Thesis Proposal

Direct Multiple View Visual Simultaneous Localization And Mapping

Hatem Alismail
halismai@cs.cmu.edu

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Robotics.

The Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

August, 2015

Thesis committee:
Dr. Brett Browning, Robotics Institute (Co-Chair)
Dr. Simon Lucey, Robotics Institute (Co-Chair)
Dr. Michael Kaess, Robotics Institute
Prof. Martial Hebert, Robotics Institute
Prof. Ian Reid, The University of Adelaide

Copyright © (2015) by Hatem Alismail. All rights reserved.
Last update: August 4, 2015 at 2:52am
Abstract

We propose a direct, featureless, Lucas-Kanade-based method as a reliable Visual Simultaneous Localization And Mapping (VSLAM) solution in challenging environments, where feature detection and precise subpixel localization may be unreliable. Current state-of-the-art direct methods have been shown to perform well on a range of challenging datasets. Nonetheless, they have been limited to relative pose estimation between two frames (Visual Odometry), and hence do not fully exploit the available information contained in a stream of imagery.

Extending direct methods to multiple views is a difficult problem because of the reliance on the brightness constancy assumption, which is seldom satisfied in robotic applications operating in unstructured scenes. Additionally — even if brightness constancy is satisfied — prior approaches to multiple view direct VSLAM either impose simplifying assumptions such as planarity of the world, or rely on the sub-optimal strategy of alternating optimization.

In this work, we propose a direct, joint optimization of the state vector over multiple views akin to the geometric Bundle Adjustment as commonly employed for optimal reconstruction from correspondences. The main difference, however, is that our framework operates on a function of image data directly without requiring precomputed, and fixed, correspondences. Instead, correspondences are estimated automatically by our algorithm as a byproduct of estimating the structure and motion.

To address the limitation of the brightness constancy assumption, we propose to borrow descriptors from feature-based methods and propose to perform direct alignment using “descriptor constancy.” This new descriptor constancy assumption is more robust with respect to appearance variations, without affecting the dimension of the state vector, or significantly increasing computational demands.

To this end, we consider binary descriptors as an illustrative case due to their compu-
tational efficiency and invariance to monotonic illumination changes. In particular, we
demonstrate the use of the Census Transform (Local Binary Patterns). To address the
non-smoothness, and discontinuity, of feature descriptors we propose a multi-channel
representation, which allows us to (i) efficiently estimate the gradient of the objective,
and (ii) to minimize the exact Hamming distance between binary descriptors using
standard nonlinear least squares optimization algorithms.

We will evaluate our proposed Direct Multiple View VSLAM on a range of synthetic
and real datasets with comparisons against state-of-the-art. We will also work towards
a deeper understanding of the descriptor constancy assumptions and explore the use of
other descriptors in the literature. Applications of descriptor constancy extend beyond
VSLAM and can be used for other correspondence problems in vision such as template
tracking and optical flow, which we will consider in our analysis and evaluation work.
# Contents

1 Introduction .......................................................... 1  
1.1 Problem & Motivation .............................................. 1  
1.2 Research Questions and Thesis Statement ..................... 3  
1.3 Expected Contributions ........................................... 5  

2 Background and Related Work ...................................... 7  
2.1 Image Alignment: Lucas-Kanade and Variations ............... 7  
2.2 Direct Visual Odometry ............................................ 12  
2.3 Direct Estimation of Structure & Motion from Multiple Views 14  
2.4 Violations of the Brightness Constancy Assumption .......... 17  
2.5 Geometric (Keypoint-based) Bundle Adjustment .............. 20  

3 Direct Disparity Space Visual Odometry ......................... 21  
3.1 Disparity space .................................................... 21  
3.2 Direct Visual Odometry .......................................... 22  
3.3 Experiments and Results ......................................... 27  
3.4 Discussion & Summary ............................................ 33  

4 Joint Multiple View Direct VSLAM ................................. 37  
4.1 Geometric BA ....................................................... 37  
4.2 Joint Multiple View Direct Estimation (LK-BA) [Proposed Work] 40  
4.3 Preliminary Results ............................................... 44  

III
Chapter 1

Introduction

1.1 Problem & Motivation

Visual Simultaneous Localization And Mapping (VSLAM) is the problem of jointly estimating the motion of the camera as well as the 3D map of the scene from a stream of overlapping images. Due to its fundamental importance as a core technology for a wide range of applications, VSLAM has been extensively studied with outstanding progress over the past few years.

In this work, we are interested in general purpose VSLAM. That is, VSLAM without imposing constraints on either the type of camera motion, or the 3D structure of the world. The state-of-the-art algorithms for this general purpose VSLAM are commonly referred to as feature-based [106, 82, 39, 164, 146, 76]. The feature-based pipeline [151, 120] abstracts away the millions of pixels in the image into a few hundreds to thousands of keypoint positions. Keypoint abstraction enhances efficiency and transforms VSLAM to a purely geometric problem. More importantly, it eases correspondence estimation across multiple wide-baseline views in order to obtain a theoretically optimal reconstruction of the camera position and scene structure using the powerful, and well-studied Bundle Adjustment (BA) approach [152]. Nonetheless, the pure geometric formulation has its limitations. For instance, when operating in challenging environments (c.f. Fig. 1.1), keypoint detection and localization become unreliable, or even impossible. More generally, it is difficult to predict when different features will work.
(a) Lack of precise feature correspondences due to decreased frequency content in the image.

(b) Real imagery from an underground mine illustrating uneven lighting and motion blur.

(c) Auto exposure effects under illumination change.

Figure 1.1: Challenging data where performance of the state-of-the-art methods is sub-optimal.
1.2 Research Questions and Thesis Statement

In this work, we aim to answer the following research question:

How to enable VSLAM in challenging environments where keypoint detection and localization may be unreliable?

Keypoint detection and precise localization are not required to perform accurate VSLAM. Recent demonstrations in direct Visual Odometry, or Lucas-Kanade (LK)-based, has shown impressive results on a variety of datasets [46, 119, 143, 81, 162, 107]. Direct algorithms [73] do not rely on the ability to precisely detect and localize geometric keypoints, which makes them more resilient to image degradations. In addition to robustness against image degradations, direct algorithms are naturally suitable for dense, or semi-dense, reconstruction of the scene since the special requirements on the Structure Tensor of pixels in the image are relaxed; all pixels with a non-vanishing gradient can be included in the optimization. Typically, the resulting reconstruction is richer than the alternative keypoint-based methods without resorting to an offline dense scene reconstruction step, or a delayed semi-dense reconstruction [113]. Based on our results [4] and results from the literature, we believe that a direct framework is a viable solution when operating with challenging data.

To this date, however, there has been no demonstration of an LK-based method capable of reducing drift over time in general scenes by exploiting longer range correspondences (i.e. beyond key frame pairs) without resorting to pose graph optimization over loops. In some applications, loops may not exist, and even when loops exist pose graph methods rely on marginalization which may degrade accuracy.

The main challenge to multiple view direct VSLAM is an efficient method of overcoming the brightness constancy assumption. The simple solution is to estimate illumination parameters alongside the structure and motion. However, including illumination parameters in the optimization problems makes the problem too large to be solved efficiently, or even reliably.

In this work, we propose a robust solution to the brightness constancy assumption without affecting the dimension of the state vector by performing the optimization over the space of feature descriptors. By efficiently overcoming illumination variations in challenging data, we aim to enable VSLAM in challenging environments, where the current state-of-the-art’s performance is suboptimal.
Motivated by the advantages of LK-based VSLAM and the prospect of illumination robust image alignment, our thesis is:

A Lucas-Kanade based approach in a direct optimization framework over multiple views is a reliable VSLAM solution in unstructured and challenging environments, where keypoint detection may be unreliable. Furthermore, we can overcome the limitations of brightness constancy without the need to explicitly model illumination variations by performing the optimization over feature descriptors (non-smooth functions of intensities).

In order to support our thesis, we must address the following challenges:

– How to perform multiple view optimization using image data directly?
– How to perform continuous gradient-based optimization using non-convex and non-smooth feature descriptors?
– What are some of the feature descriptors that are suitable for continuous gradient-based optimization, and efficient for VSLAM applications?
– Finally, how to do the above efficiently for robotic applications?

To address these challenges, we first adapt methods from geometric Bundle Adjustment to work with image data directly. We will show that this optimization is feasible as the design matrix of the linear system is sparse. We will call this algorithm Lucas-Kanade Bundle Adjustment (LK-BA).

Our demonstration of LK-BA currently relies on the brightness constancy assumption. Hence, it is sensitive to illumination change. To address the limitations of brightness constancy, we propose LK-based descriptor alignment. Using descriptors is superior to other approaches to handling appearance change. This is because: (i) descriptors do not affect the dimension of the state-vector, which becomes an important computational issue in multiple-view optimization, (ii) descriptors do not require strict assumptions on the form of illumination change, (iii) descriptors are efficient to compute, and (iv) there exists and continues to be a multitude of research for descriptor design improving descriptor invariance in different environments.

We will show that continuous gradient-based optimization over non-smooth and non-convex feature descriptors is possible via a multi-channel representation. We
demonstrate our descriptor constancy idea using the Census Transform [166] (also known as LBP: Local Binary Patterns [122]), which is one of the simplest binary descriptors. An added advantage of this multi-channel representation is that minimizing the sum of squared residuals becomes equivalent to minimizing the Hamming distance. However, when using binary descriptors, the computation of gradients and pixel differences are discrete (ternary for the former and binary for the latter). One would expect that discrete gradients are not suitable for LK (nonlinear least squares). Yet, the proposed formulation is shown to work under arbitrary appearance variations when applied to template tracking problems. In our thesis work, we plan on understanding the reasons behind good performance of binary descriptors in continuous optimization problems.

In summary, we propose the following extensions to the direct VSLAM pipeline:

1. Jointly estimating the motion and structure over multiple views.
2. Handling appearance variations using a descriptor constancy assumption, which is more robust than the brightness constancy.

1.3 Expected Contributions

Upon the completion of this thesis, we expect to contribute the following:

- Allowing VSLAM – from vision-only data – to operate more robustly in challenging environments, especially when lighting conditions considerably degrade the frequency content in the image.

- Filling the gap in direct VSLAM by extending direct methods to multiple views. In order to reduce drift, current direct VSLAM algorithms rely on loop closure [46], or resort to keypoint-based geometric BA in an online [54], or an offline fashion [98]. By extending direct VSLAM to multiple views we aim to achieve the same level of accuracy in applications where loop closure is not possible.

- Deeper understanding of descriptor alignment with Lucas-Kanade (continuous gradient-based optimization). Applications of descriptor alignment extend beyond the VSLAM problem and tackle the fundamental correspondence problem. Applications include: estimation of the Optical Flow [16], Scene Flow [157, 160], Stereo...
Template Tracking [95, 102], and Active Appearance Models [30], among others.

• Thorough experimental evaluation and comparisons with the state-of-the-art, and illustration of failure cases.

