Learned Improvements to the Visual Egomotion Pipeline

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

Learned Improvements to the Visual Egomotion Pipeline

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

The ability to estimate *egomotion* 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 information may be intermittent or unavailable. Visual egomotion estimation, also known as *visual odometry*, has become ubiquitous in mobile robotics due to the availability of high-quality, compact, and inexpensive cameras that capture rich representations of the world. To remain computationally tractable, 'classical' visual odometry pipelines make simplifying assumptions that, while permitting reliable operation in ideal conditions, often lead to systematic error. In this dissertation, we present four ways in which conventional pipelines can be improved through the addition of a learned hyper-parametric model. By combining traditional pipelines with learning, we retain the performance of conventional techniques in nominal conditions while leveraging modern high-capacity data-driven models to improve uncertainty quantification, correct for systematic bias, and improve robustness to deleterious effects by extracting latent information in existing visual data. We demonstrate the improvements derived from our approach on 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 (special thanks to Lee)
[2] especially the support and encouragement of Elyse
[3] in humouring my insatiable need for debate and banter (special thanks to Lee)

Contents

Notation

a	:	Symbols in this font are real scalars.
$\mathbf{a}$	:	Symbols in this font are real column vectors.
$\boldsymbol{a}$	:	Symbols in this font are real column vectors in homogeneous coordinates.
$\mathbf{A}$	:	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.
$\mathbf{p}_a^{cb}$	:	A vector from point b to point c (denoted by the superscript) and expressed in $\underrightarrow{\boldsymbol{\mathcal{F}}}_a$ (denoted by the subscript).
$\mathbf{C}_{ab}$	:	The 3×3 rotation matrix that transforms vectors from $\underrightarrow{\boldsymbol{\mathcal{F}}}_b$ to $\underrightarrow{\boldsymbol{\mathcal{F}}}_a$: $\mathbf{p}_a^{cb} = \mathbf{C}_{ab}\mathbf{p}_b^{cb}$.
$\mathbf{T}_{ba}$	:	The 4×4 transformation matrix that transforms homogeneous points from $\underrightarrow{\boldsymbol{\mathcal{F}}}_a$ to $\underrightarrow{\boldsymbol{\mathcal{F}}}_b$: $\boldsymbol{p}_b^{cb} = \mathbf{T}_{ba}\boldsymbol{p}_a^{ca}$.

Chapter 3

Classical Visual Odometry

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

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 two seminal tutorials on visual odometry and its more general cousin, visual SLAM, we refer the reader to Scaramuzza and Fraundorfer (2011); Cadena et al. (2016).

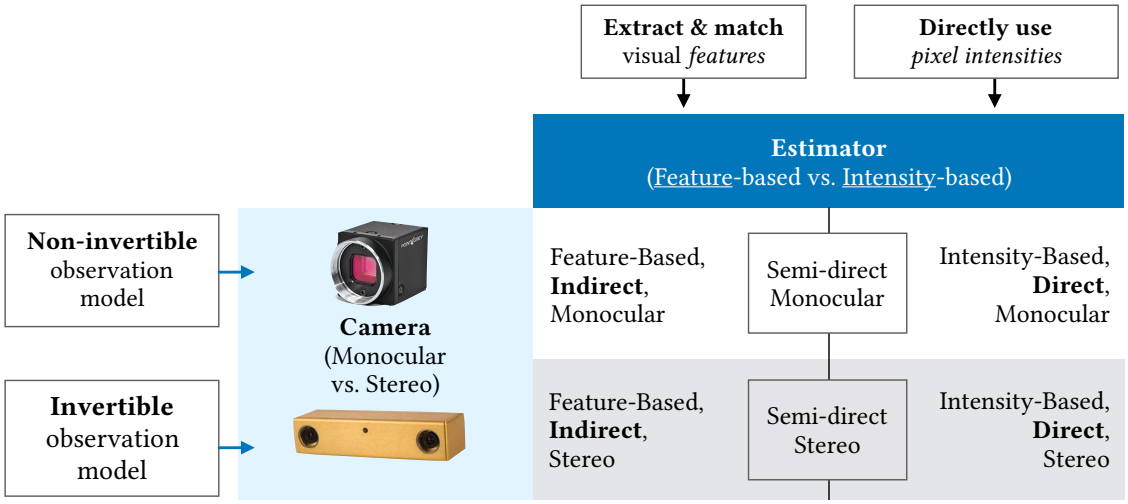


Figure 3.1: A taxonomy of different types of visual odometry.

Remark (*VO Taxonomy*). VO can be largely divided along two dimensions (Figure 3.1): (1) the type of camera used to capture images and (2) the type of data association used to compute motion estimates.

Monocular vs. Stereo Camera: Monocular VO methods (Engel et al., 2018; Tsotsos et al., 2015) 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 (Engel et al., 2018; Leutenegger et al., 2015; Cvišić and Petrović, 2015) use a stereo camera to triangulate objects with metric scale. This allows stereo VO to provide metrically-accurate 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 and by the quality of stereo matches (which can be degraded by self-similar textures, occlusions, and foreshortening effects).

Direct vs. Indirect Data Association: The second distinction is based on the type of data association used to match sequential images and infer motion. Direct methods (Engel et al., 2018; Wang et al., 2017a) make the assumption of brightness constancy, and attempt to find the egomotion estimate that *directly* maximizes the similarity of pixel intensities between images. Indirect methods (Leutenegger et al., 2015; Cvišić and Petrović, 2015), conversely, rely on image features detectors to extract a set of salient landmarks or features, and then match these landmarks across images (typically by relying on a view-invariant descriptor).

