

# Place Recognition for Mobile Robot in Changing Environments

Juan Cao

Department of Computer Science

Aberystwyth University

A thesis submitted for the degree of

*Doctor of Philosophy*

# Abstract

This thesis is concerned with the problem of place recognition for a mobile robot using an omnidirectional camera as its sole sensor modality. The problems we are faced with range from orientation estimation to loop closure detection, in the absence of any prior knowledge of position.

In order to resolve the challenging issues encountered by any appearance-based place recognition system - specifically, perceptual aliasing and variability - we first develop a quadtree-based image comparison method. In contrast to most existing methods, this method does not involve the computationally expensive step of feature or keypoint detection and description, which utilises the spatial structure property of an image to provide robustness against dynamic changes in scenes. Our algorithm is experimentally evaluated on one public dataset, and two datasets collected by ourselves in different environments, thereby demonstrating its effectiveness in handling perceptual aliasing and environment variability.

For many tasks in mobile robotics, it is crucial accurately to determine the orientation of the robot, relying on a single vision sensor. For this purpose, we propose an evaluation methodology that focuses on the ability of different image-based algorithms to establish the heading of the robot when capturing two images. Critical analysis of performance is also provided.

In addition, a quadtree-based loop closure detection method is proposed, with the intention of increasing the number of correctly-recognized revisited locations (high recall) at low false positives (high precision). The loop closure detection is performed by pairwise image compari-

son. The performance of the proposed method is evaluated using our collected dataset, which contains highly aliased images and drastic perceptual changes. The experimental results show that our method can achieve a high recall at 100% precision, and outperform other related algorithms in term of closeness to ground truth.

## Acknowledgements

First of all, I would like to express my gratitude to my supervisors Dr. Frédéric Labrosse and Dr. Hannah Dee, for their patience, motivation, enthusiasm, and guidance. They guided me to conduct research with methodological rigor. Without their continuous support, encouragement and constructive criticism, it may have been impossible for me to complete this thesis.

My sincere thanks to the Department of Computer Science of Aberystwyth University, and to Chongqing Jiaotong University in China for the scholarship which enabled me to pursue my doctoral studies.

I am grateful to the administrative staff in the Department of Computer Science for their support and hospitality. I would also like to thank colleagues from the Chongqing Jiaotong University for their kind assistance and friendship.

I received valuable assistance from my fellow students, and the staff and faculty of the Intelligent Robotics Group and the Vision, Graphics and Visualisation Group. In particular, I would like to acknowledge my indebtedness to Peter Matthew Scully, who was always open to discussion and questions, and has been an excellent officemate and friend: Marek Ososinski, who was a great help when I was just beginning to play with robots; to Colin Sauze for his support in Linux; and to Michael Clarke for his assistance in C++.

I would like to thank all of my Chinese fellow students in Aberystwyth, who made my stay all the more enjoyable.

Last, but not least, I would like to express my sincere thanks to my family: my husband Xie Tianbao, my parents Cao Huairan and Tang

Jiankun, and my brother Cao Kai, for their love, understanding, and sacrifices.

# Contents

<b>Contents</b>	<b>v</b>
<b>List of Figures</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Background and related work</b>	<b>8</b>
2.1 Visual SLAM . . . . .	9
2.2 Image features and visual vocabulary . . . . .	16
2.2.1 Image descriptors . . . . .	18
2.2.2 Visual vocabulary . . . . .	23
2.2.3 Dimensionality reduction techniques . . . . .	25
2.2.3.1 Fourier transform . . . . .	25
2.2.3.2 Principle Component Analysis (PCA) . . . . .	26
2.2.3.3 Other approaches . . . . .	27
2.3 Place recognition . . . . .	28
2.3.1 Solutions to the place-recognition problem . . . . .	30
2.3.1.1 Histograms . . . . .	30
2.3.1.2 Object-based methods . . . . .	33
2.3.1.3 Region-based methods . . . . .	38
2.3.1.4 Context-based methods . . . . .	39
2.3.2 Strategies for dealing with challenging issues . . . . .	41
2.3.2.1 Dealing with changes . . . . .	42
2.3.2.2 Disambiguating ambiguous cases . . . . .	44
2.4 Visual odometry . . . . .	46

2.5	Loop closure . . . . .	48
2.6	Quadtree structure . . . . .	52
2.7	Conclusions . . . . .	53
<b>3</b>	<b>Datasets</b>	<b>60</b>
3.1	Introduction . . . . .	60
3.2	Indoor datasets: ISL . . . . .	61
3.2.1	Acquisition platforms and procedure . . . . .	61
3.2.2	Ground truth . . . . .	62
3.2.3	The environments and examples . . . . .	63
3.3	Indoor datasets: COLD . . . . .	70
3.4	Outdoor datasets: GummyBear . . . . .	71
3.5	Outdoor datasets: New College 1 Dataset . . . . .	77
3.6	Conclusions . . . . .	79
<b>4</b>	<b>A quadtree-based method for image comparison</b>	<b>80</b>
4.1	Introduction . . . . .	80
4.2	Image distance metrics . . . . .	82
4.2.1	Euclidean distance . . . . .	82
4.2.2	Median of absolute differences . . . . .	83
4.2.3	$\chi^2$ distance . . . . .	83
4.2.4	Pearson's correlation coefficient . . . . .	84
4.2.5	Histogram intersection distance . . . . .	84
4.2.6	Earth-mover's distance . . . . .	85
4.2.7	Shannon mutual information . . . . .	86
4.3	An image comparison measure using Quadtree . . . . .	86
4.4	Quadtree and metrics . . . . .	89
4.5	Experiments and results: GummyBear dataset . . . . .	90
4.5.1	Experiment 1: Loop closure . . . . .	92
4.5.2	Experiment 2: Pinch points — nearby, but not the same place . . . . .	94
4.6	Experiments and results: New College 1 Dataset . . . . .	96
4.6.1	Experimental set-up . . . . .	97

4.6.2	Evaluation . . . . .	100
4.6.3	Experiment 1: loop closure . . . . .	100
4.6.4	Experiment 2: loop closure on noisy data . . . . .	104
4.7	Conclusions . . . . .	105
<b>5</b>	<b>An evaluation of image-based estimation techniques for robot orientation</b>	<b>107</b>
5.1	Introduction . . . . .	107
5.2	Computing robot orientation . . . . .	109
5.2.1	A feature-based method: SIFT . . . . .	109
5.2.2	Visual compass . . . . .	109
5.2.3	A quadtree-based method . . . . .	111
5.2.4	Other methods used in related publications . . . . .	111
5.2.4.1	A Discrete Fourier Transform (DFT) descriptor . . . . .	111
5.2.4.2	A Principal Components Analysis descriptor . . . . .	112
5.2.4.3	A Histograms of Oriented Gradients descriptor . . . . .	113
5.2.4.4	A Gist descriptor . . . . .	114
5.3	Experiments . . . . .	114
5.3.1	Outdoor experimental results: GummyBear dataset . . . . .	115
5.3.2	Indoor experimental results: ISL dataset . . . . .	121
5.3.3	Indoor experimental results: COLD dataset . . . . .	130
5.3.4	Indoor experimental results: COLD dataset (based on HS colour space and log transformation) . . . . .	134
5.3.4.1	Experimental results: HS colour space . . . . .	134
5.3.4.2	Experimental results: log transformation . . . . .	137
5.4	Conclusions . . . . .	139
<b>6</b>	<b>A quadtree-based method for loop closure detection</b>	<b>141</b>
6.1	Introduction . . . . .	141
6.2	Methodology . . . . .	142
6.3	Experimental results and discussion . . . . .	145
6.3.1	Evaluation loop closure accuracy . . . . .	145
6.3.2	Evaluation the proposed method . . . . .	146

## CONTENTS

---

6.3.3	Comparison with other methods . . . . .	156
6.4	Conclusions . . . . .	169
<b>7</b>	<b>Conclusions</b>	<b>170</b>
7.1	Summary of thesis . . . . .	170
7.2	Future work . . . . .	175
	<b>Appendix A</b>	<b>178</b>
A.1	Histograms for the ISL dataset . . . . .	178
A.2	PPCC coefficients . . . . .	181
	<b>References</b>	<b>183</b>

# List of Figures

3.1	(a) Catadioptric camera (b) Pioneer robot (c) Experimental environment without obstacles and (d) Experimental environment with “wall” sitting in the middle of workspace. . . . .	62
3.2	VICON-recorded robot trajectories in the $xy$ -plane in four different scenarios. . . . .	63
3.3	ISL dataset 1: 3.3(a) is a typical image from ISL dataset 1; 3.3(b) is a schematic of the whole environment; and 3.3(c) is the trajectory followed by the robot, with some annotated points in $x, y, t$ space. . . . .	64
3.4	ISL dataset 2: 3.4(a) is a typical image from dataset 2; 3.4(b) is a schematic of the whole environment; and 3.4(b) is the trajectory followed by the robot, with some annotated points in $x, y, t$ space. . . . .	66
3.5	ISL dataset 3: 3.5(a) is a typical image from dataset 3; 3.5(b) is a schematic of the whole environment; and 3.5(c) is the trajectory followed by the robot, with some annotated points in $x, y, t$ space. . . . .	67
3.6	ISL dataset 4: 3.6(a) is a typical image from dataset 4; 3.6(b) is a schematic of the whole environment; and 3.6(c) is the trajectory followed by the robot, with some annotated points in $x, y, t$ space. . . . .	69
3.7	Example images of the COLD datasets in three different lighting conditions: (a) night; (b) cloudy; and (c) sunny. The omnidirectional images are shown in the first and third rows, the corresponding unwrapped images are shown in the second and fourth rows, respectively. . . . .	71

## LIST OF FIGURES

---

3.8	Example images from the (a) FIELD, (b) CARPARK, and (c) TENERIFE datasets, respectively. The omnidirectional images are shown in the first and third rows, the corresponding unwrapped images are shown in the second and fourth rows, respectively. . . . .	72
3.9	RTK-GPS track from the “Gummy Bear” path for the FIELD dataset, with some unwrapped omnidirectional image samples. Image numbers of key positions are marked in blue. . . . .	74
3.10	The Idris robot. . . . .	75
3.11	GPS trajectory of the New College 1 Dataset. The parts of the dataset used in our experiment are indicated by the red dots. . . .	78
3.12	Example images from the New College 1 Dataset. . . . .	78
4.1	Quadtree decomposition. . . . .	88
4.2	Comparison of different metrics applied to our quadtree similarity measurement: iso-similarity point set at 50 for the left image and 90 for the right, which occurs at approximately 1,400mm on the $x$ -axis. . . . .	90
4.3	Physical distance between any pair of the images from the FIELD dataset, in meters. . . . .	91
4.4	Physical distance between any pair of the images from the CARPARK dataset, in meters. . . . .	91
4.5	Similarity between Image 0 and all images of the FIELD dataset, demonstrating the possibility of robust loop closure. . . . .	93
4.6	Left: an image pair (Image 0 and Image 1481 from the FIELD dataset). Right: visualisation of left image pair comparison using our proposed method. The similarity is 96.59%, with a threshold of 45. . . . .	93
4.7	Similarity between Image 0 and all images of the CARPARK dataset, demonstrating the possibility of robust loop closure. . . . .	93
4.8	Left: an image pair (Image 0 and Image 2119 from the CARPARK dataset). Right: visualisation of left image pair comparison using our proposed method: the similarity is 96.99%, with a threshold of 70. . . . .	94

## LIST OF FIGURES

---

4.9	Similarity between Image 143 and all images of the FIELD dataset.	94
4.10	Left: an image pair (Image 143 and Image 1162 from the FIELD dataset). Right: visualisation of left image pair comparison using our proposed method: the similarity is 82.24%, with a threshold of 45. . . . .	95
4.11	Similarity between Image 192 and all images of the CARPARK dataset. . . . .	96
4.12	Left: an image pair (Image 192 and Image 1670 from the CARPARK dataset). Right: visualisation of left image pair comparison using our proposed method: the similarity is 85%, with a threshold of 70.	96
4.13	Ground truth for the sequence between Images 150 and 450, generated at different values of $n = 5, 20$ , and $50$ . . . . .	97
4.14	Ground truth of the New College 1 Dataset. . . . .	99
4.15	(a) Similarity matrix computed using the Quadtree method; (b) Confusion matrix computed using FABMap (best viewed in magnification); (c) Distance matrix computed using BRIEF-Gist; and (d) Distance matrix computed using ABLE-P. . . . .	102
4.16	Precision-recall curves between Images 120 and 1200. . . . .	103
4.17	Precision-recall curves between Images 120 and 1900. . . . .	103
4.18	(a) Original image; (b) Original image corrupted by Gaussian noise (0, 0.01); (c) Original image corrupted by Gaussian noise (0, 0.02); and (d) Original image corrupted by Gaussian noise (0, 0.03). . .	105
4.19	(a) Average precision; (b) Best recall at 100% precision of three methods on original images and noisy images with different variance (0.01, 0.02 and 0.03) and zero mean. . . . .	106
5.1	SIFT matching example . . . . .	110
5.2	Visual compass example. The top row shows a reference image, and the bottom row shows the current image, where the dashed box indicates the column shift $\alpha$ between the two images, corresponding to the relative rotation between them. . . . .	110

## LIST OF FIGURES

---

5.3	Simplified illustration of alternative techniques for computing orientation using: (a) a fixed reference image; and (b) a moving reference image. . . . .	115
5.4	Experimental results for dataset CARPARK, with a fixed reference image. . . . .	116
5.5	Experimental results for dataset FIELD, with a fixed reference image. . . . .	117
5.6	Experimental results for dataset TENERIFE, with a fixed reference image. . . . .	118
5.7	Top: an example image pair (Image 1399: upper, Image 1436: lower) from the CARPARK dataset. Bottom: SIFT matching result. . . . .	119
5.8	Top: an example image pair (Image 0: upper, Image 570: lower) from the CARPARK dataset. Bottom: SIFT matching result. . . . .	119
5.9	Experimental results for dataset CARPARK, with a moving reference image. . . . .	120
5.10	Experimental results for dataset FIELD, with a moving reference image. . . . .	121
5.11	Experimental results for dataset TENERIFE, with a moving reference image. . . . .	122
5.12	Experimental results for dataset ISL 1, with a moving reference image. . . . .	126
5.13	Experimental results for dataset ISL 2, with a moving reference image. . . . .	127
5.14	Experimental results for dataset ISL 3, with a moving reference image. . . . .	128
5.15	Experimental results for dataset ISL 4, with a moving reference image. . . . .	129
5.16	Example images from Sunny (top row), Cloudy (middle row), and Night (bottom row) dataset of COLD datasets, (a) original images, and (b) corresponding log-transformed images. . . . .	138

6.1	Recall and precision curves, depending on the parameter of correct detection criteria, for ISL 1. The first distance metric in the caption is used for quadtree decomposition: the second, for calculating the distance between similar areas of two images applies to the following figures. . . . .	147
6.2	Recall and precision curves, depending on the parameter of correct detection criteria, for ISL 2. . . . .	148
6.3	Recall and precision curves, depending on the parameter of correct detection criteria, for ISL 3. . . . .	149
6.4	Recall and precision curves, depending on the parameter of correct detection criteria, for ISL 4. . . . .	150
6.5	Examples of loop closure detection based on the CE scheme. Images 0, 268, 523 and 779 from the ISL 3 dataset were captured at nearly the same location, but with slight offset or rotation of the camera viewpoint (see Figure 3.5(c)). . . . .	152
6.6	Examples of loop closure detection based on the CE scheme. Images 0, 266, 538 and 745 from the ISL 4 dataset were captured at nearly the same location, but with slight offset or rotation of camera viewpoint (see Figure 3.6(c)). . . . .	153
6.7	Precision and recall curves for the ISL 4 dataset. . . . .	157
6.8	Distance (in appearance space) between Image 0 and all images of the ISL 1 dataset: (a) Euclidean distance of our method; (b) Hamming distance of BRIEF-Gist; (c) Hamming distance of LDB-Gist; (d) Euclidean distance of WI-SIFT; and (e) Euclidean distance of WI-SURF. . . . .	158
6.9	Distance between Image 0 and all images of the ISL 2 dataset: (a) Euclidean distance of our method; (b) Hamming distance of BRIEF-Gist; (c) Hamming distance of LDB-Gist; (d) Euclidean distance of WI-SIFT; and (e) Euclidean distance of WI-SURF. . .	159
6.10	Distance between Image 0 and all images of the ISL 3 dataset: (a) Euclidean distance of our method; (b) Hamming distance of BRIEF-Gist; (c) Hamming distance of LDB-Gist; (d) Euclidean distance of WI-SIFT; and (e) Euclidean distance of WI-SURF. . .	160

## LIST OF FIGURES

---

6.11	Distance between Image 0 and all images of the ISL 4 dataset: (a) Euclidean distance of our method; (b) Hamming distance of BRIEF-Gist; (c) Hamming distance of LDB-Gist; (d) Euclidean distance of WI-SIFT; and (e) Euclidean distance of WI-SURF. . .	161
6.12	Precision and recall curves for the ISL 1 dataset. . . . .	165
6.13	Precision and recall curves for the ISL 2 dataset. . . . .	166
6.14	Precision and recall curves for the ISL 3 dataset. . . . .	167
6.15	Precision and recall curves for the ISL 4 dataset. . . . .	168
A.1	Histograms of the comparison scores between the image of interest (Image 0) and all the images of the ISL 1 dataset, based on (a) CC; (b) EC; (c) EE; and (d) CE methods. . . . .	178
A.2	Histograms of the comparison scores between the image of interest (Image 0) and all the images of the ISL 2 dataset, based on (a) CC; (b) EC; (c) EE; and (d) CE methods. . . . .	179
A.3	Histograms of the comparison scores between the image of interest (Image 0) and all the images of the ISL 3 dataset, based on (a) CC; (b) EC; (c) EE; and (d) CE methods. . . . .	180
A.4	Histograms of the comparison scores between the image of interest (Image 0) and all the images of the ISL 4 dataset, based on (a) CC; (b) EC; (c) EE; and (d) CE methods. . . . .	181

# Chapter 1

## Introduction

This thesis is concerned with robust long-term place recognition for autonomous mobile robots in changing environments. Specifically, this work mainly addresses two research problems: first, what image-based techniques offer good and reliable orientation estimation for robots equipped only with vision sensors; secondly, how a robot may accurately recognize a previously visited place, without any prior knowledge. In this chapter, we give a brief overview of the background and relevant research works, and the motivations behind the current work. Subsequently, the main aims and objectives of our research are described. Finally, we provide a summary of contributions and an overview of the thesis.

Autonomous robotics is a growing and increasingly popular area in both industry and academic research. Robots of different kinds and capacities from personal service robots at home to scientific planetary exploration rovers perform a variety of tasks in an intelligent and autonomous manner, potentially bringing great benefits to mankind.

A classical problem in creating an autonomous robot focuses on the ability of a robot to localize itself within a given environment, while at the same time mapping that same environment. This problem is known as Simultaneous Localization and Mapping (SLAM). It has been widely studied by the robotics communities for several decades (Chatila and Laumond [1985]; Davison [2003]; Durrant-Whyte et al. [1996]; Guivant and Nebot [2001]; Smith et al. [1987]; Williams et al. [2000]),

---

and is by now considered a relatively mature problem. However, there are still some challenges that need yet to be overcome.

Different sensor modalities have been used to provide the necessary input for SLAM solutions. Typically the Global Positioning System (GPS) has been used for localisation and navigation assistance (Thrun et al. [2006]). However, solutions based on GPS do not work well in indoor or cluttered outdoor environments, where GPS is generally less accurate or not available. To avoid the need for GPS, or any other infrastructure, a number of frameworks (Chong and Kleeman [1999]; Crowley [1989]; Guivant et al. [2000]; Rencken [1993]; Ribas et al. [2008]) have been developed that make use of active sensors (e.g., sonar and laser scanner) to acquire data. However, these active sensors are normally very heavy, expensive, and energy-hungry, and thus not suitable for some systems that must meet payload, cost, and power constraints. Examples of such systems include unmanned aerial vehicles (UAVs), autonomous underwater vehicles (AUV), Mars Exploration Rovers.

The dead-reckoning (DR) technique has long been used to provide position and orientation information. Sensors for DR include encoders, the magnetic compass, and the inertial measurement unit (IMU), among others. However, existing systems equipped with these sensors universally suffer from precision and reliability problems. Slippage of the wheels on non-smooth surfaces can cause accumulated error in position and orientation; a magnetic compass may be subject to interference from magnetic sources, such as metallic objects; and the readings from the IMU become increasingly unreliable as errors accumulate and compound over time.

By contrast, the camera as a passive sensor is an attractive alternative with many advantages, including low cost and light weight. Moreover, the camera provides a rich source of information about the environment, which enables the use of sophisticated computer vision algorithms (detection and recognition algorithms). In addition, the computational requirements of these computer vision algorithms are not a significant issue thanks to recent improvements in hardware (e.g., available parallel graphics processors and multiple CPU threads).

---

When cameras are used as the primary sensor input, solutions to such a SLAM problem are referred to as visual SLAM (vSLAM). Since 2005, intense research has been undertaken to develop a reliable, accurate, and large-scale vSLAM technique. Many techniques (Cummins and Newman [2010]; Labrosse [2007]; Maddern et al. [2014]; Mei et al. [2009]; Strasdat et al. [2010a]) that rely only on monocular or stereo visual cues have shown remarkable performance in the vSLAM problem.

Nevertheless, there is some way to go before a robust vSLAM solution can be widely employed in practice. For instance, most state-of-the-art systems require high quality camera images as input data, and assume that the world in which the robot works remains almost static in appearance (Durrant-Whyte and Bailey [2006]; Maddern et al. [2012]). These assumptions are not valid in the vast majority of real-world tasks. For real and long life operation, a robot must be able to respond to unknown or changing environments. Moreover, it is always preferable that a robot has low hardware costs.

Appearance-based place recognition is usually performed by finding matches between the current view of the robot and a set of images of previously visited locations. However, appearances are often deceiving. There are two basic factors that make the task of place recognition difficult. Firstly, in dynamic environments, the appearance of a place may change as objects move, viewpoint changes, or illumination conditions change (perceptual variability). Secondly, a number of perceptions from different parts of an environment may look similar (perceptual aliasing). Therefore, a good image comparison measure is of utmost importance to reliable completion of a place recognition task.

Some studies (Bellotto et al. [2008]; Cheng et al. [2006]; Labrosse [2007]; Magnabosco and Breckon [2013]; Williams and Reid [2010]) exploit the odometry information obtained by analyzing images taken in consecutive frames to improve the motion estimation accuracy of the robot, and thereby boost the performance of vSLAM systems. The pose (position and orientation) estimation technique, based on a sequence of acquired images, is called in robotics visual odometry (VO), or visual compass (VC), when only the orientation is desired. Clipp et al. [2010] introduced a vSLAM system that utilizes the parallelism strategy to per-

---

form visual odometry and loop closure in a relatively small scale environment. Many studies have illustrated that VO or VC allows for enhanced localization and navigation accuracy in robots, since long-term drift can be mitigated. However, these algorithms suffer from some practical limitations, which often have their roots in the explicit assumptions that there is sufficient illumination and a sufficiently large set of features to be extracted from a static, or at least partially static environment. A further assumption is that there must be enough scene overlap between consecutive frames (Scaramuzza and Fraundorfer [2011]). Laurent Kneip and Siegwart [2011] enriched a textureless scene with some sparse natural features, in order for their VO system to work properly.

Loop closure detection is one of the key challenges in a SLAM system: that is, when, or if the robot has returned to a previously visited place after a long traverse movement. This information is critical for mobile robots to maintain a global consistent map of unknown environments, and allows them to correct the accumulated errors caused by inaccurate sensor measurements. It is difficult to detect loop closure precisely using metric information, because of accumulated errors in position estimation, which tend to scale up dramatically with the dimensions of the environment. Loop closure detection using visual cues has attracted a great deal of attention in recent years. A viable solution to the loop closure problem using vision requires determining for any two images whether they have been taken from the same place.

Several successful approaches have been proposed that rely either on global appearance solutions ( Arroyo et al. [2014]; Badino et al. [2012]; Goedemé et al. [2007]; Sunderhauf and Protzel [2011]; Wu et al. [2014]), or local feature extraction ( Anati and Daniilidis [2009]; Cummins and Newman [2010]; Garcia-Fidalgo and Ortiz [2013]). Most of these frameworks are based on a visual Bag-of-Words (BoWs) strategy; data structures such as the vocabulary tree, hierarchical k-means and kd-tree are also used to speed up matching in order to cope with large scale environments. However, the BoWs method is affected by perceptual aliasing due to vector quantization, and it involves the learning of the BoWs dictionaries, whether online or offline. To avoid mismatches (false positives), some algorithms ( Angeli et al. [2008b]; Scaramuzza et al. [2010]) incorporate epipolar constraint

---

to check spatial consistency and verify candidate matchings.

When a robot is operating over a large area and within a changing environment, visual loop closure detection will become extremely challenging. For example, different places may appear the same, which may lead to erroneous loop closing and thus yield an incorrect mapping. Moreover, perceptual changes such as view-point and illumination changes, and moving objects are common in the natural environment. A comparison technique that is not robust against these changes will lead to incorrect loop closures. Even one erroneous loop closure incorporated into the map can cause catastrophic failures of estimation algorithms.

Within the context of vSLAM, the considerations about image representation and matching in the appearance-based place recognition task, and the increasing demand for high precision VO or VC systems which can extend the applicability of real time vSLAM motivated the work in this thesis.

## Research Aims and Objectives

To build a fully autonomous mobile robot that is capable of operating long-term in real environments, we must develop place recognition strategies that can handle unknown or changing environments. Our research aims to improve the capabilities of vSLAM in dynamic environments using an on-board omnidirectional camera alone. We propose to develop an image comparison method that does not rely on any artificial landmarks or natural structures within the environments, that will be robust to the changes encountered by the robot, and that can be utilized in indoor or outdoor environments. With a view to this aim, we plan to investigate how to select image-based techniques that are suitable for accurate and reliable orientation estimation. Our evaluation focuses on the ability of the techniques to estimate the relative orientation of the robot at the time when the particular images were captured. In addition, we propose to develop a novel loop closure detection technique that will enable robots to recognise reliably places that they are revisiting by matching their current view with previously stored images, without any prior position knowledge.

---

## Contributions

A summary of the contributions of this thesis is as follows:

- An extensive literature review of the most important developments in the field of vSLAM is presented. The key characteristics of some vSLAM frameworks are described and a summary table is provided, which enables quick reference to the key techniques in these approaches. In addition, a further literature review of relevant background materials and related works is provided. In particular, the performance of place recognition systems in handling challenging cases characterised by perceptual aliasing and perceptual variability has been extensively investigated.
- A novel image comparison algorithm has been proposed. We made use of the whole image as a global visual feature. In order to compensate for the weaknesses of the global feature, we combined the quadtree decomposition concept with the natural rotational invariance of the omnidirectional images. This work has been published in (Cao et al. [2012]).
- An evaluation methodology for different image-based techniques with respect to orientation estimation is introduced. Critical analysis of the performance in indoor and outdoor, static and dynamic environments are provided. This work has been presented in (Cao et al. [2013]).
- A novel appearance-based loop closure detection algorithm that focuses on tackling challenging cases (perceptual aliasing and perceptual variability) has been formulated. This method is distinct from most existing approaches, which involve the computationally expensive step of feature extraction and/or candidate verification within a probabilistic framework. Loop closure detection is achieved by matching places based on the visual distance scores between a given of pair of places, which ignores the appearance changes caused by a dynamic environment.

---

## Overview

This thesis is presented in seven chapters. Excluding the Introduction (this chapter), the thesis is divided into four main sections: background and related work (Chapters 2); datasets description (Chapter 3); major contributions (Chapters 4, 5 and 6); and conclusions and directions for future research (Chapter 7).

Chapters 2 forms the first section, which provides an overview of the most important developments in the field of vSLAM, focusing on image representation, dimensionality reduction techniques, place recognition, visual odometry, quadtree structure and loop closure detection techniques. A literature review summary table is provided at the end of this chapter. Each chapter in the dealing with the contributions also reviews more specifically related work.

Four datasets are used to evaluate the methods proposed in Chapters 4 and 6, as well as the three methods for robot orientation estimation in Chapter 5. To avoid repetition in each chapter, a detailed description of the four datasets is given in Chapter 3, which constitutes the second part of the thesis.

The third part of the thesis develops the ideas and contributions of this thesis. The main contribution is Chapter 4, which proposes a novel image comparison method to increase the robustness of image matching for visual place recognition tasks. The evaluation of this approach, and a comparison with the state-of-the-art algorithms are provided in this chapter. In Chapter 5, the performance of three methods for robot orientation estimation is evaluated, and quantitative results are provided. A novel development of the algorithmic method developed in Chapter 4 for loop closure detection is described in Chapter 6. Experimental validation and a comparison with the state-of-the-art algorithms are provided at the end of this chapter. Our conclusions and suggestions for future research are presented in Chapter 7.