In the next chapter, we provide a summary of related work. This proposal is organized as follows. In the next Chapter we provide a brief literature review of related topics. In Chapter 3 we present our direct VO algorithm designed for stereo data. This is followed by our proposed direct multiple view framework and preliminary results (Chapter 4). In Chapter 5 introduce our proposed descriptor constancy to address appearance variations. Finally, in Chapter 6 we conclude with a summary and detail our timeline for the thesis work.
Chapter 2

Background and Related Work

In this chapter, we provide a brief account of related work. We focus on the problem of parametric image alignment, direct Visual Odometry (VO), prior approaches to joint estimation of structure and motion from image data, and prior approaches to handling appearance variations.

2.1 Image Alignment: Lucas-Kanade and Variations

The goal of image alignment is to compute the deformations between one, or more, input images with respect to a fixed template/reference image such that a measure of dissimilarity between the template and the input images is minimized.

When the form of deformation between the template and input images is known, the problem is often referred to as parametric image alignment. Examples of applications of parametric image alignment include template tracking [102], corner localization [135], image registration, as well as direct Visual Odometry (VO).

In the sequel, we will provide a summary of the Lucas-Kanade (LK) algorithm [95] for parametric image alignment to introduce the notation we will use in this manuscript. For a complete exposition of LK and its variants we refer the reader to the excellent series of publications by Baker and Matthews [10] and Baker et al. [12, 11, 13, 14].
Table 2.1: Summary of common warps used in LK. For the projective warp (Homography) we show the linear 8-parameter model, but other parameterizations are possible [15]. Jacobian for the rigid-body warp depends on the specific parameterization (c.f. Eq. (3.11)). The function \( \pi (\cdot) \) denotes the projection onto the image plane.

<table>
<thead>
<tr>
<th></th>
<th>DOF</th>
<th>Input</th>
<th>Output</th>
<th>Warp</th>
<th>Jacobian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flow</td>
<td>((\theta_x, \theta_y))</td>
<td>((x, y))</td>
<td>((x', y'))</td>
<td>(w(x, \theta))</td>
<td>(\partial w / \partial \theta)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(\in) (\mathbb{R}^2)</td>
<td>(\in) (\mathbb{R}^2)</td>
<td>((x, y) + (\theta_x, \theta_y))</td>
<td>((1, 0, 0, 1))</td>
</tr>
<tr>
<td>Affine</td>
<td>((\theta_1, \theta_2, \ldots, \theta_6))</td>
<td>((x, y))</td>
<td>((x', y'))</td>
<td>((1 + \theta_1 \theta_3 \theta_5 \theta_7 \theta_8 \theta_9, \theta_2, 1 + \theta_4 \theta_6 \theta_7 \theta_8 \theta_9)) ((x, y))</td>
<td>((x, 0, y, 0, 1, 0))</td>
</tr>
<tr>
<td>Homography</td>
<td>((\theta_1, \theta_2, \ldots, \theta_8))</td>
<td>((x, y))</td>
<td>((x', y'))</td>
<td>((1 + \theta_1 \theta_2 \theta_3 \theta_4 \theta_5 \theta_6 \theta_7 \theta_8 , 1)) ((x, y))</td>
<td>((\alpha x, 0, \alpha^2 \gamma x, \alpha y, 0, -\alpha^2 \gamma y, \alpha, 0))</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rigid-body</td>
<td>((\theta_1, \theta_2, \ldots, \theta_6))</td>
<td>((X, Y, Z))</td>
<td>((x', y'))</td>
<td>(\pi (T(\theta)x))</td>
<td>(\frac{\alpha x}{\alpha x} \frac{\alpha x}{\partial \theta} \in) (\mathbb{R}^{2 \times 6})</td>
</tr>
</tbody>
</table>
2.1.1 Original LK (Forward Additive)

Given two images, a template/reference $I_0$ and an input/current image $I_1$ related via a parametric transform/warp, we desire to estimate the parameters of the warp such that a function of intensity dissimilarity between the template and input is minimized.

The warp is a geometric transform that transfers pixel coordinate from the reference image $x$ to the coordinate frame of the input image $x'$. We will write this warp as:

$$w : \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}^n$$

(2.1)

$$x' = w(x; \theta),$$

(2.2)

where $\theta \in \mathbb{R}^p$ is the vector of parameters we desire to estimate, $x \in \mathbb{R}^m$ is a pixel coordinate in the frame of the template image, and $x' \in \mathbb{R}^n$ is a pixel coordinate in the coordinate frame of the input image. Typically, pixel coordinates belong to the 2D image plane, i.e. $m = n = 2$, but other forms of warps can operate on higher dimensions depending on the problem formulation. A summary of commonly used warps is shown in Table 2.1.

The goal of LK is to estimate the vector of parameters $\theta$ such that

$$I_0(x) \equiv I_1(x').$$

(2.3)

Exact equality, however, is unattainable due to noise and outliers. If we assume equality up to an additive Gaussian noise, then minimizing the sum of squared residuals corresponds to maximizing the likelihood, which is the optimal choice given no prior knowledge. Under least squares, the minimization problem is of the form

$$\min_{\theta} \sum_{x \in \Omega_0} \|I_0(x) - I_1(w(x; \theta))\|_2^2,$$  

(2.4)

where $\Omega_0$ denotes a subset of pixel coordinates in the reference frame. This relationship is also known as the brightness constancy assumption, or brightness conservation [69].

Since the intensity value of a pixel, in general, is unrelated to the form of the warp, the optimization problem in Eq. (2.4) is nonlinear. Any nonlinear solver could be used to solve Eq. (2.4). In practice, Gauss-Newton (GN), or Levenberg-Marquardt (LM) are the algorithms of choice.

The solution proceeds by iteratively estimating a small parameter update $\Delta \theta$ in the
vicinity of a given initialization \( \theta \). A first-order Taylor series expansion about \( \Delta \theta = 0 \) yields the following linear system of equations

\[
\min_{\Delta \theta} \sum_{x \in \Omega_0} \| I_0(x) - I_1(x') - \frac{\partial I_1}{\partial \theta} \Delta \theta \|_2^2.
\] (2.5)

Differentiating Eq. (2.5) with respect to \( \Delta \theta \) results in the expression

\[
\sum_{x \in \Omega_0} \left( \frac{\partial I_1}{\partial \theta} \right)^\top \left| I_0(x) - I_1(x') - \frac{\partial I_1}{\partial \theta} \Delta \theta \right|.
\] (2.6)

The optimal solution for \( \Delta \theta \) is the critical point of the derivative (Eq. (2.6)) and obtained as the solution to the normal equations given by

\[
\sum_{x \in \Omega_0} \left( \frac{\partial I_1}{\partial \theta} \right)^\top \left( \frac{\partial I_1}{\partial \theta} \right) \Delta \theta = - \sum_{x \in \Omega_0} \left( \frac{\partial I_1}{\partial \theta} \right)^\top \left( I_0(x) - I_1(x') \right).
\] (2.7)

At the next iteration of the optimizer, parameters are updated additively:

\[
\theta \leftarrow \theta + \Delta \theta.
\] (2.8)

In the terminology of Baker and Matthews [10], the algorithm is called “Forward Additive.” Forward, because the transformation maps pixel coordinates from the coordinate frame of the template to the coordinate frame of the input. Additive, because of the vector of parameters is updated additively after each iteration of the optimization algorithm.

The LK algorithm was invented early on in computer vision [95] for the purpose of estimating correspondences between the images of a stereo pair. Applications of LK now are numerous. The algorithm was also independently developed in Photogrammetry under the name: Least Squares Matching [58, 165].

### 2.1.2 Inverse Compositional (IC)

The original LK algorithm is versatile and applicable to a variety of problems and warps. However, when the Jacobian of the warp is not constant, LK becomes computationally expensive. This is because the linearization step happens at the coordinate frame of the input image, which is warped (changes) at every iteration.
For a special set of warps, Baker and Matthews [10] devise an efficient algorithm by (conceptually) interchanging the roles of the template and input images. The warps must form a group to allow for a compositional update of the parameters. Instead of the original LK objective, Baker and Matthews [10] propose:

\[
\arg\min_{\Delta \theta} \sum_{x \in \Omega_0} \| I_0(w(x; \Delta \theta)) - I_1(w(x; \theta)) \|^2,
\]

(2.9)

with a parameter update performed using “inverse” composition:

\[
\theta \leftarrow \theta \circ (\Delta \theta)^{-1}.
\]

(2.10)

Inverting the estimated parameters after every iterations is necessary because the linearization is carried out at the coordinate frame of the template.

Performing a first-order expansion of Eq. (2.9), we obtain:

\[
\sum_{x \in \Omega_0} \| I_0(w(x; 0)) - I_1(w(x; \theta)) + \frac{\partial I_0}{\partial \theta} \Delta \theta \|^2
\]

\[
= \sum_{x \in \Omega_0} \| I_0(x) - I_1(x') + \frac{\partial I_0}{\partial \theta} \Delta \theta \|^2,
\]

(2.11)

where we assume that, without loss of generality, \( w(x; 0) \) is the identity warp. The optimal update is obtained as the stationary point of the gradient and is given by the solution to the normal equations:

\[
\sum_{x \in \Omega_0} \left( \frac{\partial I_0}{\partial \theta} \right)^\top \left( \frac{\partial I_0}{\partial \theta} \right) \Delta \theta = \sum_{x \in \Omega_0} \left( \frac{\partial I_0}{\partial \theta} \right)^\top (I_0(x) - I_1(x')).
\]

(2.13)

In IC, the Jacobian of the warp is evaluated at \( x = w(x; \theta) \), with \( \theta = 0 \):

\[
J(y; \theta) = \left. \frac{\partial I_0(y)}{\partial \theta} \right|_{y=x_0}.
\]

(2.14)

The computational saving of the IC algorithm are significant. The Jacobian of the warp, and the inverse of the (Gauss-Newton approximation to the) Hessian need only be computed once at the beginning of the algorithm. The rest of the algorithm becomes a repeated application of image differences and matrix multiplication; operations that are efficient and amenable to parallelization.
The group requirement on the set of warps is not limiting. A large majority of warps commonly used in Computer Vision form a group, such as Perspective transformations (and their subgroups, c.f. Table 2.1). Finally, IC, up to a first-order analysis, is equivalent to the original LK formulation [10].

2.1.3 Other Variations

In addition to the IC algorithm, there are two more major variations. First, is the forward compositional (FC) algorithm [136], where it is possible to pre-compute the geometric part of the Jacobian. Second, is the Efficient Second order Minimization (ESM) algorithm [22, 99]. The ESM algorithm obtains a second-order approximation of the Hessian efficiently by exploiting gradients from both the template and the input images. A recent review of LK variations can be found in the work of Crivellaro et al. [34].

