

EGOMOTION ESTIMATION USING BINOCULAR  
SPATIOTEMPORAL ORIENTED ENERGY

HAO ZHONG

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES  
IN PARTIAL FULFILMENT OF THE REQUIREMENTS  
FOR THE DEGREE OF

MASTER OF SCIENCE

GRADUATE PROGRAM IN DEPARTMENT OF COMPUTER SCIENCE  
YORK UNIVERSITY  
TORONTO, ONTARIO  
JUNE 2013

**EGOMOTION ESTIMATION USING  
BINOCULAR SPATIOTEMPORAL  
ORIENTED ENERGY**

by **Hao Zhong**

a thesis submitted to the Faculty of Graduate Studies of  
York University in partial fulfilment of the requirements  
for the degree of

**MASTER OF SCIENCE**

© 2013

Permission has been granted to: a) YORK UNIVERSITY LIBRARIES to lend or sell copies of this dissertation in paper, microform or electronic formats, and b) LIBRARY AND ARCHIVES CANADA to reproduce, lend, distribute, or sell copies of this thesis anywhere in the world in microform, paper or electronic formats *and* to authorise or procure the reproduction, loan, distribution or sale of copies of this thesis anywhere in the world in microform, paper or electronic formats.

The author reserves other publication rights, and neither the thesis nor extensive extracts for it may be printed or otherwise reproduced without the author's written permission.

# EGOMOTION ESTIMATION USING BINOCULAR SPATIOTEMPORAL ORIENTED ENERGY

by **Hao Zhong**

By virtue of submitting this document electronically, the author certifies that this is a true electronic equivalent of the copy of the thesis approved by York University for the award of the degree. No alteration of the content has occurred and if there are any minor variations in formatting, they are as a result of the conversion to Adobe Acrobat format (or similar software application).

## Examination Committee Members:

1. Richard P. Wildes
2. Minas E. Spetrakis
3. Konstantinos Derpanis
4. Joseph F.X. DeSouza

## Abstract

Camera egomotion estimation is concerned with the recovery of a camera's motion (e.g., instantaneous translation and rotation) as it moves through its environment. It has been demonstrated to be of both theoretical and practical interest. This thesis documents a novel algorithm for egomotion estimation based on binocularly matched spatiotemporal oriented energy distributions. Basing the estimation on oriented energy measurements makes it possible to recover egomotion without the need to establish temporal correspondences or convert disparity into 3D world coordinates. The resulting algorithm has been realized in software and evaluated quantitatively on a novel laboratory dataset with groundtruth as well as qualitatively on both indoor and outdoor real-world datasets. Performance is evaluated relative to comparable alternative algorithms and shown to exhibit best overall performance.

## Acknowledgements

First and foremost, I would like to thank my supervisor, Richard P. Wildes, who has been supervising my master study actively and supportively for the past two years. He continuously encouraged me when I had difficulty, provided detailed instructions on my thesis writing and devoted quite a lot of time on collaboratively working with me. Additionally, I greatly appreciate his generous financial support, which made my life much easier.

Second, special thanks to all committee members who selflessly devoted their time, kindly evaluated my work and ultimately provided valuable feedback.

I also would like to thank my labmates and friends who kindly created a comfortable environment and made my work a great pleasure. I am really grateful to their voluntary devotion while I was in need. I will carefully preserve our friendship forever.

Last but not least, great thanks to my families and relatives in China who support and encourage me materially and mentally all the time.

# Table of Contents

<b>Abstract</b>	<b>iv</b>
<b>Acknowledgements</b>	<b>v</b>
<b>Table of Contents</b>	<b>vi</b>
<b>List of Tables</b>	<b>ix</b>
<b>List of Figures</b>	<b>x</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Related work . . . . .	4
1.2.1 Monocular indirect methods . . . . .	5
1.2.2 Monocular direct methods . . . . .	6
1.2.3 Binocular indirect methods . . . . .	7
1.2.4 Binocular direct methods . . . . .	10

1.2.5	Visual odometry . . . . .	11
1.2.6	3D object motion . . . . .	11
1.2.7	Other related research . . . . .	13
1.3	Contributions . . . . .	15
1.4	Outline of thesis . . . . .	16
<b>2</b>	<b>Technical approach</b>	<b>17</b>
2.1	Spatiotemporal oriented energy background . . . . .	17
2.2	Egomotion in visual spacetime . . . . .	21
2.3	Egomotion estimation . . . . .	26
2.3.1	Basic algorithm . . . . .	26
2.3.2	Salient feature selection . . . . .	29
2.3.3	Coarse-to-fine refinement . . . . .	33
2.4	Recapitulation . . . . .	36
<b>3</b>	<b>Empirical evaluation</b>	<b>40</b>
3.1	Datasets . . . . .	40
3.1.1	Laboratory dataset . . . . .	41
3.1.2	Naturalistic datasets . . . . .	43
3.2	Algorithms compared . . . . .	47
3.3	Results . . . . .	48

3.3.1	Laboratory image results . . . . .	48
3.3.2	Natural image results . . . . .	53
3.3.3	Execution rate . . . . .	58
3.4	Discussion . . . . .	59
<b>4</b>	<b>Conclusion</b>	<b>61</b>
4.1	Summary . . . . .	61
4.2	Future work . . . . .	62
<b>A</b>	<b>SOE-based stereo matching and confidence measurement</b>	<b>64</b>
<b>B</b>	<b>Details of laboratory image acquisition</b>	<b>66</b>
B.1	Motion control platform . . . . .	66
B.2	Platform calibration . . . . .	68
<b>C</b>	<b>Revised Bruss and Horn algorithm</b>	<b>70</b>
<b>D</b>	<b>Example Sequences</b>	<b>72</b>
	<b>Bibliography</b>	<b>79</b>



## List of Tables

2.1	Image Flow Comparison . . . . .	32
3.1	Parameter Settings in Lab Dataset . . . . .	42
D.1	Example Images from the Laboratory Dataset (Part 1) . . . . .	73
D.2	Example Images from the Laboratory Dataset (Part 2) . . . . .	74
D.3	Example Images from the Naturalistic Indoor Dataset (Part 1) . . . . .	75
D.4	Example Images from the Naturalistic Indoor Dataset (Part 2) . . . . .	76
D.5	Example Images from the Naturalistic Outdoor Dataset (Part 1) . . . . .	77
D.6	Example Images from the Naturalistic Outdoor Dataset (Part 2) . . . . .	78

## List of Figures

2.1	Example 3D Spacetime Direction Vectors . . . . .	19
2.2	Example 3D Filters . . . . .	19
2.3	Stereo Camera Coordinate System . . . . .	23
2.4	Illustration of 6 DOF Egomotion Parameters . . . . .	23
2.5	Salient Feature Selection Illustration . . . . .	34
2.6	A Gaussian Pyramid . . . . .	38
2.7	System Diagram . . . . .	39
3.1	Facilities for Lab Dataset Collection . . . . .	44
3.2	Example Images in Lab, Indoor and Outdoor Datasets . . . . .	45
3.3	Representative Disparity Images and Feature Selection Results . . . . .	46
3.4	Results on Lab Dataset (Part 1) . . . . .	50
3.5	Results on Lab Dataset (Part 2) . . . . .	51
3.6	The Distributions of the Selected Feature Points under Different Thresholds . . . . .	54

3.7	Results on Lab Dataset in terms of Different Parameter Settings (Part 1) . . . . .	55
3.8	Results on Lab Dataset in terms of Different Parameter Settings (Part 2) . . . . .	56
3.9	Egomotion Estimation on Indoor and Outdoor Datasets . . . . .	57
B.1	Motion Platform Calibration (Part 1) . . . . .	67
B.2	Motion Platform Calibration (Part 2) . . . . .	69

# 1 Introduction

## 1.1 Motivation

Humans are capable of perceiving their self-motion (i.e., egomotion) and do so without conscious effort. In general, when operating in the natural world multiple sensory inputs appear to be combined to yield egomotion estimates in humans, *e.g.*, visual and proprioceptive [49]. Interestingly, however, humans also can make accurate egomotion estimates in the presence of more impoverished inputs, *e.g.*, vision only [66]. These observations motivate the research that is documented in this thesis, the design, implementation and testing of a computer vision algorithm for camera egomotion estimation.

Beyond cameras, a wide variety of technologies have been marshalled in support of egomotion estimation from a moving platform, including inertial [9] and magneto [19] sensors, the Global Positioning System (GPS) [82] and active sensing (*e.g.*, sonar [2] and lidar [17]) with or without beacons [56]. In general, all of the various potentially applicable technologies have limitations (*e.g.*, drift, limited precision,

need for line of site, expense, use of sensitive moving parts, etc.) and best results are to be expected via the combination of multiple modalities. Nevertheless, in tandem with the development of sensory integration approaches [87], it remains of interest to continue development of each technology in isolation to understand its limits and optimize its performance.

Concentration on vision-based techniques can be justified from both theoretical and practical perspectives. From the theoretical side, such studies enhance our understanding of what information is made available from images. While static cameras are capable of supporting interpretations of a viewed scene (*e.g.*, object shape and scene layout), moving cameras provide the additional possibility and challenge of recovering information about the relative motion between the sensing platform and the viewed scene.

From a practical point of view, video cameras already are commonly used to help computers and robots model and interact with the world. Successful egomotion estimation can provide vital input to a number of related processes, including 3D object modeling [69], Simultaneous Localization and Modeling (SLAM) [4] and sensor platform odometry [85]. In turn, these processes can contribute to larger systems, including mobile robots [12], vehicle guidance [89] and augmented reality [6, 7]. Further, cameras are passive, inexpensive, low-power and readily available. Overall, camera egomotion estimation is not only intriguing, it is of great utility.

Previous research has considered a variety of camera configurations for egomotion estimation, including monocular and multiocular. In this thesis, binocular cameras are preferred for the following reasons. Monocular cameras provide insufficient information to disentangle the scale of a scene’s depth and the translational component of egomotion [44]. In contrast, a calibrated binocular camera arrangement allows for such recovery [90, 64, 8, 51] and is of fundamental interest in involving the smallest number of cameras that do so. Moreover, including additional cameras beyond binocular requires more effort in configuration and calibration.

Camera egomotion estimation is already a well-defined problem in computer vision research community. In general, egomotion estimation recovers the time varying motion of a platform, typically in terms of instantaneous rotation and translation. Image-based egomotion estimation effects this recovery on the basis visual information as well as camera calibration. For the binocular case, two cameras are employed. Algorithmically, most standard approaches first find the correspondences between left and right images so as to recover disparity, which subsequently is converted to 3D scene structure via triangulation with the aid of calibration. Meanwhile, correspondences between frames at time  $t$  and  $t + 1$  also are obtained for the purpose of the recovery of image flow. Then camera egomotion is estimated based on the implied temporal correspondences of the 3D points. As discussed in Sec. 2, this general framework has a number of limitations that largely arise from

the difficulty of establishing multiple correspondences both binocularly and temporally. In response to this state of affairs, a novel approach will be pursued in the present thesis that makes use of binocularly matched orientation distributions in visual spacetime,  $(x, y, t)$  [79], to recover egomotion estimates. Since the orientation distributions capture both spatial appearance and dynamics of the projected scene in an integrated fashion [25, 91], they facilitate binocular correspondence in time varying situations [79, 80, 78]. Moreover, their joint spatial and temporal appearance properties will be shown to remove the need for explicit temporal correspondence in egomotion estimation.

## 1.2 Related work

To estimate camera egomotion, monocular, binocular (stereo-based), or multiocular (more than two cameras) algorithms have been widely studied. Generally, monocular or binocular methods are more popular. In addition, most of these algorithms can be classified into indirect methods, which require image flow as an intermediate product, or direct methods, which estimate camera egomotion directly from image measurements without the recovery of image flow.

### 1.2.1 Monocular indirect methods

For the class of monocular, indirect methods, Raudies and Neumann [74] summarized the constraints and the optimization techniques that different algorithms apply. The algorithms they consider are those estimating egomotion and depth from optical flow or parametrically defined visual motion fields. Raudies and Neumann propose that these methods can be grouped by the optimization techniques into five classes, i.e., least-squares (LSQ), fix point iteration (FP), Gauss-Newton iteration (GN), Hough transform (HT), and hierarchical grid (HG). As examples: Rieger and Lawton [75] have segregated the rotational component of the visual motion field from the translational and then applied least-squares optimization on the remaining translational part. Bruss and Horn [15] applied a fixed-point iteration optimization technique to estimating rotation and translation iteratively. Gauss-Newton iteration is used by Zhang and Tomasi [93] to optimize for translation, from which the rotation and depth can be estimated relatively easily. Moreover, Heeger and Jepson [41] showed that the nonlinear equation describing the optical flow field can be separated into three components, i.e., translational, rotational and depth components, resp. The effectiveness of their method is demonstrated by applying their algorithm on estimating these components one by one in the aforementioned order. Perrone and Stone's method [72] is a template method motivated by the hy-



pothesized function of mammalian brain areas middle temporal (MT) and medial superior temporal (MST) cortices. This approach combines a Hough transform and hierarchical grid processing to model the MT and MST operations, respectively.

Rather than attempt the recovery of precise numerical estimates of egomotion parameter values, some research has instead considered qualitative estimation or restricted itself to recovery of egomotion subcomponents (*e.g.*, the focus of expansion, FOE). Fermuller and Aloimonos [27] developed an algorithm to estimate egomotion qualitatively, which gradually reduces the space of possible solutions by checking four constraints imposed by 3D motion parameters on the normal flow field. The geometric constraint (the first considered) generates a set of possible solutions for the direction of translation and the axis of rotation, while the following three constraints further narrow down the possible space of solutions. (The exact solution is found if there is only one solution). Sinclair et al. [77] proposed an algorithm for estimating the FOE based on measurements of normal flow and tolerance constraints on angular velocity. The major application was to vehicle guidance, where it was argued the FOE alone provided useful information.

### **1.2.2 Monocular direct methods**

Other research has developed direct methods in conjunction with monocular egomotion recovery. Horn and Weldon [46] proposed what appears to be the first

direct method by considering various integrals, based on the brightness constancy constraint equation, over an image region corresponding to a single rigid object. Different integrals are proposed for solving several alternative cases, i.e., different knowns and unknowns or different constraints on the camera egomotion. For instance, if the depth is known, translation and rotation can be estimated in closed-form using a least-squares method. Alternatively, for the case of pure translation or known rotation, a least-squares method is first applied to determine translation, and then the depth is found by considering the brightness constancy constraint. Further, Hanna [35] developed another method without making any assumptions on camera motion. However, he considered parametrically defined surfaces. In this case the brightness constraint equation is first locally applied to estimate local surface parameters and then globally to recover egomotion parameters. Additional investigations of direct monocular approaches to egomotion estimation involve incorporation of Kalman filtering [42].

### **1.2.3 Binocular indirect methods**

A fundamental limitation of egomotion estimation with a single moving camera is the inherent scale ambiguity between 3D scene structure and camera translation [44]. To overcome this limitation, some researchers have addressed the problem in a different way, by using stereo cameras. Many of these algorithms share a similar

basic structure: Recover disparity between binocular views and then recover rigid motion parameters by consideration of disparity-based 3D point correspondences across time, *e.g.*, as mediated by optical flow, 2D or even 3D feature tracking. In the group of indirect methods, Badino [8] first calculated a disparity map with SSD matching and then applied the Kanade-Lucas-Tomasi (KLT) [76] tracker to track features across time and obtain image flow. Subsequently, a quaternion-based closed-form solution [45] is used to estimate camera egomotion. Similarly, the disparity image was first calculated with the zero-mean normalized cross-correlation (ZNCC) criteria in [61]. Next, good features, which had a sharper peak in the correlation surface (a surface based on the correlation score), were selected. They continued to perform the estimation of the 3D rigid body transformation using a least-squares estimation method based on a singular value decomposition (SVD) [52], similar to [37]. Weng et al. [90] added to this type of approach by including a closed-form approximate matrix-weighted least squares solution. Zhang and Faugeras [94] further contribute to this type of approach using a hypothesis and test methodology involving line segment correspondence within an extended Kalman filter (EKF) framework. Other approaches have been concerned with simultaneous egomotion estimation and motion segmentation, *e.g.*, [23] and its extension of [38].

Milella and Siegwart [64] proposed a stereo-based egomotion estimation algorithm with Iterative Closest Point (ICP) [11] as a refinement technique. Their

method, following the aforementioned procedure, first generates a dense disparity map, then selects features with the Shi-Tomasi feature detector [76] and finds potential matches between two consecutive frames via image intensity information. Additionally, the ICP technique is applied to refine the matching of the 3D features without previous knowledge of motion. Finally, Hogue and Jenkin [43] estimated 3D reef structure while simultaneously estimating stereo camera egomotion with their newly developed underwater vision sensor. After the disparity map is recovered using their stereo algorithm, the Kanade-Lucas-Tomasi (KLT) feature tracking [76] algorithm is applied to extract and track “good” features along the image sequence. Then, the least-squares rotation and translation are fitted via application of Horn’s absolute orientation method [45] and a nonlinear Levenberg-Marquardt minimization [29].

Demirdjian and Darrell [22] also calculated a disparity map but did not recover scene structure in Euclidean space. Instead, they calculated disparity motion flow (called d-motion) and build the relationship between d-motion and 3D Euclidean rigid motion. They did not recover camera egomotion explicitly; however, subsequent research did provide a closed-form solution for egomotion based on a similar disparity space analysis [24].

#### 1.2.4 Binocular direct methods

In contrast, there are fewer direct methods that make use of stereo cameras. Hanna and Okamoto [36] estimated camera egomotion directly from brightness derivatives of two or more stereo and/or motion data sets. A least-squares method with Gauss-Newton optimization [21] was employed. Also, Mandelbaum et al. [62] modeled the point matching correlation surface as a quadratic, which allows direct and explicit computation of incremental refinements for egomotion and structure using linear algebraic relations. Interestingly, this algorithm accommodates single-camera rigs and multiple-camera rigs. Stein and Shashua [83] proposed a direct egomotion estimation algorithm based on three views rather than two views. They developed a tensor brightness constraint based on the optical flow constraint equation and the geometric model of the “trilinear tensor” [5]. This “tensor brightness constraint” presented the relationship between the spatiotemporal brightness derivatives at each pixel in the image. They then proceed through a hierarchy of reduced motion models with additional assumptions, such as calibrated cameras and a small motion model. Here, it is interesting to note that Spetsakis and Aloimonos appear to have been the first to investigate the fundamentals of three view image interpretation [81].