## Chapter 2

# Background and related work

This chapter reviews the main solutions to the visual SLAM problem, mainly focusing on methods for place recognition, which is one of the fundamental tasks in visual SLAM and is typically used for localisation and loop closure. We start with a short overview of current state-of-the-art visual SLAM algorithms in Section 2.1. In order to perform SLAM tasks using visual clues, it is necessary to describe the acquired images and to be able to compare their descriptions. For this reason, a subsection (Section 2.2) is dedicated to surveying image detectors, descriptors, approaches based on Bag-Of-Words (BoW) schemes, and some dimensionality reduction techniques for image descriptors that are popular in the context of visual SLAM. We then illustrate state-of-the-art solutions to the place recognition task in Section 2.3. An overview of visual odometry, which can be used in unmanned navigation applications to recover the camera trajectory for accurate localisation, follows in Section 2.4. In Section 2.5, we review some current loop closure detection techniques that are primarily used for appearance-based SLAM systems in large-scale unknown environments. In Section 2.6, a review of methods based on quadtree structure is provided. This data structure is the core technique of our proposed algorithms. In Section 2.7 we conclude this chapter by summarising the key characteristics of some reviewed vSLAM frameworks.

---

## 2.1 Visual SLAM

In this section, we will discuss recent advances in visual SLAM. A broader survey of SLAM approaches can be found in, for example, (Bonin-Font et al. [2008]) and in (Fuentes-Pacheco et al. [2012]). There is a large body of literature address SLAM for larger environments using either monocular (Angeli et al. [2008a]; Botterill et al. [2011]; Cummins and Newman [2008a]; Davison [2003]) or stereo cameras (Kaess and Dellaert [2010]; Konolige et al. [2010]; Mei et al. [2009]; Nistr et al. [2004]).

Building a representation of the environment is an important task for a mobile robot, allowing the robot to guide itself autonomously around the surrounding space. In consequence, this problem has received significant attention in the past two decades. Next we will look at the state of existing research for map representations exploited in SLAM systems.

Classically, existing map representation studies are classified in two categories depending on whether they make use of either metric or topological maps. Approaches in the metric paradigm, such as those described in (Davison [2003]; Elfes [1989]; Grisetti et al. [2007]; Ho and Newman [2007]; Kaess and Dellaert [2010]; Montemerlo et al. [2002]; Moravec [1988]; Nistr et al. [2004]; Pinies and Tardos [2008]; Scaramuzza and Siegwart [2008]), represent environments by evenly-spaced grids for laser-scanner or sonar based SLAM.

Occupancy-grid maps were first suggested by Elfes [1989] in 1987. Each cell of the grid stores the probability that it is occupied by an obstacle. These approaches typically work well in bounded environments: however, they suffer from discretisation errors that limit the scale at which the environment can be modelled, and have high memory requirements.

Approaches in the topological paradigm, such those described in (Beeson et al. [2005]; Booij et al. [2007]; Chapoulie et al. [2011]; Choset and Nagatani [2001]; Goedemé et al. [2008]; Korrapati and Mezouar [2014]; Kuipers and Byun [1991]; Lin et al. [2013]; Neal and Labrosse [2004]; Ranganathan et al. [2006]; Remolina and Kuipers [2002]; Siagian and Itti [2009]; Sogo et al. [2001]; Wang and Yagi

---

[2012, 2013]; Weiss et al. [2007b]), represent robot environments by graphs. Nodes in such graphs correspond to distinct places or landmarks, and arcs denote connections between places. Topological maps were first introduced in 1985 as an attractive alternative to the occupancy-grid map by Chatila and Laumond [1985]. Since topological approaches usually do not require the exact determination of the geometric position, only the notions of proximity and order, this method allows robotic systems to recover better from drift and slippage phenomena. The map resolution is determined by the complexity of the environment, and less storage is required to store the nodes, compared to the large number of grid cells in occupancy grid maps. Consequently, they permit fast planning, and facilitate interfacing to symbolic planners and problem-solvers (Chatila and Laumond [1985]). However, this advantage comes with the trade-off of reduced accuracy, because of the absent metric information. The limited accuracy of topological maps thus restricts the capability of the robot for fast and safe navigation.

Recently, hybrid models that combine metric and topological information have been proposed as a promising solution to manage large-scale environments. Among others, these maps are of special interest for efficiently managing large-scale environments, and for accurate localisation. To achieve this aim, local geometric information is stored in the nodes of a graph-based global map. There are a number of SLAM algorithms that aim to create such a hierarchical map: examples include (Blanco et al. [2008]; Bosse et al. [2004]; Estrada et al. [2005]; Konolige et al. [2011]; Kouzoubov and Austin [2004]; Kuipers et al. [2004]; Siagian et al. [2014]; Tomatis et al. [2003]).

With the development of human-robot interaction, robots are gradually moving into our homes, offices, museums and other public spaces. Some traditional navigation methods depending on metric maps or topological maps will become invalid for complex, dynamic and unstructured environments. In order to perform human-like tasks alongside humans, a robot needs to have some semantic information about the entities in the environment.

Adding semantic information to environment maps is a very attractive method for improving domestic robot navigation. It is assumed that the robot is given certain knowledge about the building. Such knowledge allows the robot to recognise

---

particular areas of the building (kitchen, living room, etc.) on the current map. More recently, some authors (Astua et al. [2014]; Beeson et al. [2010]; Klasing et al. [2008]; Vasudevan and Siegwart [2008]) have reported systems in which a robot can acquire and use semantic information for navigation tasks.

In (Kosecka and Li [2004]; Lamon et al. [2003]; Vale and Ribeiro [2003]; Zivkovic et al. [2005]), a set of images that represents the environment of a robot is clustered, based on the presence of a number of automatically extracted landmarks. The method used in (Vale and Ribeiro [2003]) is only suited for image comparison techniques which are a metric function, and does not give correct results if self-similarities are present in the environment. Zivkovic et al. [2005] described an algorithm for creating a hierarchical map using graph cuts, and geometric constraints were applied to overcome self-similarities.

In (Choset and Nagatani [2001]), a generalised Voronoi diagram was constructed from laser range data to encode the topology of the environment. These early topological mapping algorithms were not probabilistic. Nowadays, various probabilistic approaches have become popular. They all rely on probabilistic inference for turning sensor measurements into maps. The popularity of probabilistic techniques arises from the fact that all the sensors for environment perception are subject to errors (i.e., measurement noise). In addition, the mapping is characterised by uncertainty. Ranganathan et al. [2006], for instance, used Bayesian inference to obtain the topological structure that best explains a set of panoramic observations, chosen out of the space of all possible topologies. A Markov Chain Monte Carlo (MCMC) algorithm was used to estimate the posterior distribution. Shatkay and Kaelbling [1997] fit Hidden Markov Models (HMMs) to the incoming sensor data, to solve the aliasing problem for topological mapping. The states of these HMMs refer to the topological nodes, between which probabilistic state transitions are identified. Other examples of HMM based work include (Gutierrez-Osuna and Luo [1996]) and (Cassandra et al. [1996]) where a second order HMM is used to model environments.

Some methods rely on the detection of loop closure to build topological maps. In these studies, probabilistic methods are also introduced to cope with the uncertainty of link hypotheses and avoid links between self-similarities. Kristopher

---

and Wesley [2005] applied Dempster-Shafer probability theory to the loop closure problem. Their robot makes a hypothesis whenever it may have revisited a place, then attempts to verify the hypothesis by continuing to traverse the environment, gathering evidence that supports or refutes the hypothesis. In their topological map, each node represents a corner, and the edges represent a sequence of behaviours to move the robot from one node to another using a wall-following strategy. Their method has the advantage that ignorance can be modelled, and no prior knowledge is needed. However, it can only be applied to sensing-limited robots in simple environments. In Goedemé et al. [2008], an agglomerative clustering algorithm is applied to a set of places, based on the visual distance, which is made proportional to the average angle difference of the matching features. Dempster-Shafer theory is then used to deal with self-similarities for each cluster. Subclusters connected with accepted hypotheses are merged into one place, while each refuted hypothesis results in the construction of a new place. After this decision, a final topological map can be built.

A mobile robot has to solve two essential problems in navigation, namely localisation (knowing where it is) and mapping (building a map of its environment). As has been pointed out by earlier researchers, the problem of localisation and mapping is a chicken and egg problem: to localize the robot based on uncertain landmark estimates, it must update landmark estimates based on noisy sensor measurements taken from the uncertain robot position. Therefore, the two problems are typically treated simultaneously (Simultaneous Localisation And Mapping). SLAM has become one of the most widely researched subfields in mobile robotics since the early 1990s, originally developed by Leonard and Durrant-Whyte [1991], building on the earlier seminal work of Smith et al. [1987]. Nowadays, SLAM can be considered a solved problem at a theoretical and conceptual level. However, SLAM for dynamic, complex and large scale environments, using vision as the only external sensor, is still an active area of research. This is referred to as visual SLAM (vSLAM). Since 2005, vSLAM has received much attention in the computer vision community because of the increasing ubiquity of cameras, and advanced computing technologies. More recently, in addition to robotics applications, vSLAM is starting to be implemented in mobile cameras and used in

---

Augmented Reality (AR), wearable computing and the automotive sector.

Probabilistic solutions to vSLAM have been studied extensively within the robotic community. These involve finding an appropriate representation for both the observation model and the motion model. Practical real-time monocular SLAM was first demonstrated by Davison [2003], using the Extended Kalman Filter (EKF) in an indoor environment. The EKF SLAM algorithm is formed by combining the robot pose and the positions of landmarks into a single state vector, and linearising the observation and motion model at each Kalman filter update. However, the EKF has a  $O(n^2)$  computational complexity per step, where  $n$  is the number of landmarks. This complexity stems from the fact that its full state EKF maintains a full  $n \times n$  covariance matrix for  $n$  landmarks, all of which must be updated even if just a single landmark is observed. Although this system is accurate and robust, it cannot be used in a large-scale environment because of the unacceptable computational overhead.

For this reason, Murphy [1999] introduced Rao-Blackwellised particle filters (RBPFs) as an effective way of solving the SLAM problem. Unlike the Kalman filter and derivatives, particle filters do not assume Gaussian noise, and are not subject to the linear hypotheses of a system. This framework has been extended subsequently by Montemerlo et al. [2002] with a view to approaching the SLAM problem with landmarks, a method termed as FastSLAM. It has the advantage that computational complexity of filter updates can be reduced to  $O(n)$  via the Rao-Blackwellisation of the filter: but the absence of an explicit full covariance matrix can make loop closing more difficult.

Sim et al. [2005] firstly presented a SLAM system based on stereo vision, combining the FastSLAM algorithm and local features of images in large-scale environments. Eade and Drummond [2006] proposed a monocular framework based on FastSLAM, which decomposes the SLAM problem into a robot localisation problem, and a separate collection of landmark estimation problems. This algorithm combines particle filtering for localisation with Kalman filtering for mapping.

An alternative technique for solving the SLAM problem is to apply algorithms used in the computer vision and photogrammetry research community for Struc-

---

ture from Motion (SFM). In general, SFM refers to the problem of recovering 3D information, such as the camera position and orientation, and the position of the landmarks (the map being composed by the set of landmarks), from a series of unordered 2D images: this is generally formulated as a computationally expensive off-line process. SFM-based techniques typically maintain the full trajectory of the camera, and use optimisation to find the best trajectory and landmark locations.

Techniques such as bundle adjustment (BA), which performs batch optimization over selected images from the live input, are generating a great deal of interest in the robotics community. It has been shown by Strasdat et al. [2010b] that optimization-based approaches provide better performance over filter-based approaches for the same computational work in purely vSLAM. BA has been used in many real-time systems as an optimisation technique for visual odometry (Nistr et al. [2004]) - which only recovers the camera trajectory, without explicitly creating a map - as well as for vSLAM (Davison [2003]; Karlsson et al. [2005]; Klein and Murray [2007]; Mouragnon et al. [2006]; Se et al. [2002]; Strasdat et al. [2010a]). All approaches mentioned above are either based on a single camera (whether forward-facing or omnidirectional) (Davison [2003]; Karlsson et al. [2005]), or multiple cameras in a stereo configuration (Nistr et al. [2004]; Se et al. [2002]).

Mei et al. [2009, 2010] presented an RSLAM system, aiming to real-time large scale SLAM based on stereo vision, which combines accurate visual odometry with constant-time large-scale mapping, appearance-based loop closure detection, and pose graph optimisation if required. Another, similar system called FrameSLAM has been developed by Konolige and Agrawal [2008]: this was further improved in (Konolige et al. [2010]) by adding a vocabulary tree to provide candidate loop closures to the RANSAC stage.

In order to allow the use of batch optimization techniques for real-time operation, Klein and Murray [2007] proposed to perform map building and localization separately, processed in parallel threads on a dual-core computer. However, this framework is not well-adapted to large scale exploration due to its high computational complexity.

---

RatSLAM is a bio-inspired single-camera SLAM system developed by Milford et al. [2004], using a computational model of the rodent hippocampus, which is distinct from other probabilistic SLAM systems presented so far. The approach uses a combination of a three-dimensional competitive attractor network and visual scene matching to form a location hypothesis. This approach was later adapted by Prasser et al. [2005] to be usable in outdoor environments, and works well on images obtained from cheap cameras. RatSLAM has successfully mapped many large-scale indoor and outdoor locations (Milford and Wyeth [2008b]), and has been combined with other approaches in order to address the challenging problem of navigation at different times of the day (Glover et al. [2010]).

Appearance-based SLAM systems augment visual localisation methods with the ability to determine whether an observation comes from a previously unvisited place.

One of the most successful algorithms is FAB-MAP (Fast Appearance-based Mapping), proposed by Cummins and Newman [2008a]. Instead of approaching the SLAM problem from a geometric perspective, FAB-MAP performs localization and mapping entirely in appearance space. A rigorous probabilistic approach to image matching has allowed FAB-MAP to be applied to a 1000km dataset with robust recognition of known places despite visual ambiguity between spatially distinct places. Maddern et al. [2011] reported an improvement to the robustness of FAB-MAP by incorporating odometric information into the place recognition process. Cadena et al. [2010] combined appearance-based place recognition with Conditional Random Fields (CRF) to filter out mismatches caused by visual ambiguity.

In a more recent line of research, Kawewong et al. [2011] presented an online and incremental appearance-based SLAM named PIRF-Nav, which can handle both perceptual aliasing and dynamic changes of places in highly dynamic environment using omnidirectional images. Maddern et al. [2012] developed a Continuous Appearance-based Trajectory SLAM (CAT-SLAM), which augments sequential appearance-based place recognition with local metric pose filtering to improve the frequency and reliability of appearance-based loop closure.

---

Milford and Wyeth [2012] presented a solution to visual navigation under weather or seasonal changes, named SeqSLAM. Instead of matching a single previously seen image given the current frame, they calculated the best candidate matches within every local navigation sequence, and then performed the localisation by recognising coherent sequences of the best candidate matches. In (Milford [2013]), the author studied the effect of the length of the matching sequences on the SeqSLAM algorithm performance. However, the SeqSLAM algorithm is based on an assumption of trajectory invariance, and is sensitive to the length of the sequence.

Recently, Maddern and Vidas [2012]; Magnabosco and Breckon [2013]; Neubert et al. [2013] proposed to solve the vSLAM problem based on both visible and thermal imaging. Thermal and visible imaging provide complementary information derived from the same scene: combining them can increase the landmark detection accuracy and the loop closure detection reliability, allowing a continuous SLAM operation across different times of day.

## 2.2 Image features and visual vocabulary

One way to characterise an image is based on extraction and description of significant points or regions. This is a widely applied technique for image retrieval and object recognition, as well as for robot localisation and loop closure detection.

Image local feature extraction consists of detection and description phases. The local feature detector serves to locate points which differ significantly from their immediate neighbourhood, while the feature descriptor captures the information in a region around these detected feature points. There is no consensus on the question of which interest point detector and descriptor are more suitable for vSLAM. Ideally, the feature detector should find salient regions in such a manner that they are repeatably detected despite modest changes in illumination, translation, orientation and scale.

Harris Corner Detector and Harris-Laplace (Harris and Stephens [1988]; Mikolajczyk and Schmid [2001]), Hessian Detector and Hessian-Laplace (Beaudet [1978];

---

Mikolajczyk and Schmid [2004]), Difference of Gaussian (DoG, SIFT Detector) (Lowe [1999, 2004]), Fast-Hessian (SURF-Detector) (Bay et al. [2008]), Center-Surround Extremas (CenSurE) (Agrawal et al. [2008]), Features from Accelerated Segment Test (FAST) (Rosten and Drummond [2006]), and Maximally Stable Extremal Region (MSER) (Matas et al. [2004]) are some prominent feature detectors that have been applied to vision-based localisation and mapping tasks. Different detectors offer different properties as required by their varying usage scenarios. For example, the Harris Corner Detector was explicitly designed for geometric stability: whereas SIFT keypoints have been shown to be robust to changes in scale, image plane rotations, illumination, and camera noise; the FAST corner detector is computationally efficient, but offers lower repeatability.

Similarly, the image descriptor should be distinctive, concise and robust to image distortions: its performance which can be compared with other descriptors with reference to a distance metric. Many methods for feature descriptions have been suggested. (See, for example, Bay et al. [2008]; Calonder et al. [2010]; Lowe [1999, 2004]; Mikolajczyk and Schmid [2005]; Rublee et al. [2011]).

Many global features have also been proposed to describe the image content. These methods use all pixels to compute a unique signature for the image. Consequently, their use is straightforward: typically, they utilize color property, textures, or a combination of both. For example, Rubner et al. [1997] proposed a Histogram search algorithm to characterise an image by its colour distribution; Menegatti et al. [2004a] applied the Discrete Fourier Transform (DFT) to build image descriptors for panoramic images; and Kunttu et al. [2004] introduced a Fourier-based descriptor presented in multiple scales for image retrieval tasks. Other examples include (Blaer and Allen [2002]; Bradley et al. [2005]; Fazi-Ersi and Tsotsos [2012]; Itti et al. [1998]; Ulrich and Nourbakhsh [2000]; Weiss et al. [2007a]; Zhou et al. [2003]).

In the rest of this section, we review some popular image descriptors that have been exploited by the robotics research community, and assign them to one of two classes: local feature descriptor, or global appearance descriptor. We also review visual vocabulary techniques that improve the efficiency of image retrieval process based on local feature description. The performance evaluation of differ-

---

ent detectors and descriptors are given in (Huynh et al. [2009]; Mikolajczyk and Schmid [2005]; Schmidt et al. [2010]; Winder and Brown [2007]).

### 2.2.1 Image descriptors

Amongst the various local feature extraction and description methods, SIFT and SURF dominate the visual descriptor choice. Both exhibit great performance under a variety of image transformation, and are thus a good choice for the first two descriptors to review.

SIFT (Scale-Invariant Feature Transform) was developed by Lowe [1999] for image feature extraction in object recognition applications. SIFT extracts features that are invariant to image scaling, rotation, and camera view-point changes. The SIFT descriptor represents local image patches around interest points characterised by coordinates in the scale space, in the form of histograms of gradient directions. The 128-dimensional SIFT descriptors have high discriminative power, while remaining robust to local variations. These characteristics make them highly suitable for localisation.

A successful example of the approach based on SIFT features was described by Se et al. [2001a,b, 2002, 2005]. They built a database map with distinctive SIFT landmarks from unmodified environments. Without any prior knowledge about its position, the robot localised itself by matching visual landmarks in the current image to a database map. In order to reduce computation time, a smaller vector containing 16 elements rather than 128 (Lowe [1999]) was used to characterise a SIFT feature. The Euclidean distance measure between the descriptors of two features was computed to check whether they were below a matching threshold. Jensfelt et al. [2006] presented a framework that was able to extract landmarks for SLAM using Harris-Laplace corner detection and a modified SIFT descriptor. The rotationally ‘variant’ SIFT descriptor was developed in order to make the landmarks matching procedure faster. This is achieved by avoiding canonical orientation at the peak of the smoothed histogram.

Currently, there are many variants that improve on the performance of the orig-

---

inal SIFT algorithm. For example, PCA-SIFT (Ke and Sukthankar [2004]) applies Principal Components Analysis (PCA) to the normalized gradient patch rather than the gradient histogram in order to get a compact descriptor. GSIFT (Mortensen et al. [2005]) integrates global texture information into the basic SIFT, while CSIFT (Abdel-Hakim and Farag [2006]) adds color invariance, and ASIFT (Morel and Yu [2009]) incorporates invariance to affine transformations. GPU-SIFT (Sinha et al. [2006]) is an implementation of SIFT for GPU (Graphics Processing Unit), and processes pixels/features in a parallel manner.

Speeded-Up Robust Features (SURF) was developed by Bay et al. [2008] and is a scale- and rotation-invariant local detector and descriptor. The main motivation for the development of SURF was to approximate the performance of SIFT while being more computationally efficient. This is obtained by using integral images, a Hessian matrix-based measure for the detector and a distribution of Haar wavelet responses for the descriptor. In the work of Valgren and Lilienthal [2008], an incremental spectral clustering (ISC) algorithm was applied to segment continuous space into topological nodes, and local feature matching was used for localisation. This work focused on robustness to seasonal changes and differing weather conditions in large scale indoor/outdoor environment. SURF variants were employed as local feature descriptors of high-resolution panoramic images. These ignore the rotational invariant characteristic of SURF. Epipolar constraint was used to improve matching performance at little extra cost.

Gradient Location and Orientation Histogram (GLOH), proposed by Mikolajczyk and Schmid [2005], is an extension of the SIFT descriptor, and also makes use of a local position-dependent histogram of gradient orientations around an interest point. It is designed to increase robustness and distinctiveness. GLOH is differentiated from SIFT in three main aspects: first, instead of the rectangular grid used in the regular SIFT, GLOH computes the descriptor over a log-polar location grid; secondly, the gradient orientation is quantised into 16 bins as opposed to 8 bins; and finally, the dimensionality of the descriptor is reduced by using principal component analysis (PCA). Consequently, GLOH results have been shown to be more distinctive, but also more expensive to compute than SIFT.

Linde and Lindeberg [2004] designed another histogram-like image descriptor,

---

referred to as high dimensional Composed Receptive Field Histograms (CRFH), which was considered an effective image description for place recognition. A CRFH is a multidimensional statistical representation of the occurrence of the responses of several image descriptors applied to the whole image. It can be computed from several types of image descriptors, such as normalized Gaussian derivatives, differential invariants (mainly the normalised gradient magnitude, the normalised Laplacian and the normalised determinant of the Hessian) and chromatic cues obtained from RGB images. Each dimension corresponds to one descriptor, and the cells of the histogram count the pixels generating similar responses under all descriptors. This approach permits the capture of various properties of the images as well as relations that occur between them.

More recently, a few lightweight feature descriptors (binary descriptors), which are targeting real-time applications processing richer data at higher rates, have attracted the attention of researchers ( Calonder et al. [2010]; Leutenegger et al. [2011]; Ortiz [2012]; Rublee et al. [2011]; Yang and Cheng [2014a]).

Binary Robust Independent Elementary Features (BRIEF) was the first binary descriptor published (Calonder et al. [2010]). It is a general-purpose feature descriptor that can be combined with arbitrary detectors. BRIEF is based on a relatively small number of intensity difference tests to represent an image patch as a binary string. Given a pair of points, if the intensity value of the first point is larger than the intensity value of the second point, the bit corresponding to this given point pair is assigned to value 1, else 0. Finally, a string of boolean values can be retrieved after intensity comparison of a number of pairs. BRIEF is robust to typical photometric and geometric image transformations, but not to viewpoint changes. It does not use an elaborate sampling pattern, the sampling scheme being based on uniform and Gaussian random sampling using different distribution parameters, determined experimentally. As with all the binary descriptors, the distance measure of BRIEF is the number of the different bits between two binary strings, which can also be computed as the sum of the XOR operation between the strings (or the number of the wrong correspondences). Such similarity measure can be computed very efficiently (much faster than the commonly used L2 norm).

---

The ORB (Oriented FAST and Rotated BRIEF, Rublee et al. [2011]) is one of the extensions of the basic concepts of BRIEF, based on the FAST detector (Rosten et al. [2010]; Rosten and Drummond [2006]). It addresses the shortcoming of the basic form of BRIEF mentioned above and improve upon it in two respects. The first improvement is increased robustness to viewpoint changes based on computing the unambiguous orientation from the FAST corner. The second improvement aspect is learned sampling pairs, achieved by using machine learning to de-correlate BRIEF features under rotational invariance. This makes the nearest neighbour search during matching less error-prone (Schmidt et al. [2013]).

BRISK is another extension of BRIEF, proposed by Leutenegger et al. [2011]. It presents some differences from both BRIEF and ORB in employing a sampling a pattern that is composed of concentric rings in which points are equally spaced. The FREAK (Fast Retina Keypoint, Ortiz [2012]) descriptor is also inspired by BRIEF. It suggests the use of a biologically-inspired retinal sampling pattern, which is also circular, but with the difference of having a higher density of points near the centre. This sampling pattern allows for the use of a coarse-to-fine approach to feature description. The first sampling pairs mainly compare points in the outer rings of the pattern, while the later pairs mainly compare points in the inner rings of the pattern. This is similar to the way in which the human eye operates. FREAK then tries to learn the pairs by maximizing variance of the pairs and taking pairs that are not correlated. Later, a cascade approach is used to further speed up the matching, allowing for faster rejection of false matches and shortening of the computation time.

The LDB (Local Difference Binary, Yang and Cheng [2014a]) descriptor follows the same basic principle as BRIEF, but using a region-based binary test instead of the single pixel method to compute the binary strings. In addition to the average intensity, the average of horizontal and vertical derivatives of equal-sized spatial regions are both compared, providing a more complete description than BRIEF. A three-level grid scheme is applied to encode the spatial structure at different scales. The LDB descriptor is obtained by concatenating the selected bits. To further enhance the distinctiveness of LDB, Yang and Cheng [2014b] adopt a bit selection scheme extended from the AdaBoost to automatically select a set of

---

salient bits. The goal of this scheme is to maximize (minimize) the Hamming distance between mismatches (matches). In addition to these local descriptors, there are ways in which to provide a global description of the information in a given scene. In the rest of the subsection, we review some popular description methods given in the literature.

The Discrete Fourier Transform (DFT) of an image can be used as a global descriptor of the scene that contains information about the dominant structural patterns, and is invariant with respect to the position of the objects. In particular, the Fourier transform of omnidirectional images exhibits the property of being invariant to image rotations, so that the orientation of the robot does not need to be taken into consideration in the matching phase.

There is another global descriptor, the Fourier-Mellin Invariant (FMI) descriptor introduced by Casasent and Psaltis [1976] that relies on the Fourier-Mellin Transform (FMT). The FMT takes advantage of properties of the Fourier and Mellin Transforms, which in combination are invariant with respect to translation, rotation and scale change. It has been applied by Bulow and Birk [2009]; Goecke et al. [2007]; Kazik and Goktogan [2011] for robot localisation purposes. Both of the above-mentioned descriptors will be revisited in Section 2.2.3.1 as data reduction techniques.

In order to mimic the human ability to immediately recognise the meaning (gist) of a scene, many researchers assume a direct mapping onto scene primitives in absence of the identity of the objects present. Oliva and Torralba [2001] proposed the Gist descriptor to address this problem. They proposed that the spatial structures of a scene can be described by several important statistic of the scene. Specifically, the Gist descriptor encodes the amount, or strength, of vertical/horizontal lines in an image, which can contribute to matching images with similar distributions of lines and textures. The Gist descriptor of an image is built from the responses of steerable filters at different scales and orientations. Several models utilising different types of Gist of a scene have been presented in mobile robotics, and this will be reviewed again in Section 2.3.1.4.

---

### 2.2.2 Visual vocabulary

Place recognition based on matching numerous local features consumes too much time for use in real-time systems. Consequently, the idea of a visual vocabulary method inspired by object recognition and text retrieval techniques built upon local invariant features has frequently been applied to this problem. The visual word vocabulary is established by clustering a large set of local features extracted from a training image corpus, in which the visual words are the cluster centers corresponding to informative regions in a image. A histogram of the frequency of visual words is used to summarize the entire image, by counting how many times each of the visual words occurs in the image. Performance in the retrieval of objects depends heavily on the distinctiveness of the vocabulary.

The first application of visual vocabulary to object retrieval in videos was conducted by Sivic and Zisserman [2003]. This idea was later extended by Nistér and Stewénus [2006] utilizing hierarchical k-means to recursively subdivide the feature space in a tree fashion, which allows the image matching to be significantly faster in a large database. Schindler et al. [2007] proposed a system for large-scale place recognition using these tree structures. Many recent appearance-based localisation and loop closure methods therefore rely on visual bags of words based on SIFT or SURF features. Wang et al. [2005] employed the idea of the visual vocabulary relating to grey images to perform global localisation. The visual vocabulary is learned off-line from SIFT descriptors using the k-means algorithm.

The visual vocabulary technique was also adopted in (Cummins and Newman [2008a]) where a principal probabilistic approach for appearance-based place recognition was proposed. The system takes into account the probabilities of features appearing together, and is able to calculate the probabilities that two images show the same place. This allows the system to recognise known places despite perceptual aliasing. A recursive Bayes estimation was used for the location estimation. The loop closure problem was considered over kilometres of travel, in which the matching between current and reference images was performed by detecting the presence or absence of features in each image from a visual vocabulary, based on quantized SURF descriptors. In this work, the generative model

---

of appearance is learned in an offline process, and the vocabulary dictionary is offline built as well, as the computational complexity can be prohibitive.