2.2 Direct Visual Odometry

Recently, with the introduction of the Kinect [168], Direct Methods [73] have resurfaced to produce robust, (semi-)dense and real-time algorithms for Visual Odometry (VO) [29, 28, 109, 46, 81, 54, 143, 93, 153, 8, 154, 108]. At their core, direct VO algorithms are an application of the Lucas-Kanade [95] (LK) algorithm with a nonlinear warp. The nonlinearity of the warp is the result of the reliance on depth as well as the perspective projections required to obtain the image of a 3D point onto a rigidly moving camera.
The LK algorithm, development, and variations are rich and versatile. In this review, we will focus on applications of LK to VO and VSLAM.

Amongst the first approaches to direct VO is the Quadrifocal warping algorithm by Comport et al. [29, 28]. The authors avoid the reliance on depth by exploiting the quadrifocal tensor between the four view of a rigidly moving stereo rig. Working with the quadrifocal tensor is complicated due to the high number of degrees of freedom. Comport et al. [29] rely on the quadrifocal tensor decomposition as two fundamental matrices and a trifocal tensor as their core warp. Recent work, however, has shown that non-linear warping with inaccurate depth can be used [46, 81, 8]. In fact, depth estimates need not be dense [4]. Even when the number of pixels used is sparse, the accumulated point cloud generates a densely populated 3D representation of the scene with enough fidelity for various robotic perception tasks as shown in Fig. 2.1.

The Microsoft Kinect has been a major force in the re-introduction of Direct Methods for VO. This is due to the availability of dense depth estimates along with RGB imagery in real-time. While using intensity-only constraints provide sufficiently accurate VO, one can also incorporate depth constraints. When using depth, however, one is presented with the challenge of having to compute a depth gradient, which may not be possible if depth estimates are sparse or compromised with large noise. A summary of the different constraints used in Direct VO is provided in Table 2.2.

Also, it is possible to use other sensors besides RGB-D in the versatile direct framework. A summary of different sensors used in direct VO is shown Table 2.3.

<table>
<thead>
<tr>
<th>Constraint type</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intensity only</td>
<td>Comport et al. [29, 28], Alismail and Browning [4], Klose et al. [83], Newcombe et al. [119], Lovegrove et al. [93], Steinbrucker et al. [143], Meilland et al. [109], Audras et al. [8], Scandaroli et al. [132]</td>
</tr>
<tr>
<td>Depth only</td>
<td>Fang et al. [50], Fang and Scherer [49]</td>
</tr>
<tr>
<td>Intensity and depth</td>
<td>Engel et al. [46, 45], Kerl et al. [81, 80], Meilland et al. [108, 109], Gutierrez-Gomez et al. [59], Tykkälä et al. [153, 154]</td>
</tr>
</tbody>
</table>

This direct approach to VO has the following advantages in contrast to feature-based: (i) More robustness in degraded scenes, (ii) virtually parameter free, and (iii) the ability to produce richer 3D reconstruction of the scene without additional computational cost.
Table 2.3: Summary of sensor type

<table>
<thead>
<tr>
<th>Sensor type</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>RGB-D</td>
<td>Steinbrucker et al. [143], Klose et al. [83], Fang et al. [50]</td>
</tr>
<tr>
<td>Monocular</td>
<td>Engel et al. [46, 45], Dafrty et al. [35], Newcombe et al. [119], Lovegrove et al. [93]</td>
</tr>
<tr>
<td>Stereo</td>
<td>Comport et al. [29, 28], Alismail and Browning [4], Omari et al. [125]</td>
</tr>
<tr>
<td>Omnidirectional</td>
<td>Meilland et al. [108], Mei et al. [105]</td>
</tr>
</tbody>
</table>

However, limitations include (i) the need for small motion between frames, as the core part of the algorithm relies on linearizing the cost function, and (ii) consistent appearance between frames, or brightness constancy.

The first limitation, namely the need for small motions, is readily addressed with modern hardware. Modern cameras operate at frame rates in excess of 60fps which produces densely sampled video for most robotic tasks. Even when inter-frame motion is not sufficiently small, one can resort to scale-space [90], which improves the basin of convergence and provides a sound method to address large motions. However, the second limitation, brightness constancy, is more challenging. In fact, illumination change is a major obstacle in efforts to extend Direct Methods over multiple views with real data.

In the following section we present a summary of previous work that attempts to tackle the issue of multiple view direct optimization and the problem of violations of the brightness constancy assumption in an LK-based framework.

2.3 Direct Estimation of Structure & Motion from Multiple Views

The power of feature-based BA arises from the ability to refine the estimate of camera position and scene structure over multiple views [152]. In this work, we aim at attaining the same level of accuracy using image data without relying on keypoint processing. Our goal is to increase the robustness of VSLAM and allow vision-only algorithms to perform in challenging environments where keypoint extraction and
accurate localization are not always possible.

Previous work on the estimation of the camera motion and scene structure over multiple views in a direct framework can be categorized into four main methods: (1) Alternating optimization of the state vector, (2) Filtering framework, (3) Simplifying assumptions, and (4) Reduction to keypoint-based geometric BA.

2.3.1 Alternating optimization

By alternating optimization we mean estimating a group of the desired variables separately while holding constant the rest. In the context of VSLAM, parameter groups include: rotation, translation and scene structure. The process is repeated for each of the parameters in turn until convergence. Examples of (direct) alternating minimization algorithms include the work of Mandelbaum et al. [100], Stein and Shashua [142], as well as Oliensis [124, 123].

The use of alternating optimization is motivated by computational efficiency. When a depth estimate is required per pixel, simultaneous estimation of all variables becomes computationally prohibitive due to the sheer number of pixels in an image.

While an alternating framework can be shown to work in some scenarios, it is difficult to examine the optimality of the solution. Also, it is challenging to characterize the convergence properties of the algorithm. In visual structure-from-motion (SFM) tasks the optimal solution is attained by joint optimization of the state vector [152]. If accuracy of the algorithm is the most important, then slightly reducing the density of reconstruction in favor of an optimal solution is recommended. In fact, this joint optimization (when performed correctly) is more efficient than other heuristics [152]. Hence, we avoid alternating frameworks due to their limitations and focus on joint estimation algorithms.

2.3.2 Filtering/recursive estimation

Filtering is a common and useful technique in Robotics and Computer Vision. In filtering approaches, previous estimates are “marginalized out” and information is summarized with a probability distribution [146].

The filtering approach (also called recursive, or casual) to direct VSLAM has been previously adopted by various researchers including Heel [66, 67] as well as Barron
and Eagleson [17]. The approach is also common in the context of multiple view stereo reconstruction, where the optimization objective is to estimate the depth per visible pixel assuming known camera motion [103, 159, 156].

Recent work by Strasdat et al. [146], however, have shown that BA is a more profitable strategy for VSLAM per unit of computation. In addition, optimality conditions in BA framework are easier to satisfy than a nonlinear filtering framework.

2.3.3 Simplifying assumptions

If one can assume the existence of certain structures in the scene, then it is possible to perform direct, joint multiple view VSLAM efficiently. The most common assumption is the existence of planar structures. The use of the planar world assumption has a rich history in robotics and has been applied early on for robot navigation tasks [117]. Recent work by Silveira et al. [139] and Mei et al. [105] demonstrate methods using multiple views to improve direct VSLAM using planes. When the scene is composed of planes, the warp becomes linear and more efficient to implement. Furthermore, planar patches can cover a large surface in the world that typically projects onto a large area in the image. When the world is mostly planar, it is also possible to perform Bundle Adjustment over multiple views as shown by Kaess [77], Ataer-Cansizoglu et al. [7], Taguchi et al. [149] and Salas-Moreno et al. [130]. One can also include blur and fold-in image degradations terms as part of the optimization [104]. Subsequently, one can process a large number pixels in the image in one go.

Of course, this assumption breaks down when the world is not composed of planar segments. Addressing general scenes with the planarity assumption becomes a challenging problem.

Another common use of planes in direct SFM is planar-parallax. If the motion of a plane (or a selection of planes) in the world has been compensated for, then we can estimate the depth of non-planar points by their parallax to a reference plane. Irani et al. [74] have demonstrated a multi-view direct method that improves on the two-frame case [131, 86]. Nonetheless, planar-parallax is not suitable for VSLAM as it relies heavily on the existence of sufficiently large planar surfaces in the world. Furthermore, the main goal of planar-parallax algorithms is the estimation of 3D structure, which is only a sub-problem of VSLAM.

Another closely related approach to direct VSLAM is the Volumetric method [118, 119].
Volumetric methods have shown excellent results for real-time dense VSLAM, albeit requiring sophisticated GPUs. Volumetric methods fundamentally rely on imposing a working volume in 3D space. This limits their applications to indoor environments, or scenes with a known range of depth. For instance, it would be challenging to apply a purely volumetric approach to unstructured outdoor scenes where many of the strong rotation constraints are located on the plane at infinity. Meilland et al. [109] provide an excellent summary of the benefits and limitation of volumetric methods as well as an approach to integrate some advantages of volumetric methods alongside a direct framework.

2.3.4 Reduction to geometric BA

Finally, one can combine the benefits of direct methods with the power of geometric Bundle Adjustment as demonstrated by Forster et al. [54]. Direct VO can be used to obtain precise estimates of camera pose in real-time. In a parallel thread, a selection of geometric keypoints are tracked, and keyframed over time in order to be used later in a sliding window geometric BA. However, this approach does not provide a solution to VSLAM in challenging environments that lack distinctive keypoints. Hence, the drift reducing properties of multiple views can only be used in scenes were keypoint processing is possible.

2.4 Violations of the Brightness Constancy Assumption

Since the seminal work of Lucas and Kanade [95] and Horn and Schunck [69] various algorithms have been developed to address the limitations of the brightness constancy assumption, especially in optical flow estimation [147]. In this section, we briefly review methods in the literature that address violations of the brightness constancy assumption. Algorithms can be categorized into the following: (1) Not relying on the brightness constancy assumption by using illumination insensitive objective functions, (2) estimating illumination variations as part of the optimization problem, (3) eliminating illumination artifacts using an image preprocessing step (pre-filtering).
2.4.1 Illumination insensitive objective

Perhaps the most intuitive solution to the brightness constancy assumption is to use an illumination insensitive objective. Such objective functions include the normalized cross correlation (NCC), and the Mutual Information (MI) [128].

Mutual Information has been successfully applied to register images from different modalities in the Lucas-Kanade framework [43, 37] as well as tracking of known 3D models [126, 32]. Similarly, maximizing the correlation coefficient has been applied to multi-modal image alignment [72], template tracking [132, 48] as well as structure-from-motion [100].

While robust cost metrics with intrinsic independence on illumination can handle challenging scenarios, their main limitations are twofold. One, accuracy of results relies heavily on an accurate approximation of the Jacobian (and Hessian) of the cost function. In many instances, the closed-form derivation of the Hessian yields a numerically non-positive definite matrix, which is not suitable for optimization. The other limitation is that we can no longer rely on least squares optimization. Instead, one must solve the problem with a general nonlinear optimization method. In template tracking problems, the number of variables is small and hence a general purpose optimizer provides satisfactory results. However, in the context of VSLAM the dimensionality of state-vector is large. In such scenarios, general purpose optimizers become slow and convergence is typically harder to attain when far from the optima.