### 1.2.5 Visual odometry

Visual odometry is closely related to egomotion estimation. In essence, visual odometry temporally integrates instantaneous egomotion estimates to obtain position and orientation estimates for the camera at any given time along this trajectory relative to some initial position. Here, a wide variety of approaches have been developed involving both single [20, 68, 86] and multiple cameras [70, 48, 53, 59, 51]. Interestingly, visual odometry in and of itself often is found to be insufficient for accurate and precise long distance traversals, a situation that can be improved significantly through incorporation of additional sensors (*e.g.*, inertial sensing) [53].

### 1.2.6 3D object motion

Complimentary to egomotion estimation, research also has addressed the estimation of 3D object motion relative to a (typically) stationary camera. Indeed, these two problems are intimately linked as they both fundamentally recover relative motion between the camera and object or scene. Here, a few representative approaches are highlighted.

Kim and Aggarwal [50] build on the usual scheme for determining 3D object motion with a stereo camera. Initially, 3D features are extracted and their correspondences are established. Then, the rigid motion parameters are computed

accordingly. Specifically, they proposed a two-pass relaxation method for matching 3D features extracted from successive depth maps. With these correspondences in hand, the 3D motion parameters are estimated by first finding a rotation matrix independent of the translation vector and then finding a translation vector given this computed rotation matrix as the solution to a system of linear equations. Lee and Kay [55] extend this type of approach by including a Kalman filter. They derived a new set of discrete Kalman filter equations, including the measurement equation and the state propagation equation. Use of the Kalman filter is shown to improve accuracy and convergence time. In addition, a method based on linear depth and brightness constraints is presented by Harville et al. for 3D pose tracking [39]. In their method, range information is first used to estimate the shape of the object and then applied to their newly derived depth constraint in imitation of and in combination with the brightness constancy assumption. They claim that the combined brightness and depth constraint equations help improve the performance compared with the use of either independently. Malassiotis and Strintzis [60] proposed a model-based algorithm for object surfaces and motion estimation. The surface and motion of the object are both modeled so that the problem is reduced to parameter optimization. In particular, object motion is first modeled using the rigid motion assumption; subsequently, non-rigid motion is estimated via appeal to finite element modeling as refinement on the initial rigid body results. Yet an-

other approach has regarded 3D shape and motion recovery in terms of optimized matching across multiple stereo images via application of dynamic programming [65].

Navab et al. [67] studied motion estimation in terms of lines. Similar to many of the already reviewed approaches, their method is based on the established stereo matches and computed optical flow. The difference is that they focused on token (line) tracking and proposed that the kinematic screw of an object can be estimated if multiple lines are available. Moreover, assuming the structure of the object is known and 3D features are extracted and tracked over the frames, Young and Chellappa [92] developed an algorithm based on detailed kinematics modeling for 3D motion parameter estimation with noisy stereo images. Their method represents various types of motion in the form of a bilinear state space model using standard rectilinear states for translation and quaternions for rotation. An Extended Kalman Filter (EKF) is applied to deal with the nonlinearities present.

### **1.2.7 Other related research**

Larusso et al. [54] presented the analysis and comparison of four popular closed-form solutions for estimating 3D rigid body transformation, which are different in the transformation representation and alternative ways of minimizing a criterion function. The comparison involves the accuracy, stability and computation time of



each algorithm.

More generally, several other methods are related to egomotion estimation, *e.g.*, in their concern for solving for the rotation and translation that align 3D data. The first set of such algorithms is targeted to solve the registration problem of 3D point sets. Arun et al. [3] proposed one of the earliest least-squares solutions to this problem in the computer vision literature. Further, Matthies and Shafer [63] argued the performance can be greatly improved by using 3D Gaussian distributions to model triangulation error, rather than scalar error as suggested in [63]. A second notable class of relevant techniques involves the Iterative Closest Point (ICP) algorithm [11] for solving the 3D registration problem. For best applicability, several limitations to the original ICP algorithm must be surmounted. Here, a particular concern is its requirement of a good initialization to avoid being trapped in a locally optimal solution. Thus, Li and Hartley [57] proposed an alternative algorithm, which improves the ICP algorithm by guaranteeing the global optimality of the solution without any initialization. Finally, bundle adjustment is a well-known optimization technique for refining a visual reconstruction to produce jointly optimal 3D structure and viewing parameters (camera position and/or calibration) estimates [88].

### 1.3 Contributions

Inspired by and building on previous work in 3D scene reconstruction and flow estimation based on spatiotemporal oriented energy distributions (SOEs) [78, 79], this thesis provides a novel approach to stereo-based egomotion estimation. Specifically, the contributions of the presented research are as follows.

- An analysis is developed that relates binocularly matched spatiotemporal oriented energy distributions to camera egomotion, as the camera traverses an otherwise rigid three-dimensional environment. Six-degree-of-freedom egomotion is encompassed as instantaneous rotational and translational velocities.
- The formal analysis is embodied in algorithmic form and implemented in software to yield a novel algorithm for camera egomotion recovery.
- The developed algorithm is evaluated both qualitatively and quantitatively, including comparison with alternative, state-of-the-art algorithms.
- A new binocular video dataset is introduced that includes groundtruth egomotion and will be made available to the community.

## 1.4 Outline of thesis

This thesis unfolds in four chapters. Chapter 1 has provided the problem overview, including motivation and discussion of related research. Chapter 2 details information about our technical approach. Following introduction of fundamental background knowledge, a novel SOE-based approach to stereo egomotion estimation is presented. Next, empirical evaluation is detailed in Chapter 3. Here, the performance of our algorithm is compared with that of representative alternative algorithms on various datasets. Finally, chapter 4 provides a summary and conclusion, as well as discussion of possible directions for future research.

## 2 Technical approach

This chapter details a theory and algorithm for spatiotemporal oriented energy (SOE) based stereo egomotion estimation. In this chapter, the proposed theory and algorithm for egomotion estimation are introduced in detail. First, requisite background material on SOE-based image representation is briefly reviewed. Second, the relationship between camera egomotion and orientation in visual spacetime,  $(x, y, t)$ , across binocular views is analyzed. Third, the proposed algorithm for egomotion estimation based on binocularly corresponding spacetime orientation measurements is developed. Finally, the developments are summarized.

### 2.1 Spatiotemporal oriented energy background

Video sequences induce very different orientation patterns in image spacetime depending on their contents. For instance, a textured, stationary object yields a much different orientation signature than if the very same object were undergoing translational motion. An efficient framework for analyzing spatiotemporal information

can be realized through the use of 3D,  $(x, y, t)$ , oriented energies [1], as shown in Fig. 2.1. These energies are derived from the filter responses of orientation selective bandpass filters that are applied to the spatiotemporal volume representation of a video stream. A chief attribute of an oriented energy representation is its ability to encompass both spatial and dynamic aspects of visual spacetime, strictly through the analysis of 3D orientation. Consideration of spatial patterns (*e.g.*, image textures) is performed when the filters are applied within the image plane. Dynamic attributes of the scene (*e.g.*, velocity and flicker) are analyzed by filtering at orientations that extend into the temporal dimension.

Spatiotemporal oriented energy measurements have been used previously for a variety of computer vision tasks; most closely related to current work are applications to optical flow [1, 40, 34] and tracking [18] as well as stereo disparity and 3D scene flow [79, 80]. While egomotion might be recovered via a regression on the recovered scene flow [80], here a more direct approach is developed that performs egomotion estimation on binocularly matched SOEs. Indeed, it appears that the approach developed in this thesis is the first to consider recovery of egomotion from measurements of spatiotemporal orientation.

For present purposes, local SOE measurements are recovered separately in the left and right streams of the binocular video via convolution with a bank of Gaussian second derivative filters,  $G_2(\hat{\mathbf{w}})$ , and their Hilbert transforms,  $H_2(\hat{\mathbf{w}})$ , which are

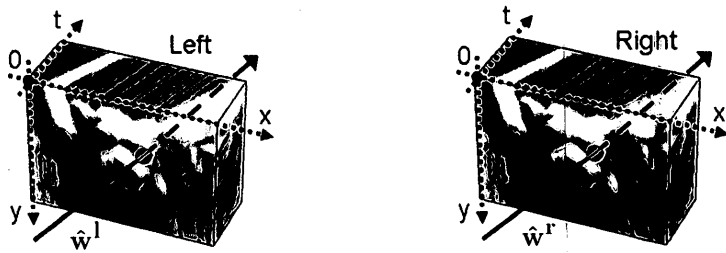


Figure 2.1: Illustration of 3D orientation in binocular visual spacetime,  $(x, y, t)$ . Example corresponding points across the left and right views are marked as black dots and their orientations,  $\hat{w}^l$  and  $\hat{w}^r$ , with red arrows. Adapted from [78].

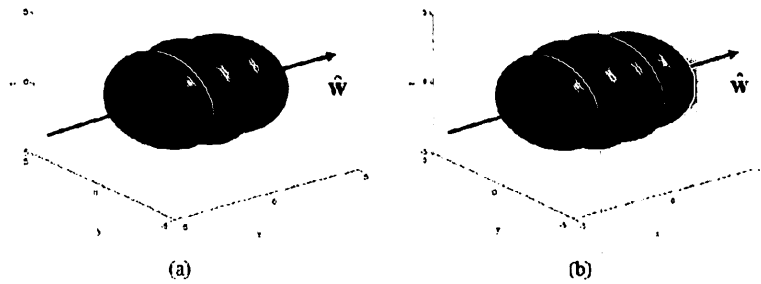


Figure 2.2: Example 3D filters. (a) and (b) are Gaussian second derivative and the corresponding Hilbert transform, respectively. Adapted from [80].

combined in quadrature to yield energy measurements

$$E(I(\mathbf{x}); \hat{\mathbf{w}}) = [G_2(\hat{\mathbf{w}}) * I(\mathbf{x})]^2 + [H_2(\hat{\mathbf{w}}) * I(\mathbf{x})]^2 \quad (2.1)$$

where  $I$  is an image,  $\mathbf{x} = (x, y, t)^\top$ , are spatiotemporal image coordinates, the unit vector  $\hat{\mathbf{w}}$  specifies the 3D direction of the filter and  $*$  is the convolution operator [30]. Example filters for G2 and H2 are shown in Fig. 2.2.

Most practical uses of energy filtering, (2.1), involve a normalization step to make responses invariant to multiplicative bias and bring response values to the uniform scale 0 to 1. The necessary operation is realized via pointwise division by the local sum of consort energies at a point

$$\hat{E}(I(\mathbf{x}); \hat{\mathbf{w}}_j) = \frac{E(I(\mathbf{x}); \hat{\mathbf{w}}_j)}{\sum_i^N E(I(\mathbf{x}); \hat{\mathbf{w}}_i) + \epsilon}, \quad (2.2)$$

with  $N$  the number of orientations that span orientation space for the order of filter employed and  $\epsilon$  a small constant to avoid division by zero when the summed energies are small. Indeed, the filter results can serve as a basis set from which energy at any other orientation can be calculated via a weighted combination. Here, since  $2^{nd}$ -order Gaussian filters and Hilbert transforms are used,  $N = 10$  is required [30], with their orientations chosen to uniformly sample 3D orientation as the normals to the faces of an icosahedron [71] with antipodal directions identified. The result of this computation is that a set of  $N$  (normalized) SOEs are available at each spacetime point,  $\mathbf{x}$ , in both the left and right image sequences.

Finally, correspondences must be established between points in the left and right image sequences to serve as input to the proposed egomotion algorithm. In general, any reliable algorithm for establishing binocular correspondence could be applied on a framewise basis to the original image sequences; for review see, *e.g.*, [14]. Here, since SOEs are available and previously have been shown useful for stereo video matching [80], that matching approach is applied to establish the needed left-right correspondences. Additional details on the operation of this algorithm are presented in Appendix A.

## 2.2 Egomotion in visual spacetime

In this subsection, a novel parameterization of 3D directions,  $\hat{\mathbf{w}}$ , in visual spacetime,  $(x, y, t)$ , is given in terms of camera egomotion parameters. To facilitate this presentation, the derivation begins by reviewing standard material on the visual motion field [44]. Let a Euclidean coordinate system,  $(\hat{\mathbf{X}}, \hat{\mathbf{Y}}, \hat{\mathbf{Z}})^\top$ , be defined at the projection centre of the left camera in a rectified binocular pair, with the optical axis and stereo baselines along the  $\hat{\mathbf{Z}}$  and  $\hat{\mathbf{X}}$ , axes, resp and the  $\hat{\mathbf{Y}}$  axis chosen to complete a right-handed coordinate system, as shown in Fig. 2.3. Under perspective projection, the image coordinates in the left system are given as  $\mathbf{x}^l = (x, y, t)^\top = (X/Z, Y/Z, t)^\top$ , with focal length set to unity for conciseness. The coordinates of a corresponding point in the right camera are then given as



$\mathbf{x}^r = (x+d, y, t)^\top$ , where  $d = B/Z$  is stereo disparity and  $B$  the baseline separation between left and right cameras.

Let egomotion of the camera be given in terms of instantaneous translational,  $\mathbf{T} = (t_x, t_y, t_z)^\top$ , and rotational,  $\Omega = (\omega_x, \omega_y, \omega_z)^\top$ , velocities with respect the centre of projection of the left camera, as shown in Fig. 2.4. Correspondingly, the 3D velocity of a point,  $\mathbf{P} = (X, Y, Z)^\top$ , relative to the camera is then

$$\dot{\mathbf{P}} = \begin{pmatrix} \dot{X} \\ \dot{Y} \\ \dot{Z} \end{pmatrix} \quad (2.3)$$

with “dot notation” used to denote temporal derivatives and which is given in vector representation as [32]

$$= -\mathbf{T} - \Omega \times \mathbf{P} \quad (2.4)$$

and which can be further expanded component-wise as

$$= \begin{pmatrix} -t_x - \omega_y Z + \omega_z Y, \\ -t_y - \omega_z X + \omega_x Z, \\ -t_z - \omega_x Y + \omega_y X, \end{pmatrix}. \quad (2.5)$$

In the usual way, the visual motion field,  $(u, v)^\top$ , which captures the perspective image projection of the relative 3D motion between a camera and 3D world, now

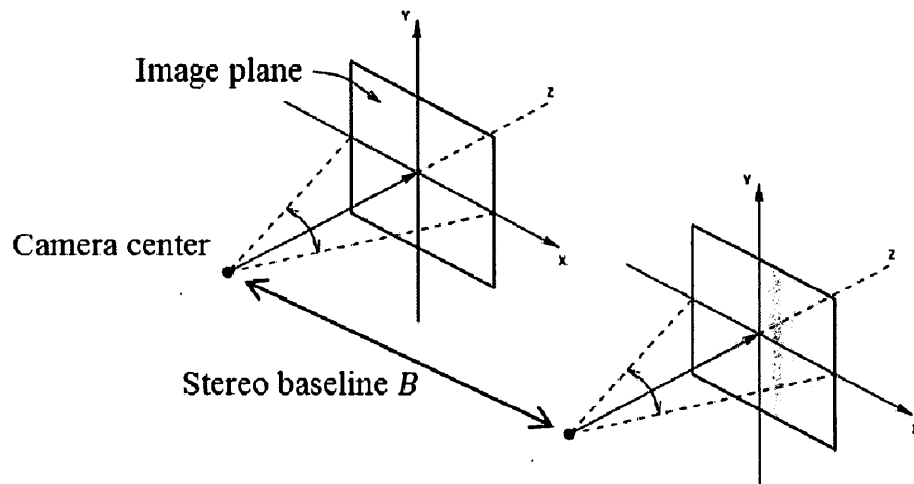


Figure 2.3: Left and right camera systems are shown in the upper left and lower right portions of the figure, resp. Perspective serves as the model of image projection. See text for details.