Filliat [2007] chose instead to described an interactive qualitative localisation system in which the visual vocabulary is learned online along with the image acquisition, in an incremental manner. Three different features, including SIFT keypoints, colour histograms and a normalised grey level histogram is extracted from images taken from a random orientation, and the corresponding words found in the dictionary. A two stage voting scheme is used to estimate the location. This process is repeated until either the quality of the vote reaches a given threshold, or a given number of images is reached. If the quality threshold has been reached, the place is then considered recognized: if no recognition is made and the limit number of images has been reached, non-recognition is considered achieved. Epipolar geometry is used to reject outliers when perceptual aliasing is present in the environment. In order to avoid exhaustive image-to-image comparisons of the visual features, the inverted index associated with the dictionary was adopted during the computation of the likelihood for the loop closure. However, using a simple linear search algorithm entailed that the size of the manageable environments was quite limited. Consequently the method was only validated for an indoors environment. Similarly, Angeli et al. [2008b] designed a simple online method to detect loop closure based on the BoWs scheme through the incremental creation of a visual vocabulary in a probabilistic framework.

Most recently, Mariottini and Roumeliotis [2011] have presented a strategy for vision-based localisation using a vocabulary tree: this allows the robot to navigate in a large-scale image map. This image map is represented as a graph, in which nodes correspond to training images, and links connect similar images. In this work, the sequence of distinctive images is exploited to disambiguate the localisation ambiguity. A place recognition system using BoWs combined with Conditional Random Fields (CRF) was proposed in Cadena et al. [2010], where CRF-Matching was applied to associate image features. An improvement to this system that considers features in the background of the image obtained was reported by Cadena et al. [2012]. When the system finds several memorised images that match the current image, the 3D information is then exploited to solve mis-

---

matches.

### **2.2.3 Dimensionality reduction techniques**

Dimensionality reduction is the process of searching for a low-dimensional manifold embedded in the high-dimensional data, and can be divided into feature selection and feature extraction. A problem that confronts many robotics applications is the large amount of data to be processed relative to limited computational resources. Therefore, there is growing demand for image descriptors that are memory-efficient, and offer rapid calculation and image matching.

#### **2.2.3.1 Fourier transform**

Several researchers have explored the use of more general dimensionality reduction techniques to represent the input image set, such as the Fourier transform decomposition of the image content into the basis functions. The Fourier coefficients of the low frequency components were used by Ishiguro and Tsuji [1996]; Yagi et al. [1998], and Menegatti et al. [2003, 2004a,b] to compute the similarity between a reference image and the current input image, which was computed from a discrete Fourier transform of an unwrapped omnidirectional image. The system can calculate the position of the robot with an accuracy that could be varied by choosing different number of Fourier components to compare in the similarity function. Specifically, a broad localization could be obtained by calculating the first few frequency components, while a more precise matching could be acquired by extending calculation to higher frequency components in the similarity function.

In the work of Ferdaus et al. [2008], colour histograms and the Fourier transform technique of image comparison were both employed for place recognition. In order to localise the mobile robot, a discrete Bayes filter was used to represent probability distributions: the training image with the highest probability value identifies the probable location of the mobile robot in the environment. Analysis of visual information was conducted in the frequency domain using the Fourier-

---

Mellin Transform (FMT) to obtain rotation, translation and scaling between consecutive images. These similarity transforms were calculated through phase correlation and used to update the rover position and heading estimates.

### **2.2.3.2 Principle Component Analysis (PCA)**

Another dimensionality reduction technique is Principal Component Analysis (PCA) invented by Pearson [1901]. PCA finds the principal components of data by calculating eigenvalues and eigenvectors of the covariance matrix. It is able to linearly project high-dimensional image descriptors onto a low-dimensional subspace, retaining only the principal image components.

Jogan and Leonardis [1999] employed an eigenspace model to build a compact representation of environments. The image set was represented as points in the eigenspace by estimating the most significant eigenvectors. The researchers used the nearest neighbour to estimate the similarity of images, and four criteria were defined to measure the recognition rate for localisation. However, the limitation of this method is that it is not sufficiently robust against occlusions and lighting changes.

The first attempt at dimension reduction for local features was PCA-SIFT, proposed by Ke and Sukthankar [2004]. The original SIFT descriptor is represented as a 128 dimensional vector: this can be reduced to 36 dimensions, by performing PCA on the gradient patches of an image.

Kröse et al. [2000] built a representation of the appearance by applying PCA to the images, and then representing places as a Gaussian density, which enables Markov localization. PCA is used for the same purpose in (Artac et al. [2002]; Gaspar et al. [2000]; Valenzuela et al. [2012]). Gaspar et al. [2000] proposed a scheme in which the greyscale omnidirectional image was compressed by building a reduced-order manifold using PCA. At place recognition time, the current image is projected onto the components of the PCA space, and a qualitative localisation is obtained by detecting the nearest neighbors. PCA has also been applied by Valenzuela et al. [2012] to reduce the SIFT and SURF feature descriptors.

---

Artac et al. [2002] implemented an incremental eigenspace model for representing panoramic images in order to allow for incremental learning and adaptation without the need to retain all the input data, mitigating the increasing demands on memory capacity and computational complexity as the number of input images increases. A similar technique was also adopted to compress image data with a view to saving memory in (Ishizuka et al. [2011]). The Euclidean distance between points is used as the measure of image similarity in the eigenspace.

### 2.2.3.3 Other approaches

While PCA is one of the most widely-used linear dimensionality reductions, this technique will not work given the scenario that data are distributed on a highly nonlinear curved surface, i.e., manifolds. A nonlinear dimensionality reduction technique called Isomap was designed by Tenenbaum et al. [2000] to preserve the neighbourhood of points in a low-dimensional manifold. Ramos et al. [2012] applied Isomap to reduce image patches to a low-dimensional space in which further statistical learning methods are then used to create a probabilistic density for each place. Place recognition is performed by computing the log-likelihood of an entire image over each place model.

Image descriptor quantisation techniques are also utilised for dimensionality reduction purposes. Generally, each number of elements of the floating-point vector is quantised so that it falls within a prescribed integer range limit. Tuytelaars and Schmid [2007] applied quantization to the SIFT descriptor: the resulting vector is only 4 bits per coordinate. Winder et al. [2009] introduced an image descriptor pipeline in which the combination of PCA and quantisation is used to compress the representation of descriptors. Chandrasekhar et al. [2009a] demonstrated a compressed histogram of gradients (CHOG) descriptor using a tree-code method.

More recently, many efficient approaches have been developed to find binary representation of high-dimensional data while maintaining their semantic similarity in the Hamming space. This is usually performed by thresholding the vectors after multiplication of the descriptors with a projection matrix, and retaining only the sign of the results. Such methods combine the effects of dimensionality re-

---

duction and binarisation, greatly hastening the matching process while requiring less memory. The similarity between descriptors can be computed very efficiently either using hash tables or efficient bit-count operations.

Torralba et al. [2008] proposed a scheme that uses Locality Sensitive Hashing (LSH) to learn compact binary codes from the Gist descriptor. Salakhutdinov and Hinton [2007] used nonlinear Neighborhood Component Analysis to binarise the Gist descriptor. Strecha et al. [2012] developed a scheme that uses hash functions to compute a binary descriptor that is robust to illumination and view-point. Hua et al. [2007] proposed an algorithm for learning local image descriptors using Linear Discriminant Analysis. Takacs et al. [2008] reduced the bit rate of SURF descriptors by using quantisation and entropy coding. Chandrasekhar et al. [2009b] addressed the compression of SIFT and SURF descriptors using transform coding. Yeo et al. [2008] used coarsely quantised random projections on SIFT descriptors to build binary hashes: descriptors are then compared using the Hamming distance between binary hashes.

## 2.3 Place recognition

Place recognition is one of the central issues in mobile robotics, determining the ability of a robot to localize itself in its environment. Vision-based place recognition methods usually consist of two procedures. Initially, images and prominent features of the environments are recorded as reference images. The reference images are labelled with some places. Afterwards, image comparisons are used to detect whether the current captured image can be associated with a known place. Recently a variety of approaches have received attention. These methods have employed either regular forward-facing cameras (Filliat [2007]; Kosecka and Li [2004]; Torralba et al. [2003]) or omnidirectional sensors (Artac et al. [2002]; Belletto et al. [2008]; Blaer and Allen [2002]; Gaspar et al. [2000]; Liu and Siegwart [2014]; Menegatti et al. [2003]; Murillo et al. [2007]; Thompson et al. [2000]; Ulrich and Nourbakhsh [2000]; Valgren and Lilienthal [2008]; Wang and Lin [2011]) to acquire images.

---

Several approaches to vision-based place recognition have been proposed during the past two decades, motivated by appearance-based approaches to object recognition. The main concerns for these methods have been: how best to represent the environment from sensor information to guarantee invariance under changes of illumination, pose, viewpoint, scale; how to achieve robustness to partial occlusion, clutter, and dynamic backgrounds; and how to ensure efficient and accurate image matching.

Some authors extracted a set of features - such as points, lines, and contours (Booi et al. [2007]; Fei-Fei and Perona [2005]) - to find correspondences between the current captured image and a reference image or set of reference images. The accuracy of these methods is highly dependent on the features used for matching, and the robustness of the feature descriptor. Other authors chose instead to use direct comparison of two images pixel by pixel, or to extract a signature from the raw images and then calculate the image similarity to perform place recognition tasks (Gross et al. [2003]; Li et al. [2000]; Pretto et al. [2010]). The disadvantage of these methods for image matching is that they require large amounts of memory and are computationally expensive. The combination of both methods provides a better solution in (Kosecka and Li [2004]; Rostami et al. [2013]).

In (Kosecka and Li [2004]), two different image descriptors and their associated distance measures were compared. The first descriptor is the gradient orientation histograms: the second is a set of local scale-invariant features. The experimental results show that the local scale-invariant features outperform orientation histograms, due to their superior discrimination capabilities and better invariance properties with respect to viewpoint changes.

Rostami et al. [2013] presented an integrated feature extraction model based on salient line segments (SLS), in which the local feature vectors are formed from the frequency of the appearance of SLSs in the finer scale, while the global features are derived from the coarser scales of the SLSs. In this model, the salient lines of an image are first obtained in four directions by applying the center-surround filter and color opponency technique. The SLS of the image patches is then extracted by creating a histogram of gradients in the receptive cells. Finally, multi-class SVM with a Radial Basis Function (RBF) kernel is used to classify the input

---

image for place recognition. However, the authors reported that their method could not deal with the presence of shadows and large occlusions.

### **2.3.1 Solutions to the place-recognition problem**

Solutions to the place recognition problems might be divided into four main types: histograms-based methods, object-based methods, region-based methods, and context-based methods.

#### **2.3.1.1 Histograms**

Histograms of various image properties (e.g. colour or image derivatives) have been widely used in appearance-based place recognition. The concept of using colour histograms as a method of matching two images was pioneered by Swain and Ballard [1991]. Colour histograms of omnidirectional images were originally utilised in (Ulrich and Nourbakhsh [2000]) to perform place recognition. They used six one-dimensional histograms for each image, three for the HLS (hue, luminance, saturation) colour bands and three for either the RGB or normalised RGB colour bands. Colour images were classified by processing each colour band separately using nearest-neighbour learning, and the results of classification from all colour bands were then combined with a simple scheme based on unanimous voting. The recognition phase was done by comparing images acquired online with the images of neighbour nodes using histogram matching on individual colour bands. Histograms were compared with Jeffreys divergence. This method is inspired by image retrieval techniques, but is more efficient because comparison is only made with images in the neighbourhood of the current location.

The work studied in (Blaer and Allen [2002]) is closely similar to that in (Ulrich and Nourbakhsh [2000]). The primary difference between the two works is that the former addresses the problem of outdoor environmental navigation involving illumination changes. In order to reduce the impact of lighting variation in uncontrolled environments, Blaer and Allen [2002] used a normalisation process on the images before histogramming them. The percentage of each colour at that

---

particular pixel, regardless of the overall intensity of that pixel, was used for histogramming.

The most commonly used histogram is the colour histogram, which is the representation of the distribution of colours values in the image. It has the advantages of rotation and translation invariance about the viewing axis. However, colour histograms can simply express the global colour information of an image, without spatial relationship. They may give ambiguous results in environments with uniform colour and luminance characteristics, which often result in high similarity values among images that are very different but exhibit similar colour histograms.

To address this shortcoming, Zhou et al. [2003] used edge density, gradient magnitude and textures in addition to colour information to set up a multidimensional histogram. The recognition step is to match a multidimensional histogram of the current image with candidate multidimensional histograms in the sample database. The Jeffrey divergence was chosen as the distance metric to evaluate the similarity between the current image and any given histogram from the database. The authors evaluated their method on an intelligent wheelchair in their lab environment, where the best percentage of correct self-localisation was 82.9%.

Blaer and Allen [2005] developed their earlier work and presented a hybrid method for localisation. Five levels of resolution for each image were used, instead of one in colour histogramming. The multiresolution histograms provided additional information about spatial relationships in the scene. First, the original image was convolved with a  $5 \times 5$  Gaussian kernel to blur it: then the blurred image was sub-sampled down to the lower resolution. The resulting multiresolution histogram is a set of five 256-bucket sub-histograms.

In (Košecká et al. [2003]), appearance in indoor environments was characterised by a simple gradient orientation histogram. In order to obtain a more robust measure, the gradient orientation histograms was computed only for the pixels with magnitude above an empirically-determined threshold. Once the features had been selected using gradient orientation histograms, the  $\chi^2$  distance metric was used to compare different features. In addition, five sub-images, including

---

one in the center and four quarters of the original image, were considered for comparison when the confidence level was below the given threshold in order to refine the classification.

Pronobis et al. [2006] modelled a visual place recognition technique based on composed receptive field histograms in combination with a large margin classifier (Support Vector Machines, SVMs) and applied this to indoor environments. High-dimensionality histogram features were used as a global image descriptor, which was computed from second order normalised Gaussian derivative filters applied to the illumination channel. The histograms consisted of six dimensions, with 28 bins per dimension.

In more recent work, spatial PACT (Principle Component Analysis of Census Transform histograms), a new representation for recognising instance (“I am in Room 113”) and categories (“I am in an office”) of places was introduced in (Wu and Rehg [2008]). PACT is a global representation that extracts the Census Transform (CT) histograms for several image patches organised in a grid and applies PCA to the resulting vector. CT is a non-parametric local transform designed for establishing correspondences between local patches, which compares the intensity values of a pixel with its eight neighbourhood pixels. A histogram of the CT values encodes both local and global information of the image.

Similarly, another interesting recent effort focuses on the classification task to distinguish places in the environment (Fazi-Ersi and Tsotsos [2012]). In this work, histograms of oriented uniform LBPs (Local Binary Patterns) are extracted from images to categorise places indoors and outdoors. Wang and Yagi [2013] proposed a new image feature, the Orientation Adjacency Coherence Histogram (OACH), to carry out coarse topological localisation. SIFT descriptors are then used for the fine localisation. The system works well in both indoor and outdoor environments.

---

### 2.3.1.2 Object-based methods

Much research on vision-based place recognition tends to focus on landmark-based approaches. Such methods rely on either artificial or natural features in order to extract information about position. Place recognition is performed by finding matches between the candidate landmarks visible in the current image and those in the database. This can be very fast and reliable if landmarks are well designed for efficient detection and well distributed in the environment. Many early approaches utilised artificial landmarks (Briggs et al. [2000]; Case et al. [2011]; Fairfield and Maxwell [2001]; Huh et al. [2006]; Sousa et al. [2009]; Yoon and Kweon [2002]), such as reflectors, ultrasonic beacons, and traffic signs, etc.. Various features have been used as natural landmarks (Asmar [2006]; Hayet et al. [2003]; Jennings et al. [1999]; Segvic and Ribaric [2001]; Thrun [1998]), such as simple features (vertical edges, corners), or characteristic objects (doors, corridors, and distinctive buildings).

The main problem in natural landmark-based systems is to detect and match characteristic features from sensory inputs. The selection of features is important, since it will determine the degree of complexity in feature description, detection, and matching. Proper selection of features will also reduce the chances for ambiguity and increase positioning accuracy.

In a sparse and indoor environment, many of the detected features correspond to corners. One system described in (Jennings et al. [1999]) used corner features and least-squares optimisation to find the transformation between the coordinate frames of the robot for cooperative robot localisation. They proposed an implementation of a multi-robot navigation system that used stereo vision in dynamic indoor environments. Segvic and Ribaric [2001] calculated the orientation of a moving robot by finding the contour of the closed corridor in which the robot was moving. Thrun [1998] and Asensio et al. [1999] used doors as their primary landmarks, since doors were regular and easily distinguishable features in their experimental environment. Their localisation algorithm is based on Markov localisation. In (Howard and Kitchen [1999]), the environment was described in terms of the location of walls and doorways, and a probabilistic localisation technique

---

was used for robot localisation. The system maintained a probability distribution over the space of all possible robot locations.

The problem of selecting salient and distinctive features from gray-scale images was addressed in (Knappek et al. [2000]). Salient features are selected with the Harris corner detector, which is robust to small changes in view point. Potential landmarks are characterised by a feature vector derived from its first and second derivatives, which are ordered by distinctiveness, the most distinctive being reserved. Recognition is then performed by nearest neighbour classification. The most distinctive landmark is that which has the largest Mahalanobis distance from all the others.

Thompson et al. [2000] described a system where localisation tasks were performed by automatically selecting good landmarks from panoramic images and places learning. Good landmarks are defined as those having good static and dynamic reliability, and that are distributed through the image. An adoption of the biologically inspired Turn Back and Look behaviour is used to evaluate potential landmarks. The landmark is represented by a  $16 \times 16$  window. Static reliability is determined by the uniqueness of the landmark in its neighbourhood. Uniform distribution is guaranteed by dividing each image into 4 patches (forward, back, left and right) and selecting the best four landmarks from each patch. Dynamic reliability is measured by the average of the static reliabilities along a test path. The landmarks with the highest dynamic reliability measure are used to represent the place. Matching is performed by a normalised correlation, to gain some robustness to illumination changes.

The combination of edge, corner and colour features was used to represent the environment locations in (Lamon et al. [2001]). Each location was denoted by a list of characters, where the letter ‘V’ characterised a vertical edge and the letters ‘A’, ‘B’, ‘C’, ..., ‘P’ represented hue bins detected by a colour patch detector. The similarity of any two strings was given by the resulting minimum energy of traversal, the value 0 referring to self-similarity.

Hayet et al. [2003] proposed a visual localisation strategy based on detection and recognition of visual landmarks that are planar quadrangular objects, such as

---

doors, windows, posters, cupboards, etc.. Homography rectification was applied to obtain an invariant representation for the PCA learning stage. Asmar [2006] developed a tree trunk recognition system which matches trees by extracting SIFT features within the borders of the trunks. This is achieved by segmenting quasi-vertical structures and choosing those structures that intersect the Ground-Sky separation line.

There are some approaches that rely on image retrieval techniques to identify the current position of the robot. These are used to find images in a given database that look similar to the given query image. Wolf et al. [2005] used an image retrieval system based on local features that are invariant with image translations and limited scale as the basis of a Monte Carlo localisation technique. Li [2006] demonstrated an approach for location recognition in indoor environments. Reduced SIFT features were extracted to represent the individual location and recognition was approached by feature matching between query and reference views. The Hidden Markov Model framework was exploited to reduce the ambiguity due to self-similarity and dynamic changes in the environment. Campos et al. [2012] described a place recognition framework in which recognition was conducted by finding the nearest neighbour among SIFT descriptors using mutual information measurement. In (Liu and Siegwart [2014]), the authors made use of the color features and geometric information that were extracted from a panoramic image to represent the environment. A Dirichlet process mixture model (DPMM) was exploited to estimate the current localization of the robot.

Natural landmarks are flexible, easy to use and cheap: however, they are also often sparse and unstable. Artificial landmarks are simple and suited for localisation and place recognition, especially in environments that are impoverished in the sense that unique natural landmarks are lacking. Artificial landmarks can be predefined, and this tends to reduce the complexity of the localisation algorithms. Researchers have used different kinds of patterns, coloured marks, 1D or even 2D barcodes, resorting to geometrical constraints and the associated techniques for position estimation. Once the landmarks are identified, the 3D position and orientation of the landmarks relative to the on-board camera can be estimated, and, consequently, the robot position and orientation relative to the landmarks.

---

A self-localisation technique based on colour pattern recognition was proposed by Yoon and Kweon [2002]. The system used colour image processing to find coloured markers, which consist of symmetrical and repetitive structures. To make each landmark distinguishable from the others, and thus to eliminate false positives for marker recognition, multiple colours having maximum distance in the chromaticity colour space were selected for each landmark.

Jang et al. [2002] made use of a pair of coloured rectangles as navigation and localisation aids. Briggs et al. [2000] used simple artificial landmarks which were made up of self-similar intensity patterns coupled with a barcode for unique identification for localisation tasks. These landmarks could be easily attached to the walls. Sousa et al. [2009] proposed a vision system to detect and identify barcodes, and to retrieve the geometric relationship between the camera and the observed markers, thereby deriving localisation information for a robot. Huh et al. [2006] addressed the localisation and navigation problem for service robots by using invisible two-dimensional barcodes on the floor surface.

In (Fairfield and Maxwell [2001]), small green plastic rings are used as landmarks. Their method projected the acquired coordinates of the landmarks in the image plane, then calculates the distances between the robot and the various landmarks. This perceived distance can be validated by comparison with the pre-stored positions of landmarks. A simple Kalman filter was integrated into the visual landmark estimation in order to correct accumulated odometry and sensor errors.

Mata et al. [2003] made use of information signs to guide a robot based on their recognition. In this system, the localisation is done by detecting 2D landmarks, including text and icons designed for human use in an office environment. More recently, Case et al. [2011] exploited text detection and recognition techniques for named location recognition, without assumptions about the language structure or spatial layout of the text. Other approaches for visual markers include using coloured poles (Sousa et al. [2005]), balls (Betke and Gurvits [1997]; Iocchi and Nardi [2000]), etc., in soccer environments.

In general, artificial landmarks are easier to detect than natural landmarks. How-

---

ever, artificial landmarks require modification of the environment. Most of the landmark-based localisation systems are tied to a specific environment: they can rarely be easily applied to different environments. For example, if ceiling lights are used as primary landmarks, the system will fail if the environment does not contain ceiling lights, or the robot does not possess a sensor that can detect them. Therefore, artificial landmarks are hardly feasible, and in any case undesirable in a large scale environment, such as an entire city.

Alternatively, other systems rely on recognition of objects that are either known a priori, or extracted dynamically (Ekvall et al. [2006]; Ranganathan and Dellaert [2007a]; Vasudevan et al. [2007]). This process depends on the objects observed and their interrelationships.

In the framework of Ekvall et al. [2006], the semantic structure of the environment in a service robot scenario was acquired automatically. The system used object recognition techniques to detect objects and build an augmented map, then used this map to perform navigation and fetching tasks. Image differences between the presence or absence of foreground objects was used to segment the objects from their background. After segmentation, visual features (gradient magnitude and Laplacian response) were extracted and used for building Receptive Field Concurrence Histograms (RFCH), which can capture more geometric information compared to a regular histogram. During the running stage, the RFCH of object hypothesis and the target object were compared using histogram intersection, resulting in a vote matrix. SIFT matching was used for final verification, giving a set of hypothesised object locations.

Vasudevan et al. [2007] put forward an object-based hierarchical probabilistic representation of space which allowed robots to be cognizant of their surroundings in a human-compatible fashion. Topological localisation was performed by conceptualising space, classifying surroundings and then performing recognition procedures. The SIFT method was used for recognising textured objects. A similar approach was adopted by Ranganathan and Dellaert [2007a]. A 3D generative model for place representation was presented, constructed using images and depth information obtained from a stereo camera. Places were represented as a set of objects, each object modelled as having a particular shape and appearance. Place

---

recognition involved finding the distribution of place labels, given the detected objects and their locations.

### **2.3.1.3 Region-based methods**

Some approaches do not use landmark objects, employing instead segmented image regions to form the signature of a location. The main problem is to perform reliable region-based segmentation, in which individual regions are robustly characterised and associated.

Shlomo [1998] described a place recognition method based on matching the image signature, which was defined as an array of measurement values derived from a portion of the original image. Reduced-size images ( $64 \times 48$  pixels) with 256 grey levels were employed to reduce the computational cost of the matching process. The input image is divided into  $n \times n$  blocks. For each block, a measurement function was applied to estimate the image properties, including dominant edge orientation, significant gradient direction, edge strength, edge density and degree of texturedness. The similarity between current image signatures and a set of signatures already stored in the database was calculated, in order to judge whether the current image could be associated to a known location. In addition, matching using multiple measurement functions conjunctively was considered: this was found to improve the recognition rate significantly.

Matsumoto et al. [1996, 1999, 2000] used a sequence of frontal views along a route which were captured at a certain interval in the training stage. Place recognition was then realised, based on the matching of the current view with the memorised view sequence. The calculation of similarity between the current view and a reference view was a simple block matching process. The views were represented by greyscale images, which were more suitable for indoor environments than for outdoor environments, where lighting condition may change drastically. In order to overcome this limitation, the stereo disparity can be used as a new type of view which is independent of changes in lighting condition. However, the disparity views were not sufficiently stable: moreover, the generation of disparity views was not fast enough for mobile robot navigation.

---

A similar approach was adopted by Hashem and Andreas [2004], here using Kernel PCA to extract features from the visual scene of a mobile robot. PCA is suitable for data generated by a Gaussian distribution. However, the distribution of natural images is highly non-Gaussian. Kernel PCA was investigated as a generalisation of PCA, which takes into account higher order correlations. In the localisation phase, the features of the current scene and the stored features were computed: the result of such a comparison giving rise to the knowledge of the position of the robot.

In (Bellotto et al. [2008]) another image matching algorithm was proposed for indoor environment place recognition. The heart of this image matching method involves dividing the scene image into several column regions, and then comparing each column with a region of a reference image stored beforehand. The measure of similarity between a slot of the scene image and a region of a stored image is based on the Normalised Correlation Coefficient. The images employed in this system are panoramic images reconstructed from snapshots: each image being made up of 12 snapshots taken at intervals of  $30^\circ$ .

#### **2.3.1.4 Context-based methods**

Context-based approaches take the whole image into account and use dimensionality reduction techniques to encode the image. The context information can be obtained from neighbouring areas of the objects (“local”) or by summarising image statistics from the image as a whole (“global”).

Contextual information approaches, such as Gist representations have become increasingly popular in the field of computer vision, since they provide rough global information, useful for many applications. The attractive features of this style of representation are that it is both memory-efficient and fast to extract. It does not contain many details about individual objects, and is not very discriminating, but it can provide sufficient information for coarse scene discrimination: e.g., indoor vs. outdoor. Moreover, such contextual information provides priors that help to disambiguate object recognition and increase the robustness of location estimation (Oliva and Torralba [2006]).

---

Oliva and Torralba [2001] proposed using the Gist descriptor to represent such spatial structures. This is built from the responses of steerable filters at different scales and orientations. Several models utilising different type of gist features of a scene have been presented.

Torralba et al. [2003] used wavelet image decomposition, each image location being represented by six orientations and four scales. To compute gist features, the resulting feature vectors were reduced from 384 dimensions to 80 dimensions using PCA. A Hidden Markov Model (HMM) was utilised to solve the localisation problem.

A similar system was described in (Siagian and Itti [2007]), where a simple context-based place recognition algorithm was proposed that combined biological centre-surround features from colour, intensity, orientation channels with visual attention situated within a segment. The gist features can only provide coarse context for localization, as they would have problems differentiating scenes when most of the background overlaps, so the saliency model was incorporated to increase the localisation resolution in this system.

The physical implementation of the model mentioned above was presented in (Siagian and Itti [2009]). A coarse localisation hypothesis was produced in the first instance by extracting the gist of a scene: then salient regions were used to refine it. The gist features and salient regions were then further processed using a Monte-Carlo localisation algorithm to allow the robot to generate its position.

Pronobis and Caputo [2007] proposed a recognition algorithm based on confidence estimation of place classification. Unlike the majority of algorithms designed to recognise pre-defined sets of environments (e.g., kitchen, corridor, etc.), this algorithm used a soft decision: that is, if the level of confidence of a single cue could not obtain a reliable decision, additional information, such as both global and local features, is to be used. A multi-dimensional statistical representation called Composed Receptive Field Histograms (CRFH) was used for the global representation, while the SIFT descriptor was exploited in order to obtain the local image representation. The classifier SVMs extended by SVM was used at the classification step, which well correlated with classification confidence.

---

Sunderhauf and Protzel [2011] presented a lightweight place recognition system based on the BRIEF-Gist descriptor. BRIEF-Gist is a simple scene descriptor based on the BRIEF descriptor introduced by Calonder et al. [2010], which encodes the whole image in a short bit string. The Hamming distance between two descriptors is used to find the single global best matching query image. BRIEF-Gist can be easily implemented, is computationally simple and does not require learning vocabulary. However, this system has a weakness shared with other appearance based place recognition systems, in that it is not robust to changes in vehicle orientation while traversing the same areas in different directions, when using the appearance of the whole scene to perform recognition.

Murillo and Kosecka [2009] demonstrated place recognition using the Gist descriptor on panoramic images in an urban environment. This descriptor is invariant with respect to traversal direction. Singh [2010] used the original Gabor-Gist descriptor in visual loop closure detection with panoramas.