More precisely, for large nonlinear optimization problems, we desire to formulate the problem as least squares if possible. Least squares problems are characterized by the small residual property. During linearization higher order derivatives of the objective vanish in comparison to the first order term ($J^TJ$). This implies a “free” and good estimate of the Hessian by only evaluating first-order partial derivatives. There exists a number of algorithms that take advantage of the special structure of least-squares problem to provide good time and convergence guarantees. In fact, solving least-squares problem can be considered a mature technology that we would like to exploit [121, ch. 10].

2.4.2 Estimating illumination variations

The most common approach to handle appearance variations is to model them explicitly as a multiplicative and an additive term. This is commonly known as gain + bias
model [11, 62, 19]. In the context of VSLAM, the (global) gain + bias model can account for illumination changes arising from the camera exposure control. This has been shown to perform well in a frame-frame direct VO from Kinect data [83]. However, in order to address more complicated appearance variations, we need to estimate the gain per pixel, or per region in the image. Hence, the main disadvantage of the gain and bias model is the explosion of the dimensionality of the state vector due to the additional photometric parameters.

Consider, for example, a multiple view optimization problem with 5000 3D points seen by 5 views. If we represent each camera with 6-vector, and each 3D point by a 3-vector, then the state vector is in $\mathbb{R}^{15030 \times 1}$. In order to make use of the gain and bias model, and without any prior assumptions on the scene, we may need to estimate a gain variable per 3D point per view and a bias term per image. This increases the state vector dimension by $5 \times 5000 + 5$ variables (assuming a gain variable per pixel and image) to become $\mathbb{R}^{40035 \times 1}$. This larger problem is now more difficult and more computationally demanding to solve.

Alternatively, one could estimate the illumination of the scene given a surface model. However, this is a “chicken-and-egg” problem as surface reconstruction is not a priori available. Other methods can be used to estimate illumination without a surface reconstruction [87, 27, 18, 44, 155] but typically rely on certain assumptions on scene, the material types, or require RGB data. Such requirements restrict the utility of a general VSLAM algorithm.

### 2.4.3 Image pre-processing techniques

Pre-processing the input images to eliminate illumination artifacts is another alternative to illumination insensitive image alignment. For example, pre-normalizing the image to be of zero mean and unit standard deviation is equivalent to a global gain and bias estimation step. However, this form of normalization can only address changes to due to camera shutter control, and fails to address other complicated variations arising from shadows and specular reflections.

Other approaches augment the registration with higher order information from the image such as first- and second-order gradient information [116, 140, 26] (or other filters [96]). Additional higher-order information improve robustness against illumination artifacts, but require fine tuning and selecting appropriate balancing weights for the intensity terms versus the other terms. Moreover, extraction of higher order information
from the image may exaggerate the sensor noise if the imagery is of low quality.

Another approach to illumination robust image alignment is using phase instead of intensity [53]. Phase is amplitude invariant. Hence, its robustness to changes in gain and bias. Nonetheless, spatial support near discontinues and occlusions is a limitation of phase-based methods [52].

Other techniques aim at applying simple filters to the image [97, 31]. However, they typically require RGB data and designed to work under natural sunlight illumination. They also “wash out” many of the useful frequencies in the image.

In optical flow estimation, Wedel et al. [161] propose the application of structure-texture decomposition to eliminate illumination artifacts from the image. The decomposition, however, is an iterative process that may not be suitable for online applications. Another risk of applying pre-filtering operations is that the filter might eliminate frequencies required to estimate certain degrees of freedom (c.f. [72]).

2.5 Geometric (Keypoint-based) Bundle Adjustment

Bundle Adjustment (BA) [152] is a well-established and mature optimization framework. Triggs et al. [152] provides an excellent mathematical coverage of the topic. Dellaert [40], and Hartley and Zisserman [64] provide a tutorial aimed for implementers. Due to its popularity in Robotics and Computer Vision, several software packages are available [2, 92]. Variations on BA with emphasis on efficiency and handling large scale problems can be found in papers by Sibley et al. [137], Kaess et al. [79, 78], Agarwal et al. [1]. We will defer the introduction of our notation to Section 4.2.

In the next section we present our completed work. The first part is the development of a direct VO algorithm for stereo data using disparity space warp function. The second is using (binary) feature descriptors to address illumination change in template tracking problems under arbitrary illumination variations.

In the next two chapter we present our approach to direct VO from stereo data (Chapter 3) and formulation of direct multiple view VSLAM (Chapter 4).
Chapter 3

Direct Disparity Space Visual Odometry

In this section we provide a summary of our Direct Disparity Space (DDS) algorithm. Additional detailed analysis and experimental results are available in a separate publication [4].

3.1 Disparity space

Consider a rectified stereo image with baseline $B$ and an upper triangular camera intrinsic matrix composed of the camera focal length $f$ and the principle point $c = (c_u, c_v)$. Without loss of generality, let the left image be the origin of the coordinate system. A point $x = (x, y, d, 1)^\top$ is an element of disparity space, where $x = u - c_u$, $y = v - c_v$ and $d = u - u_r$ is the disparity; the difference between the $u$-coordinate in the left image and its corresponding coordinate in the right image. Given this rectified stereo, the depth of an image point can be obtained with $Z = Bf/d$.

Consider two stereo pairs related via a rigid body transformation $T(\theta) \in SE(3)$ parameterized by $\theta \in \mathbb{R}^p$, where $p$ is typically 6, such that a 3D point $X = (X, Y, Z, 1)^\top$ is transformed into $X' = T(\theta)X$. This rigid-body motion relationship may be expressed in disparity space as

$$x' = \Gamma T(\theta) \Gamma^{-1} x, \quad (3.1)$$
where \( \equiv \) denotes projective equality up-to-scale, and \( \Gamma \) is a \( 4 \times 4 \) matrix that depends on the known stereo calibration and is given by

\[
\Gamma = \begin{pmatrix}
    f & 0 & 0 & 0 \\
    0 & f & 0 & 0 \\
    0 & 0 & 0 & fB \\
    0 & 0 & 1 & 0
\end{pmatrix}.
\]  

(3.2)

Demirdjian and Darrell [42] analyze the disparity space and show that it is a projective space \( \subseteq \mathbb{P}^3 \) with the important property that the measurement noise of the coordinates \( x \) is well-approximated with a Gaussian distribution.

### 3.2 Direct Visual Odometry

Let the intensity of a point \( x \) at the reference frame be given with \( I(\tilde{x}) \in \mathbb{R} \), where

\[ \tilde{x} = (x + c_u = u, y + c_v = v)\top. \]

With an abuse of notation, we will use \( I(x) := I(\tilde{x}) \).

After a rigid-body motion with \( T(\theta) \), we obtain the input image \( I'(x') \). Given an initialization \( \theta \), we seek to estimate a \( \Delta \theta \) — a small increment of pose parameters.
relating the two cameras — such that we minimize the sum of squared intensity error, or the photometric error given by

$$\Delta \theta^* = \arg\min_{\Delta \theta} \sum_{x \in \Omega} \| I'(w(x; \theta + \Delta \theta)) - I(x) \|^2,$$  \hspace{1cm} (3.3)

where $\Omega$ is a subset of pixel coordinates of interest in the reference frame, and $w(\cdot)$ is a warping function that depends on the parameter vector we seek to estimate. After every iteration, the current estimate of parameters is updated via an additive rule (i.e. $\theta \leftarrow \theta + \Delta \theta$). This process repeats until convergence, or some termination criteria have been met.

This formulation is the standard Lucas-Kanade (forward additive) algorithm [95] (Section 2.1.1). An efficient variation on the Lucas-Kanade algorithm is Baker & Matthews’ Inverse Compositional (IC) algorithm (Section 2.1.2). The IC algorithm makes two modifications to the error function that significantly improve efficiency. First, is to interchange the roles of $I$ (the reference/template image) with $I'$ (the input/current image). The other, is to compound incremental estimates using a compositional update rule instead of an additive one. Under the IC formulation we seek an update of the parameters $\Delta \theta$ such that

$$\Delta \theta^* = \arg\min_{\Delta \theta} \sum_{x \in I} \| I(w(x; \Delta \theta)) - I'(w(x; \theta)) \|^2.$$  \hspace{1cm} (3.4)

The optimization problem in Eq. (3.4) is nonlinear irrespective of the form of the warping function or the parameters. To obtain a solution, we perform a first-order Taylor expansion and arrive at the following closed form (normal equations):

$$\Delta \theta = (J^\top J)^{-1} J^\top e,$$  \hspace{1cm} (3.5)

where $J = (g(x_1)^\top, \ldots, g(x_m)^\top) \in \mathbb{R}^{m \times p}$. Here, $m$ is the number of pixels and $p = \lvert \theta \rvert$ is the number of parameters. Each $g$ is $\in \mathbb{R}^{1 \times p}$ and is given by

$$g(x) = \nabla I(x) \frac{\partial w}{\partial \theta},$$  \hspace{1cm} (3.6)

where $\nabla I = (I_u, I_v) \in \mathbb{R}^{1 \times 2}$ is the image gradient along the $u$- and $v$-directions. Finally,

$$e(x) = I'(w(x; \theta)) - I(w(x; \Delta \theta))$$  \hspace{1cm} (3.7)
is the error image.

At the next iteration of the optimization algorithm, parameters of the motion model are updated via the IC rule given by

\[ w(x, \theta) \leftarrow w(x, \theta) \circ w(x, \Delta \theta)^{-1}. \]  

(3.8)

We refer the reader to the excellent series by Baker & Matthews [10, 11] for a detailed treatment.

3.2.1 Algorithm

Given a reference image \( I \) with an associated disparity map and an input image after camera motion \( I' \), we seek to estimate the parameters of motion such that the expression in Eq. (3.4) is minimized. The warping function is given by:

\[ w : (\mathbb{R}^3 \times \mathbb{R}^6) \rightarrow \mathbb{R}^2 \]  

(3.9)

\[ w(x, \theta) = \pi (\Gamma T(\theta) \Gamma^{-1} x) + \begin{pmatrix} c_0 \\ 0 \end{pmatrix}, \]

(3.10)

where \( \pi (\cdot) \) performs homogenous division to bring back the point to Euclidean space. Finally, we add back the principle point \( c \) to obtain 2D pixel coordinates in the image plane.