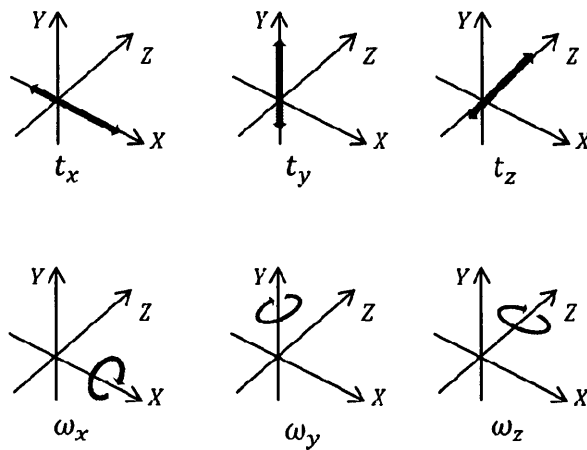


Figure 2.4: Illustration on 6 DOF egomotion parameters.

can be parameterized in terms of egomotion parameters

$$\begin{pmatrix} u(\mathbf{x}; \mathbf{T}, \Omega) \\ v(\mathbf{x}; \mathbf{T}, \Omega) \end{pmatrix} = \begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} \quad (2.6)$$

which, making use of the operative perspective model of image formation,  $(x, y) = (X/Z, Y/Z)$ , can be expanded as

$$= \begin{pmatrix} \frac{\dot{X}}{Z} - X \frac{\dot{Z}}{Z^2} \\ \frac{\dot{Y}}{Z} - Y \frac{\dot{Z}}{Z^2} \end{pmatrix} \quad (2.7)$$

and with substitution from (2.3) yielding

$$\begin{pmatrix} u(\mathbf{x}; \mathbf{T}, \Omega) \\ v(\mathbf{x}; \mathbf{T}, \Omega) \end{pmatrix} = \begin{pmatrix} \frac{1}{Z}(xt_z - t_x) + \omega_x xy - \omega_y(x^2 + 1) + \omega_z y \\ \frac{1}{Z}(yt_z - t_y) + \omega_x(y^2 + 1) - \omega_y xy - \omega_z x \end{pmatrix}. \quad (2.8)$$

Further, since binocular disparity,  $d$ , is assumed available, substituting  $\frac{1}{Z} = \frac{d}{B}$  allows for an expression that avoids explicit reference to the 3D world coordinate  $Z$ , as follows.

$$\begin{pmatrix} u(\mathbf{x}; \mathbf{T}, \Omega) \\ v(\mathbf{x}; \mathbf{T}, \Omega) \end{pmatrix} = \begin{pmatrix} \frac{d}{B}(xt_z - t_x) + \omega_x xy - \omega_y(x^2 + 1) + \omega_z y \\ \frac{d}{B}(yt_z - t_y) + \omega_x(y^2 + 1) - \omega_y xy - \omega_z x \end{pmatrix}. \quad (2.9)$$

Similarly, the visual motion field at the corresponding point in the right view is given in terms of the temporal derivative of  $(x + d, y)^\top$ , i.e.,

$$(x + d, y)^\top = (u + \dot{d}, v)^\top, \quad (2.10)$$

where

$$\delta d(\mathbf{x}; \mathbf{T}, \Omega) = \dot{d} = -B \frac{\dot{Z}}{Z^2}, \quad (2.11)$$

with  $\delta d$  simply an alternative symbol for  $\dot{d}$ , analogous to the roles of  $u, v$  for  $\dot{x}, \dot{y}$ , resp., in equation (2.8). Now, substitution of  $\dot{Z}$  from (2.3) yields

$$\delta d(\mathbf{x}; \mathbf{T}, \Omega) = d \left( \frac{1}{Z} t_z + \omega_x y - \omega_y x \right), \quad (2.12)$$

and with further substitution of  $\frac{1}{Z} = \frac{d}{B}$ , we have

$$\delta d(\mathbf{x}; \mathbf{T}, \Omega) = d \left( \frac{d}{B} t_z + \omega_x y - \omega_y x \right). \quad (2.13)$$

Finally, image spacetime,  $(x, y, t)^\top$ , directions defined in terms of the visual motion field,  $(u, v)^\top$ , and disparity flow,  $\delta d$ , at corresponding points across a binocular video sequence can be defined as follows. Let  $\hat{\mathbf{v}}^l$  and  $\hat{\mathbf{v}}^r$  be the unit direction vectors at the corresponding points in the left and right image spacetimes, resp. Then, they are parameterized in terms of egomotion parameters,  $\mathbf{T}, \Omega$ , as

$$\hat{\mathbf{v}}^l(\mathbf{x}; \mathbf{T}, \Omega) = \frac{1}{\sqrt{u(\mathbf{x}; \mathbf{T}, \Omega)^2 + v(\mathbf{x}; \mathbf{T}, \Omega)^2 + 1}} \begin{pmatrix} u(\mathbf{x}; \mathbf{T}, \Omega) \\ v(\mathbf{x}; \mathbf{T}, \Omega) \\ 1 \end{pmatrix} \quad (2.14)$$

and

$$\hat{\mathbf{v}}^r(\mathbf{x}; \mathbf{T}, \Omega) = \frac{1}{\sqrt{(u(\mathbf{x}; \mathbf{T}, \Omega) + \delta d(\mathbf{x}; \mathbf{T}, \Omega))^2 + v(\mathbf{x}; \mathbf{T}, \Omega)^2 + 1}} \begin{pmatrix} u(\mathbf{x}; \mathbf{T}, \Omega) + \delta d(\mathbf{x}; \mathbf{T}, \Omega) \\ v(\mathbf{x}; \mathbf{T}, \Omega) \\ 1 \end{pmatrix} \quad (2.15)$$

where  $u(\mathbf{x}; \mathbf{T}, \Omega)$ ,  $v(\mathbf{x}; \mathbf{T}, \Omega)$  and  $\delta d(\mathbf{x}; \mathbf{T}, \Omega)$  are given by their defining equations, (2.8) and (2.11). Significantly, it appears that these derived parameterizations of

matched orientations, (2.14) and (2.15), in terms of egomotion have not previously been presented in the literature.

## 2.3 Egomotion estimation

### 2.3.1 Basic algorithm

If a 3D,  $(x, y, t)^\top$ , spacetime direction,  $\hat{\mathbf{v}}$ , is associated with a 2D,  $(x, y)^\top$ , image flow,  $(u, v)^\top$ , then it must correspond to a minimal energy across orientations, as brightness constancy assumes uniform intensity along the direction of flow. Thus, to solve for the appropriate direction, the basis set of oriented energy measurements, (2.2), can be steered to the direction that yields minimal energy response, as parameterized by the global egomotion parameters,  $\mathbf{T}$ ,  $\Omega$ . Let oriented energy measurements for the corresponding points in left and right image spacetime be

$$\hat{E}^l(I^l(\mathbf{x}^l); \hat{\mathbf{v}}^l(\mathbf{x}^l; \mathbf{T}, \Omega)), \quad (2.16)$$

and

$$\hat{E}^r(I^r(\mathbf{x}^r); \hat{\mathbf{v}}^r(\mathbf{x}^r; \mathbf{T}, \Omega)) = \hat{E}^r(I^r(\mathbf{x}^l + \mathbf{d}); \hat{\mathbf{v}}^r(\mathbf{x}^l + \mathbf{d}; \mathbf{T}, \Omega)), \quad (2.17)$$

resp., with  $\mathbf{d} = (d, 0, 0)^\top$ , because  $\mathbf{x}^l$  and  $\mathbf{x}^r$  are in binocular correspondence. Then the matched oriented energies at a point would sum to

$$E^{stereo}(I^l(\mathbf{x}^l), I^r(\mathbf{x}^r); \mathbf{T}, \Omega) = \hat{E}^l(I^l(\mathbf{x}^l); \hat{\mathbf{v}}^l(\mathbf{x}^l; \mathbf{T}, \Omega)) + \hat{E}^r(I^r(\mathbf{x}^l + \mathbf{d}); \hat{\mathbf{v}}^r(\mathbf{x}^l + \mathbf{d}; \mathbf{T}, \Omega)), \quad (2.18)$$

with  $\hat{E}^l$  and  $\hat{E}^r$  given by (2.2) applied to the left,  $I^l$ , and right,  $I^r$ , image streams, resp. Within the developed framework, the solution of egomotion estimation now can be stated as

$$\arg \min_{\mathbf{T}, \Omega} \sum_{\mathbf{x}^l \in \mathcal{S}} E^{stereo}(I^l(\mathbf{x}^l), I^r(\mathbf{x}^l + \mathbf{d}); \mathbf{T}, \Omega) \quad (2.19)$$

with  $\mathcal{S}$  the set of image points considered in the estimation, as indexed to the left image. Due to the non-linear dependence of the objective function (2.19), on  $\mathbf{T}$  and  $\Omega$ , Gauss-Newton refinement is employed to obtain the solution. While alternative non-linear optimization methods could be employed [21], Gauss-Newton previously has proven useful in the recovery of 3D scene flow from binocular orientation measurements [78] and will be shown useful in the current context when empirical results are presented in Chapter 3. For the sake of conciseness, let

$$\begin{aligned} \mathcal{G}^l &= G_2(\hat{\mathbf{v}}^l(\mathbf{x}^l; \mathbf{T})) * I^l(\mathbf{x}^l) \\ \mathcal{H}^l &= H_2(\hat{\mathbf{v}}^l(\mathbf{x}^l; \mathbf{T})) * I^l(\mathbf{x}^l) \\ \mathcal{G}^r &= G_2(\hat{\mathbf{v}}^r(\mathbf{x}^r; \mathbf{T})) * I^r(\mathbf{x}^r) \\ \mathcal{H}^r &= H_2(\hat{\mathbf{v}}^r(\mathbf{x}^r; \mathbf{T})) * I^r(\mathbf{x}^r). \end{aligned} \quad (2.20)$$

Then, egomotion parameters are estimated in terms of the objective function, (2.19), residual

$$\mathbf{r}(\mathbf{x}; \mathbf{T}, \Omega) = \begin{pmatrix} \mathcal{G}^l \\ \mathcal{H}^l \\ \mathcal{G}^r \\ \mathcal{H}^r \end{pmatrix} \quad (2.21)$$

and Jacobian (using subscripts to denote differentiation)

$$\mathbf{J}(\mathbf{x}; \mathbf{T}, \Omega) = \begin{pmatrix} \mathcal{G}_{t_x}^l & \mathcal{G}_{t_y}^l & \mathcal{G}_{t_z}^l & \mathcal{G}_{\omega_x}^l & \mathcal{G}_{\omega_y}^l & \mathcal{G}_{\omega_z}^l \\ \mathcal{H}_{t_x}^l & \mathcal{H}_{t_y}^l & \mathcal{H}_{t_z}^l & \mathcal{H}_{\omega_x}^l & \mathcal{H}_{\omega_y}^l & \mathcal{H}_{\omega_z}^l \\ \mathcal{G}_{t_x}^r & \mathcal{G}_{t_y}^r & \mathcal{G}_{t_z}^r & \mathcal{G}_{\omega_x}^r & \mathcal{G}_{\omega_y}^r & \mathcal{G}_{\omega_z}^r \\ \mathcal{H}_{t_x}^r & \mathcal{H}_{t_y}^r & \mathcal{H}_{t_z}^r & \mathcal{H}_{\omega_x}^r & \mathcal{H}_{\omega_y}^r & \mathcal{H}_{\omega_z}^r \end{pmatrix}. \quad (2.22)$$

As defined so far, the residual, (2.21), and Jacobian, (2.22), are defined pointwise in terms of  $\mathbf{x}$ . To account for all  $n$  points in the images that are under consideration, let  $\mathbf{x}_i$  index individual points and stack the residuals, (2.21), into a single  $4n \times 1$  vector as

$$\rho(\mathbf{T}, \Omega) = \left( \mathbf{r}(\mathbf{x}_1; \mathbf{T}, \Omega)^\top, \mathbf{r}(\mathbf{x}_2; \mathbf{T}, \Omega)^\top, \dots, \mathbf{r}(\mathbf{x}_n; \mathbf{T}, \Omega)^\top \right)^\top, \quad (2.23)$$

and stack the Jacobians, (2.22), into a single  $4n \times 6$  matrix as

$$\mathcal{J}(\mathbf{T}, \Omega) = \left( \mathbf{J}(\mathbf{x}_1; \mathbf{T}, \Omega)^\top, \mathbf{J}(\mathbf{x}_2; \mathbf{T}, \Omega)^\top, \dots, \mathbf{J}(\mathbf{x}_n; \mathbf{T}, \Omega)^\top \right)^\top \quad (2.24)$$

Now, the Gauss-Newton update for egomotion parameters,  $\mathbf{T}$  and  $\Omega$ , is given as

$$\begin{pmatrix} \mathbf{T} \\ \Omega \end{pmatrix}^{k+1} = \begin{pmatrix} \mathbf{T} \\ \Omega \end{pmatrix}^k - (\mathcal{J}(\mathbf{T}, \Omega)^\top \mathcal{J}(\mathbf{T}, \Omega))^{-1} \mathcal{J}(\mathbf{T}, \Omega)^\top \rho(\mathbf{T}, \Omega), \quad (2.25)$$

with  $k$  and  $k + 1$  successive iterations.

### 2.3.2 Salient feature selection

When dealing with real-world images, feature selection can play an important role. Restricting subsequent analysis to reliable features can greatly improve an algorithm’s robustness to noise. Our feature extraction method is based on the match score map produced by the stereo matching algorithm used to provide input to the egomotion estimator, *e.g.*, [80]. We further make use of a sampling strategy to ensure selected features are reasonably distributed across the images and thereby ameliorate difficulties that arise when global egomotion parameters are estimated based on spatially biased feature selection.

#### 2.3.2.1 Confidence map

Simply relying on a “good” match score cannot guarantee that a recovered correspondence is accurate. One extreme case is to look for correspondence on a purely textureless board. Owing to the lack of pattern information, the match score goodness between any pair of points would be very large, which is substan-



tially misleading. In order to overcome this problem as well as select points with reliable disparity estimates, local extrema of curvature of the match score map (i.e., correlation surface) are employed. While a variety of approaches to feature selection might be considered, match score curvature is known to provide reliable (if conservative) indication of loci where stereo correspondence is good [26]. Curvature is calculated as the 2<sup>nd</sup> spatial derivative of the map along the horizontal axis (assuming horizontally aligned epipolar lines). This confidence map also is processed to set confidence to zero at points that are indicated as half-occluded [26], if the stereo matcher provides such information. The stereo matcher employed in the present work does indicate such points.

The derived map of match confidence can serve to focus egomotion estimation on points where stereo estimation is most accurate. Further, by indicating points where the match is well defined locally, it finds well textured points that will yield correspondingly well defined SOEs (2.2), as illustrated Fig. 2.5b. In this figure red indicates points on an image from a stereo pair where the confidence map exceeds a threshold. It is seen that these points reliably fall in well textured regions.

### **2.3.2.2 Feature selection technique**

It is not sufficient to select features purely on the basis of a confidence map. It also is critical when estimating globally defined parameters, *e.g.*, egomotion, that selected

points are chosen approximately uniformly across the image. Such a strategy helps to avoid spatially biased estimates as well as resolve potential ambiguities. Table 2.1 provides an illustrative example. In this table, we compare the ground truth image flow in the case that the camera is purely rotating around the Y-axis, or the case that the camera is purely translating along the X-axis. In each case, different positions of feature points are considered, as indicated in the table. One illustrated case corresponds to an image with evenly distributed feature points, the other with most of the feature points lying at the center of the image. As we can see, with evenly distributed feature points, it's not hard to tell the difference of the  $\omega_y$  and  $t_x$  cases. However, if the feature points are mostly gathering at the center of the image, there is little difference in terms of the image flow of the two cases, which makes the distinction practically infeasible.

To avoid this situation, we sample the match confidence map in two ways. First, non-maximum suppression is employed, which helps extract the feature points with locally maximum confidence. Briefly, given the size of the suppression window, the non-maximum suppression algorithm keeps only the local maximum within the window. However, even with such processing, it is not guaranteed that feature points would be evenly distributed. Therefore, sampling of the non-maxima suppressed confidence map is further refined. In particular, the image is gridded spatially (currently  $9 \times 12$ ) and within each grid cell a threshold on the local extrema is set

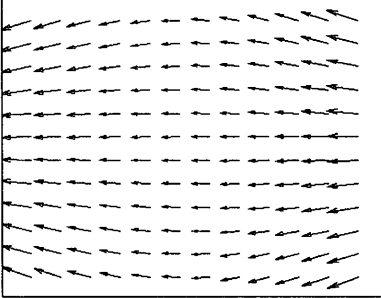
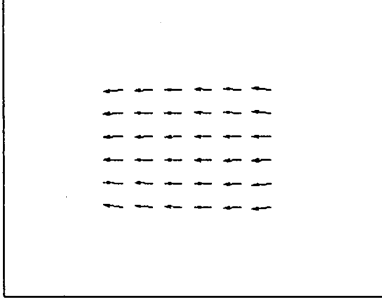
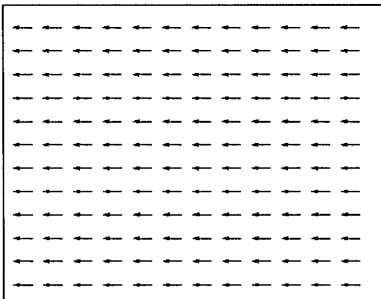
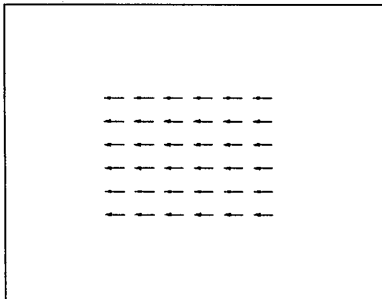
	Features evenly distributed	Features gathering at the center
$\omega_y$		
$t_x$		

Table 2.1: Image flow comparison for the case of egomotion arising from pure  $\omega_y$  vs. pure  $t_x$ . Sampling the flow across the entire image allows for the two patterns to be distinguished, while restricting the samples to the image centres makes such distinction much more difficult.

adaptively such that the number of points selected lie between specified minimum and maximum values. Example selected features are shown in Fig. 2.5b. Notice that the selected points still correspond to well textured loci that will yield correspondingly well defined SOEs (2.1). Also, gridded adaptive thresholding yields features well distributed spatially. Correspondingly, points  $\mathbf{x}_i$  that are used as input to the egomotion estimation algorithm, (2.23) and (2.24), are selecting according to the salient feature selection techniques described in this section.

### 2.3.3 Coarse-to-fine refinement

Coarse-to-fine (CTF) processing is a popular technique to help improve various algorithms, *e.g.*, stereo matching or motion estimation algorithms, so that they can tolerate larger magnitudes of disparity or image motion, *e.g.*, [73, 10]. The approach also helps algorithms avoid local minima and decreases processing time [16]. In the present case, relatively large magnitude egomotion correspondingly implies relatively large magnitude changes in visual spacetime (*e.g.*, orientation variations) and these are addressed by embedding the Gauss-Newton minimization (2.25) within a CTF refinement scheme. To be specific, incoming images are represented as (Gaussian) image pyramids, as shown in Fig. 2.6. The egomotion estimator is executed successively from the coarsest to finest levels, with each level taking the estimates from the previous level as initial conditions for its own re-



(a)



(b)



(c)

Figure 2.5: Illustration of the salient feature selection technique. (a) is the original image. (b) shows the feature candidates before post-processing to ensure relatively even distribution across the image, (c) presents the selected feature points following post-processing, which are more evenly distributed across the whole image as desired.

finement. Results of the Gauss-Newton optimization at the finest level are taken as the final answer. In the case of disparity estimation [73, 79], building image pyramids can be an essential step. In particular, the stereo algorithm that provides disparity input to the proposed egomotion algorithm makes use of coarse-to-fine pyramid processing [79]. At coarser levels the sizes of the images are smaller and so are the disparities, even while support regions aggregate over more information. During disparity estimation, initial estimates obtained at coarser levels are incrementally refined at finer levels. Similarly, for the purpose of motion estimation, the magnitude of image flow at coarser levels is smaller magnitude than at finer levels. However, rather than reducing the residual error of image disparities, image flow residual (or egomotion residual) should be addressed. Typically, the refinement procedure entails initial warping of the images at level  $l$  in the coarse-to-fine processing by the estimates at the previous level,  $l+1$ , to account for their results [10].

