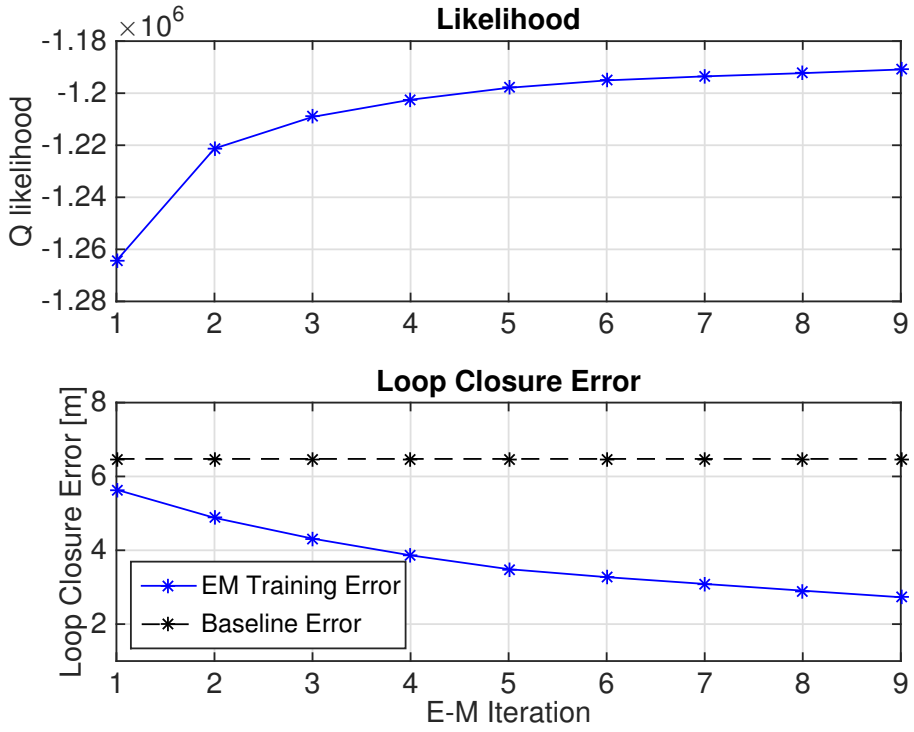


Figure 3.17: Training without ground truth using PROBE-GK-EM on a 250.2m path around the Mars Dome at UTIAS. The likelihood of the data increases with each iteration, and the loop closure error decreases, improving significantly from a baseline static M-estimator.

3.6 Summary

Predictive Robust Estimation applied two different techniques (scalar weighted covariances and the method of generalized kernel estimation) to improve on the uncorrelated and static Gaussian error models typically employed in stereo odometry. PROBE and its follow up PROBE-GK, contributed

1. a probabilistic model for indirect stereo visual odometry, leading to a predictive robust algorithm for inference on that model,

2. two different approaches to constructing the robust algorithm: one based on k-nearest neighbours, and one based on Generalized Kernel (GK) estimation,

3. a procedure for training our model using pairs of stereo images with known relative transforms, and

4. an iterative, expectation-maximization approach to train our GK model when the relative ground truth egomotion was unavailable.

Appendices

Bibliography

Agarwal, P., Tipaldi, G. D., Spinello, L., Stachniss, C., and Burgard, W. (2013). Robust map optimization using dynamic covariance scaling. In *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, pages 62–69.

Altmann, S. L. (1989). Hamilton, rodrigues, and the quaternion scandal. *Math. Mag.*, 62(5):291–308.

Barfoot, T. D. (2017). *State Estimation for Robotics*. Cambridge University Press.

Barfoot, T. D. and Furgale, P. T. (2014). Associating uncertainty with three-dimensional poses for use in estimation problems. *IEEE Trans. Rob.*, 30(3):679–693.

Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I., and Leonard, J. J. (2016). Past, present, and future of simultaneous localization and mapping: Toward the Robust-Perception age. *IEEE Trans. Rob.*, 32(6):1309–1332.

Crete, F., Dolmiere, T., Ladret, P., and Nicolas, M. (2007). The blur effect: perception and estimation with a new no-reference perceptual blur metric. In *Human vision and electronic imaging XII*, volume 6492, page 64920I. International Society for Optics and Photonics.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38.

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

Fischler, M. A. and Bolles, R. C. (1981b). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395.

Fitzgibbon, A. W., Robertson, D. P., Criminisi, A., Ramalingam, S., and Blake, A. (2007). Learning priors for calibrating families of stereo cameras. In *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, pages 1–8.

Florez, S. A. R. (2010). *Contributions by vision systems to multi-sensor object localization and tracking for intelligent vehicles*. PhD thesis.