3.1 Canonical VO Pipeline

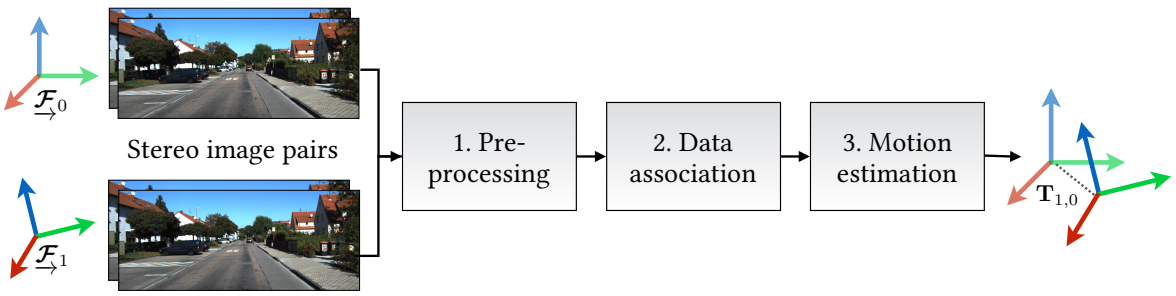
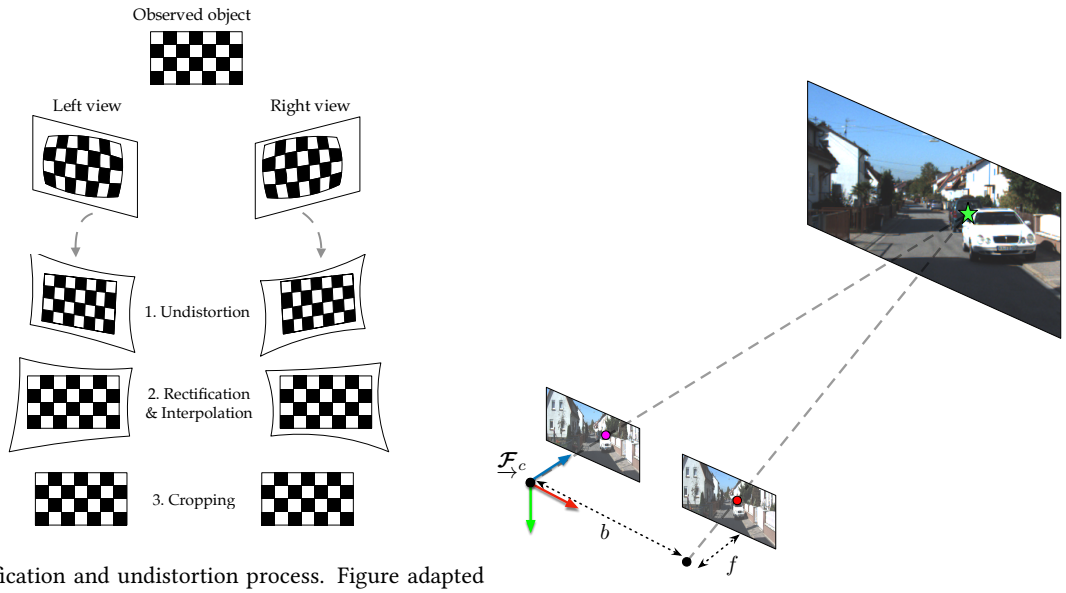


Figure 3.2: A 'classical' stereo visual odometry pipeline consists of several distinct components that have interpretable inputs and outputs.

In this dissertation, we apply our learned models to a baseline stereo, indirect visual odometry pipeline (Figure 3.2). We choose this baseline system for its computational efficiency and robustness. We briefly summarize the main components of the pipeline here.



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

(b) Ideal stereo camera.

Figure 3.3: We pre-process stereo images (left) to simulate an ideal stereo camera (right).

3.1.1 Preprocessing

The preprocessing stage consists of two major steps. First, we use a lens model to undistort each stereo image. Second, we use the stereo camera calibration (i.e., the intrinsic parameters of each camera and the extrinsic parameters defining the transform $\mathbf{T} \in \mathrm{SE}(3)$ between the two camera reference frames), to *rectify, interpolate and crop* the pair of images such that we can assume a *ideal frontoparallel stereo camera* model with a single focal length (Figure 3.3). That is, a stereo camera in which we can assume that any feature in one camera can be found in the same vertical location in the other (i.e., the epipolar line is horizontal). We assume the lens model, intrinsic and extrinsic parameters are constant and given by the dataset, or defined by a calibration process prior to data collection (e.g., given by the calibration tool detailed in Furgale et al. (2013)).

3.1.2 Data Association

Feature Extraction and Matching

Although a number of different types of indirect feature extraction and matching methods have been presented in the literature (e.g., SIFT, Lowe (1999) or ORB, Rublee et al. (2011)), we choose to use the viso2 (Geiger et al., 2011) algorithm due to its computationally efficiency (it can extract thousands of matches in milliseconds on modern hardware) and its use of a motion model that facilitates matching with sequential images. Briefly, 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 based on Sobel filter responses. Features are