Processing an entire image sequence coarse-to-fine in a batch fashion would have two undesirable implications. First, all frames would need to be warped to a single reference frame (*e.g.*, by chaining instantaneous egomotion estimates), which would entail significant error accumulation for sequences of nontrivial length. Second, it would preclude on-line operation, as no estimates would be produced until an entire sequence is acquired. To address these shortcomings, the entire

coarse-to-fine estimation scheme is realized with a temporally sliding window. The number of frames in the window is equal to the number of temporal samples (taps) considered in the spatiotemporal oriented energy filtering, (2.2). (In the current implementation, 5 taps are considered to be in accord with the filtering employed in the spatiotemporal stereo matcher that provides the input binocular correspondences [79].) This approach allows for estimates to be incrementally produced as the imagery is acquired (albeit with an initial delay to acquire one temporal window of frames) and for image warping to be limited to the number of frames in the temporal window. In the current implementation, warping always is performed with respect to the centre frame of the window. Within any temporal window, use of the central frame as reference again minimizes the length of the sequence over which warps need to be chained.

## 2.4 Recapitulation

By way of summary, Fig. 2.7 provides a flow diagram that captures the entire proposed approach to egomotion estimation. Given a temporal stream of calibrated and rectified binocular imagery, processing proceeds as follows. First, the left and right image sequences are independently filtered to extract pointwise SOE measurements, (2.2), indicated as SOEs in Fig. 2.7. Second, binocular disparity is estimated pointwise [79], shown as Stereo matching and Disparity pyramids. Third,

salient feature points are extracted, Sec. 2.3.2, indicated as Salient features selection in Fig 2.7. Fourth, the egomotion estimator is executed, (2.25), in Egomotion estimation. At the start of estimation, the egomotion parameters are initialized identically to zero; estimation ends when the residual change between iterations is below a threshold ( $10^{-6}$ ) or a maximum number of iterations (50) is reached. The entire approach is embedded within a course-to-fine refinement scheme using a temporally sliding window.



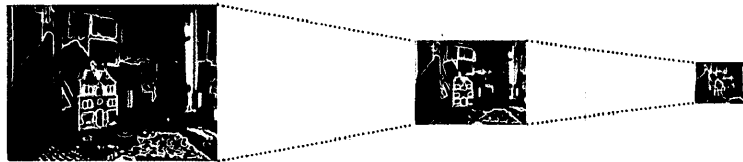


Figure 2.6: A Gaussian pyramid, with decreased resolution moving left-to-right. In course-to-fine processing, operations begin at the lowest (coarsest) resolution and proceed to the highest (finest).

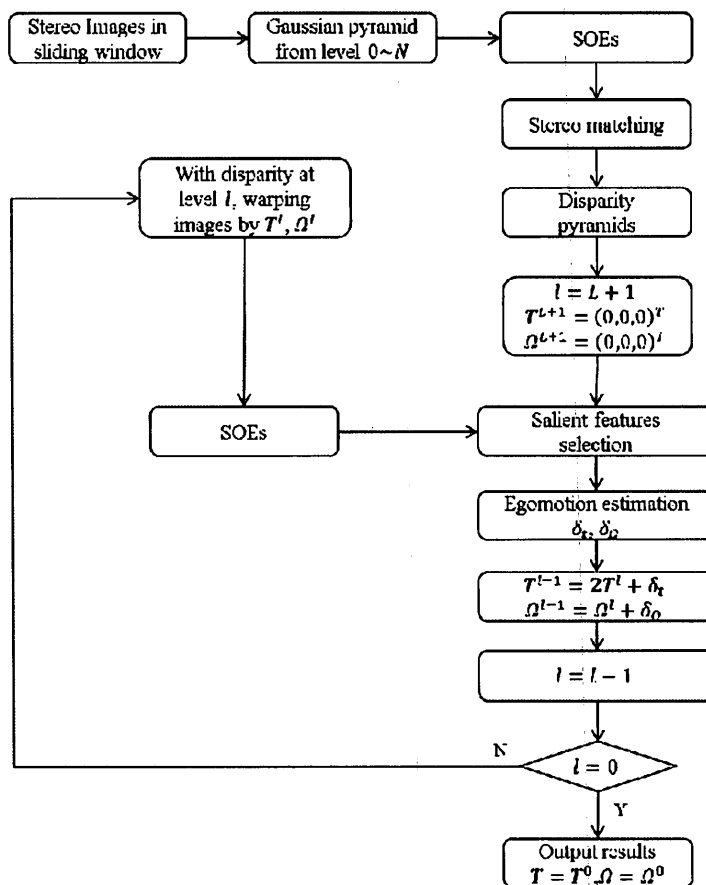


Figure 2.7: Flow diagram of camera egomotion estimation system. In this diagram,  $N$  is the maximum level in Gaussian pyramids and  $L$  is the coarsest level from which egomotion estimation is executed. This presented scheme is applied to a sliding temporal window across the entire input (binocular) image sequence. The size of the sliding window is set to the number of frames required to make SOE measurements.

### **3 Empirical evaluation**

The proposed approach to egomotion estimation, as summarized in Fig. 2.7, has been implemented in software. The software realization has been evaluated on laboratory and real-world datasets. For the sake of comparison, performance has been evaluated relative to three representative alternative egomotion estimation algorithms: two that have been implemented by the author as variants on extant approaches [24, 15] and one additional state-of-the-art algorithm with code downloaded from its author’s website [31].

#### **3.1 Datasets**

Evaluation of the proposed algorithm focuses on documenting its performance as a function of two key variables: egomotion speed and ability to perform in naturalistic scenarios. Correspondingly, two different datasets have been acquired. The first dataset was acquired in a laboratory setting with systematic variations in egomotion speed. The second dataset is acquired in real-world indoor and outdoor scenes.

### 3.1.1 Laboratory dataset

Laboratory datasets were acquired in York's Vision Lab. This calibrated facility allows for acquisition of imagery with groundtruth egomotion to support quantitative performance evaluation.

All imagery was captured with the same binocular video camera (a pair of PointGrey<sup>1</sup> Flea2 cameras) with a 6 cm stereo baseline using 75 degree horizontal field of view lenses for capture at 1024x768 spatial resolution. The same cluttered scene was viewed throughout; see Fig. 3.2. Egomotion was realized by attaching the cameras to an automated high precision motion control platform mounted on an optical bench, which also provided groundtruth readings. Fig. 3.1 provides views of the laboratory system. (See Appendix B for details of this system.)

The dataset consists of 7 videos capturing all different combinations of 3 degree-of-freedom (DOF) motion in a plane with systematic variation of velocities. Under the current notation, the parameters are given as  $t_x$ ,  $t_z$ , and  $\omega_y$ . These parameters are selected as they capture an important practical situation (ground plane motion) and due to mechanical constraints in the lab. Initially, each image sequence is acquired by advancing the motion platform incrementally based on the egomotion increments documented in Table 3.1. Following each increment, a binocular image pair is captured. Subsequently, egomotion speed is synthetically varied via temporal

---

<sup>1</sup><http://www.ptgrey.com>

Name	$t_x$ (mm)	$t_y$ (mm)	$t_z$ (mm)	$\omega_x$ (deg)	$\omega_y$ (deg)	$\omega_z$ (deg)
Lab- $t_x$	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Lab- $t_z$	0.0000	0.0000	2.0000	0.0000	0.0000	0.0000
Lab- $\omega_y$	0.0000	0.0000	0.0000	0.0000	-0.0300	0.0000
Lab- $t_x-t_z$	0.7000	0.0000	1.4000	0.0000	0.0000	0.0000
Lab- $t_x-\omega_y$	1.0000	0.0000	0.0000	0.0000	-0.0300	0.0000
Lab- $t_z-\omega_y$	0.0000	0.0000	1.4000	0.0000	-0.0300	0.0000
Lab- $t_x-t_z-\omega_y$	0.7000	0.0000	1.4000	0.0000	-0.0300	0.0000

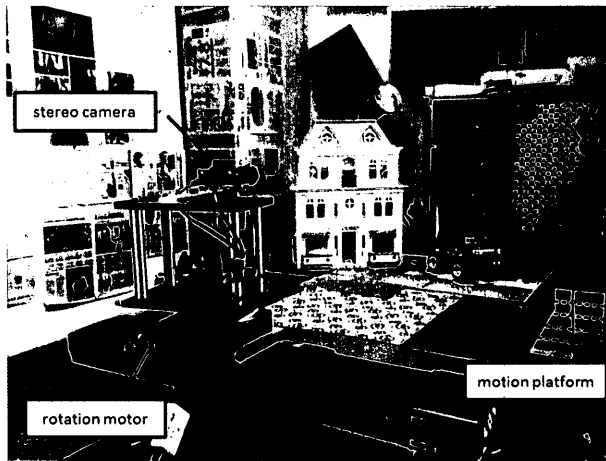
Table 3.1: Camera egomotion parameters in the different lab datasets. The various conditions are documented in the left most column. Subsequent columns in each row document the framewise increment in each egomotion parameter for the given condition. The units are millimeter for translation distance and degree for rotation angle.

subsampling of the acquired sequence. Considering the resulting image sequences as 30 frames/second videos, the subsampling yielded apparent speed increases in 15 steps for translation and rotation ranging 2.1 – 90 cm/sec. and 0.9 – 13.5 deg./sec., resp. Example images for the acquired sequences are shown in Fig. 3.2 and Appendix D.

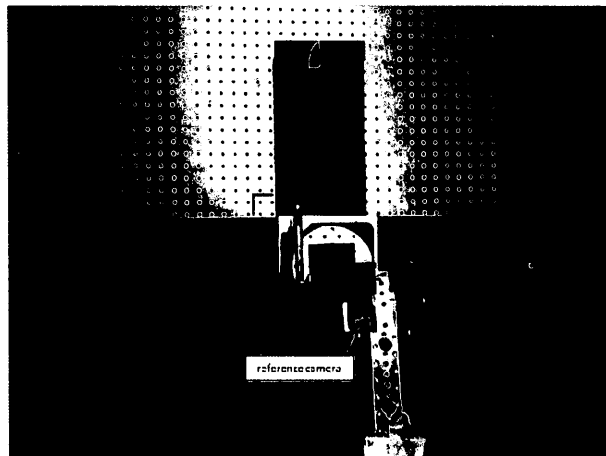
### **3.1.2 Naturalistic datasets**

To support evaluation of the egomotion estimator in more naturalistic settings, two additional datasets have been acquired. While these videos do not allow for quantitative evaluation in comparison to groundtruth, they do allow for evaluation in the presence of real-world scenes and with a wider range of egomotion parameter settings.

These naturalistic datasets were captured using the same binocular video camera used for the laboratory acquisitions. One dataset was acquired indoors in a cluttered office setting. An interesting aspect of this scene is its large areas occupied by textureless surfaces, which should challenge the algorithm. The second was acquired outdoors as the camera viewed a building exterior with foreground bushes, leaves and grass. An interesting aspect of this scene is that the wind was blowing, which causes motion beyond that arising from camera egomotion. Example images of both scenes are shown in Fig. 3.3 and Appendix D.



(a)



(b)

Figure 3.1: (a) The facilities for collecting laboratory datasets. (b) Top view of the two translational (indicated with straight, double headed arrows) and one rotational (indicated with circular arrow) motion stages.

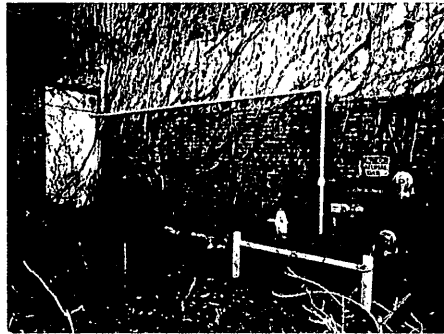
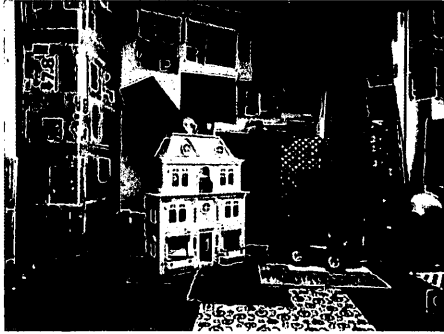


Figure 3.2: Sample left and right images in lab, indoor and outdoor datasets are shown from top to bottom, resp.



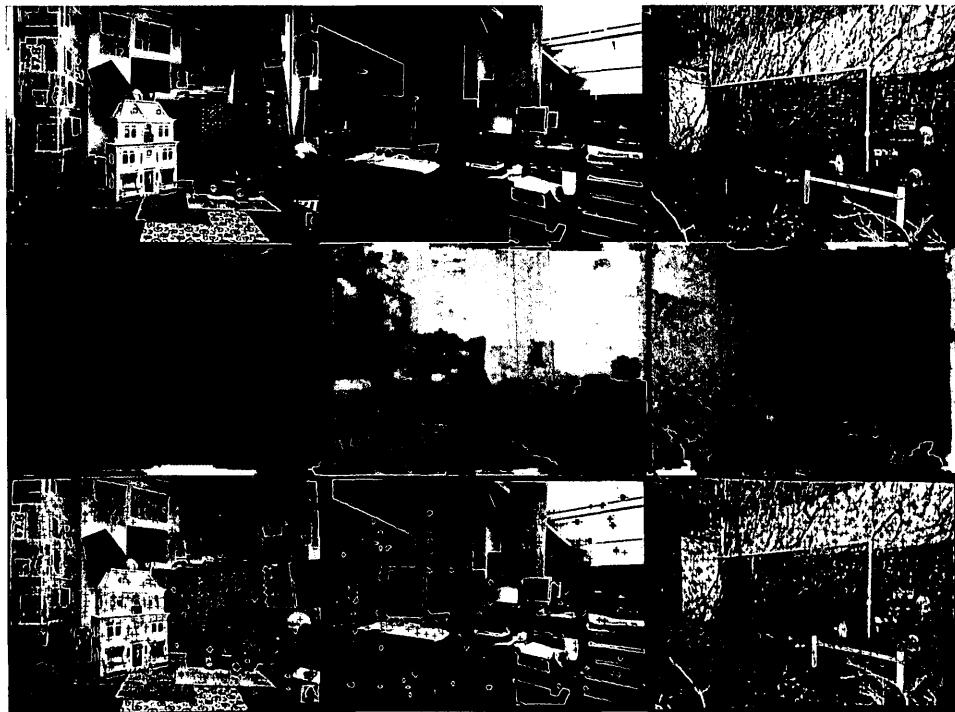


Figure 3.3: Sample images in the (left-to-right) laboratory, indoor and outdoor datasets (top row), as well as their corresponding disparity maps (middle row) and feature selection results (bottom row). Selected features are indicated as red plus signs.

In both naturalistic cases, a single binocular video sequence was acquired encompassing 6 DOF egomotion with the camera handheld. An attempt was made to move sequentially along each of the egomotion parameters, in order  $t_x$ ,  $t_y$ ,  $t_z$ ,  $\omega_x$ ,  $\omega_y$ ,  $\omega_z$ , to yield 6 temporal epochs within a single video.

### 3.2 Algorithms compared

Three alternative egomotion algorithms are considered for comparison to the proposed approach. The first, **DC**, is selected as it is an alternative that, similar to the proposed **SOE**, works without explicit projection of disparity measurements into world,  $(X, Y, Z)^T$  space, and previously outperformed such approaches [24]. This algorithm requires disparities that are matched across time. For the sake of fair comparison, the same disparity measurements and feature point selection used for the proposed approach also are used as input to **DC**. Temporal correspondences are established using the Lucas-Kanade algorithm [58] as implemented in OpenCV, with pyramids to increase capture range.

The second comparison algorithm (**BH**) is based on a classic passive navigation algorithm, proposed by Bruss and Horn [15]. While the original algorithm worked with optical flow recovered from monocular image sequences, it has been extended to work with binocular image sequences by the author to provide a better comparison with the proposed algorithm (Appendix C). As with **DC**, the needed disparities

are provided by the same algorithm used to support the proposed approach and temporal correspondences (optical flow) is recovered using the Lucas-Kanade algorithm with pyramid processing. The feature points are selected the same as for the SOE algorithm.

The third algorithm (**KGL**) is a state-of-the-art algorithm for binocular-based egomotion estimation as applied to visual odometry [51]. This approach operates by matching corner-like features across time consecutive stereo pairs. Egomotion subsequently is estimated based on trifocal tensor constraints between image triads. RANSAC [28] is used for outlier rejection and an Iterative Sigma Point Kalman Filter (ISPKF) is used for predictive filtering.

Parameter values for all three comparison algorithms were as suggested by their authors or as tuned for best performance on the present datasets.

### **3.3 Results**

#### **3.3.1 Laboratory image results**

All compared algorithms were executed on the laboratory dataset. Their instantaneous egomotion estimates were compared to groundtruth at 10 equally spaced times across each of the seven videos in the laboratory dataset video; mean and standard deviation of errors were calculated. Algorithms estimated 6 DOF egomo-

tion, even though only at most 3 were actuated. Results are plotted in Figs. 3.4 and 3.5.