Chang et al. [2010, 2011] and Siagian et al. [2014] utilised the Gist features and salient regions to solve the localisation problem in indoor and outdoor environments. Gist features that capture the dominant spatial structure of an image are used to coarsely localise the robot to within the general vicinity. Saliency is then employed to refine the location information, by recognising the more conspicuous areas in the image.

### **2.3.2 Strategies for dealing with challenging issues**

Place recognition is an open and highly challenging problem in computer vision, especially when applied to mobile robotics in changing environments. Place recognition is difficult for a number of reasons. First, finding an exact match for a previously visited place is not trivial for a robot: factors in play include potentially unreliable sensors, changes of viewpoint, and changes in the environment such as those caused by moving obstacles. Second, as the world is visually repetitive, the robot needs to be able to distinguish between different, but similar-looking places.

---

### 2.3.2.1 Dealing with changes

In order reliably to localise a mobile robot, even in dynamic environments, a variety of strategies have been proposed for resolving environment and viewpoint changes. One common solution involves strengthening the ability of the feature descriptor to cope with various changes. A body of sophisticated invariant features extracted from the images have been exploited for image matching, which include SIFT, SURF and GLOH (many more are presented in Section 2.2). Such features are represented by the vector computed from the image region localised at the interest points, which are robust to occlusion and invariant to image transformations such as scale, rotation, moderate illumination and viewpoint changes.

Some examples include the works of Castle et al. [2007]; Se et al. [2002]; Valgren and Lilienthal [2008], where SIFT or SURF feature detectors provide a rich description of the environment to match observed visual landmarks despite visual variability. Recently, the Affine-SIFT (ASIFT) algorithm was proposed by Morel and Yu [2009] to achieve full affine invariance by sampling various values for the latitude and the longitude angles in order to compute virtual views of the scene. The ASIFT algorithm was exploited to perform global localisation in (Majdik et al. [2013]), where images captured by a camera-equipped Micro Aerial Vehicle (MAV) need to be matched with images from Google Street View. In this work, the most challenging problem is severe viewpoint changes between air-level and ground-level images. The air-ground geometry of the system was used to generate virtual views of the scene, and a histogram voting scheme was applied to find the best image correspondences.

Nevertheless, feature-based methods could not successfully establish reliable correspondences if the images were captured from very different viewpoint and under the sharp illumination changes caused by direct sunlight and shadow in typical outdoor environments. Common types of features, such as corners and affine invariant regions are not fully invariant to these changes (Glover et al. [2010]; Milford [2013]). Glover et al. [2010] present an appearance-based SLAM system based on SURF feature descriptors, the system does not cope well with illumination changes over the course of a day, as the SURF features are too variable,

---

which results in the divergence of map estimate when no matches occur.

In some research works, new kinds of image descriptors are proposed, which depend on the type of captured images. Examples include a polar higher-order local auto-correlation (PHLAC) (Linåker and Ishikawa [2006]), created for the extraction of features from omnidirectional images, which is robust to noise and occlusion to some extent; Haar Invariant Features (Labbani-Igbida et al. [2011]), which is extracted by adapting Haar invariant integrals to the particular geometry and transformations of an omnidirectional camera; And a Feature Stability Histogram (FSH) (Bacca et al. [2011]), built using a voting scheme to tackle long-term SLAM in a changing environment, which stores information about the number of times each feature has been observed in each node of the topological map.

Omnidirectional images with a  $360^\circ$  field of view make it possible to create features that are invariant to the orientation of the robot. For example, various colour histogram representations were used to perform robot localisation in a series of papers (Blaer and Allen [2002]; Gonzalez-Barbosa and Lacroix [2002]; Ulrich and Nourbakhsh [2000]). The subspace of eigenvectors are computed from the original images (Artac et al. [2002]; Gaspar et al. [2000]; Kröse et al. [2000]). Fourier signatures are applied in (Ferdaus et al. [2008]; Menegatti et al. [2004b]) to represent the omnidirectional images captured for localisation.

Several publications (Möller et al. [2014]; Stürzl and Zeil [2007]) address illumination invariance through an holistic approach: that is, the entire image is utilised by resorting to pixel-by-pixel comparison techniques. These methods can be applied to low-resolution images, and do not require prior assumptions about the type of visual features to be extracted from the environment. However, preprocessing stages are required in which the images are transformed.

By way of example, we offer the work of Stürzl and Zeil [2007], in which the image differences are obtained by means of a descent in image distances (DID) model between image pairs. The preprocessing steps including subtracting the local mean, difference-of-Gaussian filtering and contrast normalization in order to make the distance measures invariant to illumination changes and shadow effects.

---

In (Möller et al. [2014]), invariance against illumination changes is accomplished by applying the pixel-wise distance measures proposed in three ways. Specifically, weak scaling invariance is obtained by finding the minimal Euclidean distance between two image columns, while strong scaling invariance is obtained by using normalised cross-correlation. Shift invariance is realised by either subtracting the mean before comparison of two image columns, or by computing the distance between edge-filtered vectors.

Maddern et al. [2014] developed the idea of an illumination-invariant colour space based on monochrome input to reduce the impacts of shadows in raw RGB images. Similar work can be found in (Alvarez-Mozos et al. [2008]; Corke et al. [2013], where a single-channel illumination-invariant imaging approach is also used to alleviate the effects of changes in illumination and shadows in the context of autonomous road vehicles.

#### **2.3.2.2 Disambiguating ambiguous cases**

In addition to the above-mentioned challenges for vision-based place recognition systems, image matching in scenes can be tricky if the environment contains few, or very similar features. Moreover, due to the limitations of the perceptual capabilities of the robot, a robot may fail to obtain enough information to distinguish reliable between two different locations that appear very similar. The problem is to overcome this perceptual aliasing, namely: the danger that the current image will match not only the corresponding location image, but also falsely match other reference images at different of other, similar locations.

Many feature-based place recognition methods may fail in environments where repeated patterns are common, as the invariant features are not sufficiently discriminating and there are many mismatches. This problem often trades off against the perceptual variability mentioned previously. Improving the robustness of the selected features to perceptual variability often leads to poor discrimination between places, and hence to perceptual aliasing. By contrast, trying to eliminate perceptual aliasing may result in increased susceptibility to perceptual variability.

---

Image-matching algorithms usually consist of two independent steps. The first involves finding a set of potentially matched pairs of interest points between two images: pruning of these matches is then performed by using geometric consistency, which keeps only correspondences consistent with epipolar constraints, or homography transformation. In the first step, some studies have made use of more suitable clustering to avoid false correspondences caused by perceptual aliasing. For example, the Fisher criterion was used in the work of Labbani-Igbida et al. [2011] to measure the separation between two classes of built signatures for robot localisation in indoor environments, providing a particularly wide separation ability for room classes.

In (Schaffalitzky and Zisserman [2003]), in order to overcome the problem of perceptual aliasing, the idea is to ignore common repetitive features. An ambiguity score is assigned to each feature, representing the number of features which match in the other image: then the ambiguity of a match is obtained by take into account the ambiguity scores of the features. The matching would be discarded if its ambiguity score is greater than six.

Other approaches fuse multiple sensors in order to have features with complementary information in the presence of adverse environments with perceptual aliasing. Zingaretti and Frontoni [2006] combined vision and sonar sensors to perform the localisation task in aliased environments. Gallegos and Rives [2010] took advantage of the metric information provided by a laser rangefinder and fused this with omnidirectional visual information. However, this technique does not take into account the problems of occlusions and illumination changes.

A wide range of place recognition systems addressed the perceptual aliasing problem using probabilistic algorithms covering Markov Localisation, Monte-Carlo Localization and Multi-Hypotheses Localization. That is the case in (Menegatti et al. [2003]) which exploited a Monte-Carlo Localisation approach to provide robust appearance-based localisation. Ranganathan and Dellaert [2007b] presented a similar model for probabilistic topological mapping based on Markov Chain Monte Carlo (MCMC) and Sequential Importance Sampling (SIS) algorithms, which incorporate previous location information (prior assumptions) into the recognition of locations to deal with perceptual aliasing. Likewise, Werner

---

et al. [2009] developed a sequential Monte Carlo SLAM technique to keep track of the belief of the position of the robot. This technique used Hausdorff distance to measure the consistency between the current view and the reference view.

Bacca et al. [2011] proposed a Bayesian filtering-based approach for robot localisation using a topological map: each topological location is assigned a probability value to restrain the degree of uncertainty. Qamar et al. [2013] addressed the perceptual aliasing problem for SLAM, employing a Fuzzy-Logic based method and a Fuzzified implementation of Scale Invariant Feature Transform (SIFT). Bellotto et al. [2008] developed a place recognition framework in which ambiguous information is solved by means of a multiple hypothesis tracking technique: the most plausible hypothesis is used for updating the location of the robot. Goedemé et al. [2007] applied Dempster-Shafer probabilistic theory to loop closing in order to avoid false links between different parts of a topological map in environments with self-similarities.

## 2.4 Visual odometry

Visual Odometry (VO) has been introduced and investigated in both the computer vision and robotics communities for some years. VO relies on the visual information from an image sequence to estimate odometry information. VO is not affected by wheel slip in uneven terrain, or other adverse conditions, and has the utmost importance in GPS-denied environments such as under water, indoors, or in the air. Methods have been proposed using both monocular cameras (Kriechbaumer et al. [2015]; Nistér and Stewénus [2006]; Tomasi and Shi [1993]) and stereo cameras (Maimone et al. [2007]; Matthies and Shafer [1987]; Moravec [1980]; Nistér and Stewénus [2006]; Olson et al. [2003]). Related work can be divided into two categories: feature-based, and appearance-based methods. Here, we review some of this work. More extensive surveys can be found in (Scaramuzza and Fraundorfer [2011]).

The earliest work on estimating the motion of a vehicle from visual imagery alone is (Moravec [1980]), where the basic algorithm identifies corner features in each

---

camera frame and estimates the depth of each feature using stereo pairs. Subsequently, potential matches are found by normalised cross correlation. Finally, motion is computed by estimating the rigid body transformation that best aligns the features at two consecutive robot positions. However, this kind of system suffers from poor accuracy and is unstable, partly because it relies on scalar models of measurement error in triangulation. Based upon this work, Matthies and Shafer [1987] used 3D Gaussian distributions to model triangulation error and incorporates the error covariance matrix of the triangulated features into the motion estimation between successive stereo pairs. The motion estimation in this work was pure translation, without considering orientation. The robot may navigate safely over short distances: however, over long distances the increasing orientation errors will lead to useless position estimation. This is extended in (Olson et al. [2003]) by incorporating an absolute orientation sensor such as a compass, a sun sensor or a panoramic camera providing periodic orientation updates, with the Förstner corner detector used as the feature detector. The results indicated that the error growth can be reduced to a linear function of the distance travelled, outperforming previous visual odometry results.

All the works reviewed above are feature-based methods. This kind of method tries to detect distinctive points or regions between consecutive image pairs. Although feature extraction can be fast, it often requires assumptions about the type of features being extracted, and natural environments can sometimes present no obvious visual landmarks, as in the case of desert or planar regions.

Some successful techniques using the whole appearance of the images have been proposed in the literature: e.g., (Bulow and Birk [2009]; Fernández et al. [2011]; García et al. [2012]; Goecke et al. [2007]; Labrosse [2006]; Milford and Wyeth [2008a]). A visual compass algorithm proposed by Labrosse [2006] provides an estimate of the heading of the robot from omnidirectional images in an incremental way. In (Goecke et al. [2007]) a Fourier-Mellin transform was applied to omnidirectional images in order to obtain a visual descriptor for the motion estimation of a vehicle. The motion of a vehicle was decomposed into a rotation and a translation component. The rotation angle estimate is taken as the median of the observed angular displacements using a mapping from camera coordinates to

---

the ground plane. In the same manner, the low frequency components of Fourier coefficients are used. Bulow and Birk [2009] proposed an improved Fourier Mellin Invariant (iFMI) descriptor, and applied this descriptor to an Unmanned Aerial Vehicle (UAV) for visual odometry to generate photo maps.

In (Fernández et al. [2011]), a single Fourier descriptor was used to represent each panoramic image obtained. When the Fourier signature has been captured in two nearby points, the relative orientation of two points will be computed using the shift theorem. Another example is (Milford and Wyeth [2008a]), in which the colour images captured from a perspective camera are first converted to greyscale images, then each pixel column is summed and normalised to form a one-dimensional array. The resulting arrays are used to extract the rotation information.

Recently, both appearance-based and feature-based methods were presented in (García et al. [2012]) to compute the motion transformation between two consecutive images incrementally. The phase information of the Fourier signature was used to compute the robot orientation, and SURF features were used to detect the interest points for image comparisons by looking for corresponding points. Kriechbaumer et al. [2015] evaluated the appearance-based and feature-based stereo visual odometry algorithms for localization of an autonomous watercraft. The feature-based technique was shown to provide accurate localization in the short term, but poor performance on the estimations of pitch and roll angles.

## 2.5 Loop closure

Appearance-based SLAM is primarily used for detecting loop closures in large-scale unknown environments, which requires determining if the current robot view matches any previously visited places, or if it should be classified as a new place. A great many techniques have been proposed to address this problem. This section reviews the state-of-the-art algorithms using appearance-only information to detect loop closure, focusing on the advances in approaches based on similarity matrices in the context of the topological paradigm, which are of greatest interest

---

in the context of this thesis.

Levin and Szeliski [2004] presented a multi-stage similarity function to address the localisation and loop closure problems. In the first stage, global colour histograms are used to obtain a first similarity score. After filtering out the worst matches, the remaining good matches are employed to compute a 3D rotation based on the first order moments of a spherical image being invariant under 3D rotation. Subsequently, Harris corners are extracted, and epipolar geometry is recovered between the remaining candidate images in a RANSAC framework. Finally, similarities between all pairs of images in the database are stored in a distance matrix (“correspondence map”). The main diagonal of the distance matrix represents the self-correspondence and correspondences between temporally neighbouring frames. An off-diagonal spot show a correspondence between two frames that are far apart. A loop closure appears as a connected sequence of off-diagonal spots in the matrix. In a similar vein, Silpa-Anan and Hartley [2005] used SIFT features combined with Harris corners to generate a visual correspondence map: this is then used for localisation and loop closure detection.

In (Valgren et al. [2006]), local features are extracted from panoramic images obtained in sequence and used to cluster the images into nodes, and then to detect loops. This technique avoids exhaustively computing the similarity matrix by using a random search guided by heuristics. In (Valgren et al. [2007]), loops are detected by exhaustive search, though the incremental spectral clustering method employed can reduce the search space when new images are processed, which implies less computation time when the similarity measure is costly to compute. In (Goedemé et al. [2007]), an invariant column segments technique, combined with rotation-reduced and colour-enhanced SIFT features, has been used to extract local regions of each image and build place representations. This is followed by agglomerative clustering of images into distinct places. Loop closures are detected using Dempster-Shäfer probabilities.

FAB-MAP (Cummins and Newman [2008a]) applied a Chow-Liu dependency tree and recursive Bayes estimation within a rigid probabilistic framework to provide loop closure information for the topological mapping system. Similarly, Angeli et al. [2009] used BayesianLCD to provide loop closure candidates for the

---

topological SLAM system.

In (Anati and Daniilidis [2009]), the author described a novel similarity measure for comparing two panoramic images. The rotational invariance with respect to changes in heading is achieved by alignment of local features projected on the horizontal plane using a dynamic programming approach. A Markov Random Field (MRF) and image similarity matrix were used to model the the probability of loop closures.

Another similar system was presented by Scaramuzza et al. [2010], in which visual loop closure detection and closing were attempted through SIFT features matching between the current image and the images in the database. The similarities between all images was calculated, and loop hypotheses generated by the five top ranked images, which will be improved by imposing geometrical verification. Finally, the loop closing optimisation will be invoked if one hypothesis passes this verification.

In order to remove the effect of repetitive structures of the environment and visually ambiguous scenes, Ho and Newman [2007] exploited a singular value decomposition of the similarity matrix. In addition, they examined an extreme value distribution to ensure the detected sequence does genuinely indicate a loop closure and to minimise false positives. A similar method was also observed in (Koch et al. [2010]) to identify loop closure sequences.

Williams et al. [2009] classified loop closures into three categories: (i) map-to-map matching methods that mainly consider geometry; (ii) image-to-image matching methods that consider only appearance; and (iii) image-to-map matching methods that use visual and metric information to perform relocation.

Three representative approaches (Clemente et al. [2007]; Cummins and Newman [2008a]; Williams et al. [2008]) selected from each category were used to compare the loop closure performance of monocular SLAM systems. Each one has its benefits and downsides: tunable parameters also affect the ultimate performance. The comparison results show that the map-to-map method cannot reliably detect loop closures when sparse maps giving inadequate information are used. The image-to-image method performs well, and could work better with extra metric

---

information. However, the image-to-map method combines appearance and geometry information and achieves the best results.

Labbe and Michaud [2013] presented an online loop closure detection algorithm for large-scale and long-term SLAM, called Real-Time Appearance-Based Mapping (RTAB-Map). This work was inspired by the work of Angeli et al. [2008a] and based on memory management mechanisms. This method caches the most recent and frequently observed locations in the main memory called working memory (WM) for loop closure detection. The rest are stored in an external memory called long-term memory (LTM).

Recently, compact global image descriptors have been popular in loop closure research, including Gabor-Gist, BRIEF-Gist, and WI-SURF ( Badino et al. [2012]; Liu and Zhang [2012]; Sunderhauf and Protzel [2011]; Wu et al. [2014]). These descriptors avoid the need to extract the keypoints, and enable rapid comparison of images.

Sunderhauf and Protzel [2011] developed a method based on the BRIEF-Gist descriptor to create a representation of the environment to solve the loop closure problem. However, loop closure cannot be detected if the images are taken at the identical place but from different points of view. Liu and Zhang [2012] applied the Gabor-Gist descriptor to detect loop closure in a Bayesian filtering scheme. A PCA projection is performed to compress the dimensionality of the descriptor in order to improve the computational efficiency. Wu et al. [2014] presented a loop closure detection framework in which a simple binary descriptor was obtained by thresholding the down-sampled images, using Otsu’s method. The similarity between the descriptors was measured by Mutual Information (MI). Arroyo et al. [2014] evaluated the performance of several global descriptors extracted from panoramic images for loop closure detection tasks. The descriptor based on LDB ( Yang and Cheng [2014a]) achieved the best performance among all the compared descriptors.

---

## 2.6 Quadtree structure

A quadtree (Samet [1984]) is a hierarchical data structure used for modeling two-dimensional objects, adapted from the binary search tree, but processing four branches at each node rather than two. The initial application of the quadtree is in image processing, with the aim of saving space and accelerating various spatial operations. This technique involves recursively dividing an image into four equally-sized quadrants, until all the pixels of each quadrant are homogeneous in colour. Quadtree has long been used for image compression (Burt and Adelson [1983]), classification and segmentation (Willsky [2002]), spatial indexing and collision detection (Jones et al. [2004]).

In the area of mobile robotics, quadtree has been frequently utilised for occupancy grid map representation and the task of path planning in order to improve location and control the trajectory of the robot. Notable examples of such application include Burgard et al. [2007]; Guivant et al. [2004]; Pirker [2010]; Shojaeipour et al. [2010]; Sujan et al. [2006]; Thorpe et al. [2005]. Moreover, many approaches have been developed for scene classification and visual localization based on the quadtree decomposition method. An early research work introduced by Kreucher and Lakshmanan [1999] addressed the problem of lane markers recognition under varying lighting and environmental conditions. A region of the scene image containing the edge-like feature is repeatedly subdivided into subquadrants until each pixel in the image has been interrogated as to whether it lies on an edge.

A scene classification method was introduced in Lazebnik et al. [2006], in which a multilayer quadtree decomposition scheme was exploited in order to obtain the spatial position information of a scene image. Firstly, a scene image was subdivided in a quadtree-like manner: then the histograms of visual words about each subimage being computed. Finally, all the histogram of visual words of all subimages at different levels were concatenated and used for representing scene images.

Going further, higher level visual recognition problems were addressed in the work of Li et al. [2010], where a three-level quadtree representation based on objects was used for scene classification tasks. Eze and Benosman [2007] proposed a visual

---

localization method for mobile robot navigation. In this method, the optimal patches of the image were generated by quadtree decomposition, from which the features could be extracted for image matching. Initially, the initial panoramic image was cut into four equal quadrants. The further division of each quadrant was determined by the quantity and homogeneity of the information present in it, such that the difference of the quantity of information between possible sub-patches is minimized.

Mei et al. [2009] developed a stereo vSLAM system in which FAST corners were detected in each frame for motion estimates. In order to achieve good tracking accuracy, these extracted FAST features should be spread throughout the whole image. To achieve this, the quadtree structure was employed to restrict the number of features in each quadrant for matching between images. The same theme of applying the quadtree subdivision technique to monocular SLAM was also proposed by Strasdat et al. [2010a].

More recently, Saudabayev et al. [2015] reported on a novel terrain classification framework utilizing an on-board time-of-flight depth sensor. A filtered depth image was recursively divided first into four equal subimages and so on, the maximum level of decomposition being four and five. The statistical data of each subimage at different levels, including minimum, maximum, mean, and standard deviation values were extracted and stored in a vector, which was then used for the terrain classification task.

## 2.7 Conclusions

This chapter outlines the development of approaches to SLAM problems using cameras as the primary method of generating observations. Despite the achievements of recent decades, there are still challenges to be faced for vSLAM systems. Although a camera provides rich information about a scene, it is vulnerable to the effects of variations in lighting, perspective changes or partial occlusion by moving objects. As noted, many researchers have examined the issue of how best to represent and match images in real world environments in order to overcome

---

the challenges mentioned above. The effectiveness and reliability of a vSLAM system depends on many characteristics, such as how the observed environment is represented, how likely it is that the system will recognise places previously visited, and how uncertainty is handled, with regard to the type of sensors used and the intended application of the robot. Table 2.7 collects relevant information about some vSLAM systems reviewed in this thesis, providing a quick reference to the key techniques in these frameworks. To generate the summaries, we focused on the aspects of the type of camera used, the type of the environment representation, the task required of the system, the type of environment used to test the performance of the system, and the details of image descriptors and detectors.

As indicated by many research works in the literature, omnidirectional (catadioptric) cameras are desirable sensors for real-time recognition of places for mobile robotics. They use lenses and mirrors to view a large area of their surroundings. The 360° view allows visual information from all sides of the robot to be acquired simultaneously. This decreases the number of images necessary to represent the environment, reduces perceptual aliasing, provides rotational invariance to the field of view, improves robustness to occlusions and matching, and hence enhances the accuracy and efficiency of place recognition. For these reasons, we choose an omnidirectional camera as the visual sensor in our research.

Global descriptors and/or local descriptors have been used to represent the environment in many popular frameworks. As is evident from existing methods, these descriptors each have their own advantages and shortcomings. Global methods compare images using all the pixels of the entire image. Although they are efficient and compact, they cannot handle severe viewpoint changes or occlusions. On the other hand, the use of local descriptors can be robust to these adverse effects. Nevertheless, these methods require pre-defined routines for feature extraction and lack of spatial information. Moreover, it is difficult to extract features robustly and correctly when the environment is cluttered or featureless. This motivated us to propose a novel image comparison method in which we consider the whole image as global visual feature and exploit the quadtree decomposition technique to capture the spatial information in an image.

---

As stated in many studies in the literature, the orientation angle of the robot directly affects the action model: it is crucial to keeping the robot moving along the expected path, driving towards the proper target destination, and maintaining vehicle safety. Therefore, in order to allow a robot to operate robustly for long periods of time, the orientation of a mobile robot must be determined properly. Accordingly, we attempt to evaluate various image-based techniques for accurate orientation estimation. Moreover, the question of how to select the frames to establish the correct relative orientation is a very important step in most VC algorithms, and is worth investigating.

On the other hand, an incorrect loop closure can be disastrous for most real-time SLAM systems, making an inconsistent map of the environment. Despite recent advances in visual loop closure research, challenges remain to improve tolerance of changes in the environment and perceptual aliasing. Therefore, it remains a worthwhile task to develop the effective and robust methods for loop closure detection.

Author	Camera	Map	Tasks	Environment	Descriptor (Detector)
Montemerlo et al. [2002]	Mono	Metric	SLAM	Outdoor	Image patches
Davison [2003]	Mono	Metric	SLAM	Indoor	Image patches (shi and Tomasi operator)
Gross et al. [2003]	OmniDir	Metric	Loc	Indoor	Image patches
Hayet et al. [2003]	Mono	Metric	Loc	Indoor	Image patches
Kořecká et al. [2003]	Mono	Topo	Loc	Indoor	Gradient orientation histogram
Menegatti et al. [2004a,b]	OmniDir	Hybrid	Loc	Indoor	Fourier Components
Nistr et al. [2004]	Stereo or Mono	Metric	VO <sup>1</sup>	Outdoor	Image patches (Harris)
Hashem and Andreas [2004]	Mono	Metric	Loc	Indoor	Kernal PCA (Edge)
Milford et al. [2004]	Mono	Metric	SLAM	Indoor	Image patches (Edge)
Bradley et al. [2005]	Mono	Topo	Loc	Outdoor	Weighted Gradient Orientation Histograms
Se et al. [2005]	Stereo	Metric	Map+Loc	Indoor	SIFT
Wang et al. [2005]	Mono	Top	Map+Loc	Indoor	SIFT (Harris)
Wolf et al. [2005]	Stereo	Metric	Loc	Indoor	Image patches
Sim et al. [2005]	Stereo	Metric	SLAM	Indoor	SIFT
Pronobis et al. [2006]	Mono	NA	PR <sup>2</sup>	Indoor	High dimensional composed receptive field histograms
Jensfelt et al. [2006]	Mono	Metric	SLAM	Indoor	SIFT (Harris)
Valgren et al. [2006]	OmniDir	Topo	Mapping	Indoor	SIFT
Eade and Drummond [2006]	Mono	Metric	SLAM	Indoors	Image patches

*Continued on next page*

Table 2.1 – *Continued from previous page*

Author	Camera	Map	Tasks	Environment	Descriptor (Detector)
Goedemé et al. [2007]	Omnicdir	Topo	Map+Loc	Indoors	SIFT+ Invariant column segment
Booij et al. [2007]	Omnicdir	Hybrid	Map+Loc	Indoors	SIFT
Filliat [2007]	Mono	Topo	Map+Loc	Indoors	SIFT
Ho and Newman [2007]	Mono	Metric	LC <sup>3</sup>	Outdoors	SIFT
Maimone et al. [2007]	Stereo	NA	VO	Outdoor	(Forstner or Harris)
Pronobis and Caputo [2007]	Mono	Topo	PR	Indoors	SIFT (Harris)+ CRFH
Vasudevan et al. [2007]	Mono	Topo	CM <sup>4</sup>	Indoors	SIFT
Weiss et al. [2007a]	Stereo	Topo	Loc	Outdoors	Weighted Grid Integral Invariant
Valgren and Lilienthal [2008]	Omnicdir	Topo	Mapping	In+Out	SIFT
Angeli et al. [2008b]	Mono	Topo	SLAM	In+Out	SIFT+Colour Histogram
Bellotto et al. [2008]	Mono	Topo	Loc	Indoors	Image patches
Cummins and Newman [2008a]	Mono	Topo	SLAM	Outdoors	SURF
Eade and Drummond [2008]	Mono	Topo	LC	Outdoors	SIFT
Milford and Wyeth [2008a]	Mono	Topo	SLAM	Outdoors	Image patches
Scaramuzza and Siegwart [2008]	Omnicdir	Metric	VO	Outdoors	Image patches
Takacs et al. [2008]	Mobile Phone	NA	AR <sup>5</sup>	Outdoors	SURF
Werner et al. [2008]	FPGA	Topo	SLAM	Indoors	Colour Histogram
Pinies and Tardos [2008]	Mono	Metric	Mapping	Outdoors	Image patches (Harris)
Bulow and Birk [2009]	Mono	NA	VO	Outdoors	Fourier-Mellin Invariant (FMI)

*Continued on next page*

Table 2.1 – *Continued from previous page*

<b>Author</b>	<b>Camera</b>	<b>Map</b>	<b>Tasks</b>	<b>Environment</b>	<b>Descriptor (Detector)</b>
Murillo and Kosecka [2009]	Omnicam	Topo	PR	Outdoors	Gist
Siagian and Itti [2009]	Mono	Topo	Map+Loc	Outdoors	Gist+SIFT
Mei et al. [2009]	Stereo	Hybrid	SLAM	Outdoors	SIFT
Angeli et al. [2009]	Mono	Hybrid	SLAM	In+Out	SIFT+Color Histogram
Cadena et al. [2010]	Stereo	Topo	PR	In+Out	SURF
Chang et al. [2010]	Mono	Topo	Loc	In+Out	Gist+Saliency
Comport et al. [2010]	Stereo	NA	VO		Image patches
Cummins and Newman [2010]	Omnicam	Topo	SLAM	Outdoors	SURF
Koch et al. [2010]	Omnicam	Topo	Map+Loc	Indoors	Image patches
Scaramuzza et al. [2010]	Omnicam	Metric	PR	Outdoors	SIFT
Singh [2010]	Mono	Topo	LC	Outdoors	Gist
Strasdat et al. [2010a,b]	Mono	Metric	SLAM	Outdoors	SURF (FAST)
Konolige et al. [2010]	Stereo	Hybrid	SLAM	In+Out	SAD (STAR+FAST)
Mei et al. [2010]	Stereo	Hybrid	SLAM	Outdoors	SIFT
Glover et al. [2010]	Mono	Hybrid	SLAM	Outdoors	SURF
Botterill et al. [2011]	Mono	Topo	SLAM	In+Out	Image patches (FAST)
Kaess and Dellaert [2010]	Multiple cameras	Metric	SLAM	Indoors	Image patches
Konolige et al. [2010]	Stereo	Hybrid	SLAM	In+Out	FAST+SAD
Mariottini and Roumeliotis [2011]	Mono	Metric	Loc	In+Out	Image patches
Kawewong et al. [2011]	Omnicam	Topo	SLAM	In+Out	PIRF (SIFT)
Maddern et al. [2011]	Omnicam	Hybrid	SLAM	Outdoors	SURF

*Continued on next page*

Table 2.1 – *Continued from previous page*

Author	Camera	Map	Tasks	Environment	Descriptor (Detector)
Cadena et al. [2012]	Stereo	Topo	SLAM	In+Out	SURF
García et al. [2012]	Stereo	NA	VO	Indoors	SURF
Ramos et al. [2012]	Mono	Topo	PR	In+Out	Image patches
Maddern et al. [2012]	Omnidir	Hybrid	SLAM	Outdoors	SURF
Fazi-Ersi and Tsotsos [2012]	Mono	Topo	PR+PC <sup>6</sup>	Indoors	Histogram of Oriented Uniform Patterns (HOUP)
Milford [2013]	Mono	Topo	SLAM	Outdoors	Image patches
Rostami et al. [2013]	Mono	Topo	PR	Outdoors	Salient Line Segments (SLS)
Labbe and Michaud [2013]	Webcam	Metric	LC	In+Out	SURF
Wang and Yagi [2013]	Mono	Topo	Loc	In+Out	Orientation Adjacency Coherence Histogram (OACH)+SIFT
Lin et al. [2013]	Mono or Omnidir	Topo	PR	In+Out	Extended-HCT
Magnabosco and Breckon [2013]	Mono (Cross-spectral)	Metric	SLAM	Outdoors	SURF
Siagian et al. [2014]	Mono	Hybrid	Map+Loc	In+Out	SIFT+Saliency

<sup>1</sup>VO: Visual Odometry<sup>2</sup>PR: Place Recognition<sup>3</sup>LC: Loop Closure<sup>4</sup>CM: Cognitive Mapping<sup>5</sup>AR: Augmented Reality<sup>6</sup>PC: Place Classification

# Chapter 3

## Datasets

### 3.1 Introduction

This chapter provides a description of all the datasets that have been used to evaluate the proposed algorithms in this thesis, including the environments, the robot platforms, and the cameras employed during acquisition.