In order to use a direct approach, we need to compute an analytic expression of the Jacobian with respect to the parameters around the identity \( \theta = 0 \). Using the Lie algebra parameterization of rigid transformations, i.e. a twist, \( \theta = (\omega_x, \omega_y, \omega_z, \nu_x, \nu_y, \nu_z)^\top \in \mathbb{R}^6 \), we obtain the Jacobian of the warping function in Eq. (3.10) per point \( x \) as [68]:

\[
\nabla I \frac{\partial w}{\partial \theta} \bigg|_{\theta=0} = g(x) = \begin{pmatrix} -fI_y + \alpha y/f \\ fI_x - \alpha x/f \\ yI_x - xI_y \\ \beta I_x \\ \beta I_y \\ \alpha \beta/f \end{pmatrix}^\top \in \mathbb{R}^{1 \times 6},
\]

(3.11)
where $\nabla I = (I_x, I_y)$ is the image gradient, $x = c_u - u$, $y = c_v - v$, $d = u - u_r$, with

$$\alpha = xI_x + yI_y, \quad \text{and} \quad \beta = d/B.$$  

For $m$ pixels, we stack the values of Eq. (3.11) into an $m \times 6$ matrix and obtain an update of parameters $\Delta \theta$ by solving the normal equations in Eq. (3.5). After every iteration the pose estimate is updated using the inverse compositional update rule given by Eq. (3.8).

### 3.2.2 Robustness

The least squares optimization (Eq. (3.5)) is sensitive to outliers. In order to obtain a robust estimate we replace the squared error with a robust cost function. Choice of the robust function is rather arbitrary and can only be determined experimentally [167].

We experimented with several cost functions and found Tukey’s bi-weight [21] to perform the best. This is possibly due to suppressing high residuals instead of only reducing their influence. The bi-weight function for a residual $r_i \in \mathbb{R}$ and parameter/cutoff threshold $\tau \in \mathbb{R}$ is given by

$$\rho(r; \tau) = \begin{cases} 
(1 - (r_i/\tau)^2)^2 & \text{if } |r_i| \leq \tau; \\
0 & \text{otherwise.} 
\end{cases} \tag{3.12}$$

The cutoff threshold $\tau$ is set to 4.6851 to obtain a 95% asymptotic efficiency of the normal distribution. The threshold assumes normalized residuals with unit deviations. For this purpose, we use a robust estimator of standard deviation. For $m$ observations and $p$ parameters, the robust standard deviation is given by:

$$\hat{\sigma} = 1.4826 \left[ 1 + 5/(m - p) \right] \text{median} |r_i|.$$

The constant 1.4826 is used to obtain the same efficiency of least squares under Gaussian noise, while $[1 + 5/(m - p)]$ is used to compensate for small data [167]. In practice, $m \gg p$ and the small data constant vanishes.

In summary, given a list of residuals $r = (r_1, \ldots, r_m)^\top$, where each residual is given by:

$$r_i = I'(w(x_i; \theta)) - I(w(x_i; \Delta \theta)). \tag{3.14}$$

25
we compute a robust estimate of the standard deviation $\hat{\sigma}_r$ using Eq. (3.13) and compute the weight per residual as $w_i = \rho (r_i / \hat{\sigma}_r)$. By concatenating the weights into an $m \times m$ diagonal matrix $W$, we may obtain an estimate of the parameters at every iteration by solving with the following weighted normal equations:

$$\left( J^T W J \right) \Delta \theta = J^T W e. \quad (3.15)$$

### 3.2.3 Pixel selection

Traditionally, direct methods are associated with the concept of dense, or semi-dense, algorithms that make use of all possible pixel information. Intuitively, the use of as many as possible data points could increase robustness. However, the large number of pixels typically used in direct algorithms incur a high computational cost necessitating implementation on parallel architectures such as high-end GPUs.

The literature on pixel selection for direct pose tracking is sparse. In the case of optical flow, the seminal work of Shi & Tomasi [135] introduced a feature selection method based on the "textureness" of the patch surrounding the pixel. The textureness score is obtained by analyzing the Eigenvalues of the design matrix, which is composed of image gradients.

For pose tracking applications, Dellaert & Collins [41] propose a method that selects pixels that constrain each degree-of-freedom (DOF) the most. A known motion prior is required, however. Meillard et al. [108] propose an alternative based on recursively sorting each dimension of the Jacobians and greedily keeping elements with the highest magnitude. Simpler methods include discarding pixels with a gradient magnitude smaller than a fixed threshold [83].

In this work, we show that direct camera tracking need not be dense. By reducing the number of pixels, the algorithm runs in real-time on a single CPU core without compromising accuracy or robustness. Our pixel selection is based on the feature “binning/bucketing” idea common to feature-based methods [120], where the image is virtually split into a grid/buckets, and a certain number of pixels with strong cornerness score is kept in each bucket.

In our case, the influence of a pixel correlates, to an extent, with its gradient magnitude. For example, a pixel with no gradient does not contribute to the optimization as its contribution to the Jacobian in Eq. (3.11) vanishes. Hence, we perform our pixel
selection using the gradient magnitude as a substitute for the cornerness map. Pixels with a local gradient magnitude maxima in a neighborhood of $3 \times 3$ pixels are used for pose estimation. In contrast to feature-based methods, we do not enforce a maximum number of features per grid cell.

### 3.2.4 Additional implementation details

Our algorithm does not require elaborate parameter tuning or specialized heuristics. The only tunable parameters are the stereo algorithm parameters, which depend on the dataset. We use a basic block matching stereo (as implemented in OpenCV\(^1\)). Stereo parameters include, SAD window size and disparity range. To address large motions (and speed up the convergence rate) the algorithm is implemented in a scale space pyramid. We do not scale down the disparity image to avoid interpolation across occlusion boundaries. Instead, disparities for coarser levels of the pyramid are interpolated from the disparity map computed at the finest level using nearest-neighbor interpolation. Each level of the pyramid is scaled by a factor of $\frac{1}{2}$ of the previous image size and smoothed with a Gaussian filter prior to downsampling with bilinear interpolation. Convergence is determined if the norm of the estimated parameters is less than $1 \times 10^{-6}$, change in parameters is less than $1 \times 10^{-8}$ or a maximum number of iterations is reached. The maximum number of iterations was set to 300 on the finest level, and 50 for all other levels.

### 3.3 Experiments and Results

In this section, we evaluate the performance of our algorithm on different datasets including outdoor and indoor environments. For the outdoor datasets we process all frames. That is, we do not perform any keyframing, even when the robot is stationary. We also do not perform any global optimization/bundle adjustment or make use of other sensors. In the following, we will call our algorithm DDS: Direct Disparity Space.

\(^1\)Using MATLAB R2013b [http://www.mathworks.com/help/vision/ref/disparity.html](http://www.mathworks.com/help/vision/ref/disparity.html)
3.3.1 KITTI data

We evaluate the performance of our algorithm on the KITTI odometry benchmark [57] in comparison to two open source algorithms targeted for robotic applications: (1) VISO2 [56] and (2) FOVIS [70]. We use both algorithms with the authors’ default parameters, which perform well.

Results on KITTI data are summarized in Fig. 3.2. Our algorithm’s average translation error is 2.35% and the average rotation error is 0.0058°/m, which are accurate for a frame-frame method without pose initialization. In particular, our rotation error is close to, and sometimes better, than some multi-frame methods on the KITTI benchmark. The main sources of error appear in estimating the translation of the camera at high vehicle speeds. High speed driving produces larger baseline between images and violates the small motion assumption.

Interestingly, rotation error for both FOVIS and ours (DDS) are better than VISO2. This is potentially due to more accurate rotation estimation results when using image intensities directly. In fact, most of VISO2’s rotation drift appears to be in roll estimates and consequently camera height. We hypothesize that this is related to the small vertical FOV of the camera. In contrast, direct methods are able to better exploit this reduced FOV by not relying on the accuracy subpixel feature localization.
### Table 3.1: Wean hall data summary

<table>
<thead>
<tr>
<th>focal length</th>
<th>baseline</th>
<th>distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\approx 3.88$ mm</td>
<td>0.12 m</td>
<td>$\approx 294$ m</td>
</tr>
</tbody>
</table>

#### 3.3.2 Wean hall (indoor data)

This dataset was collected with a Bumblebee2 stereo color camera of resolution $640 \times 480$ px$^2$ at $\approx 30$ Hz [5]. A summary of the data is shown in Table 3.1. The camera was mounted on a LAGR robot. For ground truth, we use a 2D estimate of robot pose using calibrated wheel odometry combined with an accurate gyroscope. This is an approximate ground truth, but it is reasonable as the indoor environment is flat. The camera’s raw output is a Bayer pattern, which we interpolate to a color image prior to use for motion estimation. The Bayer pattern causes a reduction in resolution in comparison to native grayscale output.

The dataset features strong specular reflections on the ground, lack of texture in some areas as well as high frequency repetitive texture in others. An example is shown in Fig. 3.3. The robot was driven at an average speed of $\approx 0.7 \text{ m s}^{-1}$.

Due to the high frame-rate of the camera, we implement a keyframing strategy based on the magnitude of the estimated motion. The pose of each non-keyframe is initialized with the current estimate of pose until the motion magnitude is large enough. Upon keyframing, we reset the pose initialization to the identity. For the results shown here, we keyframe when the estimated translation magnitude is 30 cm, or when any of the estimated rotation angles exceeds 5°. Results are shown in Fig. 3.3 and Fig. 3.4.

#### 3.3.3 Dense 3D reconstruction

We demonstrate a simple scheme to obtain sufficiently dense 3D reconstruction using our algorithm. The output of our algorithm after every keyframe consists of an estimate of the camera pose, as well as the set of disparity space point and their IRLS weights upon convergence.

$^2$Data available on [http://www.cs.cmu.edu/~halismai/](http://www.cs.cmu.edu/~halismai/)

$^3$See [http://www.nrec.ri.cmu.edu/projects/lagr/](http://www.nrec.ri.cmu.edu/projects/lagr/). Our robot is slightly modified to use a single stereo camera titled towards the ground, and equipped with an accurate fiber optic gyro from KVH, model # DSP-3000.
Figure 3.3: Example images from the Wean hall dataset.

Figure 3.4: Top view of estimated path for ground truth (—), VISO2 (---) and DDS (——).
After every keyframe, we select points with Tukey weights from the third quartile (75 percentile) and with range of at most 30 m. We triangulate the points, and project them to the world coordinates using the current estimate of the keyframe’s pose. As the pixel selection scheme is not based on features, the selected points over multiple frames do not correspond to a single 3D point space. Hence, the overlap between frames will consist of mostly distinct 3D points that produce a dense reconstruction of the environment.

Examples of our reconstruction are shown in Fig. 3.1, and Fig. 3.5. Note, disparity maps were obtained via block matching and include a large amount of noise and outliers. The 3D reconstruction results indicate the accuracy of the method over a short sequence of frames as well as robustness against outliers.

### 3.3.4 Visual odometry from stereo thumbnails

We also evaluate the robustness of the algorithm using low resolution images (178 × 54) and compare it against VISO2 using full resolution images (1241 × 376). Results are shown in Fig. 3.6. Even with the low stereo resolution our algorithm outperforms VISO2 and remains accurate and robust.
Figure 3.6: Result from KITTI Seq. 02. Ground truth is shown in ( ), VISO2 path is in ( ), and DDS is in ( ). The ( ) indicates the start of the sequence, and ( ) indicates the final location. Our results are generated from an image of size $178 \times 54$, while VISO2 results are generated from the full resolution $1241 \times 376$. 
3.3.5 Pixel selection

Not all pixels contribute equally to the cost function; only a small number of pixels contribute towards the error function [41]. The simplest approach to pixel selection is to discard pixels with gradient magnitude less than a pre-specified fixed threshold. The rational is that a pixel with zero gradient does not contribute to the error function. However, the mere magnitude of the gradient is not a sufficient predictor of a pixel’s performance. For instance, restricting the optimization to pixels with a high gradient magnitude might bias the solution in undesirable ways.