For the pure  $t_x$  case, it is seen that **SOE** exhibits smaller error than the alternatives on the actuated  $t_x$ , essentially 0 error on the  $\Omega$  parameters and small error on the nonactuated  $t_y, t_z$ . **KGL** also shows small errors, but with a tendency to oscillate about 0 as speed varies. **BH** is comparable to **SOE** on all parameters except  $t_x$  and  $\omega_y$ , where it performs more poorly and slightly worse than **KGL** overall. **DC** is weakest, with error increasing at higher speeds for  $t_x, \omega_y$  and  $\omega_z$ . For the pure  $t_z$  case, all algorithms do well in yielding close to 0 error for the  $\Omega$  parameters. However, differences are apparent on **T**: **SOE** and **KGL** show similar small errors on the nonactuated  $t_x, t_y$ , but **SOE** shows better performance on  $t_z$  until at highest speeds it is equaled by **KGL**. **BH** performs similarly to **SOE** and **KGL** on **T** except  $t_z$ , where it shows increasing error and variance with speed. **DC** shows a marked increase of error for  $t_x$ , as speed increases. For pure  $\omega_y$ , all algorithms show small errors, but with **KGL** again oscillating.

For combined  $t_x, t_z$ , **SOE** has smallest errors for all nonactuated parameters and  $t_x$ . At lower speeds, it also shows smallest errors for  $t_z$ , but is equaled by **KGL** at higher speeds. **BH** is the second best and is comparable with **SOE** on most all parameters except on parameter  $t_y$  and  $\omega_y$  where it drifts to higher error rates. **KGL** generally is third smallest in error, but continues to oscillate as speed

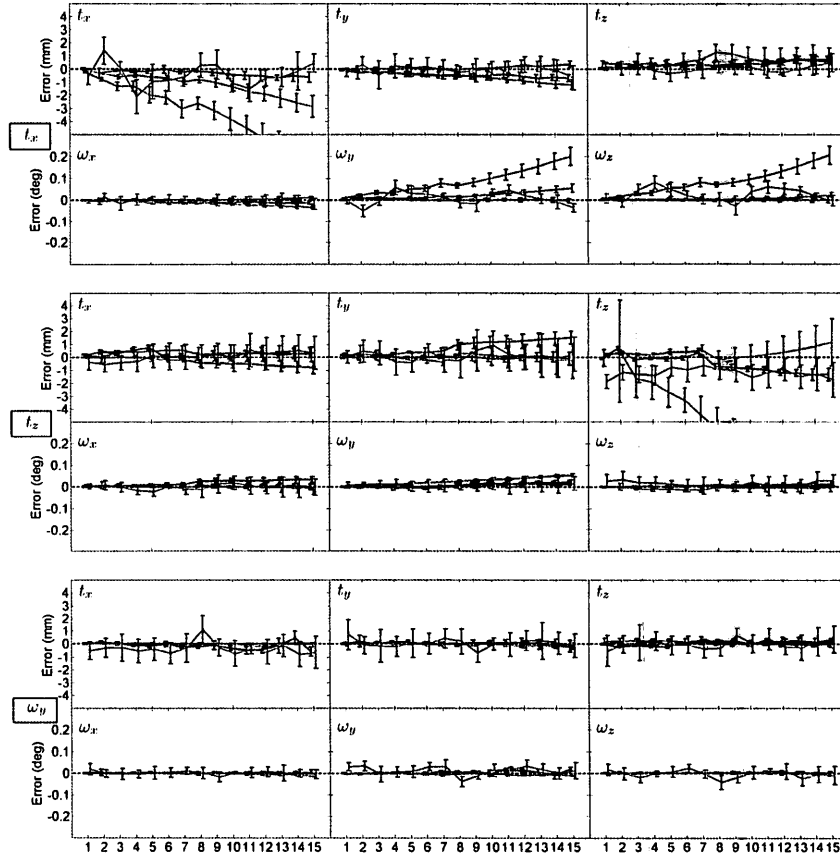


Figure 3.4: Results on laboratory dataset. Top-to-bottom are grouped error plots as actual egomotion is purely  $t_x, t_z, \omega_y$ , resp. Subplots show error mean and standard deviation for indicated parameters along the ordinate as speed increases along the abscissa. Blue, green, cyan and red denote results for **SOE** (proposed), **DC**, **BH** and **KGL**, resp. See text (Sec.3.1.1) for details of how the 15 levels along the abscissa correspond to egomotion speeds.

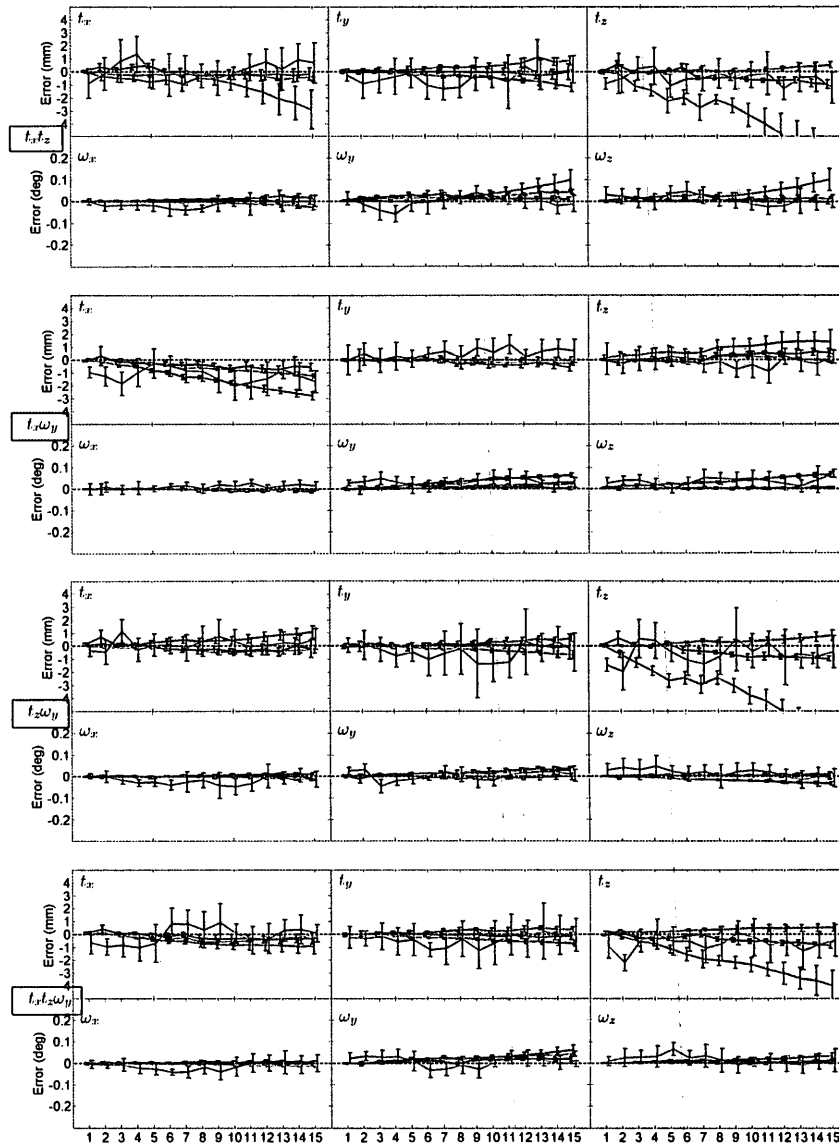


Figure 3.5: Results on laboratory dataset, part 2. Top-to-bottom are grouped error plots as actual egomotion is  $t_x t_z$ ,  $t_x \omega_y$ ,  $t_z \omega_y$ ,  $t_x t_z \omega_y$ , resp. Format otherwise same as Fig. 3.4.

varies. **DC** continues its trend of increased error with increased speeds. Combined  $t_x, \omega_y$  shows **SOE** with smallest error on all parameters. **BH** is slightly worse than **SOE**, but generally competitive. **KGL** again has the third smallest error (but still oscillating with changes in speed). **DC** also shows small errors, but larger than the alternatives. Combined  $t_z, \omega_y$  shows **SOE** and **BH** with generally smallest error rates, **KGL**'s tendency to oscillate about 0 error particularly pronounced (*e.g.* on  $t_y$  and  $t_z$ ) and **DC** outperforming **KGL**, except on  $t_x$  and  $t_z$ . Finally, combined  $t_x, t_z, \omega_y$  again shows **SOE** and **BH** with smallest error on all **T** parameters. **KGL** achieves similar error to **SOE** on  $\Omega$  and on  $t_z$  at highest speeds, but still is plagued by oscillation, especially on **T** errors. **DC** performs somewhat better than **KGL**, except on  $t_z$ .

For the purpose of testing how sensitive our algorithm is to different parameter settings, we set different global thresholds on the confidence map for salient feature point selection. The reason for altering this particular parameter is that it is the one that mostly influences the feature selection. Therefore, we set the threshold variously to 0.05, 0.10, 0.20, 0.30 and 0.40. Example images to illustrate the distribution of the selected feature points are shown in Fig. 3.6. The egomotion estimation results are presented in Fig. 3.7 and 3.8. As shown, the egomotion estimates vary little while the threshold is not greater than 0.20. For thresholds in excess of 0.2, performance notably decreases. Consideration of Fig. 3.6 shows

that the features selected for thresholds of 0.30 and 0.40 have become not only sparse, but also unevenly distributed. Recalling the discussion of the importance of having features selected relatively evenly across an image for egomotion estimation (Sec. 2.3.2.2), it becomes apparent why the algorithm is failing at such extreme selection thresholds. Overall these parameter variations indicate that the algorithm is stable with respect to the key variable of feature selection, provided it results in an even distribution of features.

### 3.3.2 Natural image results

All four compared algorithms were executed on both the indoor and outdoor naturalistic datasets. These datasets do not support comparison to groundtruth. Instead, the numerical values of the instantaneous egomotion estimates are plotted as time series in Fig. 3.9. The vertical lines in the plots indicate the six temporal epochs during which individual egomotion parameters were actuated, in order  $t_x, t_y, t_z, \omega_x, \omega_y, \omega_z$ .

Indoors, all algorithms sequentially increase/decrease their estimates reasonably as the  $\mathbf{T}$  parameters are actuated/deactuated. For  $\Omega$ , qualitatively correct estimates also are shown, as rotation is performed about each axis first in one direction and then back. A similar pattern of results is shown for outdoors. In both cases, all algorithms tend to show slight nonzero responses to parameters that the



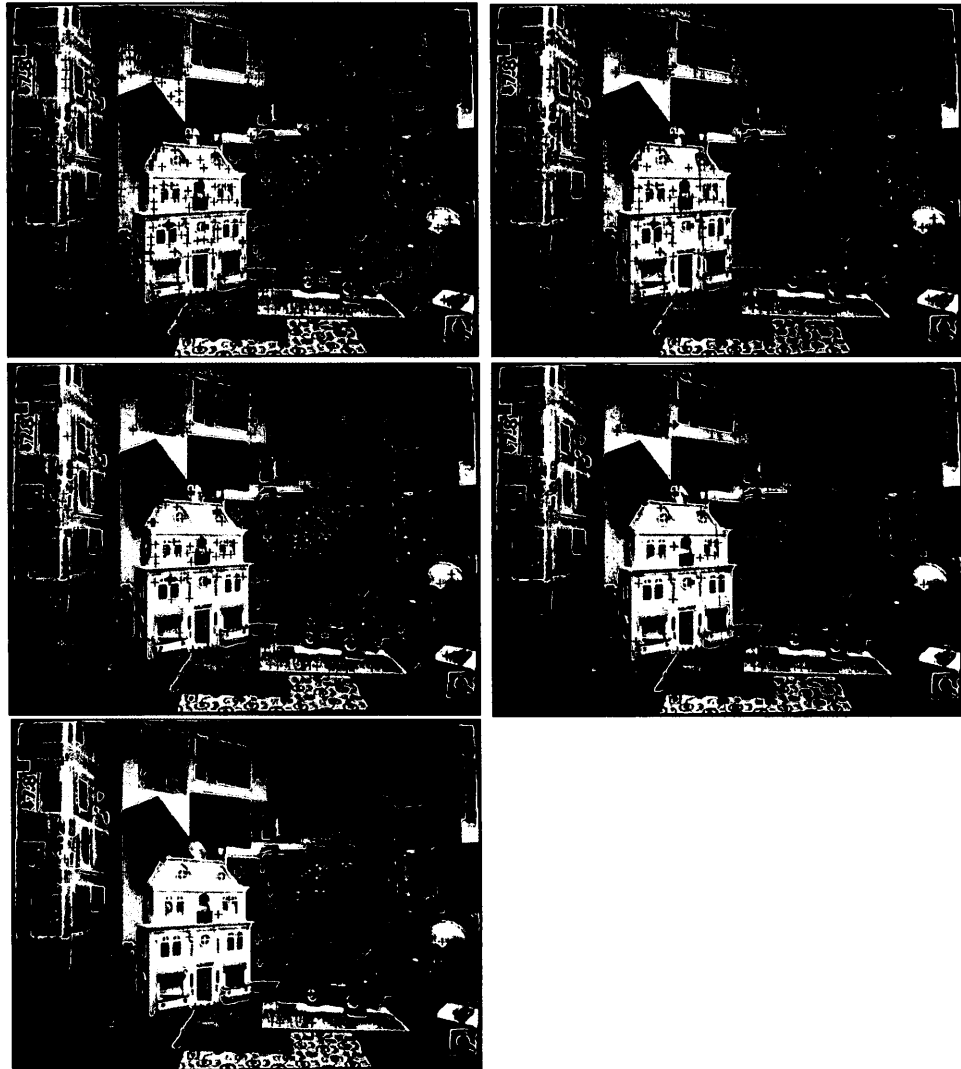


Figure 3.6: The distributions of the selected feature points based on the confidence threshold, left-to-right, top-to-bottom the threshold is set to 0.05, 0.10, 0.20, 0.30 and 0.40.

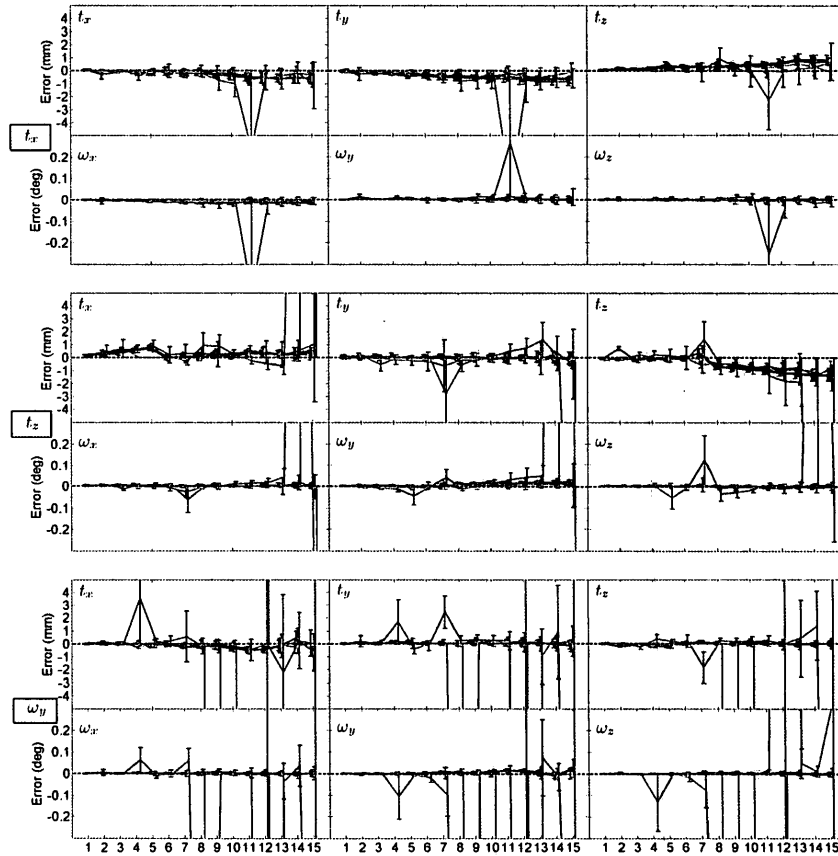


Figure 3.7: Results on lab dataset as the key parameter setting on feature selection is varied (part 1). Top-to-bottom are grouped error plots as actual egomotion is purely  $t_x$ ,  $t_z$ ,  $\omega_y$ , resp. Green, blue, cyan, red and magenta denote results as feature selection threshold is set 0.05, 0.10, 0.20, 0.30, 0.40, resp. See text for details.