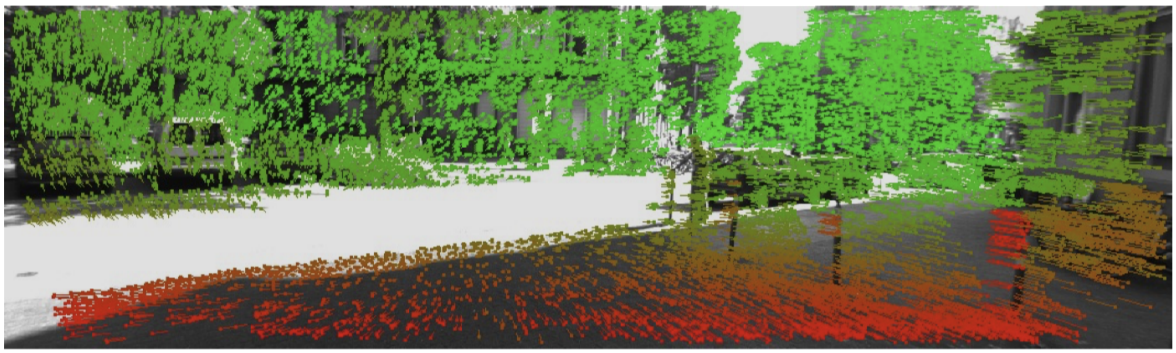


Figure 3.4: Feature tracking using VISO2, taken from Geiger et al. (2011). Colours correspond to depth.

matched across a stereo-pair and forward in time with a consistency check to ensure that a single feature exists in all four images in two stereo camera poses.

Ideal Stereo Camera Model

We model each feature extracted by viso2 as a three-dimensional point landmark that can be expressed (in homogeneous coordinates) in the camera frame, $\underrightarrow{\mathcal{F}}_c$, as $\boldsymbol{p}_{i,c} \in \mathbb{P}^3$. Our ideal stereo camera model, $\mathbf{f}(\cdot)$, projects $\boldsymbol{p}_{i,c}$ into image space coordinates as

$$\mathbf{y}_{i,c} = \begin{bmatrix} u_l \\ v_l \\ d \end{bmatrix} = \mathbf{f}\left(\boldsymbol{p}_{i,c}\right) = \mathbf{f}\left(\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}\right) = \mathbf{M}\frac{1}{z}\boldsymbol{p}_{i,c}, \tag{3.1}$$

where

$$\mathbf{M} = \begin{bmatrix} f & 0 & c_u & 0 \\ 0 & f & c_v & 0 \\ 0 & 0 & 0 & fb \end{bmatrix}. \tag{3.2}$$

Here, $\{c_u, c_v\}$, f, and b are the principal points, focal length and baseline of the stereo camera respectively (computed through intrinsic and extrinsic calibration) and $d \triangleq u_l - u_r$ is the *disparity* of the feature. Note that in this formulation, the stereo camera frame is in the left optical centre. Given $\mathbf{y}_{i,c}$, we also define the inverse operation, $\mathbf{f}^{-1}(\cdot)$ (triangulation) as:

$$\boldsymbol{p}_{i,c} = \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \mathbf{f}^{-1}\left(\begin{bmatrix} u_l \\ v_l \\ d \end{bmatrix}\right) = \begin{bmatrix} \frac{b}{d}(u_l - c_u) \\ \frac{b}{d}(v_l - c_v) \\ \frac{b}{d}f \\ 1 \end{bmatrix}. \tag{3.3}$$

Outlier Rejection

Given N_t feature *tracks* (i.e., the image locations of matching features across time), $\{\mathbf{y}_{i,c_0}, \mathbf{y}_{i,c_1}\}_{i=1}^{N_t}$, we filter out any *outliers* by applying three-point random sample consensus algorithm (RANSAC, Fischler and Bolles (1981)) based on an analytic solution to the six degree-of-freedom motion (Umeyama, 1991) (refer to Appendix B for more details).

3.1.3 Maximum Likelihood Motion Solution

Finally, we compute the rigid-body transform between two stereo camera frames using maximum likelihood estimation. We define the rigid-body transform, $\mathbf{T}_t \in \mathrm{SE}(3)$, to be the rigid-body transform between two subsequent stereo camera poses, $\underrightarrow{\mathcal{F}}_{c_0}$ and $\underrightarrow{\mathcal{F}}_{c_1}$,

$$\mathbf{T}_t = \mathbf{T}_{c_1 w} \mathbf{T}_{c_0 w}^{-1}, \tag{3.4}$$

where $\underrightarrow{\mathcal{F}}_w$ is a predefined world frame. For each track, $(\mathbf{y}_{i,c_0}, \mathbf{y}_{i,c_1})$, we define an error function, $\mathbf{e}_{i,t}(\mathbf{T}_t, \mathbf{y}_{i,c_0}, \mathbf{y}_{i,c_1})$, that relates the rigid transform to these stereo feature matches. For notational clarity, we will refer to this term as simply $\mathbf{e}_i(\mathbf{T}_t)$ with the dependence on the track implied. Next, we assume that these errors are corrupted by zero-mean independent Gaussian noise with the covariance, $\mathbf{\Sigma}_{i,t}$;

$$\mathbf{e}_i(\mathbf{T}_t) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{i,t}). \tag{3.5}$$

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

$$\mathbf{T}_t^* = \underset{\mathbf{T} \in \mathrm{SE}(3)}{\operatorname{argmax}} \prod_{i=1}^{N_t} p(\mathbf{e}_i(\mathbf{T}_t)) = \underset{\mathbf{T} \in \mathrm{SE}(3)}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{N_t} \mathbf{e}_i(\mathbf{T}_t)^\mathsf{T} \mathbf{\Sigma}_{i,t}^{-1} \mathbf{e}_i(\mathbf{T}_t). \tag{3.6}$$