We make use of four datasets: one is an open-access indoor environmental dataset collected by Ullah et al. [2007], which is named COLD; the second is an openly available outdoor environmental dataset, New College 1 dataset released by Smith et al. [2009]; the two remaining were acquired by ourselves in indoors and outdoors environments; these were named ISL, and GummyBear, respectively. The New College 1 and GummyBear datasets have been utilised for validation of the proposed image comparison method. All of the datasets except the New College 1 dataset have been used to evaluate the various image-based techniques for the robot orientation estimation task, and the indoor dataset (ISL) and New College 1 dataset have also been used for evaluation of the loop closure detection methods.

---

Table 3.1: Characteristics of ISL datasets

Dataset	Frames	Length	Rate	Notes
ISL 1	679	40m	30fps	Static, no objects within the workspace
ISL 2	766	40m	30fps	Static, objects in the middle of the workspace
ISL 3	780	40m	30fps	Static objects in the middle of the workspace, a moving object present
ISL 4	752	40m	30fps	Static objects in the middle of the workspace, two moving objects present

---

## 3.2 Indoor datasets: ISL

This dataset has been captured in our laboratory. It consists of four sub-datasets captured from four different scenarios containing repetitive structures, people wandering around and moved objects. These four datasets feature significant numbers of repeated loop closures in both static and dynamic environments. The sequences contain 679, 766, 780 and 752 images, respectively. For ease of precessing, every omnidirectional image with a size of  $200 \times 200$  pixels was unwrapped into a panoramic view with a size of  $360 \times 40$  pixels. The unwrapping is performed by scanning the pixels along the radial lines with one degree increment, and eliminating the pixels that do not correspond to the environment, such that each panoramic image has a horizontal angular resolution of 1 pixel per degree. The characteristics of the datasets are described in Table 3.1. This dataset has been used for evaluation of various orientation estimation algorithms and the proposed loop closure detection algorithm. The evaluation results will be presented in Chapter 5 and 6.

### 3.2.1 Acquisition platforms and procedure

A Pioneer robot was instructed to drive along roughly rectangular closed loops from one end of the experimental area to the other and then back to the starting position. Each sub-dataset collection consists of a journey around the laboratory consisting of three loops. A catadioptric system consisting of a digital colour

---

camera pointed upwards looking at a hyperbolic mirror (see Fig 3.1 (a)) was used to capture image sequences of  $200 \times 200$  pixels resolution at 30fps. Note that the location and the appearance of the local scene were synchronously captured as the robot moved through its workspace.

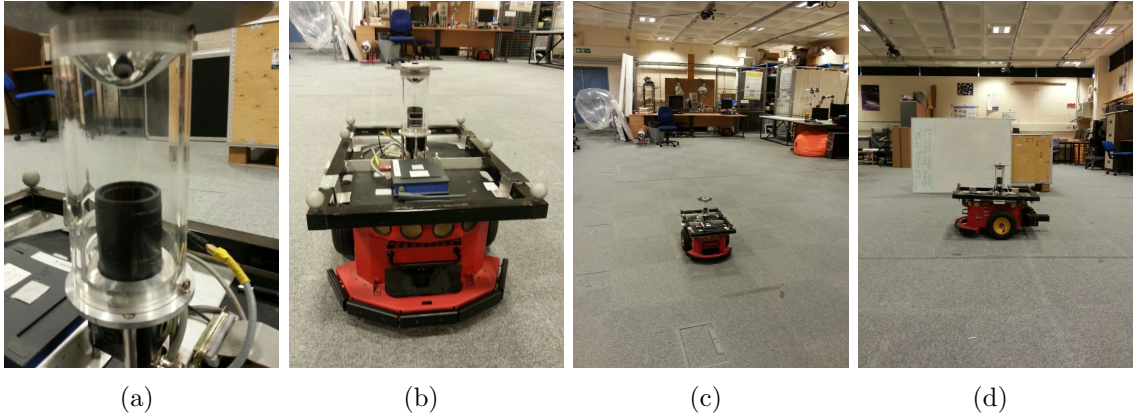


Figure 3.1: (a) Catadioptric camera (b) Pioneer robot (c) Experimental environment without obstacles and (d) Experimental environment with “wall” sitting in the middle of workspace.

### 3.2.2 Ground truth

The ground truth information was captured by a VICON motion tracking system, which provides the position ( $x, y$  and  $z$ ) and orientation ( $yaw, pitch$  and  $roll$ ) of the robot at 30Hz with an accuracy on the order of millimetres. Seven cameras, outfitted with infrared (IR) optical filters and an array of IR LEDs, were mounted on the ceiling. Six IR reflective markers were attached asymmetrically and rigidly to the robot (see Fig 3.1 (b)). The cameras emit infrared light that is reflected by the markers attached to the robot. The VICON software constructs a three-dimensional representation of the markers using the images taken from the seven cameras and triangulation with the known camera positions, from which it then derives the pose of the robot. A detailed description of this system may be found in (<http://users.aber.ac.uk/hoh/CS390/512ViconSWManual.pdf>).

---

### 3.2.3 The environments and examples

The dataset collection area is an approximately  $4m \times 5m$  indoor environment. Figure 3.2 depicts the trajectories of the robot in four various scenarios, in which the robot was driven around three closed loops following almost the same path, where starting points correspond to red points, and green arrows indicate the driving direction.

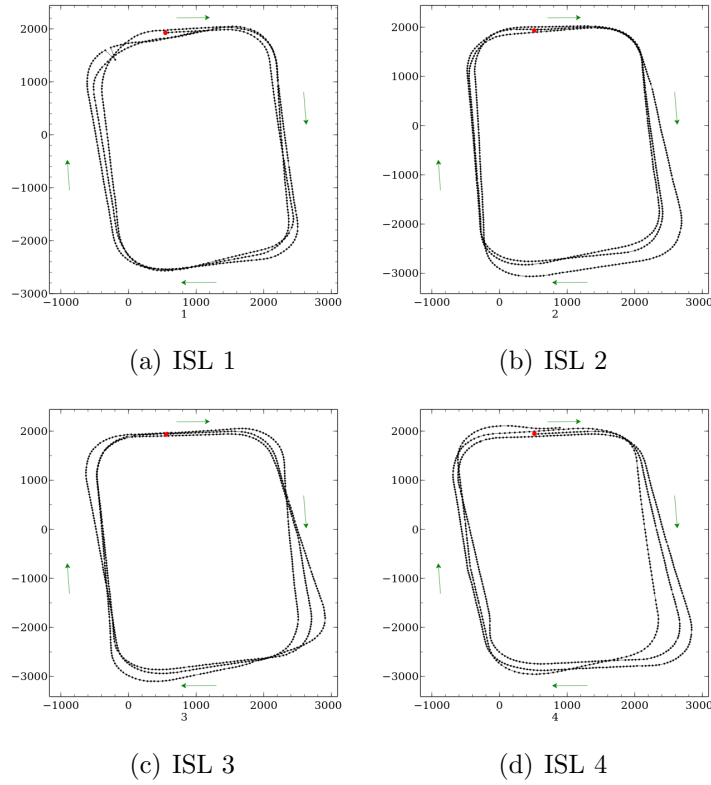


Figure 3.2: VICON-recorded robot trajectories in the  $xy$ -plane in four different scenarios.

#### Scenario 1:

In this scenario, as is common in indoor office environments, there were many duplicated objects (e.g., tables, chairs, monitors, etc.) around the experimental area, but no objects within the workspace. There were no obstacles present in

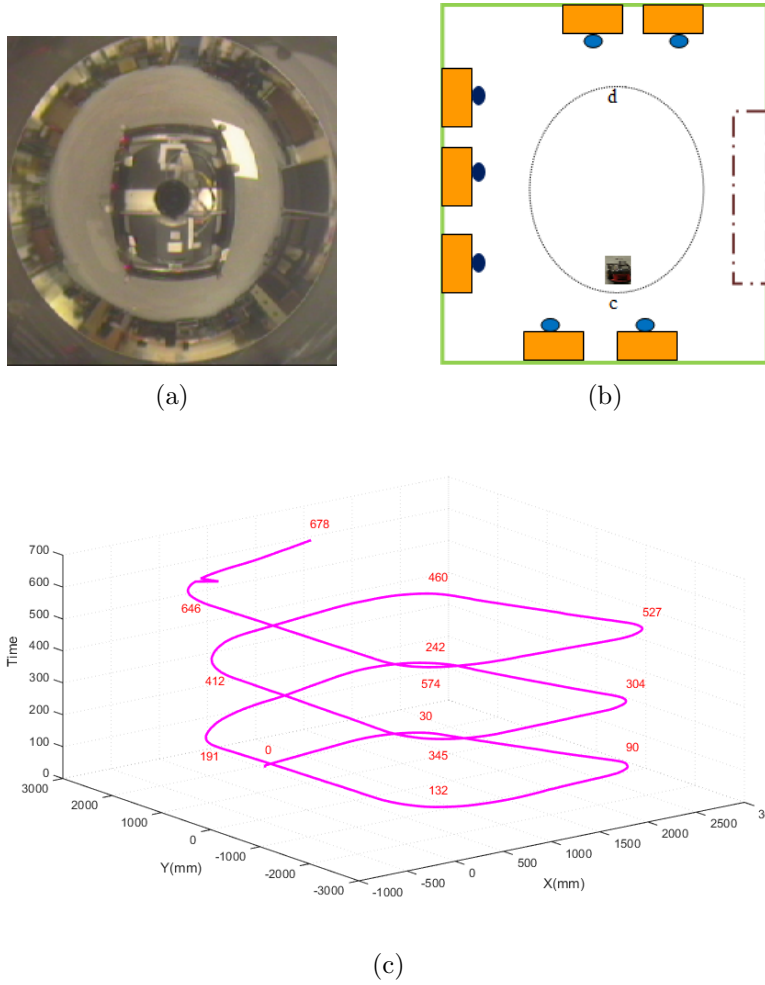


Figure 3.3: ISL dataset 1: 3.3(a) is a typical image from ISL dataset 1; 3.3(b) is a schematic of the whole environment; and 3.3(c) is the trajectory followed by the robot, with some annotated points in  $x, y, t$  space.

---

the workspace during the data capture. Therefore, this dataset is representative of a static and self-similar environment. Figure 3.2(a) shows the 2D trajectories followed by the robot. Figure 3.3 shows an example image from this scenario, the experimental environment, and the trajectory followed by the robot, with some annotated points in  $x, y, t$  space. It is important to note that there is a sudden jump around point 646 in the trajectory (See Figure 3.3(c)). This might be due to the fact that some reflective markers were not correctly identified when the robot was driven near the border of the capture space, which produced an incorrect tracking result. We have manually corrected the trajectory to avoid bias for the experimental evaluations.

### **Scenario 2:**

In the second dataset, a wooden box and a white board are introduced, standing side by side, which forms a “wall” in the middle of the workspace. Due to the existence of the “wall”, and the height of the wall above the vertical field of view of the robot during the experiment, from the perspective of the robot within the experimental area the wall creates two different places, one on each of its sides.

Figure 3.2(b) shows the 2D trajectories followed by the robot. Figure 3.4 shows a representative image in this scenario, the experimental environment, and the trajectory followed by the robot, with some annotated points in  $x, y, t$  space.

### **Scenario 3:**

In a more realistic scenario, a robot has to be able to deal with environmental changes after a long term traverse: for example, a object can move, change its shape and size, or even disappear. Due to the low placement of the camera on our robot, the projection of moving people in the image is small. Therefore, in order to produce obvious image variability in this dataset, we specially designed a scene with changes that involve the appearance and disappearance of a prominent object. In this case, a bean bag began to appear when the robot was close to completing its first loop, and then disappears from the field of view of the robot.

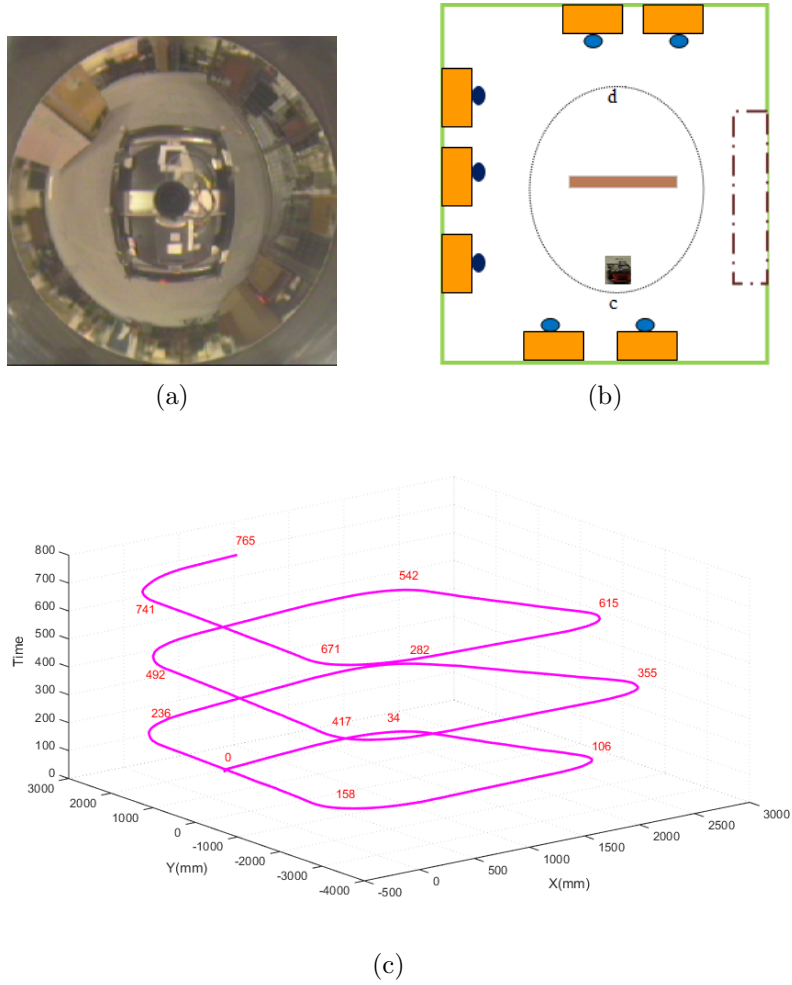


Figure 3.4: ISL dataset 2: 3.4(a) is a typical image from dataset 2; 3.4(b) is a schematic of the whole environment; and 3.4(b) is the trajectory followed by the robot, with some annotated points in  $x, y, t$  space.

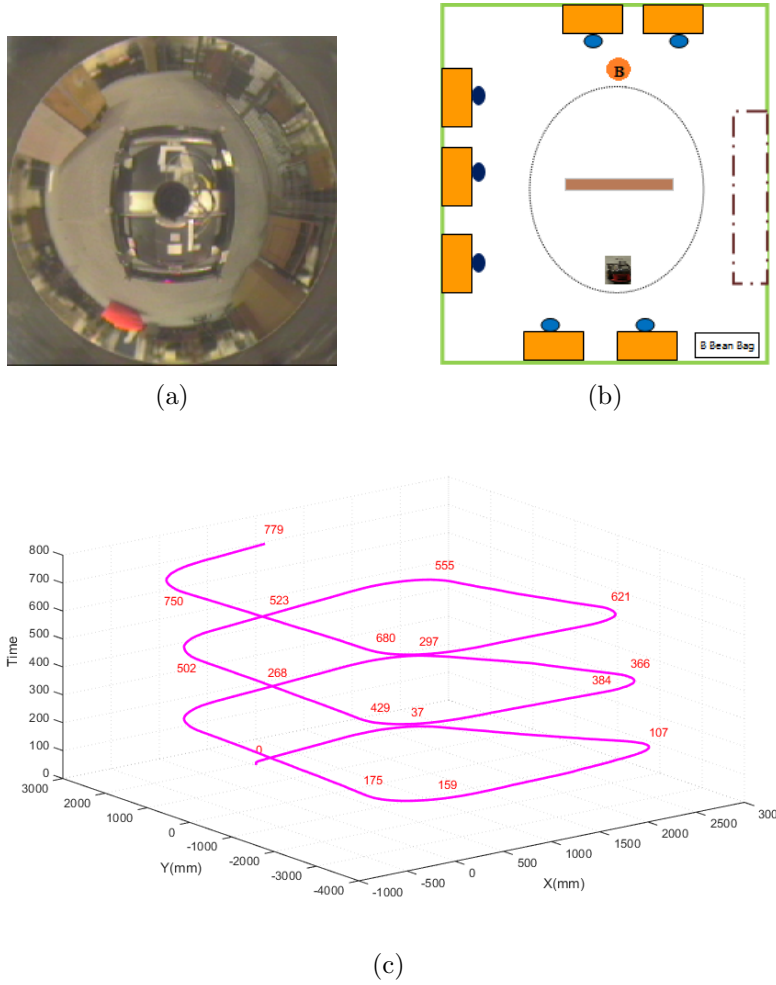


Figure 3.5: ISL dataset 3: 3.5(a) is a typical image from dataset 3; 3.5(b) is a schematic of the whole environment; and 3.5(c) is the trajectory followed by the robot, with some annotated points in  $x, y, t$  space.

---

Figure 3.2(c) shows the 2D trajectories followed by the robot. Figure 3.5 shows a typical image from this scenario, the experimental environment, and the trajectory followed by the robot, with some annotated points in  $x, y, t$  space. A bean bag was placed near point 268 (see Figure. 3.5(c)) when the robot was travelling toward the end of the first loop, and then was removed as the robot travelled toward the end of the second loop (point 523). Note that points 268, 523 and 779 are almost the same positions, but on different loops (first, second and third loops, respectively). Specifically, the bean bag is in sight from frame 159 to 384 in the sequence of this dataset: the robot is closest to the bean bag at point 266 (frame 266), at which point the robot is approximately 0.5 metres from the bean bag.

#### **Scenario 4:**

The fourth dataset is characterised by larger environmental changes. A great variability in appearance was introduced by a person crouched down beside the robot, in addition to a bean bag being added during the second lap data collection process. Note that points 266, 528 and 745 are almost the same positions, but on different loops (first, second and third loops, respectively). The objects are in sight from frame 156 to 393 in the sequence of this dataset, the robot is closest to the objects at point 268 (frame 268), where there is approximately 0.5 metres between the robot and the objects. In a robot configuration in which only cameras are available, identifying loop closure in this scenario can be very challenging. Accordingly, this scenario was considered suitable for evaluating the robustness of the algorithms against dynamic changes. Figure 3.2(d) shows the 2D trajectories followed by the robot. Figure 3.6 shows a typical image grabbed in this scenario, the experimental environment, and the the trajectory followed by the robot, with some annotated points in  $x, y, t$  space.

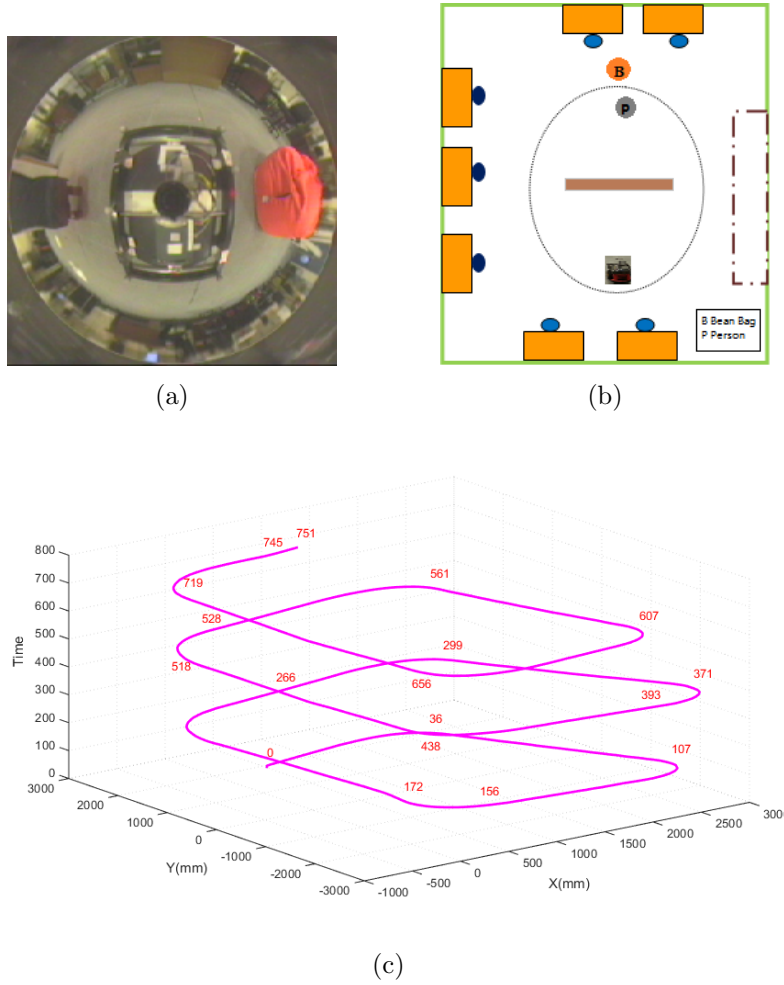


Figure 3.6: ISL dataset 4: 3.6(a) is a typical image from dataset 4; 3.6(b) is a schematic of the whole environment; and 3.6(c) is the trajectory followed by the robot, with some annotated points in  $x, y, t$  space.

---

### 3.3 Indoor datasets: COLD

The COLD database is a publicly available dataset (Ullah et al. [2007]). The name COLD is an acronym, which stands for COsy (Cognitive systems for Cognitive Assistants) Localization Database. The database consists of three separate datasets acquired in three different indoor environments across Europe (Saarbruecke, Freiburg and Ljubljana). Perspective and omnidirectional image sequences were recorded using three different mobile robot platforms. Laser range scans and odometry data were also collected for most of the sequences. The acquisition process is repeated under a variety of weather and illumination conditions (sunny, cloudy and night) and across a time span of two to three days. Dynamic elements, such as people wandering around, and missing or newly added objects, were introduced into the scenes.

The COLD database has already been used in the literature (Campos et al. [2012]; Liu and Siegwart [2014]; Wang and Lin [2011]) for evaluating the robustness of vision-based place recognition systems against different kinds of variations (introduced by illumination variations and human activity). In our work, the Freiburg sub-dataset (omnidirectional sequence A) is used to validate our proposed method. The mobile robot Pioneer-3 was used as a robot platform with an omnidirectional camera mounted about 91cm above the ground plane. The dataset was collected at the rate of 5 frames per second while the robot navigated through five different functional areas; a printer area, a corridor, two-person office, a stairs area, and a bathroom. The resolution of an omnidirectional image is  $640 \times 480$  pixels, which we unwrapped to  $360 \times 40$  pixels in order to enable fair comparison of the experimental results with our other datasets. Ground truth for position ( $x, y$  coordinates) and orientation of the robot was acquired using an odometry sensor.

Figure. 3.7 shows typical images from COLD database under different weather and illumination conditions. More detailed information about the COLD database may be found online (<http://www.cas.kth.se/COLD/>).

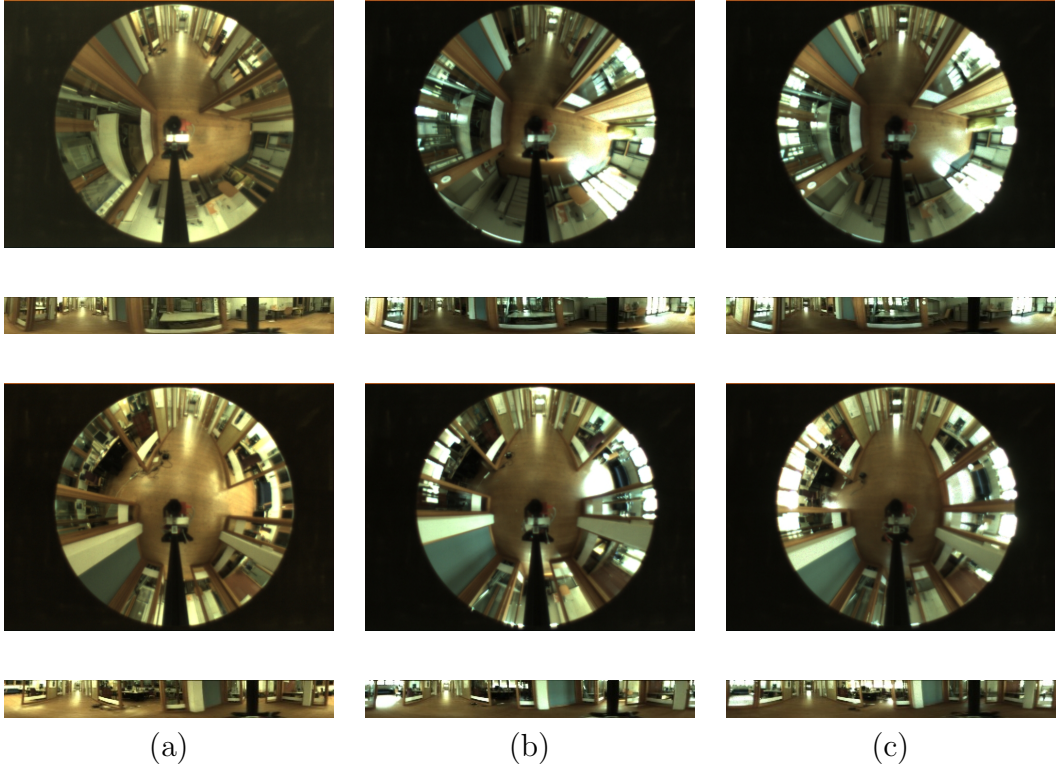


Figure 3.7: Example images of the COLD datasets in three different lighting conditions: (a) night; (b) cloudy; and (c) sunny. The omnidirectional images are shown in the first and third rows, the corresponding unwrapped images are shown in the second and fourth rows, respectively.

### 3.4 Outdoor datasets: GummyBear

This dataset contains three sub-datasets acquired in three different outdoor dynamic environments: FIELD, CARPARK, and TENERIFE. Each sub-dataset consists of a sequence of images acquired along a “Gummy Bear” path (see Figure 3.9) by our four-wheel drive, four-wheel steering, electric vehicle *Idris* (see Figure 3.10).

The carefully designed path shown in Figure 3.9 has the appearance of a “Gummy Bear” in profile, and provides many curves and sets of image pairs that are challenging for visual robot localisation. For example, the “ear” region contains a sequence of images on a tight curve: and there are pinch points (at the “neck”

and “knees”), where the robot is quite close to where it has been before, but is clearly not in the same place (e.g., images 143 and 1162 might be expected to be similar). The path finishes at the start point, but with Idris rotated through  $90^\circ$ .

We steered the robot through the environment and collected GPS signal and image data along its trajectory. Test images were captured as the robot was moving, by an omnidirectional camera approximately one and a half meters above the ground surface. Some example images from these datasets are shown in Figure 3.8, while the characteristics of the datasets are described in Table 3.2.