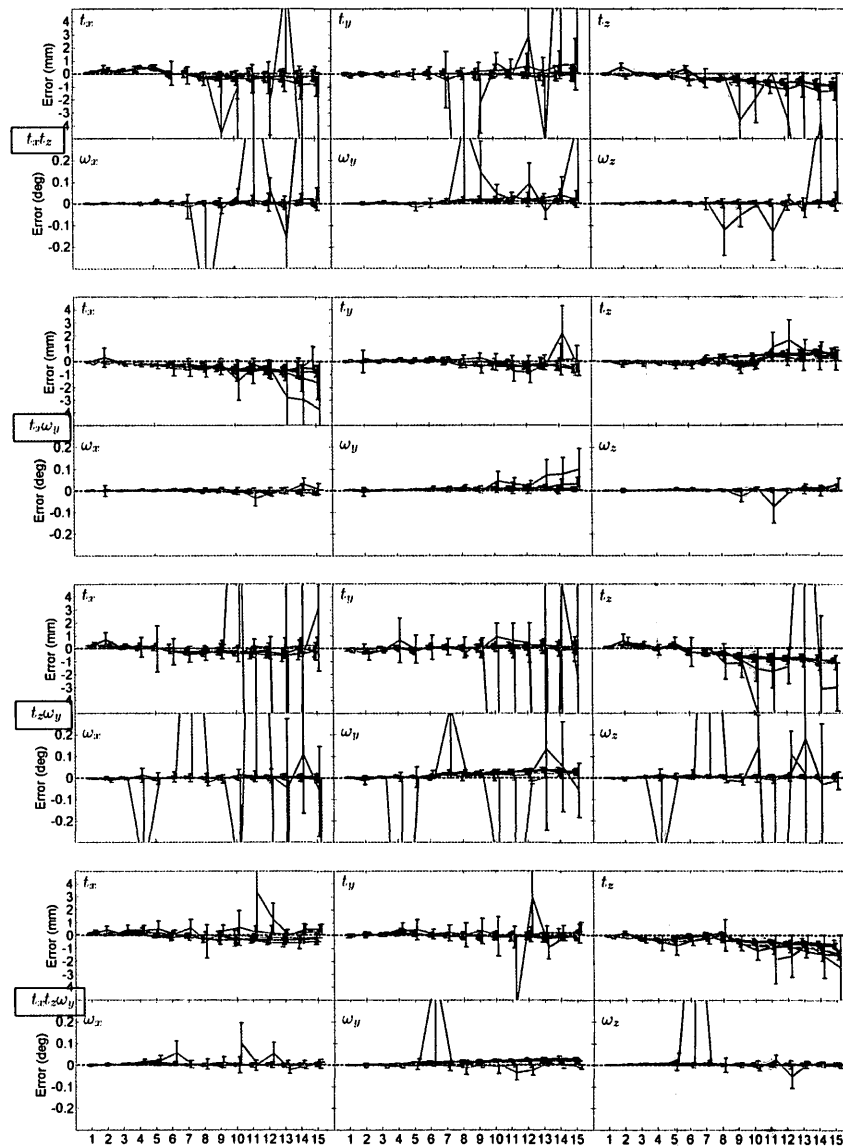
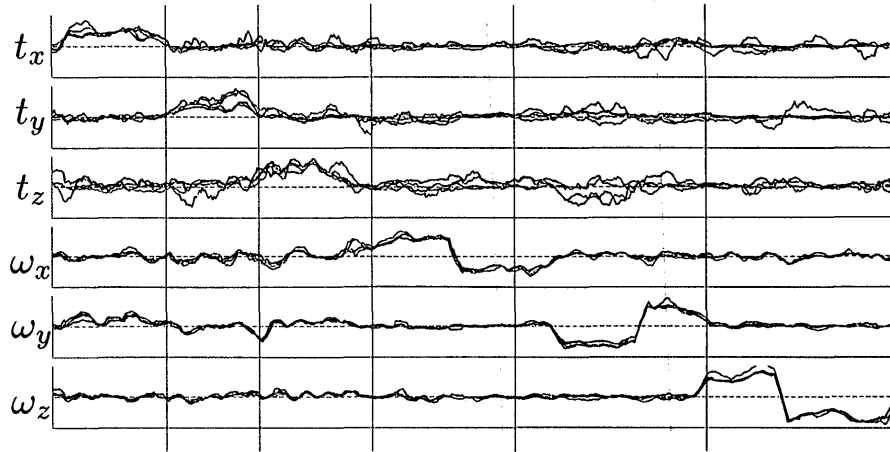
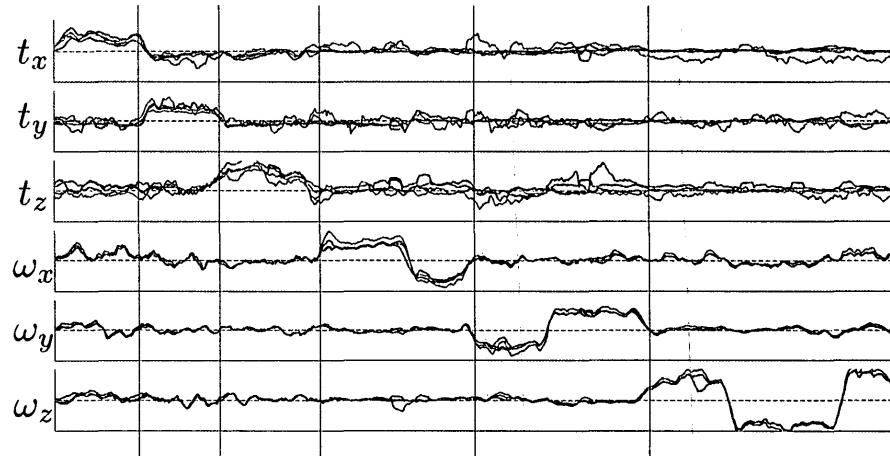


Figure 3.8: Results on lab dataset as the key parameter setting on feature selection is varied (part 2). Top-to-bottom are grouped error plots as actual egomotion is  $t_x t_z$ ,  $t_x \omega_y$ ,  $t_z \omega_y$ ,  $t_x t_z \omega_y$ , resp. Format otherwise same as Fig. 3.7



(a)



(b)

Figure 3.9: Estimated egomotion parameter values vs. time for indoor (top) and outdoor (bottom) datasets. Algorithm colour coding as in Fig. 3.4. See text for details.

camera operator attempted not to actuate. Given the general agreement between these estimates, they are likely due to the difficulty of holding the camera still along certain axes while actuating on another. Nevertheless, it appears that **SOE** gives more stable estimation across time and in better accord with the input videos than the alternatives, especially in the outdoors. For example, the greater tendency of **KLG** to oscillate about 0 for **T** during  $\Omega$  actuation as well as oscillation in its estimates of **T** during **T** actuation is not apparent in the video. Similarly, **DC**'s tendency to provide relatively pronounced responses to  $t_z$  during  $\Omega$  activation in the outdoor case does not appear to correspond to what is seen in the video. **BH** is performing as well as **SOE** most of the time, except sometimes following the trends of **DC**. For example, **BH** and **DC** both show variations in  $\omega_y$  and  $\omega_z$  during their actuations, which are not apparent in the captured video. In particular, **BH** and **DC** both show a bump or sink which cannot be observed while viewing the video taken with the camera handheld. Further, when **SOE** deviates from smoothness the video suggests its estimates follow the actual egomotion (*e.g.*  $t_z$  responses during  $t_y$  actuation indoors, where the operator inadvertently also actuated  $t_z$ ).

### 3.3.3 Execution rate

Our algorithm has been realized in C++ for execution on a PC with 3.40GHz processor and 16.0GB RAM. The execution time varies with the image size and

pyramid levels. For example, working only at the base (i.e., finest) pyramid level with images of size  $512 \times 384$  execution of the entire egomotion estimation algorithm for a pair of binocular images takes 84.17 milliseconds, beyond the time required for SOE filtering and stereo matching. Significantly, previous research has shown that both SOE filtering and stereo matching can be done in real-time, *e.g.*, [79]. Thus, overall the entire approach has potential for real-time applications. Finally, in all experiments the algorithm was found to converge in no more than 50 Gauss-Newton iterations and it never diverged.

### 3.4 Discussion

The results in comparison to groundtruth in the laboratory setting show that **SOE** exhibits best performance relative to three alternative algorithms. **BH** is the second best performer and can sometimes even equal that of **SOE**. Overall, however, its performance is demonstrably worse than that of **SOE**, *e.g.*, in its greater tendency to diverge at higher speeds. **KGL** is third best and tends for its error rates to oscillate with increased velocity. **DC** shows weakest performance, especially at higher speeds. These tendencies may underline the difficulty of establishing reliable temporal correspondences as egomotion (and hence image displacement) increases, a challenge **SOE** avoids by not requiring correspondences across time. Results on the natural imagery indicate that all algorithms perform qualitatively correctly,

with **SOE** showing somewhat more consistent estimates across time. Temporal consistency may result from the benefits of using spatiotemporal orientation analysis, which integrates more temporal information at a given instant (*e.g.* due to underlying filter support). Further, when **SOE** results do deviate from temporal smoothness, they appear to correspond to actual non-smooth variations in the ego-motion parameter values. Finally, run-time of the algorithm suggests potential for real-time deployment with further optimization and/or hardware realization.

## 4 Conclusion

### 4.1 Summary

In this thesis, we have presented a novel binocular camera egomotion estimation algorithm based on spatiotemporal oriented energy (SOE) distributions. Its fundamental theory, design, implementation and testing have been documented in detail. Highlights of the presented research are as follows.

- The relationship between binocularly matched spatiotemporal orientation distributions and camera egomotion has been explicitly analyzed and presented. It appears that this relationship has not been presented previously.
- Based on the developed analysis, a novel algorithm for camera egomotion estimation has been developed. The algorithm inputs binocularly matched measurements of spatiotemporal oriented energies and outputs estimates of camera egomotion as instantaneous translation and rotation. The algorithm does not require explicit temporal correspondences nor backprojection of binocular



correspondences into world,  $(X, Y, Z)$ , coordinates.

- The algorithm has been implemented in software and empirically evaluated both qualitatively and quantitatively on a variety of datasets. The datasets include laboratory data with groundtruth and real-world indoor and outdoor data.
- In comparison to a variety of representative alternative egomotion algorithms, the proposed approach yields best overall performance.

## 4.2 Future work

In the light of the work that has been described in this thesis, several directions for future work can be considered, as follows.

First, it is of interest to extend the algorithm so that it is better applicable to estimating egomotion when objects in the viewed scene are moving independently of the camera. Along those lines, one way to proceed is to make use of a robust estimation framework (*e.g.*, RANSAC [28] or a Hough transform [47]). Second, it is of interest to embed the egomotion estimator within a predictive filter to make additional use of the temporal history of the process. Here, Kalman [42] and particle [33] filters would be good candidates for consideration. Third, it would be interesting to embed the egomotion estimator within a larger system for sensor

platform odometry. Such developments could include integration with additional sensors (*e.g.*, inertial sensors). Finally, the current implementation works off-line. It would be of interest to reimplement the current approach and any extensions in real-time, *e.g.*, via GPU realizations. Real-time performance is not only relevant to enabling a wide range of applications (*e.g.*, visual odometry for mobile robots and other vehicles), but also to facilitate extensive testing.

## A SOE-based stereo matching and confidence measurement

Disparity information is a prerequisite for the proposed approach to egomotion estimation. Here, we apply Sizintsev and Wildes’s work [79] to recover the disparity map. There are basically two reasons why we choose this algorithm. First, it is a state-of-the-art disparity estimation algorithm and the accuracy of the proposed egomotion estimator will depend on the input disparity accuracy. Notably, the performance of the algorithm on various datasets has been demonstrated to yield superior performance to a variety of alternative algorithms [79]. Moreover, it is a point-wise algorithm and generates dense disparity maps, without any scene rigidity assumption. Second, this algorithm is also based on spatiotemporal oriented energy distributions. Thus, it is of interest to determine how well an entirely SOE-based approach, both disparity and egomotion estimation, can perform.

As described in [79], a binocular match constraint between corresponding orientations in visual spacetime can be specified in terms of a spatiotemporal epipolar

constraint [80]

$$\hat{\mathbf{w}}^r = H\hat{\mathbf{w}}^l, \quad \text{where } H = \begin{pmatrix} 1 + h_1 & h_2 & h_3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad (\text{A.1})$$

where  $\hat{\mathbf{w}}^l$  and  $\hat{\mathbf{w}}^r$  represent corresponding  $(x, y, t)$  orientations in binocular (left and right, resp.) visual spacetime. Here,  $h_1, h_2, h_3$  encapsulate the effects of binocular viewing, 3D motion between the cameras and scene as well as 3D scene structure [80]. The resulting stereo matching algorithm minimizes the sum of squared errors across all  $m$  oriented energy measurements (2.1) as

$$\sum_{m=1}^M \varepsilon_m^2(\mathbf{x}^l, \mathbf{x}^r) = \sum_{m=1}^M [E^r(I^r(\mathbf{x}^r); \hat{\mathbf{w}}_m^r) - E^l(I^l(\mathbf{x}^l); \hat{\mathbf{w}}_m^l)]^2, \quad (\text{A.2})$$

where the notational convention is adopted that

$$E^r(I^r(\mathbf{x}^r); \hat{\mathbf{w}}_m^r) = \hat{E}(\mathbf{x}; \hat{\mathbf{w}}) \quad (\text{A.3})$$

as applied to the right image,  $I^r$ , and  $E^l$  is correspondingly defined. Combined with (A.1), we have

$$\sum_{m=1}^M \varepsilon_m^2(\mathbf{x}^l, \mathbf{x}^r) = \sum_{m=1}^M [E^r(I^r(\mathbf{x}^r); \frac{H\hat{\mathbf{w}}_m^l}{\|H\hat{\mathbf{w}}_m^l\|}) - E^l(I^l(\mathbf{x}^l); \hat{\mathbf{w}}_m^l)]^2. \quad (\text{A.4})$$

The above error function (A.4) is minimized by setting the corresponding gradient with respect to  $\mathbf{h} = [h_1, h_2, h_3]^T$  to zero and subsequently solving for  $\mathbf{h}$  [80]. Readers are encouraged to refer to [80] for more details.

## B Details of laboratory image acquisition

### B.1 Motion control platform

The facilities for collecting the laboratory dataset are shown in Fig. 3.1a. A Newport optical bench serves as the base. At the bottom-left part of the image, two translational motion platforms are stacked perpendicularly to provide  $t_x$  and  $t_z$  translation for the cameras. On top of the upper translational platform, a rotation platform is mounted, which allows the camera to rotate around the  $Y$ -axis. The two PointGray Flea2 cameras are mounted on a steel plate, which further attaches them atop the rotational platform. Notably, the controllers for all 3 motion platforms are programmable to yield precisely controlled movements. The top view (Fig. 3.1b) illustrates how the camera can be translated and rotated. As shown in the figure, the viewed scene is constructed with various objects to guarantee sufficient depth variation, including a house model, vehicle model and lamp. In addition, to provide texture, a part of the table is covered by a patterned cloth and newspapers are posted on the wall.

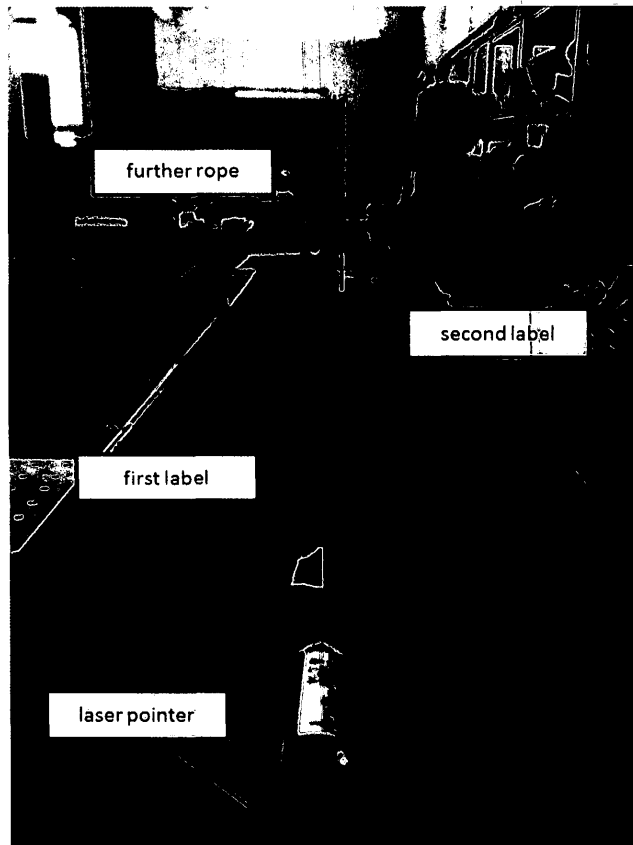


Figure B.1: The laser pointer is placed along the centre of the translational stage with the aid of labels affixed to the stage. Its distant projection indicates where the rope along the same line must be positioned. The process is repeated to position a second, nearer rope along the same line.

## B.2 Platform calibration

For present purposes, motion platform calibration is concerned with aligning the translational stages with the camera  $X$  and  $Z$  axes and ensuring that the rotation occurs about the camera's centre of projection and orthogonal to the  $X$  and  $Z$  axes.

We applied a simple but efficient method for adjusting the location of the reference camera. Note that the two translational stages are mounted perpendicularly using the attachment plates provided by the manufacturer. Similarly, the rotational stage is affixed to the upper translational stage using attachment plates so that its axis of rotation is perpendicular to the lower stages.

Camera alignment is accomplished via ensuring that points taken at the reference (left) image plane centre, the centre of the translational stage that serves as the  $Z$ -axis and a distant third point all lie along a line. To facilitate the alignment, two ropes are hung so that they both lie above the line along the centre of the  $Z$ -axis translational stage. A laser pointer is used to align the ropes, as shown in Fig. B.1. With the two ropes aligned, the camera is adjusted by translating it along its  $X$  and  $Z$  axes with the use of micro adjustment stages such that the two aligned ropes always image as overlapped at the (left) image centre, even as the motion controller actuates translation along the  $Z$ -axis and rotation about the  $Y$ -axis. The procedure is illustrated in Fig. B.2.