To illustrate this, we run our algorithm using all available pixels with an absolute gradient magnitude greater than a threshold. The average performance on the KITTI benchmark training data is illustrated in Fig. 3.7.

As shown in the figure, including all possible pixels is suboptimal. Similarly, selecting pixels with a very high gradient magnitude is suboptimal as well. A good threshold for the tested optics and the benchmark environment is within $10\%$ of the image dynamic range.

3.4 Discussion & Summary

Scene points on the plane at infinity are independent of camera translation and can be used for rotation estimation and calibration. In contrast to other work, our algorithm can make use of points at infinity without special handling. This is particularly useful for outdoor applications and we believe leads to improved rotation estimates.

We did not observe a need to use sophisticated stereo matching algorithms. Indeed, our stereo matching is very straightforward SAD block matching with limited disparity range resolution. Enhanced stereo may improve the accuracy and/or convergence speed at the expense of more computation time for stereo. Improvements are not guaranteed as (semi) global stereo methods may over smooth the estimated disparities. This issues remains to be experimentally validated.

In this work we dealt with the problem of pose estimation only (camera tracking). Two important improvements are possible. The first would be structure/disparity refinement. We can include disparity refinement in the same pose tracking framework by using observations from the right image. Another possibility is modeling disparities with some surface representation (c.f. [139]). The second important improvement
Figure 3.7: Average performance on KITTI training data with different absolute gradient magnitude cutoff thresholds. The input images are converted to floating point prior to computing the gradients, and their range is kept $\in [0, 255]$. Detailed evaluation plots are shown in Fig. 3.8.
Figure 3.8: Detailed performance on KITTI benchmark for various absolute gradient magnitude cutoff thresholds. See Fig. 3.7 for a summary.
is integrating information from multiple frames in a bundle adjustment/filtering framework. This, in fact, is necessary to reduce drift over long sequences. Both improvements are good avenues of future work. This is the topic of Chapter 4.

In this chapter, we presented a direct framework for visual odometry using a warping function in disparity space (DDS). The algorithm is shown to be efficient, robust and accurate even with low resolution images. Experiments illustrate the applicability of the algorithm to various environments with little to no manual parameter tuning. Finally, we have also shown that direct camera tracking can achieve accurate and robust performance while using only a fraction of the image data via a simple pixel selection strategy.

A limitation of direct methods is the reliance on the brightness constancy assumption. The next section presents our work on using feature descriptors to address appearance variations.
Chapter 4

Joint Multiple View Direct VSLAM

In this chapter, we present our first proposed contribution: a framework for the joint estimation of structure and motion (SFM) directly from image using multiple view. We will refer to this as Lucas-Kanade Bundle Adjustment (LK-BA). Our goal to allow Direct algorithms to reduce drift in situations where loop closure, or graph optimization are not possible.

Prior to the introduction of our proposed Lucas-Kanade BA, we briefly review Geometric BA for VSLAM in order to introduce notation.

4.1 Geometric BA

The goal of geometric BA is to obtain an optimal estimate of the camera pose, the world’s structure and possibly calibration parameters from a set of overlapping images. BA is a large nonlinear geometric parameter estimation problem. Due to the special nature of variable interactions, BA structure is sparse. Exploiting the sparsity structure of BA problems is key to efficient implementation. In fact, for large scale problems, dense linear algebra becomes intractable.

The first step in BA, as with all other optimization problems, is defining the objective function. The step is also known as modeling. In geometric BA, the optimal choice, or the Gold Standard [64], is the reprojection error, which corresponds to a physically motivated measurement. Given m cameras (pose parameters) \( \{\theta_i\}_{i=1}^m \) and n world points, \( \{X_j\}_{j=1}^n \) with associated measurements/observations \( x_{ij} \), the reprojection error
is the Euclidean distance between the measured pixel coordinates and their predictions by the camera projection model. The reprojection error is given by

\[ e_{ij}(x_{ij}; \theta_i, X_j) = \delta_{ij} \|x_{ij} - \tilde{x}_{ij}\|_2, \]  

(4.1)

where \(\tilde{x}_{ij}\) is the predicted feature position given pose and structure parameters as provided by the camera projection model \(\pi(\cdot)\)

\[ \tilde{x}_{ij} = \pi(\theta_i, X_j). \]  

(4.2)

The function \(\delta_{ij}\) is an indicator with a non-zero value when the \(j\)-th world point \(X_j\) is visible in the \(i\)-th view.

Given a number of cameras, \(m > 2\), and a number of world points \(n\), the goal of BA is to estimate the parameters of pose and structure that minimize the sum of squared reprojection errors. Let \(\Theta\) indicate the joint state vector,

\[ \Theta = (\theta_1^\top, \theta_2^\top, \ldots, \theta_m^\top, X_1^\top, X_2^\top, \ldots, X_n^\top)^\top, \]  

(4.3)

we can write the optimization problem as

\[ \Theta^* = \arg\min_{\Theta} \sum_{i=1}^m \sum_{j=1}^n e_{ij}^2(x_{ij}; \Theta). \]  

(4.4)

The optimal solution starts with an estimate of parameters, \(\Theta\), in the vicinity of the local minima and iteratively seeks to estimate an incremental update of a small magnitude \(\Delta \Theta\). The problem now becomes, given \(\Theta\), find a \(\Delta \Theta\) such that

\[ \Delta \Theta^* = \arg\min_{\Delta \Theta} \sum_{i=1}^m \sum_{j=1}^n e_{ij}^2(x_{ij}; \Theta + \Delta \Theta). \]  

(4.5)

Due to the large problem, and the intractability of computing the Hessian, an estimate at each iteration is obtained using a suitable optimization algorithm. Common solvers include: Levenberg-Marquardt (LM), Gauss-Newton (GN), and DogLeg.

For GN-type algorithms, first-order expansion of Eq. (4.5) yields the following lineariza-
tion

\[ \sum_{i=1}^{m} \sum_{j=1}^{n} |e_{ij}(x_{ij}; \Theta) + \delta_{ij} J(\Theta) \Delta \Theta|^2, \]  

(4.6)

where \( J(\Theta) \) is the Jacobian of the reprojection error in Eq. (4.1) evaluated at the current estimate of parameters \( \Theta \).

**Sparsity**

BA problems are typically large involving hundreds, or thousands of cameras, as well as thousands of world points. However, we observe that the Jacobian vanishes unless the \( j \)-th point is visible in the \( i \)-the view. Hence, the resulting linear system of equations is sparse, with a pattern similar to the own shown in Fig. 4.1.

**Error modeling**

So far, we have assumed a squared loss. When using a squared loss model, we assume that errors in feature localization are distributed according to a Normal (Gaussian) distribution. In which case, minimizing the sum of squares maximizes the likelihood. If feature localization errors are not Normally distributed, an appropriate distribution must be used to guarantee the optimality of the estimate [152, 64]. However, in the limit of many observations, errors tend to be normally distributed.

Another important consideration is outlier rejection. Robust estimation frameworks are designed to tolerate a certain degree of errors [144], but eliminating outliers from the outset in a screening stage is usually important to improve the solution’s efficiency and precision.

Ensuring an accurate model for the distribution of errors, as well as an explicit outlier screening process are important for a good performing BA as feature correspondences are fixed. The best performance we hope to achieve is defined using those fixed correspondences. Hence, any errors or biases in correspondence estimation will transfer directly to the estimated solution. In fact, the effect of feature localization errors can be significant [112]. Addressing such errors is a challenging problem. Feature localization covariances can mitigate some errors, but not always [47, 150].
BA in VSLAM

The structure of BA in robotic VSLAM is even more special. Images in typical sequences are collected sequentially with limited amount of overlap between views. For example, a 3D world point may be visible in only a few views before its track is lost. Hence, BA problems are implemented in a sliding window fashion over keyframes [85, 148]. Keyframes are usually selected to maximize the baseline between views to better constrain the refinement of pose with forward camera motion. The selection follows various heuristics relying on statistics from feature tracking and VO pose estimation [120, 145, 84].

Given the special structure of BA in VSLAM, where feature points cannot be tracked for a long time, an accurate model of keypoint localization error becomes important. Recently, Badino et al. [9] has empirically demonstrated that keypoint localization errors are better modelled with a Laplace distribution, then the commonly employed Gaussian. Taking the error distribution into account to correct localization errors over time considerably improves the performance of VSLAM – even without an explicit BA step. This approach is currently the state-of-the-art in the KITTI VO benchmark.

Of course, it is not possible to use geometric BA in situations where feature localization is a difficult task. Next, we present our proposed algorithm, which is a variant of BA that does not rely on (fixed) correspondences.

4.2 Joint Multiple View Direct Estimation (LK-BA) [Proposed Work]

The formulation of BA is not restricted to the geometric reprojection error. In principle, any suitable error metric may be used. Here, we propose to go back to the raw images and use functions of intensity as the error model. Afterall, raw intensities are the fundamental elements of an image (and feature-based methods).

Conceptually, our optimization framework attempts to address the question: Given an initial estimate of camera pose and scene structure (state vector), as well as pixel locations of interest with an indication of which frames they might be visible in, how to refine the initial estimate of the state vector such that the correct correspondences are found? This is explained in detail next.
Referring to Fig. 4.1 we consider a collection of \( n \) 3D points in the world coordinate system \( \{ X_j \}_{j=1}^n \) imaged by \( m \) calibrated cameras parameterized by a vectorial representation of pose \( \{ \theta_i \in \mathbb{R}^6 \}_{i=1}^n \). We assume that we are given an appropriate initialization of the 3D structure and camera pose, such that the projection of the \( j \)-th point onto the \( i \)-th view is given by the following warp:

\[
^i x_j = w(X_j; \theta_i) = \pi(T(\theta_i)X_j),
\]

where \( \pi(\cdot) \) is the projection onto the image plane using the intrinsic camera matrix followed by homogeneous division. The \( 4 \times 4 \) matrix \( T(\theta) \) is the rigid-body transform corresponding to the parameter vector \( \theta \).

We obtain this initialization from our previous visual odometry work [4], but other sources of initialization are suitable. Initialization of the 3D structure is obtained with a correlation-based stereo algorithm (block matching [24]).

There are three main steps in the algorithm:

1. Pixel selection,
2. Putative track establishment and outlier rejection, and
3. Multiple view optimization

The steps are summarized next.