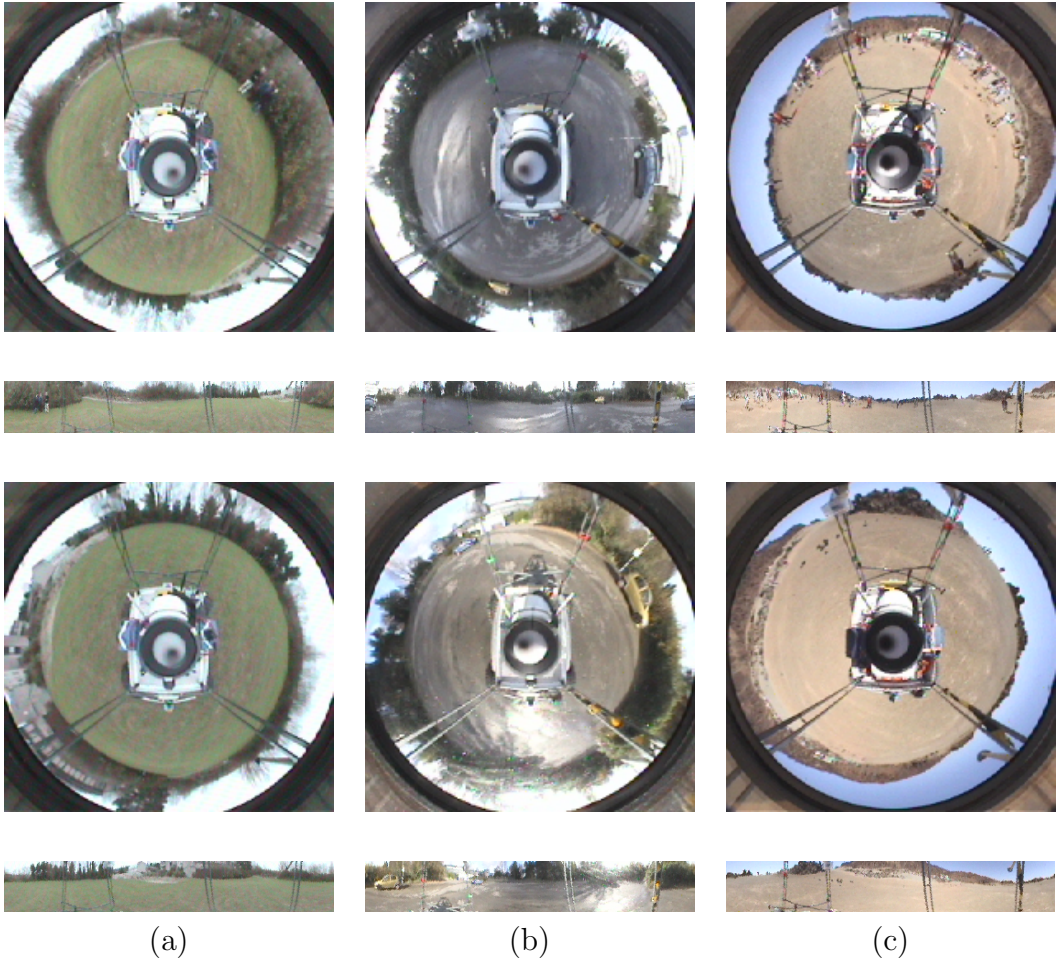


Figure 3.8: Example images from the (a) FIELD, (b) CARPARK, and (c) TENERIFE datasets, respectively. The omnidirectional images are shown in the first and third rows, the corresponding unwrapped images are shown in the second and fourth rows, respectively.

---

Table 3.2: Characteristics of GummyBear dataset

Dataset	Frames	Length	Rate	Notes
FIELD	1525	60m	6Hz	Flat but rough surface, can see about 50 m
CARPARK	2101	60m	8Hz	Flat, can see 30 m, light changes, moving objects
TENERIFE	2156	60m	8Hz	Bumpy, can see 100 m, moving objects

---

The FIELD dataset was collected in a field-type area, with some buildings in sight, but consisting mainly of trees and grass. The CARPARK dataset was captured in a carpark with trees around, where few cars were parked (and one moved) and some parts of the ground were wet with rain, providing challenging reflections and shadows. The TENERIFE dataset was obtained at the El Teide National Park, Tenerife. Its flat landscape, with fine textures of volcanic sand, pebbles and occasional rocky outcrops is similar to those encountered on the surface of Mars. Some tourists were walking around during the data acquisition process.

The position of the robot was estimated using a real-time kinematic (RTK) GPS system (Hofmann-Wellenhof et al. [1997]). A mobile RTK-GPS unit was mounted on the robot to receive correction signals over the internet from a GPS base station. Data was logged and post-processed to measure the position of the robot. The RTK technique was invented in the early 1990s, which was used to eliminate or reduce the error sources derived from satellite-based positioning systems, such as GPS. The basic concept is to estimate the position of the mobile unit relative to the base station using differenced carrier phase observations. This carrier-based measurement is more precise than pseudo-code measurements, allowing for centimeter-level positioning accuracy. We have assumed that it is valid to treat the orientation derived from the RTK GPS data as ground truth for orientation estimation, as the relative accuracy of orientation estimate is better than the accuracy goal of one degree. However, the accuracy and reliability to be achieved depends on several factors, including satellite availability, baseline length, and sufficient redundancy of GPS observations.

From Figure 3.9, we can see that there are some data gaps occurring along the

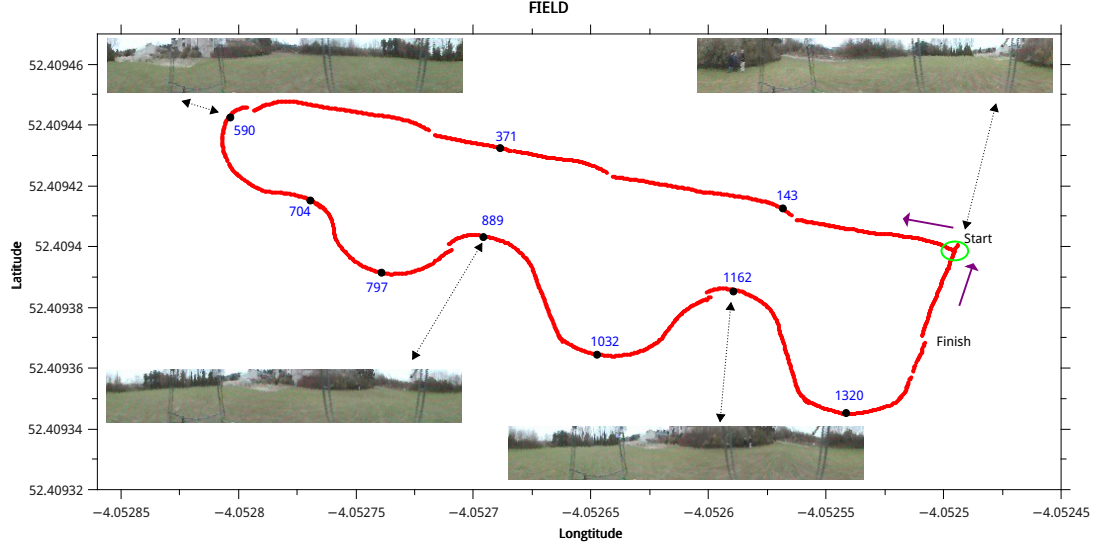


Figure 3.9: RTK-GPS track from the “Gummy Bear” path for the FIELD dataset, with some unwrapped omnidirectional image samples. Image numbers of key positions are marked in blue.

GPS path. The reason for this is that phase data of satellites is missing in places, probably due to carrier signal obstruction by objects, or other tracking problems. For this reason, the Kalman filter was used to smooth the glitches in our GPS data. To allow comparison, absolute GPS heading is converted to relative bearing by subtracting the absolute heading of the starting point of the trajectory, and changed to a range between 0 and 360. This is used as ground truth for orientation estimation.

The Kalman filter was originally proposed by Kalman and Bucy for estimating the state of a dynamic system from a series of incomplete and/or noisy measurements (Kalman [1960]; Kalman and Bucy [1961]). It is an efficient recursive filter, which assumes that the best estimate of the current state is the last known state. The state of a system is estimated in a way that minimises the mean-square error. The filter is modelled by the following two equations:



Figure 3.10: The Idris robot.

$$x_{k+1} = Ax_k + Bu_k + w_k \quad (3.1)$$

$$y_{k+1} = Cx_k + z_k \quad (3.2)$$

where  $A$ ,  $B$  and  $C$  are state transition, control input and measurement matrices, respectively,  $x_{k+1}$  is the state vector of the system at time  $k+1$ ,  $u_k$  is the known input vector at time  $k$ ,  $y_{k+1}$  is the measured output vector at time  $k+1$ ,  $w_k$  is a process noise and  $z_k$  is a measurement noise. To simplify the derivation of the Kalman filter, we assume that the  $w_k$  and  $z_k$  follow the normal distribution with covariance  $Q_k$  ( $w_k \sim N(0, Q_k)$ ) and  $R_k$  ( $z_k \sim N(0, R_k)$ ) respectively, and that they are statistically independent. It should be noted that the control input  $u$  and matrix  $B$  are ignored in our case, as the motor of the robot was instructed to move forward.

The Kalman filter algorithm can be split into two different stages: prediction, and updating. In the prediction stage, the new state is being predicted: a new

---

covariance is also being calculated, the equations being given by:

$$\hat{x}_k = Ax_{k-1} + Bu_k \quad (3.3)$$

$$\widehat{P}_k = AP_{k-1}A^T + Q \quad (3.4)$$

where  $\hat{x}_k$  is the predicted state estimation at the actual time step,  $\widehat{P}_k$  is the predicted estimate covariance matrix,  $P_{k-1}$  is the updated estimated covariance at the previous time step, and  $x_{k-1}$  is the updated state estimation from the previous time step. Next, in the update stage, the current state estimation is revised using the Equation 3.5, and the updated estimated covariance matrix is also calculated, using the Equation 3.6.

$$x_k = \hat{x}_k + K_k(y_k - C\hat{x}_k) \quad (3.5)$$

$$P_k = (I - K_kC)\widehat{P}_k, \quad (3.6)$$

where

$$K_k = \widehat{P}_kC^T(C\widehat{P}_kC^T + R)^{-1}, \quad (3.7)$$

$y_k$  is the measurement at the actual time step,  $K_k$  is the Kalman gains matrix, and  $I$  is the identity matrix.

These two stages are conducted alternately and repeated recursively until filtering ends, given an initial estimated state. In our context, we assumed that the robot was moving over a planar ground at constant speed, so the  $z$  coordinate for the robot was ignored. The state vector of the robot was simply presented by its position coordinates,  $x$  and  $y$ . The process noise covariance matrix  $Q$  and the measurement noise covariance matrix  $R$  are tuning parameters: we started with some reasonable initial estimate, and then tuned  $Q$  and  $R$  experimentally.

---

### 3.5 Outdoor datasets: New College 1 Dataset

The New College 1 dataset, published by Smith et al. [2009], was intended for use within the mobile robotics community. It was acquired on a 2.2km-long route during a wheeled mobile robot trip through different areas of the New College campus. The dataset comprises 8127 panoramic images captured by a five view LadyBug panoramic camera. The total length of the acquisition sequence is 44 minutes, and the frame rate is 3Hz. Each of the panoramic images consists of five single images, each of  $384 \times 512$  pixels resolution. The dataset presents a dynamic outdoor environment with multiple loop closures, including moving people and changing illumination.

Figure 3.12 illustrates some examples from the dataset. Apart from the camera images, odometry, laser scanner and GPS data were recorded at the same time. An overview of the trajectory constructed from GPS data is provided in Figure 3.11. Unfortunately, we can see (Figure 3.11) that the route is not smooth and intact. This is because the GPS data were not always available during the acquisition due to instances of lost connections with the satellites.

For the experiments detailed in Chapter 4, we picked a sequence of panoramic images (Images 120 . . . 1900) from the dataset for our evaluation. In Figure 3.11, the red dots indicate the position of tested images. These images were collected when the robot was driven around three laps of the circular area, then traversed through a short tunnel to another area, before returning to the previously visited area moving in the opposite direction. In consequence, they are suitable for testing multiple loop closures detection, especially evaluating the robustness to very different camera views (traversal directions). More details about the New College 1 dataset are available on the dataset website (<http://www.robots.ox.ac.uk/NewCollegeData/>).

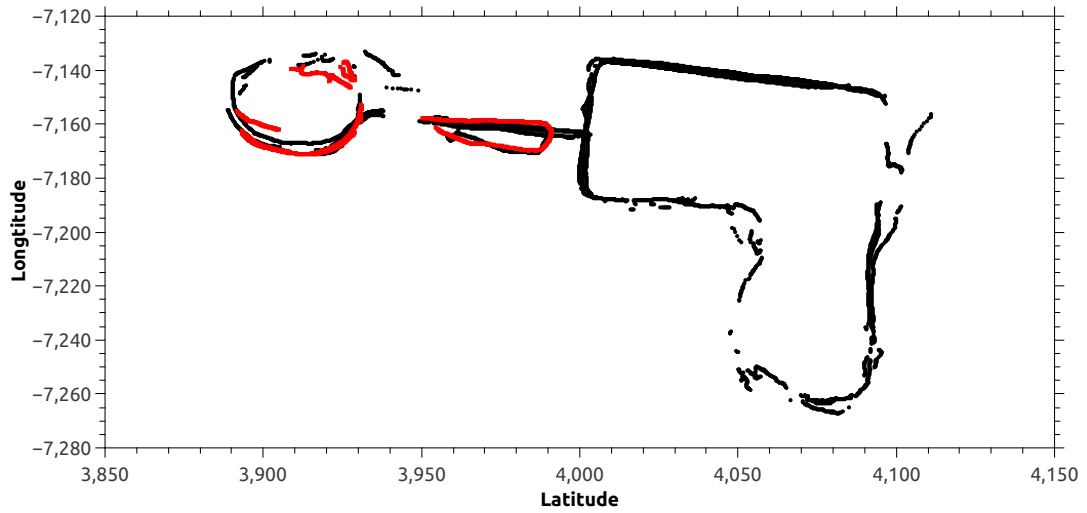


Figure 3.11: GPS trajectory of the New College 1 Dataset. The parts of the dataset used in our experiment are indicated by the red dots.



Figure 3.12: Example images from the New College 1 Dataset.

---

## 3.6 Conclusions

This chapter has provided a description of four datasets used to evaluate our proposed algorithms in this thesis. All datasets were collected using either an omnidirectional camera, or a five-view panoramic camera, in static or changing indoor and outdoor environments. The GummyBear dataset was captured in field-like, car park and Mars-like surroundings. The ground truth was provided by an RTK GPS. This dataset has been used for performance evaluation of our quadtree comparison algorithm (see Chapter 4), as well as for comparison of three methods for estimating robot orientation (see Chapter 5). The ISL dataset was collected from a laboratory environment, and the ground truth was obtained from a VICON motion capture system. This dataset has been used to test our loop closure detection method (see Chapter 6). An open dataset COLD captured under various weather and illumination conditions (sunny, cloudy and night) has also been used to evaluate our orientation estimation method (see Chapter 5). A publicly available dataset New College 1, was recorded in a natural outdoor environment with people moving within the scene. More importantly, the dataset contains loop closures traversed in opposite directions when the robot returns, making it ideal for testing loop closure detection algorithms, and accordingly it has been exploited to verify the effectiveness of the proposed method (see Chapter 4.6).

## Chapter 4

# A quadtree-based method for image comparison

### 4.1 Introduction

In robotic navigation it is common to use a range of sensors, such as laser, sonar, and GPS, to determine the position of a robot. However, since digital cameras have become more affordable, more research has been devoted to navigation using visual cues. Cameras can provide richer sensory input for better place discrimination, but with this richness comes noise and irrelevant data. Appearance-based place recognition consists in the main of two procedures. The first involves recording and storing images or prominent features of the environment: this is the pre-training stage. The robot must then localise itself by matching the current view with the stored reference images or features. As one would expect, establishing matches between observation and expectation is the most difficult step. Often, this requires a search that can be usefully constrained by prior knowledge and by knowledge of uncertainties about the robot, such as different possible robot headings. Therefore, the choice of a similarity measure between two images is the key issue in place recognition tasks.

A great deal of work has been carried out on appearance-based place recognition:

---

however, it is difficult to prevent mismatching completely even if state-of-the-art place recognition techniques are used. Among many methods in appearance-based place recognition, FAP-MAP, introduced in (Cummins and Newman [2008b]) addressed place recognition as a recursive Bayesian estimation problem, which adopted distinctive and invariant local features, such as SURF and MSER, and the BoWs method for computing image similarity. However, FAP-MAP requires off-line training on a suitable dataset, and extracting and detecting local features are usually time-consuming.

Recently, several binary descriptors that encode the image with a compact binary string, and whose similarity can be computed very quickly by the Hamming distance, have been shown to be very efficient in performing the task of place recognition. In (Sunderhauf and Protzel [2011]), an appearance-based place recognition system based on BRIEF-Gist descriptors was proposed, combining the BRIEF descriptor with the holistic representation of Gist. The BRIEF-Gist feature offers several important advantages, such as robustness to low quality and blurred images, smaller storage requirements, and faster processing. However, a place recognition system based on the BRIEF-Gist descriptor (and other, similar algorithms) suffers from the disadvantage that it is not invariant to traversal direction. In order to overcome this problem, Arroyo et al. [2014] presented a framework that divides each panoramic image into sub-panoramas and builds the binary descriptor around the center of sub-panoramas. A panoramic image is then represented by a concatenation of a set of binary strings. Subsequent matching between two images is based on cross-correlation between sub-panoramas of image pairs.

To achieve a robust image similarity measure between two panoramic images for place recognition, we use the concept of quadtree decomposition, combined with a number of standard image distance measures and involved standard three-colour (RGB) spaces, to create a novel image similarity method, which is robust to perceptual aliasing (the images we tested are mostly made of repetitive features) and can cope with the appearance of new objects in the robot environment without prior information. In addition, our method can detect loop closure on the basis of image matches, which is essential for reliable navigation. Quadtrees not only provide a noise resistant, fast, and easy to use comparison method, but they

---

also allow us to identify those image regions that genuinely represent changes within the environment. Our method is successfully validated on the Gummy-Bear and New College 1 datasets, and compared against FAB-MAP, BRIEF-Gist, and ABLE-P.

The rest of this chapter is organised as follows. Several image difference measurements are reviewed briefly in Section 4.2. The principles of our proposed image similarity measure are described in Section 4.3. A comparison of this method applied to different kinds of metrics is presented in Section 4.4. Section 4.5 and 4.6 detail the experiments undertaken, and report results. Finally, Section 4.7 concludes the chapter, and outlines possible future improvements.

## 4.2 Image distance metrics

Image distance metrics are methods that can quantitatively evaluate the similarity/dissimilarity between two images, or two image regions. Considerable efforts have been made to define distance metrics, and methods used thus far include Euclidean, city-block, earth mover, Mahalanobis, chi-square, Pearson’s correlation coefficient, tangent distance, histogram intersection and many more. In this section, we briefly introduce some of these metrics to determine which best suits our application. The reader may refer to (Goshtasby [2012]) for a comprehensive survey of similarity/dissimilarity measures.

### 4.2.1 Euclidean distance

The Euclidean distance has been one of the most commonly-used metrics in computer vision, due to its efficiency and effectiveness (Duda et al. [2001]). It measures the distance between two images by calculating the square root of the sum of the squared differences of corresponding pixels in images.

One advantage of this metric is that the distance is a sphere around the centroid (smoothness). It also has the advantages of being continuously differentiable, and fast to compute. However, this distance measure suffers from high sensitivity to

---

small deformations in images, because it does not take into account the spatial relationships between pixels. Moreover, it can be over-sensitive to variations in lighting conditions.

### 4.2.2 Median of absolute differences

The median of absolute differences (MAD) may be used to measure the dissimilarity between two images. Instead of the squares of the difference between the corresponding pixels used in Euclidean distance, MAD involves calculating the absolute intensity differences of corresponding pixels in images, sorting the absolute differences, and choosing the middle value as the dissimilarity measure (Duda et al. [2001]). Compared with the Euclidean distance, MAD has the advantage of robustness to occlusion and impulse noise.

### 4.2.3 $\chi^2$ distance

The  $\chi^2$  distance is also called the weighted Euclidean distance. It differs from the Euclidean distance in that each square is now weighted by the inverse of the average proportions, so that the distributional equivalence can be satisfied. The  $\chi^2$  distance between two images is given by:

$$\chi^2(I_i, I_j) = \frac{1}{2} \sum_{k=1}^{h \times w} \sum_{l=1}^c \frac{(I_j(k, l) - I_i(k, l))^2}{I_j(k, l) + I_i(k, l)}, \quad (4.1)$$

where  $I_i(k, l)$  and  $I_j(k, l)$  are the  $l^{\text{th}}$  colour component of the  $k^{\text{th}}$  pixel of images  $I_i$  and  $I_j$ , respectively.

---

#### 4.2.4 Pearson's correlation coefficient

Pearson's correlation coefficient, first presented in Pearson [1896], is also a useful tool for image comparison. It is given by the following equation:

$$\rho(I_i, I_j) = \frac{\sum_{k=1}^{h \times w} \sum_{l=1}^c (I_i(k, l) - \bar{I}_i)(I_j(k, l) - \bar{I}_j)}{\sqrt{\sum_{k=1}^{h \times w} \sum_{l=1}^c (I_i(k, l) - \bar{I}_i)^2 \sum_{k=1}^{h \times w} \sum_{l=1}^c (I_j(k, l) - \bar{I}_j)^2}}, \quad (4.2)$$

where  $\bar{I}_i$  and  $\bar{I}_j$  are the mean intensity of image  $I_i$  and  $I_j$ .

The correlation coefficient  $\rho$  value ranges from 1 for two images are identical, to -1 for two images are completely anti-correlated. Value zero indicates two images are completely uncorrelated (Huntington [1919]). The correlation coefficient subtracts the mean intensity from the intensity of each pixel, limiting the bias in image intensities. Additionally, the scale normalization is performed by dividing the inner product of the normalized intensities by the standard deviation of intensities in each image. Therefore, this metric is well-suited to comparing images taken under different illumination conditions, and it is invariant to a linear transformation of either image  $I_i$  and/or image  $I_j$ . However, it has the disadvantage of being computationally expensive. Another problem is that if one of the images has constant, uniform intensity,  $\rho$  is undefined due to division by zero.

#### 4.2.5 Histogram intersection distance

In many applications, histograms are used as representations of images. To compare two images, we can compute the similarity between their histograms. The histogram intersection distance was proposed by Swain and Ballard [1991], and has been widely used for image retrieval, object recognition and classification

---

tasks due to its simplicity and effectiveness. It is defined as:

$$I_{HI}(h_i, h_j) = \sum_{k=1}^n \min(h_i(k), h_j(k)), \quad (4.3)$$

where  $h_i(k)$  and  $h_j(k)$  represent the  $k^{\text{th}}$  bin of histogram  $h_i$  and  $h_j$ , respectively and  $m$  is the number of bins of the histogram .

The histogram encodes an image by the distribution of colours, and discards all spatial information. This makes histogram intersection distance invariant to object position and orientation changes. However, this is at the cost of limited discriminating power.

#### 4.2.6 Earth-mover's distance

The Earth-Mover's Distance (EMD) (Rubner et al. [2000]) is an important, perceptually meaningful metric between histograms. The EMD between two histograms is defined as the solution of the transportation problem from linear optimization (LP).

Specifically, the EMD is computed by finding the minimum cost required to transform one histogram into the other. Given two histograms  $X$  and  $Y$ , the EMD is defined by the following equation:

$$EMD(X, Y) = \frac{\sum_{i=1}^m \sum_{j=1}^n d_{ij} f_{ij}}{\sum_{i=1}^m \sum_{j=1}^n f_{ij}}, \quad (4.4)$$

where  $f_{ij}$  denotes the flows. Each  $f_{ij}$  represents the amount transported from the  $i^{\text{th}}$  supply to the  $j^{\text{th}}$  demand; ground distance  $d_{ij}$  represents the distance between bin  $i$  and bin  $j$  in the histograms, chosen according to the task at hand. The normalisation factor is the total flow, defined as:  $\sum_{i=1}^m \sum_{j=1}^n f_{ij} = \min(\sum_{i=1}^m X_i, \sum_{i=1}^m Y_j)$ , which is introduced in order to avoid favouring smaller signatures.

---

EMD has been successfully used for image retrieval, shape matching and image registration. However, it has an empirical time complexity between  $O(n^3)$  and  $O(n^4)$ , where  $n$  is the number of bins in histograms. This high computation cost is still a major hurdle to using EMD for some applications.

#### 4.2.7 Shannon mutual information

Mutual information (MI) is a measure of statistical dependence. The concept of MI was introduced by Shannon [1949] and later generalized by Gelfand and Yaglom [1959]. MI has important uses in communication theory. It was firstly used as a similarity measure for multi-modal gray scale image registration by Viola [1995]. The MI between two images is based on their marginal and joint/conditional entropies.

The MI information for two images is defined as:

$$MI(X, Y) = \sum_{i=0}^{255} \sum_{j=0}^{255} p_{ij} \log_2 \frac{p_{ij}}{p_i p_j}, \quad (4.5)$$

where  $p_{ij}$  is the joint probability that corresponding pixels in image  $X$  and  $Y$  have intensities  $i$  and  $j$ , respectively;  $p_i$  and  $p_j$  are the probability of intensity  $i$  and  $j$  appearing in image  $X$  and  $Y$ .

Shannon MI is a powerful measure for determining the similarity between multi-modal images. However, it is sensitive to noise, and is relatively computationally expensive, as density estimation is more time-consuming than a simple correlation calculation.

### 4.3 An image comparison measure using Quadtree

We are interested in the spatial structure properties of an image rather than detailed textural information. Unlike the majority of current image comparison methods, which use feature extraction and matching for place recognition,

---

our approach is a direct pixel-wise comparison of two images incorporating the quadtree concept. Quadtrees provide a fast and easy-to-use comparison method that improves robustness to noise. The comparison process can be mapped to a top-down built quadtree.

```

// Base case
begin
   $I_n \leftarrow I_{new}$ ;
   $I_r \leftarrow I_{ref}$ ;
  // Calculate the distance (in appearance space) between  $I_n$  and
   $I_r$ 
  Dist=distance( $I_n, I_r$ );
  if Dist > THRESHOLD then
    BuildQuadTree(rootNode);
  else
    Quadtree building stopped;
  end
end

// Quadtree building
BuildQuadTree(Node *n)
begin
  if  $n \rightarrow dist > THRESHOLD$  and  $n \rightarrow size > MIN$  then
    // Break image or patch into 4 patches
    for  $n=0$  to 3 do
      nodeIn=BuildNode( $n \rightarrow child[i], n, i$ );
      BuildQuadTree(nodeIn);
    end
  end
end
end

```

**Algorithm 1:** Pseudocode representation of the image comparison algorithm using quadtree.

Our method is a recursive operation. The principle idea behind the method is given in the pseudocode in Algorithm 1. It starts with two images which are to be compared. The first step is to calculate the complete image distance (in appearance space), employing one of the image metrics described in the previous section: this forms the root node of the tree. Next, if the images distance saved

in the root node is above a given threshold, the two images are each divided into four quadrants of identical size. If this is not the case, the comparison comes to a halt, as the two images are deemed similar. To what extent the two images are actually similar is of course influenced by the chosen threshold. For each non-similar quadrant of the two compared images, further recursive quadrant comparison is performed. This recursive operation continues until either two quadrants are judged sufficiently similar, or the resulting quadrants are too small.

Figure 4.1(a) is a visualisation of recursive image comparison and Figure 4.1(b) the corresponding tree-based representation. Figure 4.1(a) shows that the decomposition into sub regions provides us not only with robustness to noise, but also with an indication of the locations of visual change between image pairs. In addition, it should be noted that there has been camera motion between the two images, but that the only difference detected is the presence of the car, demonstrating robustness to small changes. In Figure 4.1(b), the root of the tree corresponds to the comparison of the two original images. Circles represent internal nodes of the tree, and leaf nodes correspond quadrants that are either similar or too small.

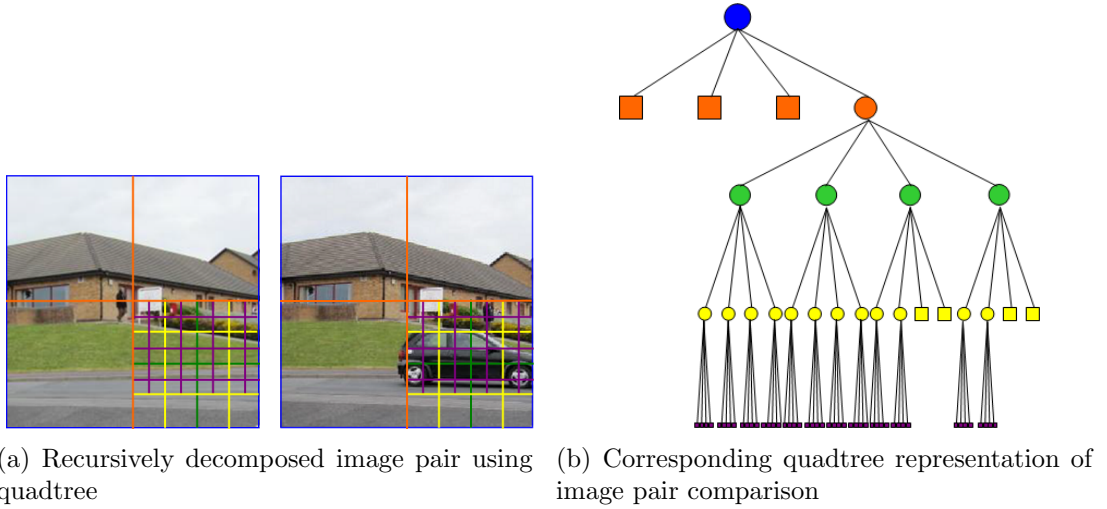


Figure 4.1: Quadtree decomposition.