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 $\boldsymbol{p}_{i,c_1} = \mathbf{f}^{-1}(\mathbf{y}_{i,c_1})$, and then define a three-dimensional error,

$$\mathbf{e}_i(\mathbf{T}_t) = \mathbf{D}(\boldsymbol{p}_{i,c_1} - \mathbf{T}_t \boldsymbol{p}_{i,c_0}) \in \mathbb{R}^3, \tag{3.7}$$

where $\mathbf{D} = \begin{bmatrix} \mathbf{1}_{3\times 3} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{3\times 4}$ converts homogenous coordinates into Euclidean coordinates.
We can then follow Maimone et al. (2007) and assume each stereo projection is corrupted by additive Gaussian noise,

$$\mathbf{y}_{i,c} \sim \mathcal{N}(\bar{\mathbf{y}}_{i,c}, \mathbf{R}_{i,c}), \tag{3.8}$$

and compute a density on the error function itself through first order noise propagation. This gives the density

$$\mathbf{e}_i(\mathbf{T}_t) \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_{i,t}\right), \tag{3.9}$$

where

$$\mathbf{\Sigma}_{i,t} = \mathbf{D}\mathbf{G}_{i,c_1}\mathbf{R}_{i,c_1}\mathbf{G}_{i,c_1}^{\mathsf{T}}\mathbf{D}^{\mathsf{T}} + \mathbf{D}\mathbf{T}_t\mathbf{G}_{i,c_0}\mathbf{R}_{i,c_0}\mathbf{G}_{i,c_0}^{\mathsf{T}}\mathbf{T}_t^{\mathsf{T}}\mathbf{D}^{\mathsf{T}}, \tag{3.10}$$

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

$$\mathbf{e}_i(\mathbf{T}_t) = \mathbf{y}_{i,c_1} - \mathbf{f}(\mathbf{T}_t\mathbf{f}^{-1}(\mathbf{y}_{i,c_0})), \tag{3.11}$$

and assume the following simple noise model

$$\mathbf{e}_i(\mathbf{T}_t) \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_{i,t}\right) = \mathcal{N}\left(\mathbf{0}, \mathbf{R}_{i,t}\right), \tag{3.12}$$

where the subscript t in $\mathbf{R}_{i,t}$ indicates that this measurement covariance refers to the reprojection error that involves a temporal track.

Importantly, Sibley et al. (2007) show that using reprojection error results in less biased estimates for long-range stereo triangulation (when compared to point cloud error). Consequently, we favour this latter formulation in the large majority of our work (the one exception being the initial work on isotropic PROBE described in Appendix A).

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 (3.6). We follow Section 2.4 and use the left perturbations $\delta\boldsymbol{\xi}^{\ell} \in \mathbb{R}^6$:

$$\mathbf{T}_t = \mathrm{Exp}\left(\delta\boldsymbol{\xi}^{\ell}\right)\mathbf{T}_t^{(n)} \approx (\mathbf{1} + \delta\boldsymbol{\xi}^{\wedge})\mathbf{T}_t^{(n)}. \tag{3.13}$$

where we have dropped the perturbation superscript for brevity. This allows us to transform Equation (3.6) into a linear least squares objective in $\delta\boldsymbol{\xi}$:

$$\mathcal{L}(\delta\boldsymbol{\xi}) = \frac{1}{2}\sum_{i=1}^{N_t}\left(\mathbf{e}_i - \mathbf{J}_i\delta\boldsymbol{\xi}\right)^{\mathsf{T}}\mathbf{\Sigma}_i^{-1}\left(\mathbf{e}_i - \mathbf{J}_i\delta\boldsymbol{\xi}\right) \tag{3.14}$$

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,

$$\delta \boldsymbol{\xi}^\star = \left(\sum_{i=1}^{N_t} \mathbf{J}_i^\mathsf{T} \boldsymbol{\Sigma}_i^{-1} \mathbf{J}_i \right)^{-1} \sum_{i=1}^{N_t} \mathbf{J}_i^\mathsf{T} \boldsymbol{\Sigma}_i^{-1} \mathbf{e}_i. \tag{3.15}$$

Given $\delta \boldsymbol{\xi}^\star$, we can update the operating point using the constraint-sensitive update

$$\mathbf{T}_t^{(n+1)} = \operatorname{Exp}(\delta \boldsymbol{\xi}^\star)\, \mathbf{T}_t^{(n)}, \tag{3.16}$$

and iterate until convergence. See Appendix B for more details and an analytic expression for $\mathbf{J}_i$. There are many reasonable choices for both the initial transform $\mathbf{T}_t^{(0)}$ and for the conditions under which we terminate iteration. For most visual odometry applications, it suffices to initialize the estimated transform to identity, and iteratively perform the update given by Equation (3.16) until we see a relative change in the squared error of less than one percent after an update.

3.2 Robust Estimation

Since Equation (3.14) assigns cost values that grow quadratically with measurement error, it is very sensitive to outlier measurements that may persist through RANSAC. A common solution to this problem is to replace the quadratic loss 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[1], and many variants exist. Each uses a re-weighting function, $\rho(\cdot)$, to define robust least squares (RLS) objective,

$$\mathbf{T}_{\text{RLS}}^* = \underset{\mathbf{T} \in \text{SE}(3)}{\operatorname{argmin}} \sum_{i=1}^{N} \rho\left(\sqrt{\mathbf{e}_i{}^\mathsf{T} \boldsymbol{\Sigma}_i^{-1} \mathbf{e}_i} \right) = \underset{\mathbf{T} \in \text{SE}(3)}{\operatorname{argmin}} \sum_{i=1}^{N} \rho(\epsilon_i) = \underset{\mathbf{T} \in \text{SE}(3)}{\operatorname{argmin}}\, \mathcal{L}_{\text{RLS}}(\mathbf{T}), \tag{3.17}$$