### 4.2.1 Pixel Selection

In our framework, we detect pixel locations of interest at every frame and extract a fixed patch that we will designate as the template/reference patch. Currently, pixel selection is performed by detecting the local maxima of the absolute gradient magnitude in the left stereo image over a \( 3 \times 3 \) neighborhood (c.f. [4] for details). We select a subset of those relatively high frequency content pixels based on their score/weight that we obtain from an M-Estimator framework during the VO step. The threshold is set to be above \( 80\% - 90\% \) confidence from the estimator, which results in \( 5 - 10 \) thousand illegible pixels per frame. To reduce bias in the systems, pixels are selected to cover the entire the field of view of the camera. We note, we select only integer pixel locations; there are no feature extraction, or localization step.
The reference patch is chosen to be in the first frame it was detected in. In the future, we expect to revisit this decision and select a reference patch in an image frame with the least amount of perspective distortion, in a spirit similar to reference patches selection in multi-view stereo [55].

Note, having a fixed patch is important. Otherwise, the optimization problem is under-constrained, and the resulting linearized system of equations will become singular.

4.2.2 Putative track establishment and outlier rejection

The next step in our pipeline is to verify that a selected pixel is indeed visible in a list of other views. The aim of this step is to eliminate outliers and occluded pixels prior to the optimization as an explicit screening step. We perform this by projecting the 3D point corresponding to the selected reference patch onto the desired view given the pose initialization. Then, we measure the strength of correlation against the reference patch using the Zero-mean Normalized Cross Correlation (ZNCC) over a $5 \times 5$ patch. We found that this simple test eliminates occluded pixels and erroneous putative correspondences. ZNCC threshold is fixed to the value of 0.8. This steps usually rejects a few hundred pixels.

The output of this step indicates if we should initialize a putative “track” for this
Table 4.1: Putative correspondences. The ZNCC score indicates if a certain view should be added to the list of potential views for correspondence search. For example, for patch \( p_2 \) its reference/fixed frame was assigned to frame 0. Putative tracking check indicates that the patch may be visible in frames 2 and 3 as indicated by X, while it may be not visible in frame 1 as indicated by 0.

<table>
<thead>
<tr>
<th>Patch id.</th>
<th>Ref. frame #</th>
<th>View #</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1 2 3</td>
</tr>
<tr>
<td>( p_1 )</td>
<td>0</td>
<td>- X X X</td>
</tr>
<tr>
<td>( p_2 )</td>
<td>0</td>
<td>- 0 X X</td>
</tr>
<tr>
<td>( p_3 )</td>
<td>1</td>
<td>X - X 0</td>
</tr>
<tr>
<td>( p_4 )</td>
<td>2</td>
<td>X X - 0</td>
</tr>
</tbody>
</table>

selected pixel in the given view. This is illustrated in Table 4.1.

4.2.3 Multiple view optimization

Given the list of putative tracks, we perform direct multiple view estimation jointly over structure and camera parameters. More precisely, we desire to solve the following nonlinear least squares problem:

\[
\arg\min_{\Theta} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathcal{E}_{ij}(\Theta = (\theta_1, \theta_2, \ldots, \theta_m, X_1, X_2, \ldots, X_n)\top) \tag{4.9}
\]

\[
\mathcal{E}_{ij}(\Theta) = \frac{1}{2} \delta_{ij} \| p(I_*(x_j)) - p(I_i(w(X_j; \theta_i))) \|_2^2, \tag{4.10}
\]

where \( \delta_{ij} \) indicates if the \( j \)-the point projects onto the \( i \)-th image,

\[
\delta_{ij} = \begin{cases} 
1 & \text{if } X_j \text{ projects onto camera } \theta_i; \\
0 & \text{otherwise.}
\end{cases}
\]

The expression \( I_*(x) \) denotes the intensity of the reference patch at pixel location \( x = (x, y)\top \), while the expression \( p(\cdot) \) denotes extracting a rectangular patch of a certain size from the image at the specified pixel location. This reference patch remains “fixed” and does not depend on the estimated parameters.
Referring to the objective for a single block of parameters in Eq. (4.10). Let $\theta_i$ and $X_j$ be an appropriate initialization of the parameters. We seek to find an update on the parameters $\Delta \theta_i$ and $\Delta X_j$ that minimizes the squared residual given by

$$E(\Delta \theta_i, \Delta X_j) = \frac{1}{2} \|p(I_i(x_j)) - p(I_i(w(X_j; \theta_i + \Delta \theta_i)))\|^2_2. \quad (4.11)$$

A first-order Taylor expansion of the objective in Eq. (4.11) about $\Delta \theta_i = 0$ and $\Delta X_j = 0$ yields:

$$E(\Delta \theta_i, \Delta X_j) \approx \frac{1}{2} \|p(I_i(x_j)) - p(I_i(w(X_j; \theta_i))) - \frac{\partial p}{\partial \Theta} \Delta \Theta\|^2_2, \quad (4.12)$$

where $\Theta = (\theta_i^T, X_j^T)^T$ is the state vector. The solution to the linearized system of equations in Eq. (4.12) is obtained by solving the normal equations. The process is repeated as necessary until convergence, or a maximum number of iterations is reached.

We note here that the Jacobian of the problem in sparse admitting a form similar to the Jacobian of a geometric BA problem [152]. This is the main reason we are able to fit this large optimization in memory and be able to solve it efficiently. In our implementation, the nonlinear optimization is solved using a trust-region algorithm (Levenberg-Marquardt [89, 101]). The sparse normal equations are solved with Sparse Cholesky factorization [38, ch. 4].

### 4.3 Preliminary Results

An example result is illustrated in Fig. 4.2. In this example, we use a $3 \times 3$ image patch and perform the optimization over a sliding window of three frames. To improve robustness, we use the Huber norm [71] instead of the squared norm.

We are able to obtain a significant improvement over frame-frame VO in an outdoor scenario with large inter-view baseline. Furthermore, the optimization was performed under the brightness constancy assumption. The only form of smoothness of the objective is provided by the $3 \times 3$ patch of raw image intensities. However, in more challenging scenarios where an image patch is not enough to smooth out appearance changes the algorithm does not improve on the VO baseline initialization.

This leads us to our next proposed work in: Handling appearance variations using feature descriptors.
Figure 4.2: Preliminary results of our direct multiple view VSLAM algorithm without modeling appearance variations. The improved result is on par with the state-of-the-art (c.f. Badino et al. [9]).
Chapter 5

Descriptor Constancy with BitPlanes

5.1 Introduction

Feature descriptors, such as SIFT [94], SURF [20] and HOG [36] among others [110] have revolutionized research in object detection, recognition and the correspondence problem in general. Part of the power of feature descriptors lies in their invariance with respect to scale change and moderate appearance variations. While descriptors were designed to work with sparsely sampled keypoint locations, recent work has shown the benefit of using densely sampled features descriptors. Fei-Fei and Perona [51] as well as Bosch et al. [23] have applied dense SIFT for scene classification with compelling results. Liu et al. [91] have taken the idea farther by developing an image alignment framework based on SIFT to find dense correspondences between large baseline views. More recently, Bristow and Lucey [25] have demonstrated the Lucas-Kanade gradient-based alignment algorithm on densely sampled descriptors with applications to image matching across object categories.

In this work, we are interested in using feature descriptors as a means to address appearance variations for multiple view direct VSLAM. To this end, we consider binary feature descriptors due to their efficiency and invariance to monotonic illumination variations. More specifically, we consider the Census Transform [166] (also known as LBP: Local Binary Patterns [122]) as an illustrative case.

The computation of the Census descriptor is illustrated in Fig. 5.1, which consists of a series of comparisons in a local neighborhood. By definition, the Census Transform
Figure 5.1: Illustration of the Census Transform (CT). A $3 \times 3$ patch is centered around a pixel with intensity value equals to 42 (Fig. 5.1a). Each bit in the Census digit is the result of comparing the center pixel with its neighbors (Fig. 5.1b). The final Census signature is obtained by reading the output of the comparisons in a certain order (Fig. 5.1c). If we read the bits in row-major order starting at the top left corner, then the Census signature is 11000011, which has the value 195 when converted to a decimal. An example of the CT on a real image is shown in Fig. 5.2.

![Figure 5.2: Example of the Census transform applied to a real image.](image)
is invariant to monotonic appearance changes, which is a desirable property when working with real robotic data. A similar class of descriptors known as Local Binary Patterns (LBP) have been developed by Ojala et al. [122]. More recent variations on the binary testing idea has resulted in other descriptors that include BRISK [88], FREAK [3] and ORB [129], which have been commonly used in VSLAM due to their efficiency and distinctive power [114, 65].

The Census transform is a popular tool in illumination robust optical flow. Stein [141] has applied a modified version of the Census transform for optical flow computation targeted for automotive applications. Recently, a theoretical analysis of why the Census transform is good for robust optical flow was conducted by Hafner et al. [60, 61]. Performance of the Census Transform is illustrated in Fig. 5.3 and Fig. 5.5.

5.1.1 The Hamming Distance in Nonlinear Least Squares

To exploit the full power of binary descriptors we must match them under a norm designed for binary vectors, such as the Hamming distance [63]. This poses a problem as the Hamming distance is non-smooth. Let a and b be two bit strings of length B. Their Hamming distance is defined as the number of coordinates of mismatched bits:

\[ d_H(a, b) = \sum_{i=1}^{B} \delta(a_i \neq b_i), \]  

(5.1)

where \( \delta(\nu) \) equals 1 iff \( \nu \neq 0 \). Consider for example the two 8-bit strings \( a = [1, 1, 0, 0, 1, 0, 0, 0] \) and \( b = [1, 1, 0, 0, 0, 1, 0, 0] \), their distance under the Hamming norm is equal to 2 as they differ at two locations. However, if we treat them as decimals, then \( a = 100 \), and \( b = 98 \), and their distance under the squared Euclidean norm is equal to 4. Hence, simply using the bit strings as decimals for the purposes of minimizing the squared distance is not equivalent to minimizing their Hamming distance.

Notwithstanding, if we compute the squared Euclidean distance by expanding the decimal digit into a binary vector we obtain the same result as the Hamming distance,

\[ \sum_{i=1}^{b} |a_i - b_i|^2 = \sum_{i=1}^{b} \delta(a_i - b_i). \]  

(5.2)
Figure 5.3: Evaluation of different cost functions without illumination change (autocorrelation). The input image is shown in Fig. 5.4. Each of the cost function is evaluated over a grid of $x$- and $y$-translations. The minima of all functions correspondences with the ground truth translation of $(0, 0)$ pixels. Please note, the ZNCC is negated to become a dissimilarity measure.
Figure 5.4: Image with illumination change given by $255 \left( \frac{\alpha I + \beta}{255} \right)^\gamma$, where $\alpha$ is a multiplicative gain, $\beta$ is an additive bias and $\gamma$ is used for Gamma correction. Registration results for different cost functions are shown in Fig. 5.5 with the Census Transform having the sharpest minima (and the least affected by interpolation artifacts.

This, of course, is valid when the vector-space is a binary alphabet \( i.e., a, b \in \{0, 1\}^B \).