Forster, C., Pizzoli, M., and Scaramuzza, D. (2014). SVO: Fast semi-direct monocular visual odometry. In *Proc. IEEE Int. Conf. Robot. Automat.(ICRA)*, pages 15–22. IEEE.

Furgale, P. (2011). *Extensions to the Visual Odometry Pipeline for the Exploration of Planetary Surfaces*. PhD thesis.

Furgale, P., Rehder, J., and Siegwart, R. (2013). Unified temporal and spatial calibration for multi-sensor systems. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1280–1286.

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: The KITTI dataset. *Int. Journal Robot. Research (IJRR)*.

Geiger, A., Lenz, P., and Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Geiger, A., Ziegler, J., and Stiller, C. (2011a). StereoScan: Dense 3D reconstruction in real-time. In *Proc. Intelligent Vehicles Symp. (IV)*, pages 963–968. IEEE.

Geiger, A., Ziegler, J., and Stiller, C. (2011b). StereoScan: Dense 3D reconstruction in real-time. In *Proc. IEEE Intelligent Vehicles Symp. (IV)*, pages 963–968.

Geman, S., McClure, D. E., and Geman, D. (1992). A nonlinear filter for film restoration and other problems in image processing. *CVGIP: Graphical models and image processing*, 54(4):281–289.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, pages 73–101.

Irani, M. and Anandan, P. (2000). About direct methods. In *Vision Algorithms: Theory and Practice*, pages 267–277. Springer.

Kerl, C., Sturm, J., and Cremers, D. (2013). Robust odometry estimation for RGB-D cameras. In *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, pages 3748–3754.

Lucas, B. D. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'81, pages 674–679, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

MacTavish, K. and Barfoot, T. D. (2015). At all costs: A comparison of robust cost functions for camera correspondence outliers. In *Proc. Conf. on Comp. and Robot Vision (CRV)*, pages 62–69.

Maimone, M., Cheng, Y., and Matthies, L. (2007). Two years of visual odometry on the mars exploration rovers. *J. Field Robot.*, 24(3):169–186.

Peretroukhin, V., Clement, L., Giamou, M., and Kelly, J. (2015a). PROBE: Predictive robust estimation for visual-inertial navigation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'15)*, pages 3668–3675, Hamburg, Germany.

Peretroukhin, V., Clement, L., Giamou, M., and Kelly, J. (2015b). PROBE: Predictive robust estimation for visual-inertial navigation. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, pages 3668–3675.

Peretroukhin, V., Clement, L., and Kelly, J. (2015c). Get to the point: Active covariance scaling for feature tracking through motion blur. In *Proceedings of the IEEE International Conference on Robotics and Automation Workshop on Scaling Up Active Perception*, Seattle, Washington, USA.

Peretroukhin, V., Vega-Brown, W., Roy, N., and Kelly, J. (2016). PROBE-GK: Predictive robust estimation using generalized kernels. In *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, pages 817–824.

Scaramuzza, D. and Fraundorfer, F. (2011a). Visual odometry [tutorial]. *IEEE Robot. Autom. Mag.*, 18(4):80–92.

Scaramuzza, D. and Fraundorfer, F. (2011b). Visual odometry [tutorial]. *IEEE Robot. Automat. Mag.*, 18(4):80–92.

Sola, J. (2017). Quaternion kinematics for the error-state kalman filter. *arXiv preprint arXiv:1711.02508*.

Solà, J., Deray, J., and Atchuthan, D. (2018). A micro lie theory for state estimation in robotics.

Sünderhauf, N. and Protzel, P. (2007). Stereo odometry: a review of approaches. *Chemnitz University of Technology Technical Report*.

Tsotsos, K., Chiuso, A., and Soatto, S. (2015). Robust inference for visual-inertial sensor fusion. In *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, pages 5203–5210.

Umeyama, S. (1991). Least-Squares estimation of transformation parameters between two point patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(4):376–380.

Vega-Brown, W., Bachrach, A., Bry, A., Kelly, J., and Roy, N. (2013). CELLO: A fast algorithm for covariance estimation. In *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, pages 3160–3167.

Vega-Brown, W. and Roy, N. (2013). CELLO-EM: Adaptive sensor models without ground truth. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, pages 1907–1914.

Vega-Brown, W. R., Doniec, M., and Roy, N. G. (2014). Nonparametric bayesian inference on multivariate exponential families. In *Advances in Neural Information Processing Systems 27*, pages 2546–2554.

Zhang, G. and Vela, P. (2015). Optimally observable and minimal cardinality monocular SLAM. In *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, pages 5211–5218.