where we have defined $\epsilon_i \triangleq \sqrt{\mathbf{e}_i{}^\mathsf{T} \boldsymbol{\Sigma}_i^{-1} \mathbf{e}_i}$ (and dropped the t subscript for clarity). The basic idea with M-estimation is to use a $\rho(\cdot)$ that reduces the influence of large ϵ below that of the quadratic $\rho(\epsilon) = \frac{1}{2}\epsilon^2$. There are several examples of such functions, including,

$$\rho(\epsilon) = \begin{cases} \frac{c^2}{2} \log\left(1 + \frac{\epsilon^2}{c^2}\right) & \text{Cauchy,} \\ \frac{1}{2} \frac{\epsilon^2}{c^2 + \epsilon^2} & \text{Geman-McClure (Geman et al., 1992),} \\ \begin{cases} \frac{\epsilon^2}{2} & \text{if } \epsilon < c \\ c\epsilon - \frac{c^2}{2} & \text{if } \epsilon \geq c \end{cases} & \text{Huber (Huber, 1964).} \end{cases} \tag{3.18}$$

[1] M, for *maximum-likelihood-type* since they generalize the basic maximum likelihood solution (Barfoot, 2017).

where the constant c can be set with reference to asymptotic efficiency relative to a unit Gaussian (Holland and Welsch, 1977). To solve Equation (3.17), it is common in the literature to apply the technique of *iteratively reweighted least squares* (IRLS) (Holland and Welsch, 1977). To do this, we define a new non-linear least squares minimization problem,

$$\mathbf{T}^*_{\text{IRLS}} = \underset{\mathbf{T}\in\text{SE}(3)}{\text{argmin}} \frac{1}{2}\sum_{i=1}^{N} \mathbf{e}_i{}^\mathsf{T}\,\mathbf{M}_i\mathbf{e}_i = \underset{\mathbf{T}\in\text{SE}(3)}{\text{argmin}}\, \mathcal{L}_{\text{IRLS}}(\mathbf{T}) \tag{3.19}$$

where we define these new weights, $\mathbf{M}_i$, based on an *influence function*, $\psi(\cdot)$ as

$$\mathbf{M}_i = \frac{1}{\epsilon_i}\underbrace{\left.\frac{\partial\rho}{\partial\epsilon}\right|_{\epsilon_i}}_{\psi(\epsilon_i)}\boldsymbol{\Sigma}_i^{-1}, \tag{3.20}$$

and solve it using the Gauss-Newton approach presented in Section 3.1.3. We claim that upon convergence, $\mathbf{T}^*_{\text{IRLS}}$ will also minimize Equation (3.17). To see why, consider that

$$\frac{\partial\mathcal{L}_{\text{RLS}}}{\partial\delta\boldsymbol{\xi}} = \sum_i^N \frac{\partial\rho}{\partial\epsilon_i}\frac{\partial\epsilon_i}{\partial\mathbf{e}_i}\frac{\partial\mathbf{e}_i}{\partial\delta\boldsymbol{\xi}} = \sum_i^N \frac{1}{\epsilon_i}\frac{\partial\rho}{\partial\epsilon_i}\,\mathbf{e}_i{}^\mathsf{T}\,\boldsymbol{\Sigma}_i^{-1}\frac{\partial\mathbf{e}_i}{\partial\delta\boldsymbol{\xi}}, \tag{3.21}$$

where he have used the fact that $\dfrac{\partial\epsilon_i}{\partial\mathbf{e}_i} = \frac{1}{\epsilon_i}\,\mathbf{e}_i{}^\mathsf{T}\,\boldsymbol{\Sigma}_i^{-1}$. Now using our definition of $\mathbf{M}_i$, we can write,

$$\frac{\partial\mathcal{L}_{\text{RLS}}}{\partial\delta\boldsymbol{\xi}} = \sum_i^N \mathbf{e}_i{}^\mathsf{T}\underbrace{\frac{1}{\epsilon_i}\frac{\partial\rho}{\partial\epsilon_i}\boldsymbol{\Sigma}_i^{-1}}_{\mathbf{M}_i(\mathbf{T})}\frac{\partial\mathbf{e}_i}{\partial\delta\boldsymbol{\xi}} = \sum_i^N \mathbf{e}_i{}^\mathsf{T}\,\mathbf{M}_i(\mathbf{T})\frac{\partial\mathbf{e}_i}{\partial\delta\boldsymbol{\xi}}, \tag{3.22}$$

where we have made the dependence on $\mathbf{T}$ explicit. We could potentially proceed to set this gradient to $\mathbf{0}$ and attempt to solve for an optimal update $\delta\boldsymbol{\xi}$. However, due to $\mathbf{M}_i(\mathbf{T})$, this may be difficult in general. Instead, we note that if we evaluate $\mathbf{M}_i(\mathbf{T})$ at the current operating point, $\mathbf{T}^{(n)}$, (i.e., we *re-weight* the loss) we are then left with the equivalent normal equations that solve $\dfrac{\partial\mathcal{L}_{\text{IRLS}}}{\partial\delta\boldsymbol{\xi}} = \mathbf{0}$.