---

## 4.4 Quadtree and metrics

In this section we discuss details of implementation, including the choice of image distance metric to use within the quadtree algorithm, and the choice of the threshold. To use our system for robot localisation it is important that images that are spatially close together in the real world have a similarity score reflecting this proximity, and it is the choice of comparison and threshold which determines this feature of our system.

To determine which distance metric is appropriate for localisation, we first plot comparison curve charts. These show the value of the metric for image pairs taken in different physical locations, by moving a robot slowly along a straight line path and taking an image every 10 centimetres. We compare an image from the middle of this sequence with all other images, and we seek a measurement that is a) smooth and b) not too “steep”. The graphs in Figure 4.2 are a sample of these comparison curve charts. The three curves in each chart represent Euclidean distance,  $\chi^2$  distance, and Pearson’s correlation coefficient incorporated within the quadtree similarity measurement, respectively. In order to produce a fair comparison between the three different measures (with thresholds on different scales) we define an iso-similarity point for each test: this sets the threshold for quadtree decomposition such that the three distance measures produce identical similarity measures. You can see these iso-similarity points clearly in the two graphs in Figure 4.2. You can also see that for low thresholds of Euclidean distance and  $\chi^2$  distance (Euclidean distance: 21,  $\chi^2$  distance: 1) and higher threshold of Pearson’s correlation coefficient (0.78), our quadtree measure is sensitive to small displacements (left image, Figure 4.2), but that with higher thresholds of Euclidean distance and  $\chi^2$  distance (Euclidean distance: 43,  $\chi^2$  distance: 5.1) and lower threshold of Pearson’s correlation coefficient (0.43), we are able to determine similarity between images on a broader scale (right image, Figure 4.2).

Briefly summarising our tests, we can see that Euclidean distance,  $\chi^2$  distance and Pearson’s Correlation Coefficient can behave in much the same way when we find a threshold that defines an iso-similarity point. Between the two graphs given here,

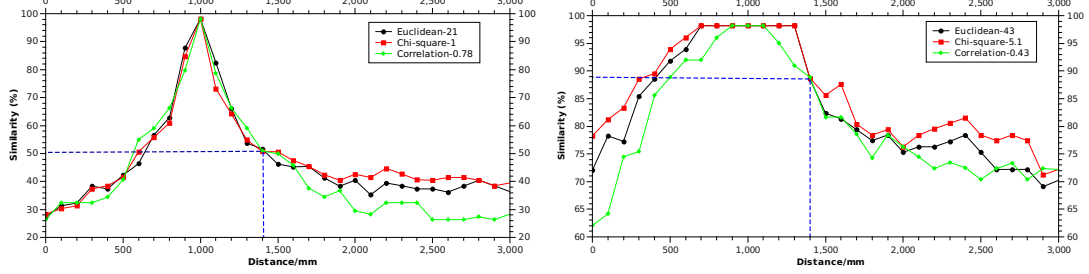


Figure 4.2: Comparison of different metrics applied to our quadtree similarity measurement: iso-similarity point set at 50 for the left image and 90 for the right, which occurs at approximately 1,400mm on the  $x$ -axis.

the similarity of the iso-similarity point increases from 50% to 90%. Pearson’s correlation coefficient seems to be the most sensitive to small displacements (has a narrower peak) and it is also the most computationally-intensive metric we have considered. Euclidean distance and  $\chi^2$  distance are both fast and easy to compute, with the comparison results showing little difference between them. For the sake of simplicity, for the rest of this paper we will present results from Euclidean distance only.

## 4.5 Experiments and results: GummyBear dataset

In this section we examine the effectiveness of our image comparison method. Two different experiments are conducted on the FIELD and CARPARK sub-datasets of the GummyBear dataset. These two sub-datasets represent different challenges: a self-similar environment in FIELD, and shadows and ground reflections due to water in the CARPARK (more details about this dataset have been given in Section 3.4). In our experiments we used a collection composed of every 10th image taken from the dataset. The threshold (THRESHOLD in Algorithm 1) was set to 45, 40, 35 and 30 for the FIELD dataset, and 70, 65, 60 and 55 for the CARPARK dataset. Please note that these values represent the appearance distance in their corresponding distance metrics. Compared images are aligned by horizontally shifting them, column by column, until a maximum of similarity is obtained. This is to compensate for the change in heading of the robot during

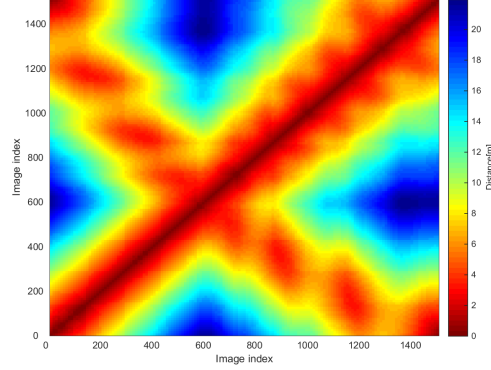


Figure 4.3: Physical distance between any pair of the images from the FIELD dataset, in meters.

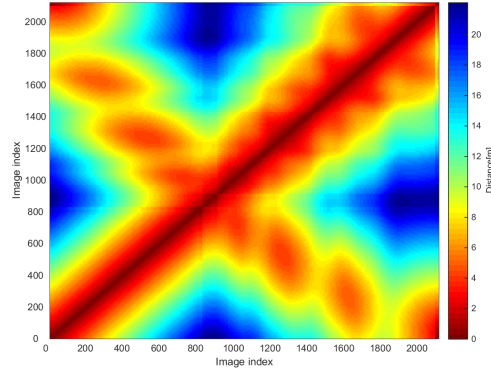


Figure 4.4: Physical distance between any pair of the images from the CARPARK dataset, in meters.

the capture.

Figure 4.3 and Figure 4.4 are the visualisations of physical distances between any pair of images from the FIELD and CARPARK datasets, respectively. The distance in meters is calculated using RTK GPS coordinates. The use of a colour spectrum from warm reds to cool blues maps the distance values from low to high. These provide the ground truth upon which the proposed image comparison method can be visually evaluated. For example, the  $i$ -th column corresponds to the physical distance between the locations of capture of the  $i$ -th image and all others. The main diagonals are minimal, as the distance of a location itself is zero. As stated

---

in Chapter 3, the Gummy Bear path is a loop, and the robot returns to the starting position in the loop. It can be observed in Figures 4.3 and 4.4 that the last column of the first row shows dark red, illustrating the loops.

### 4.5.1 Experiment 1: Loop closure

The first experiment concerns the problem of *loop closure*; this is the ability of a robot to realize when it has been in a particular place before. However, a robot is unlikely to return to the same pose when it revisits a previous place. The GummyBear dataset was obtained by driving the robot along a closed loop and returning to the initial location in the loop with a different orientation: this fact can be utilized to test the performance of our algorithm, especially its robustness to changes in robot orientation.

Given a similarity threshold (THRESHOLD in Algorithm 1), the similarity measures between the start image and all other images in the path can be calculated. Figure 4.5 shows the results for the FIELD dataset, and Figure 4.7 shows the results for the CARPARK dataset. As we can see, the similarity scores are increased towards the end of the path, when the robot has come round to the same place. This shows that we are able to determine places where the robot has been before, even when the robot is on uneven ground and at a different orientation.

Image 0 and Image 1481 from the FIELD dataset were captured at the same location with the camera rotated clockwise by about 90 degrees between images. The visualisations of the aligned quadtree representation on the unwrapped panoramic images is given in Figure 4.6. This intuitively demonstrates that the proposed method finds a correct horizontal shift between the two images. Moreover, this shows that whilst there are some small differences between the two images, it is clear that our method is able to determine loop closure within a reasonable tolerance, and that the main causes of dissimilarity in this experiment are the frame around the camera on Idris (which appear as vertical black lines on the images). Future work may involve handling features such as this in a pre-processing stage.

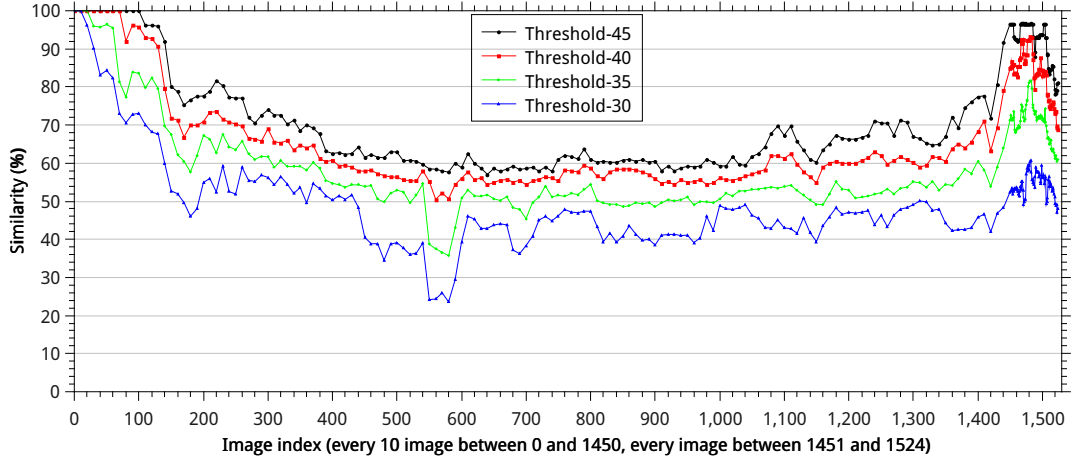


Figure 4.5: Similarity between Image 0 and all images of the FIELD dataset, demonstrating the possibility of robust loop closure.



Figure 4.6: Left: an image pair (Image 0 and Image 1481 from the FIELD dataset). Right: visualisation of left image pair comparison using our proposed method. The similarity is 96.59%, with a threshold of 45.

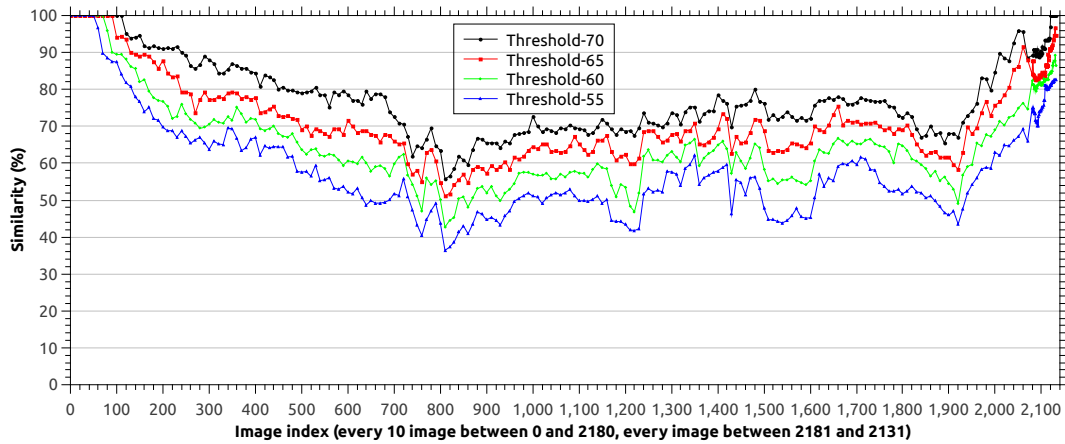


Figure 4.7: Similarity between Image 0 and all images of the CARPARK dataset, demonstrating the possibility of robust loop closure.



Figure 4.8: Left: an image pair (Image 0 and Image 2119 from the CARPARK dataset). Right: visualisation of left image pair comparison using our proposed method: the similarity is 96.99%, with a threshold of 70.

Figure 4.8 shows the visualisations of the aligned quadtree representation on the unwrapped panoramic Image 0 and Image 2119 from the CARPARK dataset. As expected, our method estimates the correct horizontal alignments between them. In this case, the primary difference between the two images is the orientation: the small differences are due to the overexposed area of Image 2119.

#### 4.5.2 Experiment 2: Pinch points — nearby, but not the same place

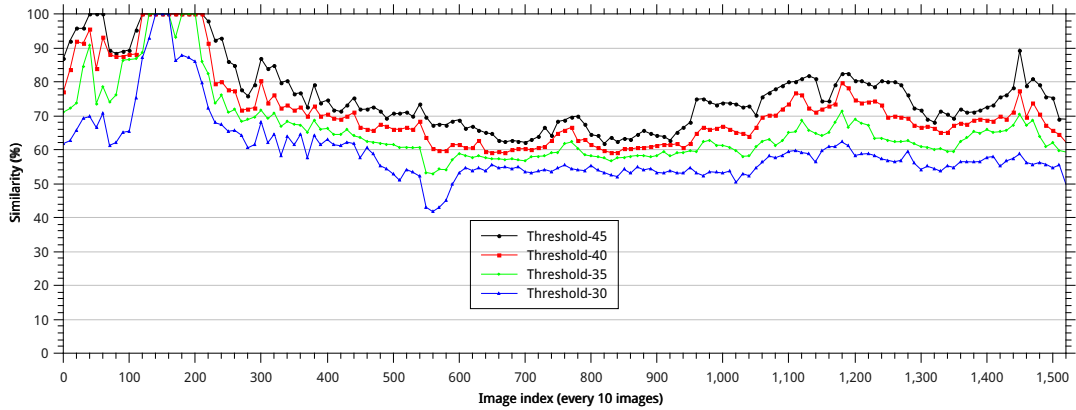


Figure 4.9: Similarity between Image 143 and all images of the FIELD dataset.

In this experiment we investigate the robustness of our quadtree similarity measure to small displacements. Usually, images taken at closely adjacent positions are likely to be very similar. By choosing a comparison image close to one of the “Gummy Bear” pinch points we can see whether it would be possible to determine when we are close to this image on the return trip. For robust visual robot



Figure 4.10: Left: an image pair (Image 143 and Image 1162 from the FIELD dataset). Right: visualisation of left image pair comparison using our proposed method: the similarity is 82.24%, with a threshold of 45.

navigation this is a key ability; if we need to work out how to get to a particular place, it is important that we are able to work out when we are close to it.

Figure 4.9 shows the similarity measure between Image 143 (on the “back” of the “Gummy Bear”, close to a pinch point) and all others of the FIELD dataset. From this we can see a rise in similarity around 1170, as expected, which shows that our measure can determine when we are near a particular target destination. It is noted that higher similarity scores also occur around 1450. According to the ground truth, it is true that Image 1450 and Image 143 are taken from nearby locations.

Figure 4.10 shows the quadtree visualisation for the pairs of spatially close images (Image 143 and Image 1162) from the FIELD dataset. This demonstrates that there are differences between these two places (which is true), but that there is still a high similarity between them from an appearance perspective.

Figure 4.11 shows the similarity measure between Image 192 and all others of the CARPARK dataset. As we can see, higher similarity scores are visible around 1300 and 1600. we can observe from the Figure 4.4 that yellow and red area representing the smaller distance around the 1300th and 1600th row from the 192nd column.

Figure 4.12 shows the quadtree visualisation for the image pair 192:1670 from the CARPARK dataset. Look closely at the image pair, we find that the two images were captured from different but nearby locations, and under different illumination conditions: Image 192 was taken under strong sunlight, while Image 1670 was collected while a cloud was covering the sun. In addition, the frame

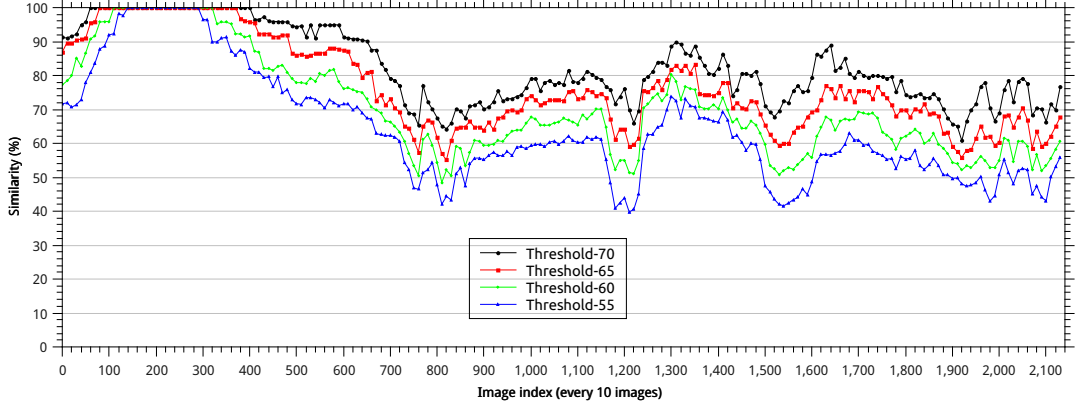


Figure 4.11: Similarity between Image 192 and all images of the CARPARK dataset.



Figure 4.12: Left: an image pair (Image 192 and Image 1670 from the CARPARK dataset). Right: visualisation of left image pair comparison using our proposed method: the similarity is 85%, with a threshold of 70.

around the camera on Idris is again a cause of the dissimilarity between the two images. As the result shows, the proposed method could detect these changes and still give a high similarity score.

## 4.6 Experiments and results: New College 1 Dataset

In this section, we conducted a quantitative evaluation of loop closure detection performance using our proposed method, and then compared it with three state-of-the-art frameworks (FAB-MAP, BRIEF-Gist, and ABLE-P). In particular, we compared the performance of the algorithms for loop closure detection tasks in two different situations: unidirectional loop closures, and a mixture of

---

unidirectional and bidirectional loop closures. We also evaluated the loop closure detection performance of our proposed method in the presence of noise, and compared it with two state-of-the-art frameworks (BRIEF-Gist and ABLE-P).

#### 4.6.1 Experimental set-up

The publicly available New College 1 dataset (Smith et al. [2009]) was used to evaluate each approach. The New College 1 dataset has been covered previously, in Section 3.5. We chose a sequence from the dataset between Images 120 and 1900 for evaluations.

The dataset does not provide direct information about loop closure, and the GPS data provided are not completely reliable, so we manually generated the loop closure ground truth. We assumed that the robot moved at a constant speed, and initially conducted some tentative experiments on ground truth generation. The best matches for every  $n$ th image were obtained by visual inspection, and linear interpolation was performed within these fixed matches. Different values of  $n$  ( $n = 5, 20, 50$ , respectively) have been used to produce the ground truth for the sequence between Image 150 and 450, in order to investigate whether changing the interval (value  $n$ ) has any significant effect on the loop closure ground truth.

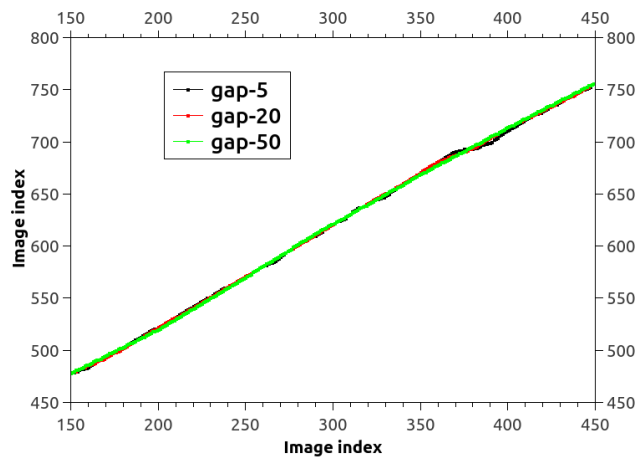


Figure 4.13: Ground truth for the sequence between Images 150 and 450, generated at different values of  $n = 5, 20$ , and  $50$ .

---

Figure 4.13 shows the ground truth for the sequence between Images 150 and 450. Please note that we only created the ground truth from the second traversal. As we can see, the obtained ground truth using different value  $n$  basically coincide, except around Image 380-390 ( $X$ -axis) and Image 693-702 ( $Y$ -axis) where the diagonals have slightly inconsistent slope with  $n = 5$  and 20. Looking more closely at the sequence between Images 150 and 450, we found that the second traversal of the loop is not a direct overlap of the first. The Image 693-703 from the second traversal are not taken at exactly the same location and the same viewpoint compared to the Image 380-390 from the first traversal, in this case, it is hard to label the loop closure ground truth unambiguously. Therefore, a match within a margin of 10 frames is considered a true positive event for loop closure detection in the following evaluation,

We found, as a result of these tests, that we could not distinguish visually between the ground truths generated by the different interval values. In consequence, we chose to generate the loop closure ground truth for the sequence between Images 120 and 1900 manually, using a 20-frame interval, following Sunderhauf and Protzel [2011]. Figure 4.14 shows the loop closure ground truth between Images 120 and 1900 (only the lower triangle is shown), where the red off-diagonals indicate the locations where loops are closed. It is interesting to see that the right-side diagonals in the top left of the matrix correspond to the unidirectional loop closures, while the left-side diagonals correspond to the bidirectional loop closures.

In the case of FAB-MAP, an open source implementation developed by Glover et al. [2012] called OpenFABMAP was used for testing. We used the default parameter settings and chose part of the New College 1 dataset as the training dataset. Note that the tested images in our experiments are not included in the training dataset. The final result is a confusion matrix, representing the probability of loop closure.

BRIEF-Gist was implemented using the C++ language and the BRIEF descriptor provided by OpenCV (Bradski [2000]). Firstly, a panorama is divided into five equally sized sub-panoramas, and each sub-panorama is downsampled to  $64 \times 64$  pixels. Secondly, the center of each sub-panorama is chosen as the keypoint, and

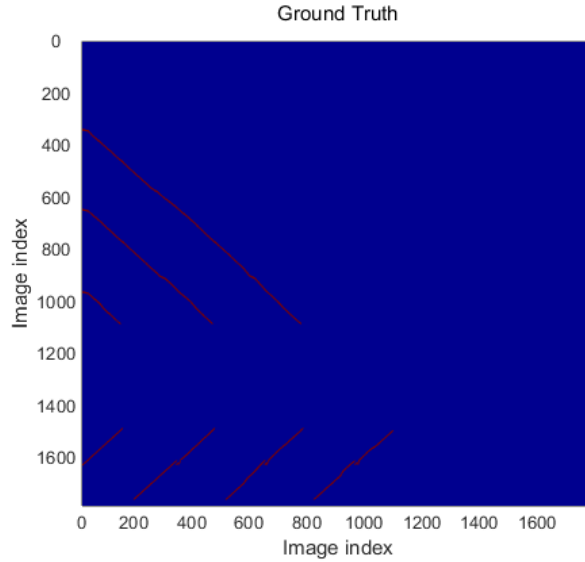


Figure 4.14: Ground truth of the New College 1 Dataset.

a BRIEF descriptor with length 32 bytes is computed for each of sub-panoramas separately. The final descriptor of the panoramas is obtained by concatenating the five descriptors of the sub-panoramas. The similarity between images is then measured by the Hamming distance between their descriptors.

ABLE-P was also implemented following the guidelines in the original paper (Arroyo et al. [2014]). We used the LDB descriptor provided by the authors on their website (Yang and Cheng [2014a]), and kept the default parameter settings for the LDB descriptor. The implementation process is similar to that of the BRIEF-Gist, except for the matching strategy and the number of sub-panoramas. Each panoramic image is split into six sub-panoramas rather than five, as this yields much better results (following the advice from the author). The similarity between the two images is measured by the minimum Hamming distance of the different LDB strings, which are obtained from the six possible alignments of the six sub-panoramas of the image.

---

### 4.6.2 Evaluation

We use traditional precision and recall metrics to evaluate the performance of all methods. Precision is defined as the ratio of the number of true-positive loop closure detections to the total number of detections. Recall is defined as the ratio of the true-positive loop closure detections to the loop closures in the ground-truth. The area under the precision-recall curve, known as the average precision, is also used to evaluate the overall performance of all algorithms tested. A match  $(S_c, S_p)$  is considered as a true positive detection if the distance to the ground truth is less than 10 frames in either direction.

For testing our proposed method, we first downsampled the images to  $360 \times 40$  pixels, then calculated the similarity score between the current scene  $S_c$  and the previous scene  $S_p$ : if their similarity is higher than a threshold  $T_s$ , we consider these two images to correspond to a loop closure event. For FAB-MAP, the confusion matrix obtained shows the probability that the current scene  $S_c$  and the previous scene  $S_p$  comes from the same place. In order to make the results of the different algorithms comparable, we normalized the distance matrices  $D$  derived by the BRIEF-Gist and ABLE-P algorithms with respect to the maximum distance, making each value range from 0 to 1. In cases of BRIEF-Gist and ABLE-P, if the normalized distance between the current scene  $S_c$  and the previous scene  $S_p$  is lower than a threshold  $T_s$ , the match indicates loop closure.

### 4.6.3 Experiment 1: loop closure

We picked every 20th frames from the sequence (Images 120 ... 1200) as the current robot view ( $S_c$ ) and performed unidirectional loop closure detection between these images against all images in the sequence. Given the frame rate of 5Hz and the velocity of the robot of  $0.8m/s$  (Smith et al. [2009]), every twenty frames correspond to a time interval of roughly 4 seconds and a distance of approximately 3 meters. This is meaningful for loop closure detection according to the size of the explored environment. By using the thresholds  $T_s$  evenly distributed in the range  $[0,1]$  with step size of 0.01, we obtained the precision-recall curves presented in

---

Figure 4.16. In a similar way, we conducted the unidirectional and bidirectional loop closure detections on the sequence of images (Images 120 ... 1900). In both cases, we ignored 50 images immediately before and after the current view to avoid matching images taken within a short time of each other.

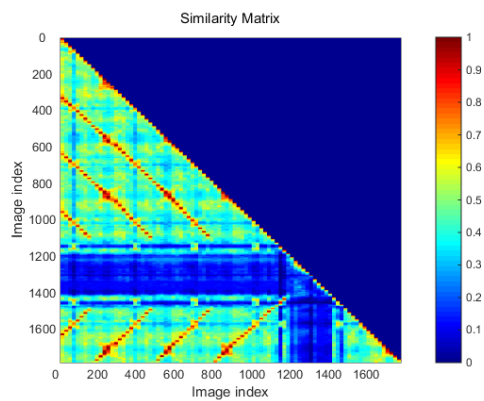
The visualizations of experimental results are shown in Figure 4.15 (only the lower triangles are shown), and the precision-recall curves presented in Figure 4.17. Table 4.1 summarises the average precision and the best recall rates at precisions of 100% on two sequences, using all four algorithms.

Table 4.1: Average precision (AP), and best recall (R) at 100% precision.

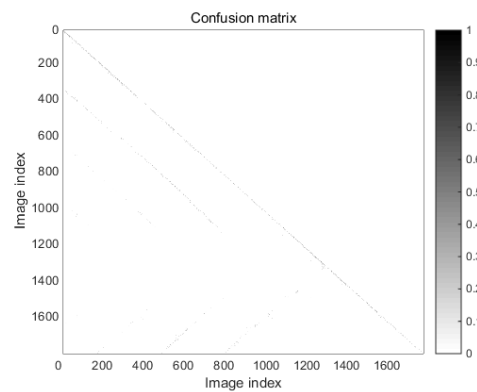
Algorithm	Images 120 ... 1200		Images 120 ... 1900	
	AP	R	AP	R
Our method	99.48%	88.00%	97.35%	69.00%
FAB-MAP	64.60%	9.00%	58.89%	8.00%
BRIEF-Gist	99.27%	83.00%	63.98%	48.00%
ABLE-P	98.82%	80.00%	93.06%	39.00%

As can be seen in Figure 4.15, the Quadtree method and FAB-MAP present high similarity and probability values for real loop closures, while BRIEF-Gist and ABLE-P present low distance values, which are shown as diagonals. From Figure 4.15 we also can see that bidirectional loop closures can be recognized by our method, FAB-MAP and ABLE-P, revealed as left-side diagonals between Images 1500 and 1900, while BRIEF-Gist fails to identify them.