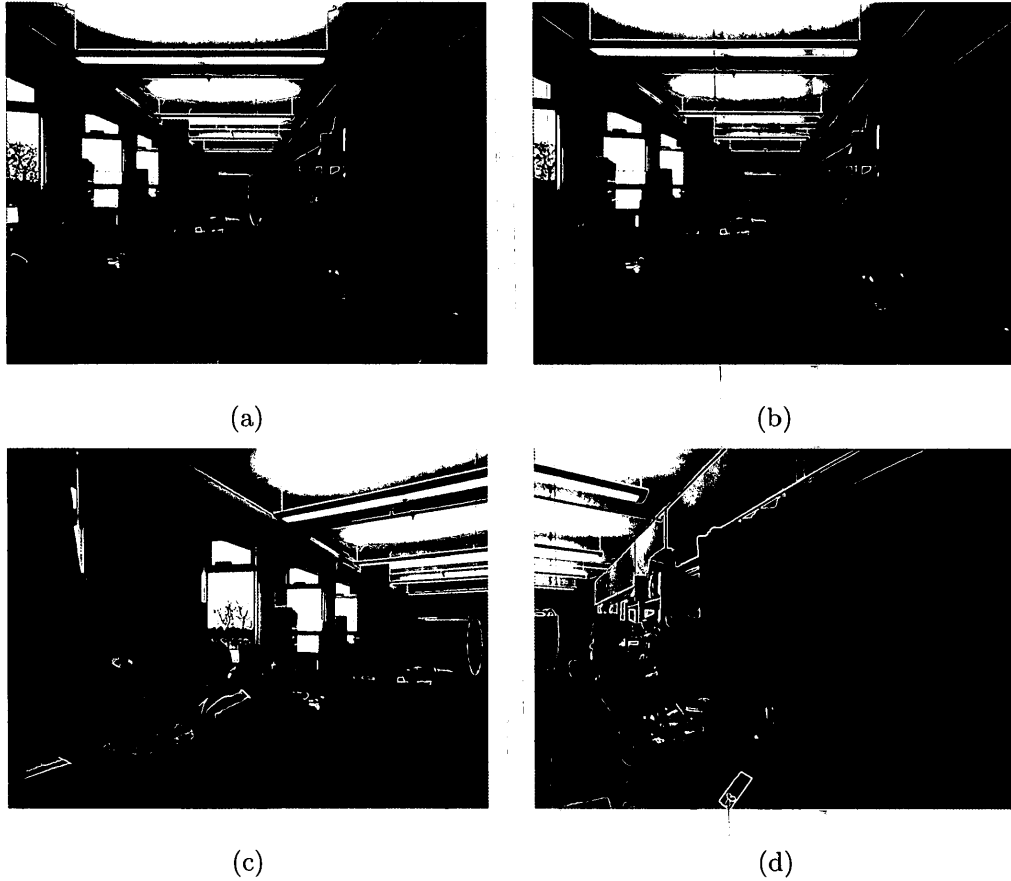


Figure B.2: Illustration of how to calibrate the motion platforms. The two ropes are viewed from the reference camera in difference cases. (a) and (b) show views when the camera is translated by the motion controller from its farthest to nearest extents (resp.) along the  $Z$ -axis. (c) and (d) are views while the reference camera is rotated to the extreme left and right. Since the two ropes are overlapped in all these 4 cases, the location of the reference camera is confirmed to be at the center of the rotation motor and  $Z$ -translation is aligned with the camera optical axis.



## C Revised Bruss and Horn algorithm

This appendix documents a novel binocular extension of the classic Bruss and Horn egomotion algorithm [15]. The algorithm requires as input optical flow and temporal differences of disparity, i.e., disparity flow. Optical flow is recovered via the OpenCV implementation of the Lucas-Kanade algorithm operating over image pyramids [58, 13]. Disparity is recovered using the same disparity estimator used to provide input to the proposed **SOE** algorithm, [78]. The needed temporal disparity differences are calculated by subtracting disparity estimates that are brought into correspondence across time by the optical flow field. Features are selected for input to the algorithm in the same fashion as used for the **SOE** algorithm, described in Section 2.3.2.

Let  $\hat{u}^i$ ,  $\hat{v}^i$  and  $\hat{\delta}d^i$  be the input flow of pixel  $i$ , where  $i = 1, 2, \dots, N$  (the number

of valid feature points). Recall from the equations (2.8) and (2.11), ideally we have

$$\begin{aligned}\hat{u}^i &= u^i(\mathbf{T}, \Omega), \\ \hat{v}^i &= v^i(\mathbf{T}, \Omega), \\ \hat{\delta d}^i &= \delta d^i(\mathbf{T}, \omega).\end{aligned}\tag{C.1}$$

with the left hand sides parametric representations of the flow in terms of egomotion parameters. Correspondingly, we consider an error measure of the form

$$F^i(\mathbf{T}, \Omega) = (\hat{u}^i - u^i(\mathbf{T}, \Omega))^2 + (\hat{v}^i - v^i(\mathbf{T}, \Omega))^2 + (\hat{\delta d}^i - \delta d^i(\mathbf{T}, \Omega))^2\tag{C.2}$$

to capture the discrepancy between the observed and modeled flow under the current egomotion estimates. Accordingly, we can obtain the estimation of  $\mathbf{T}, \Omega$  by minimizing

$$\sum_{i=1}^N F^i(\mathbf{T}, \Omega)\tag{C.3}$$

with respect to the egomotion parameters.

Solution is had via standard methods for solving least-squares problems [84]: Differentiate the objective, (C.3), with respect to each of the egomotion parameters, set each of the resulting six equations to zero and rearrange to isolate the desired (egomotion) parameters.

## D Example Sequences

In this appendix, we provide example image sequences derived from the laboratory and naturalistic datasets that are used in empirical evaluation of the developed approach to egomotion estimation. In all cases, the left image of the binocular dataset is shown. For indication of the difference between left and right views as well as the estimated disparity, see Fig. 3.2 and 3.3 in the main text.

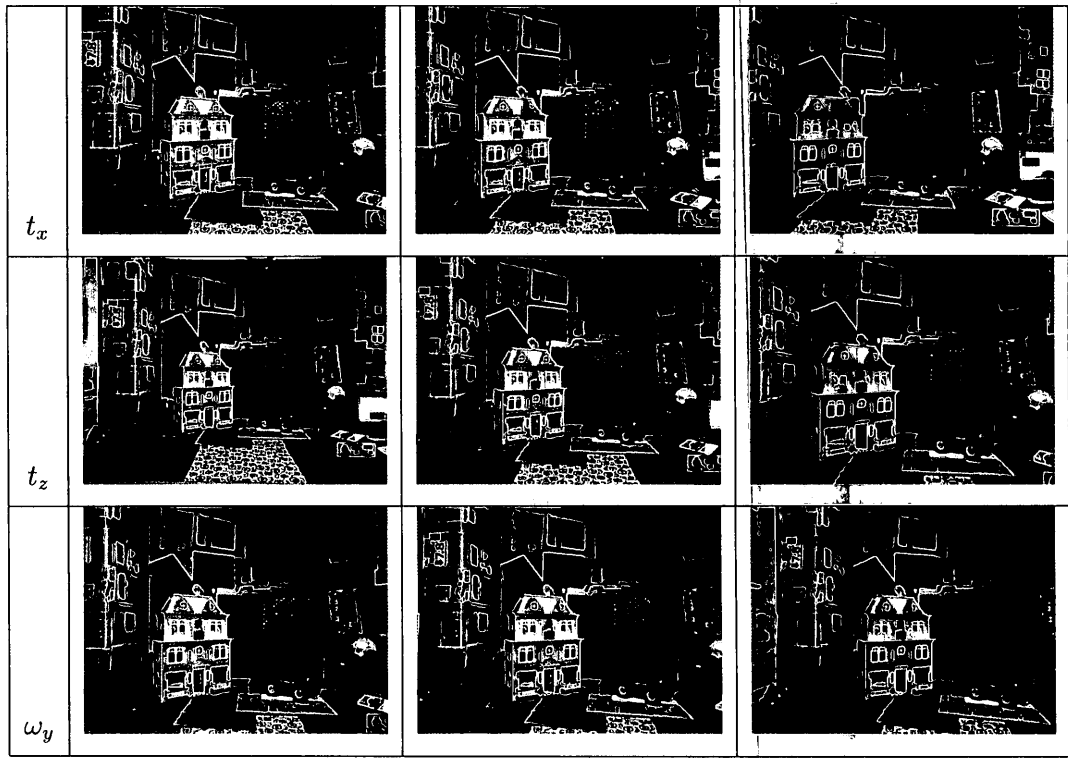


Table D.1: Example images from the laboratory dataset (Part 1). The labels in the left most column document the actuated egomotion that is the subject of each row. The three images in each row were taken from near the beginning, middle and end of the labelled sequence.

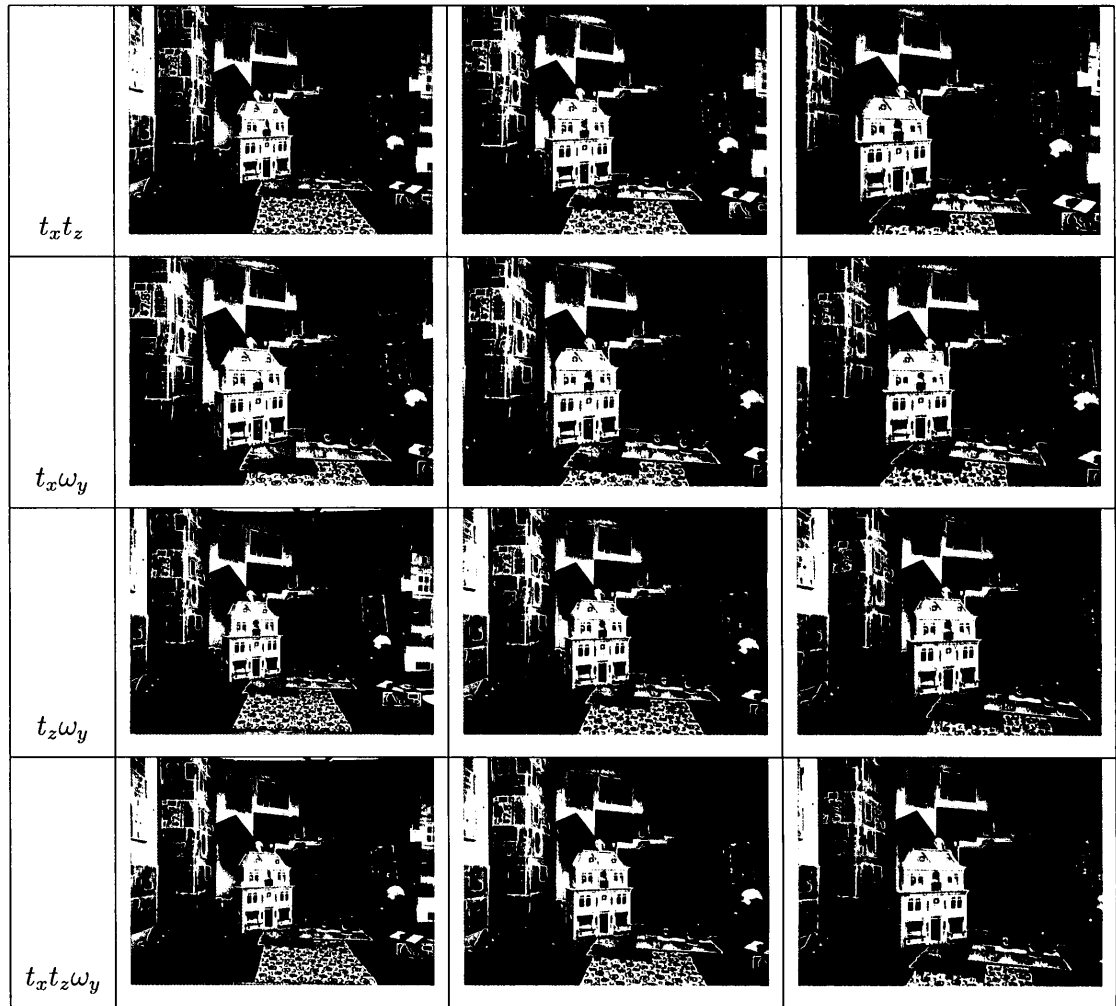


Table D.2: Example images from the laboratory dataset (Part 2). The labels in the left most column document the actualized egomotion that is the subject of each row. The three images in each row were taken from near the beginning, middle and end of the labelled sequence.

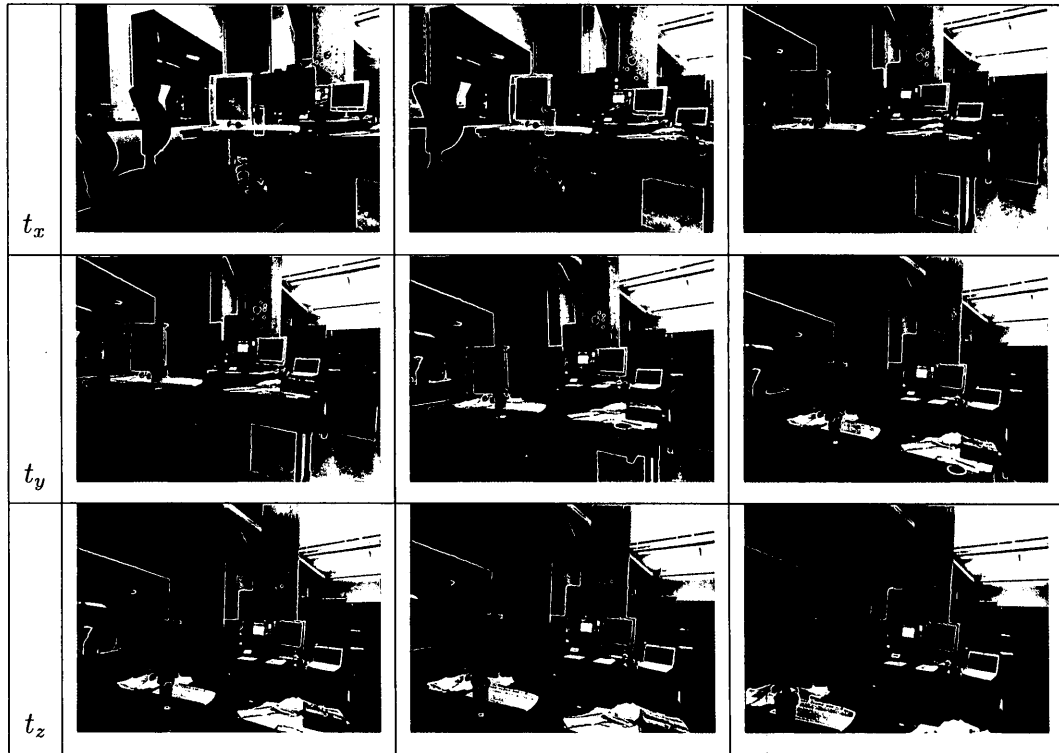


Table D.3: Example images from the naturalistic indoor dataset (Part 1). The labels in the left most column document the actuated egomotion that is the subject of each row. The three images in each row were taken from near the beginning, middle and end of the labelled sequence.

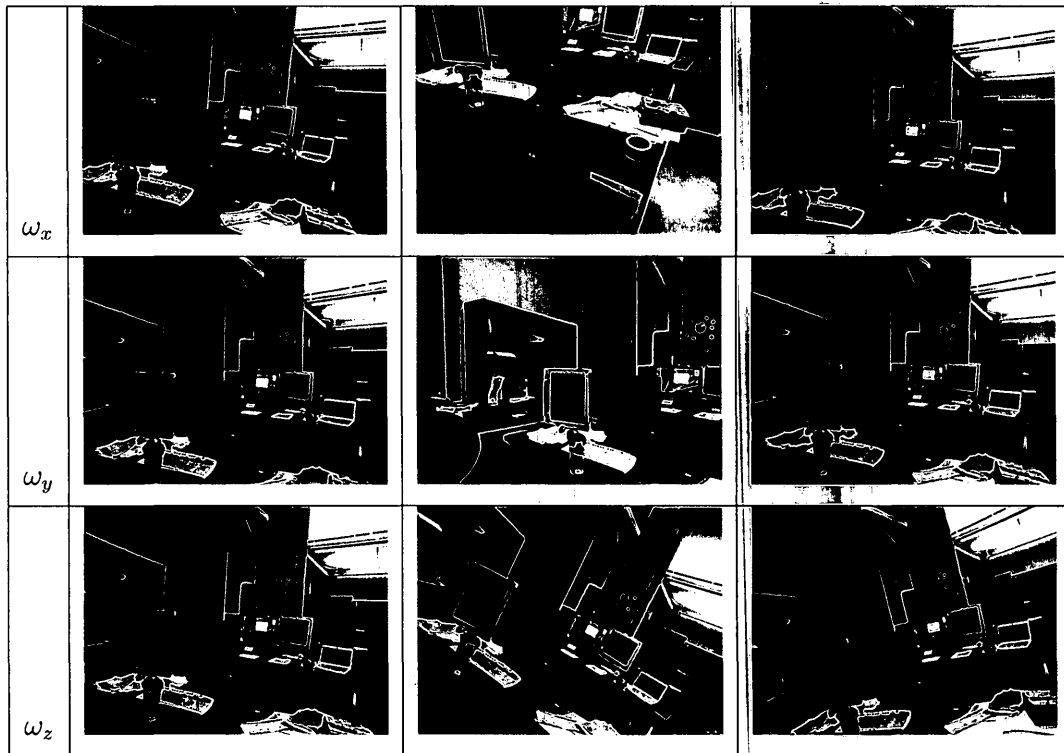


Table D.4: Example images from the naturalistic indoor dataset (Part 2). The labels in the left most column document the actuated egomotion that is the subject of each row. The three images in each row were taken from near the beginning, middle and end of the labelled sequence.

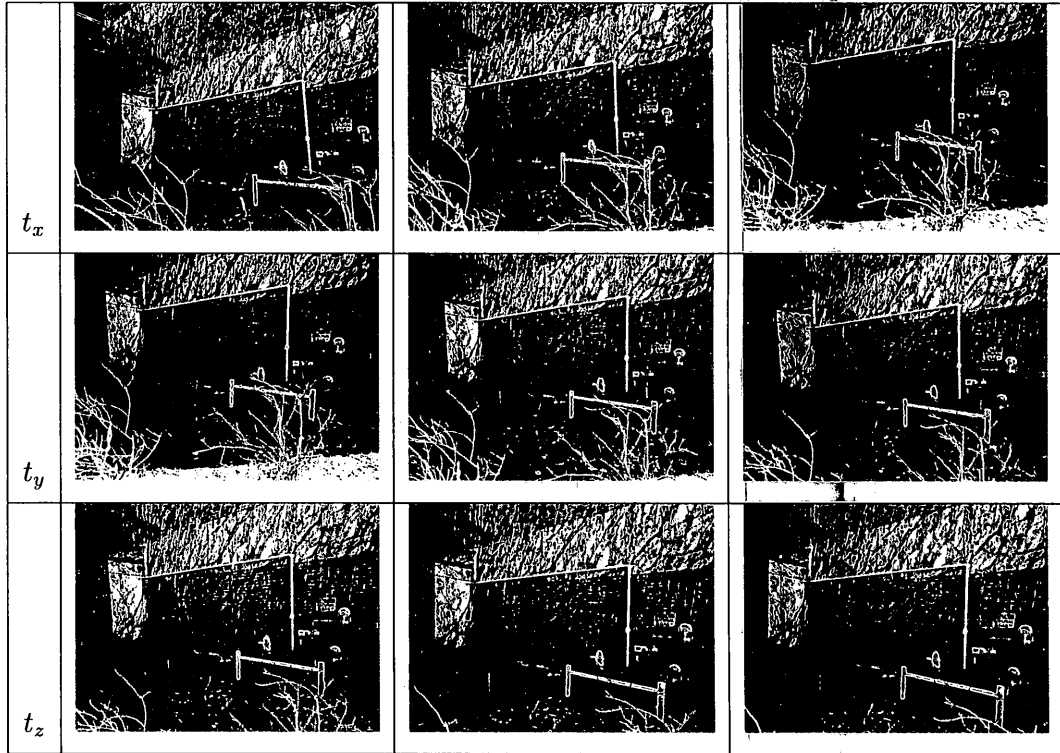


Table D.5: Example images from the naturalistic outdoor dataset (Part 1). The labels in the left most column document the actuated egomotion that is the subject of each row. The three images in each row were taken from near the beginning, middle and end of the labelled sequence.