Furthermore, upon convergence, our solution to the iteratively re-weighted problem $\mathbf{T}^{(n)} = \mathbf{T}^*_{\text{IRLS}}$ will also minimize the robust objective Equation (3.17), since we must have that,

$$\left.\frac{\partial\mathcal{L}_{\text{IRLS}}}{\partial\delta\boldsymbol{\xi}}\right|_{\mathbf{T}^*_{\text{IRLS}}} = \left.\frac{\partial\mathcal{L}_{\text{RLS}}}{\partial\delta\boldsymbol{\xi}}\right|_{\mathbf{T}^*_{\text{IRLS}}} = \mathbf{0}. \tag{3.23}$$

3.3 Outstanding Issues

Finally, we summarize (Table 3.1) several limitations of the canonical visual odometry pipeline we have presented in this chapter. Namely, the issues of efficiency, systematic bias and homoscedastic uncertainty. For each limitation, we will show in this dissertation that we can build a learned model that addresses it for a particular environment (defined by the training data).

Table 3.1: Issues that can be addressed by learned models.

Synopsis	Addressed by
Computational efficiency: classical VO pipelines face a difficult-to-optimize trade-off between using all of the information contained within image while still remaining computationally tractable.	PROBE (Chapter 4) , DPC (Chapter 6) , Sun-BCNN (Chapter 5), HydraNet (Chapter 7)
Systematic bias: 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 (Chapter 6), HydraNet (Chapter 7)
Homoscedastic uncertainty: 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.	DPC (Chapter 6) , Sun-BCNN (Chapter 5), HydraNet (Chapter 7)

Appendices

Bibliography

Agarwal, S., Mierle, K., et al. (2016). Ceres solver.

Albertz, J. (2007). A look back. *Photogrammetric Engineering & Remote Sensing*, 73(5):504–506.

Alcantarilla, P. F. and Woodford, O. J. (2016). Noise models in feature-based stereo visual odometry.

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

Amos, B. and Kolter, J. Z. (2017). OptNet: Differentiable optimization as a layer in neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 136–145. PMLR.

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.

Brachmann, E. and Rother, C. (2018). Learning less is more-6d camera localization via 3d surface regression. In *Proc. CVPR*, volume 8.

Brachmann, E. and Rother, C. (2019). Neural- Guided RANSAC: Learning where to sample model hypotheses. In *ICCV*.

Bruss, A. R. and Horn, B. K. (1983). Passive Navigation. *Computer Vision, Graphics, and Image Processing*, 21(1):3–20.

Byravan, A. and Fox, D. (2017). SE3-nets: Learning rigid body motion using deep neural networks. In *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, pages 173–180.

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.

Carlone, L., Rosen, D. M., Calafiore, G., Leonard, J. J., and Dellaert, F. (2015a). Lagrangian duality in 3D SLAM: Verification techniques and optimal solutions. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 125–132.

Carlone, L., Tron, R., Daniilidis, K., and Dellaert, F. (2015b). Initialization techniques for 3D SLAM: A survey on rotation estimation and its use in pose graph optimization. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4597–4604.

Censi, A. (2007). An accurate closed-form estimate of icp's covariance. In *Proceedings 2007 IEEE international conference on robotics and automation*, pages 3167–3172. IEEE.

Cheng, Y., Maimone, M. W., and Matthies, L. (2006). Visual odometry on the mars exploration rovers - a tool to ensure accurate driving and science imaging. *IEEE Robot. Automat. Mag.*, 13(2):54–62.

Clark, R., Wang, S., Wen, H., Markham, A., and Trigoni, N. (2017). Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem.

Clement, L., Peretroukhin, V., and Kelly, J. (2017). Improving the accuracy of stereo visual odometry using visual illumination estimation. In Kulic, D., Nakamura, Y., Khatib, O., and Venture, G., editors, *2016 International Symposium on Experimental Robotics*, volume 1 of *Springer Proceedings in Advanced Robotics*, pages 409–419. Springer International Publishing, Berlin Heidelberg. Invited to Journal Special Issue.

Costante, G., Mancini, M., Valigi, P., and Ciarfuglia, T. A. (2016). Exploring representation learning with CNNs for Frame-to-Frame Ego-Motion estimation. *IEEE Robotics and Automation Letters*, 1(1):18–25.

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.

Cvišić, I. and Petrović, I. (2015). Stereo odometry based on careful feature selection and tracking. In *Proc. European Conf. on Mobile Robots (ECMR)*, pages 1–6.

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.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Comput. Vision and Pattern Recognition, (CVPR)*, pages 248–255.

DeTone, D., Malisiewicz, T., and Rabinovich, A. (2016). Deep image homography estimation.

Duan, Y., Chen, X., Houthooft, R., Schulman, J., and Abbeel, P. (2016). Benchmarking deep reinforcement learning for continuous control. In *Proc. Int. Conf. on Machine Learning*, ICML'16, pages 1329–1338.

Durrant-Whyte, H., Rye, D., and Nebot, E. (1996). Localization of autonomous guided vehicles. In *Robotics Research*, pages 613–625. Springer.

Eisenman, A. R., Liebe, C. C., and Perez, R. (2002). Sun sensing on the mars exploration rovers. In *Aerosp. Conf. Proc.*, volume 5, pages 5–2249–5–2262 vol.5. IEEE.

Engel, J., Koltun, V., and Cremers, D. (2018). Direct Sparse Odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3):611–625.

Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pages 363–370. Springer.

Fischler, M. and Bolles, R. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Comm. 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., Carlone, L., Dellaert, F., and Scaramuzza, D. (2015). IMU preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation.

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. and Barfoot, T. D. (2010). Visual teach and repeat for long-range rover autonomy. *J. Field Robot.*, 27(5):534–560.

Furgale, P., Carle, P., Enright, J., and Barfoot, T. D. (2012). The devon island rover navigation dataset. *Int. J. Rob. Res.*, 31(6):707–713.

Furgale, P., Enright, J., and Barfoot, T. (2011). Sun sensor navigation for planetary rovers: Theory and field testing. *IEEE Trans. Aerosp. Electron. Syst.*, 47(3):1631–1647.