To achieve the same effect of the Hamming distance in a gradient-based continuous optimization framework, we propose to represent each coordinate of the descriptor as a channel in a multi-channel image. When working with binary values, we obtain a multi-channel binary representation. We call this representation “BitPlanes” as shown in Fig. 5.6.

In addition to obtaining the Hamming distance in a continuous framework, the multi-channel representation simplifies linearizing the objective in the descriptor space. Consider the following squared residuals between two descriptors in an LK framework:

$$r(\Delta \theta) = \| \phi \left( I_0(x) \right) - \phi \left( I_1(w(x; \theta + \Delta \theta)) \right) \|^2_2,$$

where $\phi(\cdot)$ is the descriptor (e.g. 128-dimensional SIFT). A first-order Taylor expansion about $\Delta \theta = 0$ is of the form

$$\| \phi \left( I_0(x) \right) - \phi \left( I_1(w(x; \theta)) \right) - \frac{\partial \phi}{\partial \theta} \Delta \theta \|^2_2,$$

where the expression $\frac{\partial \phi}{\partial \theta}$ denotes the partial derivative of the descriptor with respect to
Figure 5.5: Evaluation of different cost functions with illumination change (autocorrelation). The input images are shown in Fig. 5.4. Each of the cost function is evaluated over a grid of x- and y- translations. The Census Transform (CT) is the only dissimilarity measure capable of locating the true image displacement. In fact, the CT is barely affected in comparison to the case without illumination change as shown in Fig. 5.3. The SSD is close to the true minima. However, the cost surface is highly non-convex.
Figure 5.6: Illustration of bit-planes. Each bit of the Census Transform digit is split into a plane in a multi-channel image. This allows us to compute the Hamming distance between each bit using the ordinary Euclidean distance suitable for continuous optimization.

the vector of parameters (the Jacobian).

We remark here that interesting descriptors do not have a closed-form derivative and estimating the Jacobian via finite difference is an expensive operation for large descriptors. In contrast, when applying the registration per-channel, finite differencing could be utilized to efficiently estimate the gradient.

Separating the computation of the descriptor’s gradient per-coordinate is equivalent to computing the gradient over all channels simultaneously if each coordinate of the descriptor is independent (uncorrelated). This may, or may not, be the case depending on the descriptor. Even if descriptor coordinates are correlated, there may be an advantage of “exploding” the representation in this fashion. This is something we propose to investigate in the future.

5.2 Related Work

The use of the Census Transform (CT) in optical flow has been demonstrated by various authors. Stein [141] develops an efficient algorithm for optical flow using a modified CT targeted for automotive applications. Efficiency is achieved by computing
correspondences using lookup tables. The estimated flow, however, is sparse. Müller et al. [115] demonstrates an illumination robust optical flow algorithm using the CT for dense optical flow estimation. However, the paper does not provide details on computing gradients of the Hamming distance. Rashwan et al. [127], Mohamed et al. [111] presents an algorithm for optical flow using the CT (LBP) and small HOG descriptors with compelling results on the KITTI benchmark [57]. An evaluation of the data term in variational optical flow has been recently conducted by Vogel et al. [158] showing that the CT provides reliable illumination invariance.

The idea of a multi-channel Lucas-Kanade has been previously explored for various applications. Sevilla-Lara and Learned-Miller [134] use a multi-channel representation for tracking. Each channel is composed of a simple nonparametric statistic of the intensity values of the input image (histogram). Recently, Crivellaro and Lepetit [33] have proposed a multi-channel representation for 3D tracking of a known object model. Their representation consists of applying a nonlinear function to varying local jets (Gaussian filters) of the image.

The recent work by Antonakos et al. [6] is the closest to ours. Their multi-channel representation is composed of each coordinate value of the HOG descriptor [36]. They apply this representation to Active Appearance models and obtain compelling results. The main differences to our work are two fold. One, we use the Census Transform, a binary descriptor, instead of HOG. The other, we have found that re-computing the multi-channel representation after each warp iteration produced better results. Our conclusions, however, might be specific to the small neighborhood used in computing the Census Transform.

5.3 Experiments and Results

We applied the multi-channel descriptor alignment idea using the Census transform (bit-planes) on the task of planar template tracking under arbitrary illumination variation. An example of results is shown in Fig. 5.7. Performance of tracking with bit-planes is superior to tracking under the brightness constancy assumption. Performance is also on par with other specialized methods that aim to explicitly estimate appearance variations [138]. The added advantage of BitPlanes is that the state vector dimension is not affected. The added computational cost of evaluating the cost function per-channel outweighs nonlinear optimization intricacies with higher dimensional state vectors. In addition, the discrete nature of the Census Transform allows us to employ specific
software optimizations for faster processing.

![Figure 5.7: Experiments with 8-DOF template tracking using data from Silveira and Malis [138]. Top row shows tracking results using intensity only. Bottom rows show our proposed tracking in descriptor space with BitPlanes. Towards the end of the sequence specular reflections cause intensity-based tracking to fail. In contrast, tracking with descriptor remain robust.]

5.4 Descriptor Constancy for Illumination Invariant LK-BA [Proposed Work]

In Chapter 4, we have demonstrated a good result on one of the challenging sequences from KITTI data [57]. However, the brightness constancy assumption remains an integral component of the work. This makes the system susceptible to brightness variations.

To address this limitation, we plan to use the descriptor constancy assumption introduced in this chapter. Descriptor constancy with the binary Census Transform (CT) has shown good results in the problem of template tracking under arbitrary illumination variations. We expect to demonstrate similar results applied to VSLAM in our LK-BA framework.

5.4.1 Research questions

This binary multi-channel approach raises the following interesting research questions:

- What are the benefits of the multi-channel representation?
• Is it theoretically justifiable to perform the optimization over separate binary channels?

A partial answer to the first question (benefits) is that by using the multi-channel representation we are able to use the Hamming distance in a standard least squares optimization framework. Further investigation of the other benefits, or possible disadvantages, is a good avenue of future work.

The answer to the latter question (justifiability) warrants further investigation. A partial answer is to examine the cost surface of BitPlanes. In Fig. 5.8 we show the cost surface for a range of $x$- and $y$- translations. We observe that both the Census cost and the BitPlanes cost are locally quadratic in the vicinity of the true solution. Hence, the use of gradients on Bitplanes appears to be justified. In fact, the cost is quadratic per separate (binary) channel as shown in Fig. 5.9.

![Input image](image1.png) ![Census cost surface](image2.png) ![Bitplanes cost surface](image3.png)

**Figure 5.8:** Cost surface for both Census cost, and BitPlanes cost are locally quadratic in the vicinity of the solution.
Figure 5.9: Cost surface per binary channel in BitPlanes.
Chapter 6

Timeline & Summary

6.1 Summary of Proposed Work

In summary, we propose two contributions to enable VSLAM in challenging environments:

1. **Lucas-Kanade Bundle Adjustment (LK-BA)**, where parameters of the structure and motion are obtained such that a function of image intensities in a small patch is minimized. The framework does not require correspondences, instead correspondences are estimated as a by product of pose and structure estimation.

2. **Descriptor constancy** to address appearance variations over multiple views. This approach to handling illumination variations is robust and does not affect the dimension of the state vector.

Research problems we plan to address include:

- Selection of the reference patch in LK-BA. We envision that selecting a patch with the least perspective distortion and the least image noise will improve the system’s accuracy.

- Parametrization of the structure parameters in LK-BA. Given the fixed patch constraint, we might be able to obtain the same accuracy by parameterizing points by their inverse depth instead of a 3-vector. This should also improve the computational efficiency of the algorithm.
• There are various parameters in the systems that we plan to evaluate, which include:
  – Size of the optimization sliding window in LKBA
  – Number of pixels per view
  – Selection of pixels per view
  – Size of the local window per patch, or descriptor neighborhood support
  – Implementation details pertaining to the sparse linear solver

• The implementation of LKBA can be improved. The main computational bottleneck is due to having to store all images in memory. We plan on improving the memory efficiency of the current implementation. Similarly, our implementation of the Census Transform template tracking does not make use of the fact that resulting BitPlanes are binary. Making use of this fact is something we plan to consider.

• Descriptor constancy is a promising approach to robust correspondence estimation. We plan on working towards a better understanding of theory as well as the application to other correspondence estimation problems, such as template tracking and optical flow.

• We also plan on experimenting with other descriptors in the literature. We aim that by having a better understanding of descriptor constancy, our descriptor selection will be more principled and informed.

• Finally, we plan on demonstrating results of our system in challenging data where current state-of-the-art fails. For this purpose, we plan on using data from an underground tunnel characterized by poor lighting conditions, motion blur, as well as other image degradations.

6.2 Timeline

A proposed timeline for the thesis work is shown in Fig. 6.1
**Figure 6.1: Proposed thesis timeline**
Bibliography


appearance-based SLAM for RGB-D sensors. In Australasian Conf. on Robotics and Automation, 2011. 2.2, 2.2


[40] Frank Dellaert. Visual SLAM Tutorial: Bundle Adjustment, 2014. 2.5


[56] Andreas Geiger, Julius Ziegler, and Christoph Stiller. StereoScan: Dense 3d Reconstruction in Real-time. In Intelligent Vehicles Symposium (IV), 2011. 3.3.1


65


[68] Berthold Horn. Robot vision. MIT press, 1986. 3.2.1

[69] Berthold KP Horn and Brian G Schunck. Determining optical flow. Artificial intelligence, 17(1):185–203, 1981. 2.1.1, 2.4

[70] Albert S Huang, Abraham Bachrach, Peter Henry, Michael Krainin, Daniel Maturana, Dieter Fox, and Nicholas Roy. Visual odometry and mapping for autonomous flight using an RGB-D camera. In International Symposium on Robotics Research (ISRR), pages 1–16, 2011. 3.3.1


66


67

[83] Sebastian Klose, Philipp Heise, and Alois Knoll. Efficient compositional approaches for real-time robust direct visual odometry from RGB-D data. In IEEE/RSJ Int’l Conf. on Intelligent Robots and Systems, 2013. 2.2, 2.3, 2.4.2, 3.2.3


[92] Manolis Lourakis and Antonis Argyros. The design and implementation of a generic sparse bundle adjustment software package based on the levenberg-


[95] Bruce D. Lucas and Takeo Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision (DARPA). In Proc. of the 1981 DARPA Image Understanding Workshop, pages 121–130, April 1981. 1.3, 2.1, 2.1.1, 2.2, 2.4, 3.2


72


[152] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle Adjustment – A Modern Synthesis. LNCS, January 2000. 1.1, 2.3, 2.3.1, 2.5, 4.1, 4.2.3

[153] Tommi Tökkälä, Cédric Audras, and Andrew I Comport. Direct iterative closest point for real-time visual odometry. In ICCV 2011 Workshops, 2011. 2.2, 2.2


[160] Andreas Wedel, Clemens Rabe, Tobi Vaudrey, Thomas Brox, Uwe Franke, and Daniel Cremers. Efficient Dense Scene Flow from Sparse or Dense Stereo Data. In European Conf. on Computer Vision. Springer Berlin Heidelberg, 2008. 1.3