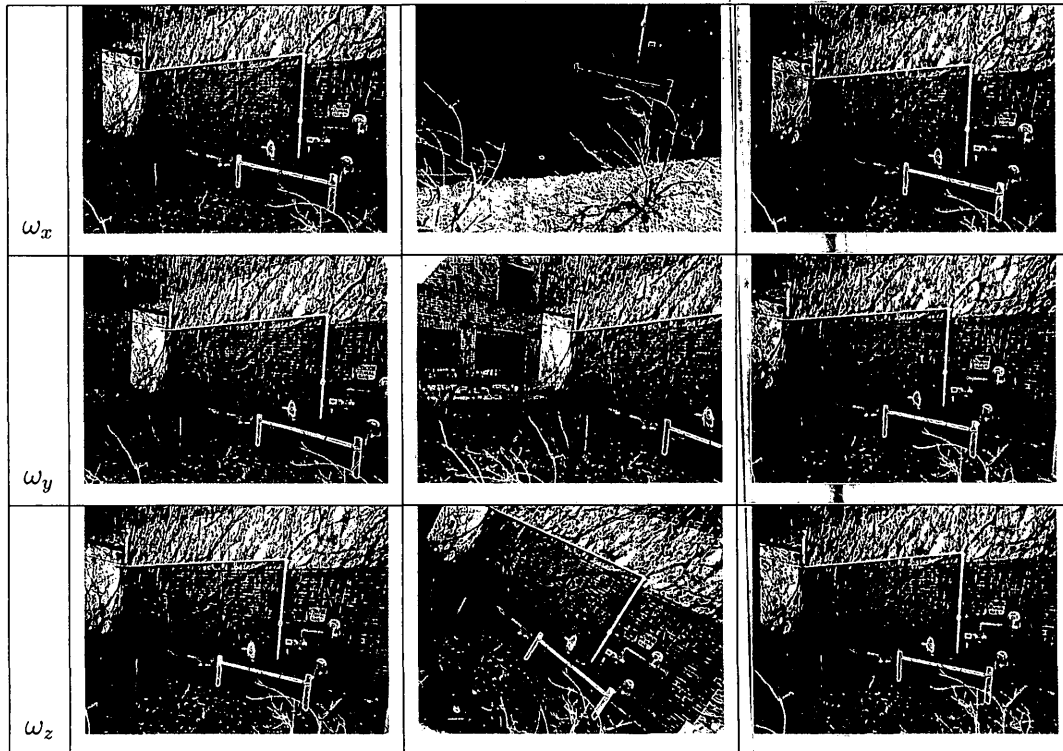


Table D.6: Example images from the naturalistic outdoor dataset (Part 2). The labels in the left most column document the actuated egomotion that is the subject of each row. The three images in each row were taken from near the beginning, middle and end of the labelled sequence.

## Bibliography

- [1] E. H. Adelson and J. R. Bergen. Spatiotemporal Energy Models for the Perception of Motion. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, 2(2):284–99, 1985.
- [2] H. Akbarally and L. Kleeman. A sonar sensor for accurate 3d target localization and classification. In *Proceedings of the International Conference on Robotics and Automation*, pages 3003–3008, 1995.
- [3] K. Arun, T. Huang, and S. Blostein. Least-squares Fitting of Two 3-D Point Sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(5):698–700, 1987.
- [4] J. Aulinas, Y. Petillot, J. Salvi, and X. Lladó. The SLAM problem: a survey. In *Proceedings of Conference on Artificial Intelligence Research and Development*, pages 363–371, 2008.
- [5] S. Avidan and A. Shashua. Novel view synthesis in tensor space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1034–1040, 1997.
- [6] R. Azuma. A Survey of Augmented Reality. *Presence: Teleoperators and Virtual Environments*, 4:355–385, 1997.
- [7] R. Azuma and Y. Baillet. Recent Advances in Augmented Reality. *IEEE Computer Graphics and Applications*, 21:34–47, 2001.
- [8] H. Badino. A robust approach for ego-motion estimation using a mobile stereo platform. *Proceedings of International Conference on Complex Motion*, pages 198–208, 2007.
- [9] N. M. Barbour, J. Elwell, R. H. Setterlund, and G. Schmidt. Inertial instruments: Where to now? In *Proceedings of the AIAA Guidance, Navigation and Control Conference*, pages 10–12, 1992.

- [10] J. R. Bergen, P. Anandan, K. Hanna, R. Hingorani, and J. Hanna. Hierarchical model-based motion estimation. In *Proceedings of the European Conference on Computer Vision*, pages 237–252, 1992.
- [11] P. Besl and N. McKay. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.
- [12] J. Borenstein, L. Feng, and H. R. Everett. *Navigating Mobile Robots: Systems and Techniques*. A. K. Peters, Ltd., Natick, MA, 1996.
- [13] G. Bradski. The opencv library. *Dr. Dobb's Journal of Software Tools*, 25(11):120–126, 2000.
- [14] M. Brown, D. Burschka, and G. Hager. Advance in computational stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8):993–1008, 2003.
- [15] A. Bruss and B. Horn. Passive navigation. *Computer Vision, Graphics, and Image Processing*, 21(1):3–20, 1983.
- [16] P. P. Burt. Smart sensing within a pyramid vision machine. In *Proceedings of the IEEE*, pages 1006–1015, 1988.
- [17] J. B. Campbell. *Introduction to remote sensing*. Guilford Press, New York City, NY, 2002.
- [18] K. Cannons, J. Gryn, and R. P. Wildes. Visual tracking using a pixelwise spatiotemporal oriented energy representation. In *Proceedings of the European Conference on Computer Vision*, pages 511–524, 2010.
- [19] M. J. Caruso. Applications of magnetoresistive sensors in navigation systems. *Society of Automotive Engineers (SAE) Transactions*, 106:1092–1098, 1997.
- [20] P. Corke, D. Strelow, and S. Singh. Omnidirectional visual odometry for a planetary rover. In *Proceeding of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2–7, 2004.
- [21] G. Dahlquist and A. Bjork. *Numerical Methods*. Prentice-Hall, Englewood Cliffs, NJ, 1974.
- [22] D. Demirdjian and T. Darrell. Motion Estimation from Disparity Images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 213–218, 2001.

- [23] D. Demirdjian and R. Horaud. MotionEgomotion Discrimination and Motion Segmentation from Image-Pair Streams. *Computer Vision and Image Understanding*, 78(1):53–68, 2000.
- [24] K. Derpanis and P. Chang. Closed-form linear solution to motion estimation in disparity space. *Proceedings of the IEEE Intelligent Vehicles Symposium*, pages 268–275, 2006.
- [25] K. G. Derpanis and R. P. Wildes. Spacetime texture representation and recognition based on a spatiotemporal orientation analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(6):1193–1205, 2012.
- [26] G. Egnal, M. Mintz, and R. P. Wildes. A stereo confidence metric using single view imagery with comparison to five alternative approaches. *Image and Vision Computing*, 22(12):943–957, 2004.
- [27] C. Fermler, Y. Aloimonos, and Y. Aloimonos. Qualitative egomotion. *International Journal of Computer Vision*, 15:7–29, 1993.
- [28] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.
- [29] B. P. Flannery, W. H. Press, S. A. Teukolsky, and W. H. Vetterling. *Numerical recipes in C*. Cambridge University Press, Cambridge, England, 2002.
- [30] W. Freeman and E. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991.
- [31] A. Geiger, J. Ziegler, and C. Stiller. StereoScan: Dense 3d reconstruction in real-time. *Proceedings of the IEEE Intelligent Vehicles Symposium*, pages 963–968, 2011.
- [32] H. Goldstein. *Classical mechanics*. Addison-Wesley Pub. Co., Boston, MA, 1980.
- [33] N. Gordon, D. Salmond, and A. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, 140:107–113, 1993.
- [34] G. Granlund and H. Knutsson. *Signal Processing for Computer Vision*. Kluwer Academic Publisher, Dordrecht, The Netherlands, 1995.

- [35] K. Hanna. Direct multi-resolution estimation of ego-motion and structure from motion. In *Proceedings of the IEEE Workshop on Visual Motion*, pages 156–162, 1991.
- [36] K. Hanna and N. Okamoto. Combining stereo and motion analysis for direct estimation of scene structure. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 357–365, 1993.
- [37] R. Haralick, H. Joo, and C. Lee. Pose estimation from corresponding point data. *IEEE Transactions on Systems, Man and Cybernetics*, 19(6):1426–1446, 1989.
- [38] R. I. Hartley and P. Sturm. Triangulation. *Computer Vision and Image Understanding*, 68(2):146–157, 1997.
- [39] M. Harville, A. Rahimi, T. Darrel, G. Gordon, and J. Woodfill. 3D Pose Tracking with Linear Depth and Brightness Constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 206–213, 1999.
- [40] D. J. Heeger. Model for the extraction of image flow. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, 4:1455–1471, 1987.
- [41] D. J. Heeger and A. D. Jepson. Subspace methods for recovering rigid motion I: Algorithm and implementation. *International Journal of Computer Vision*, 7(2):95–117, 1992.
- [42] J. Heel. Dynamic motion vision. *Robotics and Autonomous Systems*, 6(3):297–314, 1990.
- [43] A. Hogue and M. Jenkin. Development of an Underwater Vision Sensor for 3D Reef Mapping. In *Proceeding of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5351–5356, 2006.
- [44] B. Horn. *Robot vision*. MIT Press, Cambridge, MA, Cambridge, MA, 1986.
- [45] B. Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, 4(4):629–642, 1987.
- [46] B. Horn and E. Weldon. Direct methods for recovering motion. *International Journal of Computer Vision*, 2(1):51–76, 1988.

- [47] P. V. Hough and B. W. Powell. A method for faster analysis of bubble chamber photographs. *Il Nuovo Cimento*, 18(6):1184–1191, 1960.
- [48] A. Howard. Real-time stereo visual odometry for autonomous ground vehicles. In *Proceeding of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3946–3952, 2008.
- [49] I. P. Howard and W. B. Templeton. *Human spatial orientation*. John Wiley & Sons Ltd, Wiley, NY, 1966.
- [50] Y. Kim and J. Aggarwal. Determining object motion in a sequence of stereo images. *IEEE International Journal of Robotics and Automation*, 3(6):599–614, 1987.
- [51] B. Kitt, A. Geiger, and H. Lategahn. Visual odometry based on stereo image sequences with ransac-based outlier rejection scheme. *Proceedings of the IEEE Intelligent Vehicles Symposium*, pages 486–492, 2010.
- [52] V. Klema and A. Laub. The singular value decomposition: Its computation and some applications. *IEEE Transactions on Automatic Control*, 25(2):164–176, 1980.
- [53] K. Konolige, M. Agrawal, and J. Sola. Large-scale visual odometry for rough terrain. In *Proceedings of the International Symposium on Robotics Research*, pages 201–212. 2007.
- [54] A. Larusso, D. Eggert, and R. Fisher. A Comparison of Four Algorithms for Estimating 3-D Rigid Transformations. In *Proceedings of the British Machine Vision Conference*, pages 24.1–24.10, 1995.
- [55] S. Lee and Y. Kay. A Kalman Filter Approach for Accurate 3-D Motion Estimation from a Sequence of Stereo Images. *CVGIP: Image Understanding*, 54(2):244–258, 1991.
- [56] J. J. Leonard and H. F. Durrant-Whyte. Mobile robot localization by tracking geometric beacons. *IEEE Transactions on Robotics and Automation*, 7(3):376–382, 1991.
- [57] H. Li and R. Hartley. The 3D-3D Registration Problem Revisited. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–8, 2007.

- [58] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.
- [59] M. Maimone, Y. Cheng, and L. Matthies. Two years of visual odometry on the Mars exploration rovers. *Journal of Field Robotics*, 24:169–186, 2007.
- [60] S. Malassiotis and M. Srinivas. Model-based joint motion and structure estimation from stereo images. *Computer Vision and Image Understanding*, 65(1):79–94, 1997.
- [61] A. Mallet, S. Lacroix, and L. Gallo. Position estimation in outdoor environments using pixel tracking and stereovision. In *Proceedings of the International Conference on Robotics and Automation*, pages 3519–3524, 2000.
- [62] R. Mandelbaum, G. Salgian, and H. Sawhney. Correlation-based estimation of ego-motion and structure from motion and stereo. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 544–550, 1999.
- [63] L. Matthies and S. Shafer. Error Modeling in Stereo Navigation. *IEEE International Journal of Robotics and Automation*, 3(3):239–248, 1987.
- [64] A. Milella and R. Siegwart. Stereo-Based Ego-Motion Estimation Using Pixel Tracking and Iterative Closest Point. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 21–21, 2006.
- [65] M. Mozerov, V. Kober, and T. Choi. Improved motion stereo matching based on a modified dynamic programming. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 501–505, 2000.
- [66] K. Nakayama. Biological image motion processing: a review. *Vision Research*, 25(5):625–60, 1985.
- [67] N. Navab, R. Deriche, O. D. Faugeras, and I. S. Antipolis. Recovering 3D motion and structure from stereo and 2D token tracking cooperation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 513–516, 2004.
- [68] D. Nistér, O. Naroditsky, and J. Bergen. Visual odometry for ground vehicle applications. *Journal of Field Robotics*, 23(1):3–20, 2006.
- [69] J. Oliensis. A Critique of Structure-from-Motion Algorithms. *Computer Vision and Image Understanding*, 80(2):172–214, 2000.

- [70] C. Olson, L. Matthies, H. Schoppers, and M. Maimone. Robust stereo ego-motion for long distance navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 453–458, 2000.
- [71] P. Pearce and S. Pearce. *Polyhedra Primer*. Van Nostrand Reinhold, New York, 1978.
- [72] J. A. Perrone and L. S. Stone. A model of self-motion estimation within primate extrastriate visual cortex. *Vision Research*, 34(21):2917–38, 1994.
- [73] L. Quam. Hierarchical warp stereo. In *Proceedings of the DARPA Image Understanding Workshop*, pages 149–155, 1984.
- [74] F. Raudies and H. Neumann. A review and evaluation of methods estimating ego-motion. *Computer Vision and Image Understanding*, 116(5):606–633, 2012.
- [75] J. H. Rieger and D. T. Lawton. Processing differential image motion. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, 2(2):354, 1985.
- [76] J. Shi and C. Tomasi. Good features to track. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994.
- [77] D. Sinclair, A. Blake, and D. Murray. Robust estimation of egomotion from normal flow. *International Journal of Computer Vision*, 13(1):57–69, 1994.
- [78] M. Sizintsev. *On 3D Spacetime Oriented Energy Representation For Spatiotemporal Stereo and Motion Recovery*. PhD thesis, York University, 2013.
- [79] M. Sizintsev and R. P. Wildes. Spatiotemporal oriented energies for spacetime stereo. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1140–1147, 2011.
- [80] M. Sizintsev and R. P. Wildes. Spatiotemporal stereo and scene flow via stequel matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(6):1206–19, 2012.
- [81] M. E. Spetsakis and J. Y. Aloimonos. Structure from motion using line correspondences. *International Journal of Computer Vision*, 4(3):171–183, 1990.
- [82] J. J. Spiker. *The Global Positioning System: Theory and Application*. AIAA, Reston, VA, 1996.



- [83] G. Stein and A. Shashua. Model-based brightness constraints: on direct estimation of structure and motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9):992–1015, 2000.
- [84] G. Strang. *Introduction to Applied Mathematics*. Wellesley-Cambridge Press, Wellesley, MA, 1986.
- [85] N. Sünderhauf and P. Protzel. Stereo Odometry - A Review of Approaches. Technical report, Chemnitz University of Technology, 2007.
- [86] J.-P. Tardif, Y. Pavlidis, and K. Daniilidis. Monocular visual odometry in urban environments using an omnidirectional camera. In *Proceeding of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2531–2538, 2008.
- [87] S. C. Thomopoulos. Sensor integration and data fusion. In *Proceedings of the Intelligent Robotics Systems Conference*, pages 178–191, 1990.
- [88] B. Triggs, P. F. Mclauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle Adjustment - A Modern Synthesis. In *Proceedings of Vision Algorithms: Theory and Practice*, pages 298–372, 2000.
- [89] I. F. Vis. Survey of research in the design and control of automated guided vehicle systems. *European Journal of Operational Research*, 170(3):677–709, 2006.
- [90] J. Weng, P. Cohen, and N. Rebibo. Motion and structure estimation from stereo image sequences. *IEEE Transactions on Robotics and Automation*, 8(3):362–382, 1992.
- [91] R. P. Wildes and J. Bergen. Qualitative spatiotemporal analysis using an oriented energy representation. In *Proceedings of the European Conference on Computer Vision*, pages 768–784, 2000.
- [92] G. Young and R. Chellapa. 3-d motion estimation using a sequence of noisy stereo images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):710–716, 1988.
- [93] T. Zhang and C. Tomasi. Fast, robust, and consistent camera motion estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 164–170, 1999.

- [94] Z. Zhang and O. Faugeras. Estimation of displacements from two frames obtained from stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(12):1141–1156, 1991.