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.

Gal, Y. (2016). *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge.

Gal, Y. and Ghahramani, Z. (2016a). Bayesian convolutional neural networks with Bernoulli approximate variational inference. In *Proc. Int. Conf. Learning Representations (ICLR), Workshop Track*.

Gal, Y. and Ghahramani, Z. (2016b). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proc. Int. Conf. Mach. Learning (ICML)*, pages 1050–1059.

Gallego, G. and Yezzi, A. (2015). A compact formula for the derivative of a 3-D rotation in exponential coordinates. *J. Math. Imaging Vis.*, 51(3):378–384.

Garg, R., Carneiro, G., and Reid, I. (2016). Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *European Conf. on Comp. Vision*, pages 740–756. Springer.

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: The KITTI dataset. *Int. J. Rob. Res.*, 32(11):1231–1237.

Geiger, A., Ziegler, J., and Stiller, C. (2011). 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.

Glocker, B., Izadi, S., Shotton, J., and Criminisi, A. (2013). Real-time rgb-d camera relocalization. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 173–179.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Grewal, M. S. and Andrews, A. P. (2010). Applications of kalman filtering in aerospace 1960 to the present [historical perspectives]. *IEEE Control Syst. Mag.*, 30(3):69–78.

Gustafsson, F. and Gustafsson, F. (2000). *Adaptive filtering and change detection*, volume 1. Citeseer.

Haarnoja, T., Ajay, A., Levine, S., and Abbeel, P. (2016). Backprop KF: Learning discriminative deterministic state estimators. In *Proc. Advances in Neural Inform. Process. Syst. (NIPS)*.

Handa, A., Bloesch, M., Pătrăucean, V., Stent, S., McCormac, J., and Davison, A. (2016). gvnn: Neural network library for geometric computer vision. In *Computer Vision – ECCV 2016 Workshops*, pages 67–82. Springer, Cham.

Hartley, R., Trumpf, J., Dai, Y., and Li, H. (2013). Rotation averaging. *Int. J. Comput. Vis.*, 103(3):267–305.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Holland, P. W. and Welsch, R. E. (1977). Robust regression using iteratively reweighted least-squares. *Communications in Statistics - Theory and Methods*, 6(9):813–827.

Hu, H. and Kantor, G. (2015). Parametric covariance prediction for heteroscedastic noise. In *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Syst. (IROS)*, pages 3052–3057.

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

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proc. ACM Int. Conf. Multimedia (MM)*, pages 675–678.

Julier, S. J. and Uhlmann, J. K. (2007). Using covariance intersection for slam. *Robotics and Autonomous Systems*, 55(1):3–20.

Kelly, J., Saripalli, S., and Sukhatme, G. S. (2008). Combined visual and inertial navigation for an unmanned aerial vehicle. In *Proc. Field and Service Robot. (FSR)*, pages 255–264.

Kendall, A. and Cipolla, R. (2016). Modelling uncertainty in deep learning for camera relocalization. In *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, pages 4762–4769.

Kendall, A. and Cipolla, R. (2017). Geometric loss functions for camera pose regression with deep learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6555–6564.

Kendall, A. and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision?

Kendall, A., Grimes, M., and Cipolla, R. (2015). PoseNet: A convolutional network for Real-Time 6-DOF camera relocalization. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2938–2946.

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.

Kingma, D. P. and Ba, J. (2017). Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*. arXiv: 1412.6980.

Ko, J. and Fox, D. (2009). Gp-bayesfilters: Bayesian filtering using gaussian process prediction and observation models. *Autonomous Robots*, 27(1):75–90.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 6402–6413. Curran Associates, Inc.

Lalonde, J.-F., Efros, A. A., and Narasimhan, S. G. (2011). Estimating the natural illumination conditions from a single outdoor image. *Int. J. Comput. Vis.*, 98(2):123–145.

Lambert, A., Furgale, P., Barfoot, T. D., and Enright, J. (2012). Field testing of visual odometry aided by a sun sensor and inclinometer. *J. Field Robot.*, 29(3):426–444.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551.

Lee, S., Purushwalkam, S., Cogswell, M., Crandall, D., and Batra, D. (2015). Why M heads are better than one: Training a diverse ensemble of deep networks.

Leutenegger, S., Lynen, S., Bosse, M., Siegwart, R., and Furgale, P. (2015). Keyframe-based visual–inertial odometry using nonlinear optimization. *Int. J. Rob. Res.*, 34(3):314–334.

Levine, S., Finn, C., Darrell, T., and Abbeel, P. (2016). End-to-end training of deep visuomotor policies. *J. Mach. Learn. Res.*

Li, Q., Qian, J., Zhu, Z., Bao, X., Helwa, M. K., and Schoellig, A. P. (2017a). Deep neural networks for improved, impromptu trajectory tracking of quadrotors. In *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, pages 5183–5189.

Li, R., Wang, S., Long, Z., and Gu, D. (2017b). UnDeepVO: Monocular visual odometry through unsupervised deep learning.

Liu, K., Ok, K., Vega-Brown, W., and Roy, N. (2018). Deep inference for covariance estimation: Learning gaussian noise models for state estimation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1436–1443. IEEE.

Lowe, D. (1999). Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157 vol.2.

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.

Ma, W.-C., Wang, S., Brubaker, M. A., Fidler, S., and Urtasun, R. (2016). Find your way by observing the sun and other semantic cues.

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.

Maddern, W., Pascoe, G., Linegar, C., and Newman, P. (2016). 1 year, 1000 km: The oxford RobotCar dataset. *Int. J. Rob. Res.*

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.

Mayor, A. (2019). *Gods and Robots*. Princeton University Press.

McManus, C., Upcroft, B., and Newman, P. (2014). Scene signatures: Localised and point-less features for localisation. In *Proc. Robotics: Science and Systems X*.

Melekhov, I., Ylioinas, J., Kannala, J., and Rahtu, E. (2017). Relative camera pose estimation using convolutional neural networks. In *Proc. Int. Conf. on Advanced Concepts for Intel. Vision Syst.*, pages 675–687. Springer.

Melkumyan, A. and Ramos, F. (2011). Multi-kernel gaussian processes. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

Moravec, H. P. (1980). Obstacle avoidance and navigation in the real world by a seeing robot rover. Technical report, Stanford Univ CA Dept of Computer Science.

Oliveira, G. L., Radwan, N., Burgard, W., and Brox, T. (2017). Topometric localization with deep learning. *arXiv preprint arXiv:1706.08775*.

Olson, C. F., Matthies, L. H., Schoppers, M., and Maimone, M. W. (2003). Rover navigation using stereo ego-motion. *Robot. Auton. Syst.*, 43(4):215–229.

Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. (2016). Deep exploration via bootstrapped DQN. In *Proc. Advances in Neural Inform. Process. Syst. (NIPS)*, pages 4026–4034.

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., and Kelly, J. (2015b). 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., Clement, L., and Kelly, J. (2017). Reducing drift in visual odometry by inferring sun direction using a bayesian convolutional neural network. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'17)*, pages 2035–2042, Singapore.

Peretroukhin, V., Clement, L., and Kelly, J. (2018). Inferring sun direction to improve visual odometry: A deep learning approach. *International Journal of Robotics Research*, 37(9):996–1016.

Peretroukhin, V. and Kelly, J. (2018). DPC-Net: Deep pose correction for visual localization. *IEEE Robotics and Automation Letters*, 3(3):2424–2431.

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.

Peretroukhin, V., Wagstaff, B., and Kelly, J. (2019). Deep probabilistic regression of elements of SO(3) using quaternion averaging and uncertainty injection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'19) Workshop on Uncertainty and Robustness in Deep Visual Learning*, pages 83–86, Long Beach, California, USA.

Punjani, A. and Abbeel, P. (2015). Deep learning helicopter dynamics models. In *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, pages 3223–3230.

Ranftl, R. and Koltun, V. (2018). Deep fundamental matrix estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 284–299.

Richter, C. and Roy, N. (2017). Safe visual navigation via deep learning and novelty detection.

Rosen, D. M., Carlone, L., Bandeira, A. S., and Leonard, J. J. (2019). SE-Sync: A certifiably correct algorithm for synchronization over the special euclidean group. *Int. J. Rob. Res.*, 38(2-3):95–125.

Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. R. (2011). Orb: An efficient alternative to sift or surf. In *ICCV*, volume 11, page 2. Citeseer.

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

Schonberger, J. L., Hardmeier, H., Sattler, T., and Pollefeys, M. (2017). Comparative Evaluation of Hand-Crafted and Learned Local Features. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6959–6968, Honolulu, HI. IEEE.

Sibley, G., Matthies, L., and Sukhatme, G. (2007). Bias reduction and filter convergence for long range stereo. In *Robotics Research*, pages 285–294. Springer Berlin Heidelberg.

Sobel, D. (2005). *Longitude: The true story of a lone genius who solved the greatest scientific problem of his time*. Macmillan.

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.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Sünderhauf, N., Shirazi, S., Dayoub, F., Upcroft, B., and Milford, M. (2015). On the performance of ConvNet features for place recognition. In *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Syst. (IROS)*, pages 4297–4304.

Sunderhauf, N., Shirazi, S., Jacobson, A., Dayoub, F., Pepperell, E., Upcroft, B., and Milford, M. (2015). Place recognition with ConvNet landmarks: Viewpoint-robust, condition-robust, training-free. In *Proc. Robotics: Science and Systems XII*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proc. IEEE Conf. Comput. Vision and Pattern Recognition, (CVPR)*, pages 1–9.

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 *Proc. Advances in Neural Information Proc. Syst. (NIPS) 27*, pages 2546–2554.

Wang, R., Schworer, M., and Cremers, D. (2017a). Stereo DSO: Large-Scale Direct Sparse Visual Odometry with Stereo Cameras. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3923–3931, Venice. IEEE.

Wang, S., Clark, R., Wen, H., and Trigoni, N. (2017b). DeepVO: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2043–2050.

Warren, R. (1976). The perception of egomotion. *Journal of Experimental Psychology: Human Perception and Performance*, 2(3):448.

Wilson, A. G. and Ghahramani, Z. (2011). Generalised wishart processes. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 736–744. AUAI Press.

Yang, F., Choi, W., and Lin, Y. (2016). Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In *Proc. IEEE Int. Conf. Comp. Vision and Pattern Recognition (CVPR)*, pages 2129–2137.

Yang, N., Wang, R., Stueckler, J., and Cremers, D. (2018). Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In *European Conference on Computer Vision (ECCV)*. accepted as oral presentation, arXiv 1807.02570.

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

Zhou, B., Krähenbühl, P., and Koltun, V. (2019). Does computer vision matter for action? *Science Robotics*, 4(30).

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014). Learning deep features for scene recognition using places database. In *Advances in Neural Inform. Process. Syst. (NIPS)*, pages 487–495.

Zhou, T., Brown, M., Snavely, N., and Lowe, D. G. (2017). Unsupervised learning of depth and Ego-Motion from video. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6612–6619.