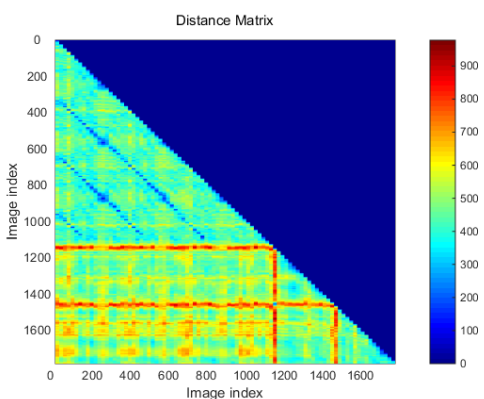
It is essential to avoid false positives for a loop closure detection algorithm, as they have the potential to corrupt the map generated in vSLAM tasks. It can be observed from Figures 4.16 and 4.17, that all approaches can obtain high rates of correct detection and reach 100% in precision. In both cases, our method achieved the highest recall of 88% and 69% at 100% precision, respectively, followed by the BRIEF-Gist, which obtained the maximum recall of 83% and 48%, respectively, and ABLE-P, which achieved the best recall of 80% and 39%, respectively. FAB-MAP ranked last, with highest recall values of only 9% and 8%, respectively. Nevertheless, this is consistent with the results presented in the original paper (Newman et al. [2009]), where the best recall rate is slightly less than 10%. It



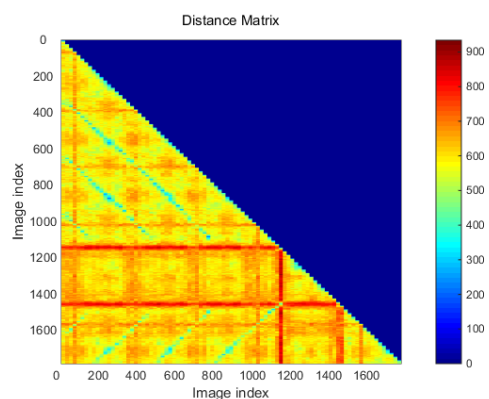
(a) Similarity matrix computed using Quadtree method



(b) Confusion matrix computed using FABMap



(c) Distance matrix computed using BRIEF-Gist



(d) Distance matrix computed using ABLE-P

Figure 4.15: (a) Similarity matrix computed using the Quadtree method; (b) Confusion matrix computed using FABMap (best viewed in magnification); (c) Distance matrix computed using BRIEF-Gist; and (d) Distance matrix computed using ABLE-P.

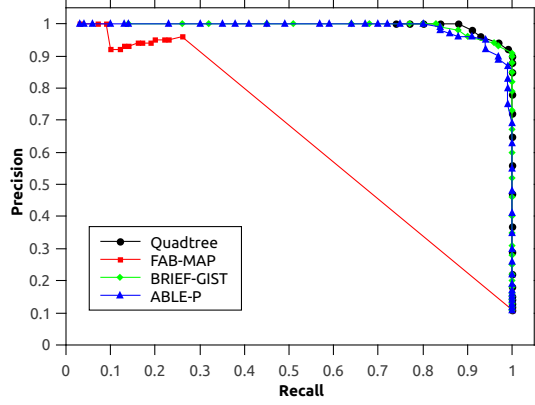


Figure 4.16: Precision-recall curves between Images 120 and 1200.

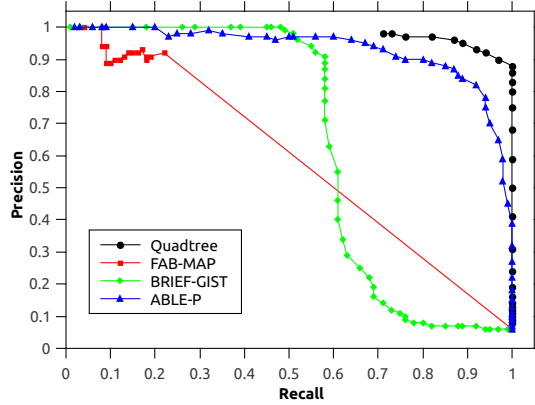


Figure 4.17: Precision-recall curves between Images 120 and 1900.

should be noted that our results, based on part of the New College 1 dataset, may not be representative of those obtained from the complete dataset: consequently, comparison might not be valid.

As seen from Table 4.1, in terms of the average precision, the accuracy of the proposed method, BRIEF-Gist and FAB-MAP is comparable on the sequence of images (Images 120 ... 1200). The performance of our method is a little better than that of the other two methods, and an accuracy of 99.48% is obtained. The worst performance is shown by FAB-MAP, which only achieved an accuracy of 64.6%. The reasons why FAB-MAP demonstrated inferior performance to the

---

other approaches may include a shortage of training data. It is also possible that the default feature detection type and thresholds used might not work well for the chosen dataset.

Overall, the proposed approach achieves a competitive trade-off between precision and recall, and produces a rather good performance in both unidirectional and bidirectional loop closure detection tasks. However, this is dependent on a specific dataset, and real-world use would require users to tune the parameters for each case.

#### 4.6.4 Experiment 2: loop closure on noisy data

In order to evaluate the noise robustness of the proposed method for the loop closure task, we used natural images (New College 1 dataset) corrupted by synthetic noise. We also compared the proposed method with the BRIEF-Gist and ABLE-P methods under different levels of noise. Loop detection results for this experiment were obtained by running the same loop closure detection detailed in Subsection 4.6.3. The noisy images were generated by adding Gaussian noise to the sequence of images (Images 120 ... 1200) using the Matlab function *imnoise*, with the variance parameter set to 0.01, 0.02 and 0.03. Figure 4.18 illustrates a sample image (a) from the New College 1 dataset, and three variant images (b, c, d) which have been corrupted by Gaussian noise, with mean 0 and different variances (0.01, 0.02 and 0.03, respectively).

Table 4.2 and Figure 4.19 sum up the average precision and best recall at 100% precision results, depending on the noise variance, as well as the original results for comparison. It will be found from Table 4.2 and Figure 4.19 that the performance of our method degrades slightly as the level of noises increases. The average precision and maximum recall values decrease from 99.48% and 88% to 99.05% and 78%, respectively. The BRIEF-Gist method demonstrated similar performance to the proposed method, indicating these two methods are robust enough when encountering noisy images. The ABLE-P method demonstrated the worse performance, with the highest recall value dropping dramatically from 80% to 22%. This implies that the LDB descriptor adopted by the ABLE-P method is



Figure 4.18: (a) Original image; (b) Original image corrupted by Gaussian noise (0, 0.01); (c) Original image corrupted by Gaussian noise (0, 0.02); and (d) Original image corrupted by Gaussian noise (0, 0.03).

more sensitive to Gaussian noise. Overall, the proposed method achieves better performance under low level noise with variance of 0.01, while the BRIEF-Gist method slightly outperforms the proposed method with higher levels of noise, at variance of 0.02 and 0.03.

Table 4.2: Average precision (AP), and best recall (R) at 100% precision, for original images and images corrupted by Gaussian noise at different variances.

Algorithm	Original results		<i>variance</i> = 0.01		<i>variance</i> = 0.02		<i>variance</i> = 0.03	
	AP	R	AP	R	AP	R	AP	R
Our method	99.48%	88.00%	99.28%	86.00%	99.09%	78.00%	99.05%	78.00%
BRIEF-Gist	99.27%	83.00%	99.04%	81.00%	99.05%	82.00%	99.07%	83.00%
ABLE-P	98.82%	80%	92.03%	22.00%	93.78%	29.00%	94.24%	32.00%

## 4.7 Conclusions

In this chapter we have presented an image similarity measure for robot place recognition based on the concept of quadtree decomposition. Quadrees not only provide a noise-resistant, fast, and easy-to-use comparison method, but also allow us to identify those image regions that genuinely represent changes within the environment.

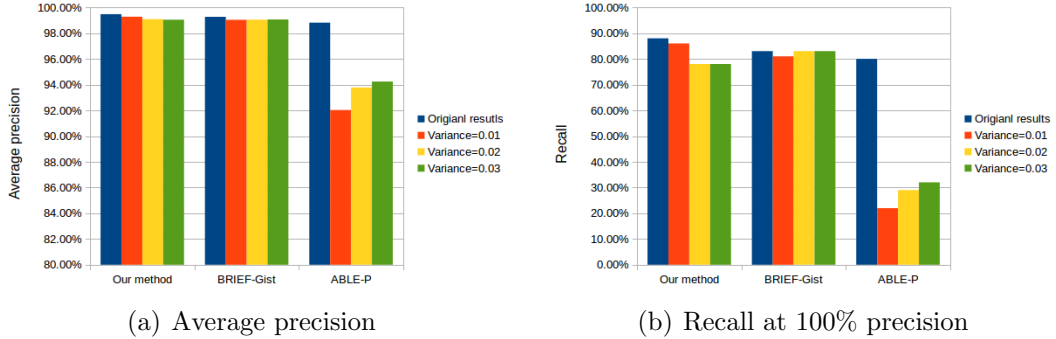


Figure 4.19: (a) Average precision; (b) Best recall at 100% precision of three methods on original images and noisy images with different variance (0.01, 0.02 and 0.03) and zero mean.

To demonstrate the effectiveness of our method, we conducted the experiments with two datasets captured in two different outdoor environments. The results of the experiments indicate that such image similarity methods can handle perceptual aliasing and achieve a high recall while maintaining 100% precision in loop closure detection tasks, even if the same place is seen from a different orientation and the images are corrupted by Gaussian noise. We compare the proposed method with other three methods for loop closure detection without additive noise. The experimental results illustrates that the performance of our proposed method is superior to those of the other three methods in terms of recall. We also compare the proposed method with two other methods for loop closure detection under various levels of additive Gaussian noise. The experimental validation shows that the proposed method is comparable to that of the BRIEF-Gist method when processing images contaminated by Gaussian noise.

However, our method is a direct pixel-to-pixel comparison between the images, and so it might prove to be sensitive to changes in illumination. Nonetheless, the experimental results have been promising so far, and suggest that our method provides a reasonable similarity measurement between image pairs, and can be applied in robot orientation estimation and loop closure detection with multiple revisits and different camera viewpoints. The further development of this method will be presented in Chapter 5 and Chapter 6.

## Chapter 5

# An evaluation of image-based estimation techniques for robot orientation

### 5.1 Introduction

When mobile robots move, one of the basic problems to be solved is that the robot must know its orientation as accurately as possible. Various solutions to the problem have been proposed, using visual cues. We can categorise these solutions into two main groups: feature-based, and appearance-based. Feature-based methods try to detect distinctive and robust points, or regions, between consecutive images, while appearance-based methods concentrate their efforts on the information extracted from the pixel intensity, the whole image being represented by a single descriptor, without local feature extraction. The change in orientation between frames is then computed by aligning the features, or images, using a calibration of the projection onto the image plane.

Among these solutions, many methods rely on optical flow, or local image features to establish the spatial relationship between two images. However, these methods are generally sensitive to the systematic errors caused by intrinsic and extrinsic

---

calibrations. Tracking feature points is a challenging situation in an omnidirectional vision-based system, since images obtained from hyperbolic quadratic mirror surfaces are highly distorted. On the other hand, some methods visually describe the environment locations based on global descriptors. These descriptors are normally very fast to compute and compact, simplifying the image matching process. A few frameworks using various global descriptors have been reviewed in Chapter 2. An interesting example is the work of Payá et al. [2014], which compared four global descriptors in order to resolve the robot pose estimation and mapping problems using omnidirectional information. The four descriptors in question were based on the Discrete Fourier Transform (DFT), Principal Component Analysis (PCA), Histograms of Oriented Gradients (HOG), and Gist of scenes, respectively. The relative orientation of the robot is then computed from the comparison between the global image descriptors.

In this chapter, we aim to address the question “What image based techniques are best for orientation estimation?” We do this by comparing appearance-based methods such as the visual compass (Labrosse [2006]), our proposed quadtree method (Cao et al. [2012]) and feature matching techniques such as SIFT (Scale-invariant feature transform, Lowe [2004]). In order to make our comparison of these methods thorough, we measure their performance on our collected outdoor dataset (GummyBear ) and indoor dataset (ISL), as well as on the open dataset (COLD-Freiburg omnidirectional sequences A). We also make a direct comparison with experimental results in (Payá et al. [2014]). In this comparison, we restrict our attention to the orientation estimation task.

The remainder of this chapter is organised as follows. The next section (Section 5.2) addresses the compared methods in order to evaluate their relative performance in estimating robot orientation. Section 6.3 offers evaluation results on the GummyBear, ISL, COLD and log-transformed COLD datasets. The final section concludes our study and states its findings.

---

## 5.2 Computing robot orientation

In this section, we describe the compared methods for orientation estimation. These techniques include a feature-based method, our proposed method, the visual compass, and those employed in (Payá et al. [2014]).

### 5.2.1 A feature-based method: SIFT

In this method, SIFT features (Lowe [2004]) were extracted from our panoramic images and then used to align these images. The method was implemented by ourselves using the C language and the SIFT descriptor provided by the authors on their website (<http://www.cs.ubc.ca/~lowe/keypoints/>). The orientation estimation operates as follows: the interest points are first detected from a pair of images using a scale-space difference-of-Gaussians approach. Each detected interest point is characterized by a SIFT descriptor, which is a histogram of gradient orientation within the subregion around the interest point, and contains 128 elements. Euclidean distance was used to compute the difference between SIFT features, and an acceptance ratio of 0.6 was chosen for matching similar interest points between two images. Two features are matched if their distance in feature space is less than 0.6 times the distance of the second closest feature. In order to obtain a reliable solution in the presence of outliers, matches were also filtered using a Gaussian distribution to model the feature displacements, cutting off the matches if displacements are one standard deviation away from the mean. Once features have been matched, the two images are aligned by computing the average horizontal displacement over all the features. Figure 5.1 shows the feature correspondences between two images, using the method described above.

### 5.2.2 Visual compass

The implementation of the visual compass is provided by Labrosse [2006]. This method is based on a linear search for the minimum of the difference function. The Euclidean distance in image space was used to measure the similarity between

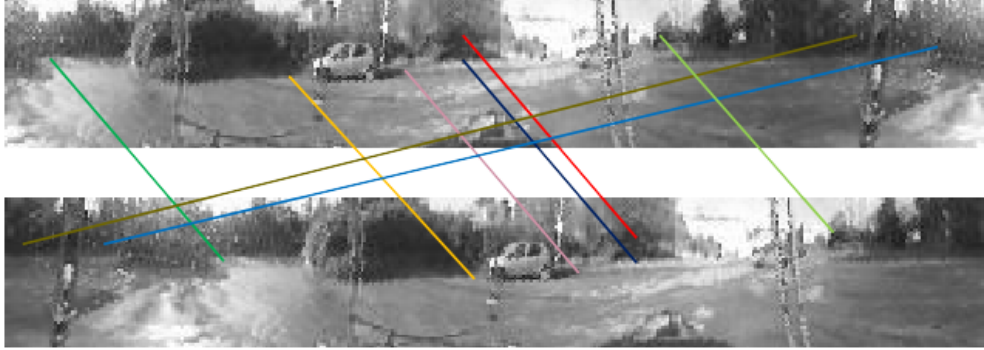


Figure 5.1: SIFT matching example

images. The relative rotation between a pair of successive panoramic images is obtained by finding the best match (minimum difference) between a reference image and a column-wise shift of the current image. The column shift corresponds to the orientation change between two images (see Figure 5.2). The orientation estimation is done from a moving reference image: the decision on when to change it being made using a measure of difference between images. This offers a compromise between accumulating error and comparing similar images to get a better estimation of the change in orientation. It should be noted that only the parts of the images that correspond to the front and back of the moving robot are used in the matching process.



Figure 5.2: Visual compass example. The top row shows a reference image, and the bottom row shows the current image, where the dashed box indicates the column shift  $\alpha$  between the two images, corresponding to the relative rotation between them.

---

### 5.2.3 A quadtree-based method

The core technique we use for orientation estimation is an image similarity measure method based on the quadtree decomposition combined with a number of standard image distance measures, which has been presented in Chapter 4. We calculated the orientation change between current and reference images by shifting the columns of current image leading to the maximum similarity between the two images. The column shift gives the orientation difference when working with panoramic images.

### 5.2.4 Other methods used in related publications

Four global appearance descriptors are applied to represent the omnidirectional images in (Payá et al. [2014]), which involves the study of viability of these descriptors for map building and localisation tasks. The experiments conducted consist of two phases: learning, and validation. In the learning phase, the descriptors for each image in the database are computed to build the map. In the validation phase, the descriptor of the current image captured by the robot is generated, and then compared with all the descriptors in the map in terms of Euclidean distances: the distance vector obtained is then sorted in ascending order. The nearest neighbour in the database is defined as the image with the minimum distance, which is used to estimate the present position of the robot. Once the robot has been localized in the map, the orientation can be calculated by comparing the descriptors of the current image with the nearest neighbour. Next, we will briefly investigate four descriptors, and how the orientation of the robot is computed using these descriptors.

#### 5.2.4.1 A Discrete Fourier Transform (DFT) descriptor

In a Fourier domain, each point represents a particular frequency contained in the image. An image in the spatial domain can be transformed into the frequency domain by taking the Discrete Fourier transform. The DFT descriptor describes

---

the appearance of a scene using the Fourier coefficients of the low frequency components, called the Fourier signature (FS), which is acquired by the following steps: first, the one-dimensional DFT of every row of the panoramic image is calculated, then the frequency components are stored in a matrix, line by line. Only a subset of the columns in the matrix is retained: this corresponds to the lower spatial frequencies, and functions as a signature for the image. This matrix can be decomposed into a magnitude matrix and a phase matrix.

The motion of the robot can be separated into translation and rotation components. When using a panoramic image to represent the environment of the robot, the rotational component of the motion corresponds to a horizontal shift in the image. One of the most important properties of the DFT is that the horizontal shifts between two panoramic images cause only a linear phase shift, and no magnitude shift, when working on each row of the images. As a result, the relative orientation of the robot can be estimated by computing the phase shift between two DFT descriptors. The FS configurable parameter is the number of the Fourier coefficients saved as the signature: this parameter is used to control the computational cost and accuracy.

#### **5.2.4.2 A Principal Components Analysis descriptor**

Principal Components Analysis (PCA) finds the principal components of data by calculating the eigenvalues and eigenvectors of the covariance matrix. Each image can be treated as a vector: PCA is able to linearly project a high-dimensional image onto a low-dimensional subspace, retaining only the principal image components, as mentioned earlier in Section 2.2.3.2. However, the standard PCA descriptor obtained is not robust to robot orientation changes. In order to remedy this weakness, Jogan and Leonardis [2000] proposed a representation which simulates all possible rotations (for example  $N$ ) for the robot when collecting one image at each location. First, a set of  $N$  artificially-rotated images is created from each original panoramic image, which generates  $N$  data vectors per original image. The rotational PCA projection is then performed for  $N$  data vectors, forming the final representation of each location. Rotating the image is equivalent

---

to phase shifting its principal component coefficients: this fact can be used for orientation estimation by simulating the projections of all the rotations ( $N$ ). It is worth noting that in (Payá et al. [2014]), due to the tremendous computational and memory burden when processing the whole dataset, only 200 images were chosen to carry out the experiments. The variable parameters are the numbers of artificial image rotations.

#### **5.2.4.3 A Histograms of Oriented Gradients descriptor**

A Histograms of Oriented Gradients (HOG) descriptor was introduced by Dalal and Triggs [2005]. The essential thought behind this technique is that the appearance and shape of local objects in an image can be characterized by the distribution of intensity gradients, or edge directions. The implementation of these descriptors relies on the following stages: the image is first divided into small cells, which can be either rectangular or radial; for each cell, a histogram of oriented gradients over the pixels of the cell is accumulated; a histogram over a larger region (block) is accumulated; and the normalisation is computed over all of the cells in this block, introducing better invariance to illumination and shadowing. The final descriptor is represented by the combination of this set of histograms.

In (Payá et al. [2014]), HOGs of panoramic images are built in both the horizontal and vertical directions, yielding two descriptors. The first descriptor is obtained by dividing the panoramic image into horizontal cells and accumulating the histograms, while the second is built by dividing the panoramic image into vertical cells with overlapping. The orientation can be calculated by comparison of vertical block descriptors between the test image and the nearest image in the map. The variable parameters of the descriptor for orientation estimation are the numbers of horizontal cells.

---

#### 5.2.4.4 A Gist descriptor

The study in (Oliva and Torralba [2001, 2006]) shows that humans have an ability to rapidly recognize and understand the meaning (“gist”) of complex visual scenes, where the gist refers to the structural information about the scene layout, and provides a low dimensional representation of a scene. A Gist descriptor was originally proposed in (Oliva and Torralba [2001]): this was also termed the Spatial Envelope of a scene.

The procedure for building a Gist descriptor is as follows. First, the image is convolved with a set of Gabor filters ( $k$ ) at different orientations and scales, producing  $k$  feature map; Each feature map is then divided into  $N \times N$  non-overlapping blocks, and the average of feature values in each block is computed. Finally, the averaged values of all feature maps are concatenated, resulting in a Gist descriptor. Dimension reduction is sometimes performed, using PCA.

In (Payá et al. [2014]), the descriptor of the panoramic image was constructed around the Gist concept, which promotes invariance against rotations on the ground plane. A low-pass Gaussian filter was employed to downsample the image and an image pyramid was then built, describing image properties at different orientations and scales. The technique used to divide the images into a number of blocks was similar to that used in the HOG method: i.e., both horizontal and vertical division. Two descriptors are obtained after the blockification process: these are used for localisation and orientation estimations, respectively, as in HOG. The variable parameter of the descriptor for orientation estimation is the number of orientations per scale used by the Gabor filter.

### 5.3 Experiments

In this section, experimental results from testing different methods for orientation estimation are shown. Three datasets have been used for testing: these are the GummyBear, ISL and COLD datasets. For convenience, the visual compass, the quadtree based method and the SIFT feature-based method are abbreviated as

---

VC, QT and SIFT, respectively.

### 5.3.1 Outdoor experimental results: GummyBear dataset

For both the QT and SIFT methods we estimate the orientation in two ways. The first uses the first image as a reference, from which the orientation is calculated. The second uses a moving reference image, and accumulates changes in orientation. In the second instance we present results skipping a fixed number of images between pairs. Figure 5.3 illustrates the procedure of computing orientation based on these two techniques. The VC method uses a moving reference with automatically adjusted skips: therefore, it is compared to the methods using a fixed reference image. We compared the results of all methods with ground truth data from post-processed RTK GPS data. Table 5.1 gives quantitative results for all cases.

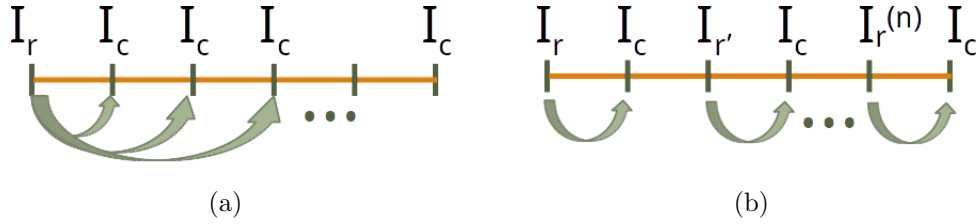


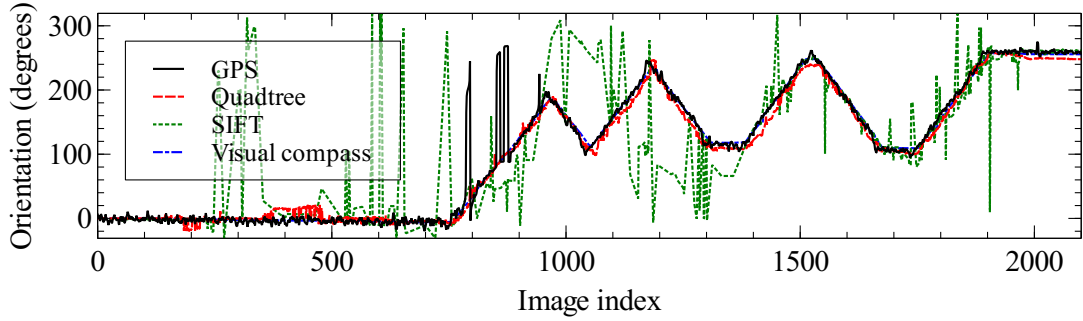
Figure 5.3: Simplified illustration of alternative techniques for computing orientation using: (a) a fixed reference image; and (b) a moving reference image.

Figures 5.4 to 5.6 show the results for the three methods with a fixed reference image for QT and SIFT (the VC method uses a moving reference image, but this is internal to the method and not exposed). These show that both appearance-based methods perform well and consistently for the whole path of the robot. The feature-based method performs well when the images are close to the reference image, but poorly when separated by many frames in which no features were found to match. This might be due to a lack of matched SIFT features, since distortions are unavoidably introduced by the parabolic mirror and the relatively low resolution of the images, as we can see in the following examples. Indeed, the orientation could not be calculated at all for many frames using the SIFT

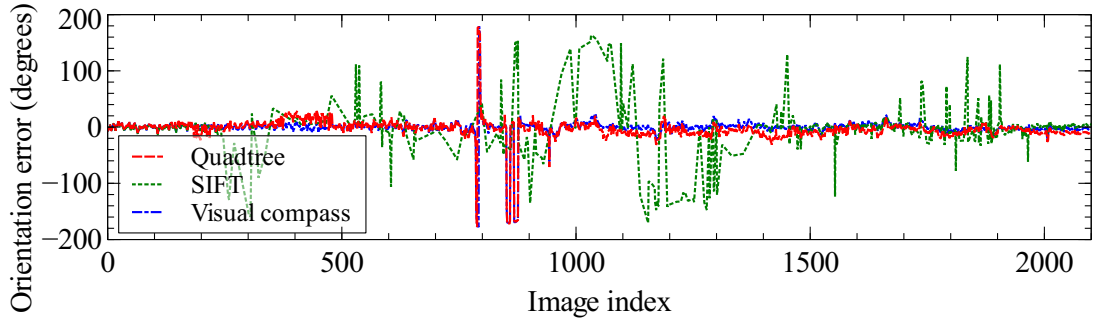
Table 5.1: Mean Error, Mean Absolute Error and Standard Deviation Error for the GummyBear dataset (F: FIELD; C: CARPARK; T: TENERIFE; <sub>f</sub>, <sub>5</sub> and <sub>20</sub> following SIFT and QT, respectively, indicate orientation estimation based on a fixed reference, and a moving reference with the corresponding skip value.)

Method	Mean Error			MAE			SD		
	F	C	T	F	C	T	F	C	T
VC	<b>4.63</b>	-1.33	-24.47	<b>8.21</b>	<b>6.17</b>	28.77	<b>10.43</b>	19.47	33.12
SIFT <sub>f</sub>	10.61	-3.57	<b>-2.89</b>	15.45	18.62	<b>10.94</b>	21.70	40.52	21.63
SIFT <sub>5</sub>	-35.28	-48.82	-69.55	40.63	53.53	71.28	39.68	47.38	62.52
SIFT <sub>20</sub>	-17.01	48.63	-5.36	84.71	85.25	84.03	105.30	96.87	107.69
QT <sub>f</sub>	16.00	<b>-0.76</b>	-11.59	20.54	11.36	15.39	25.30	<b>16.60</b>	<b>20.88</b>
QT <sub>5</sub>	-38.42	-55.79	9.70	43.10	59.15	55.86	49.73	59.32	78.76
QT <sub>20</sub>	5.90	5.66	-10.00	<b>8.67</b>	11.74	18.66	<b>11.10</b>	23.38	22.39

method: 36% in the case of CARPARK, 62% for FIELD, and 67% for TENERIFE.

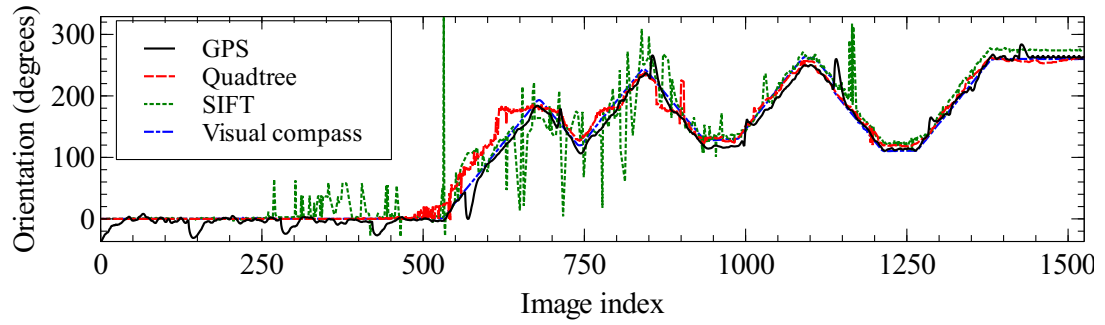


(a) Orientation estimation and ground truth

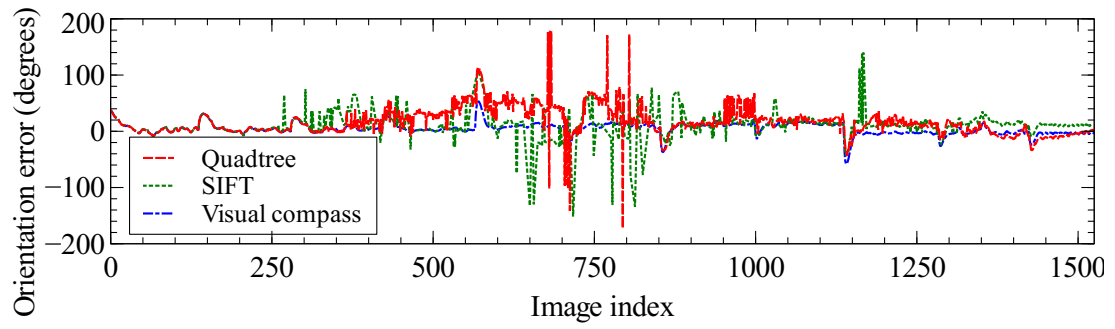


(b) Orientation error from ground truth

Figure 5.4: Experimental results for dataset CARPARK, with a fixed reference image.

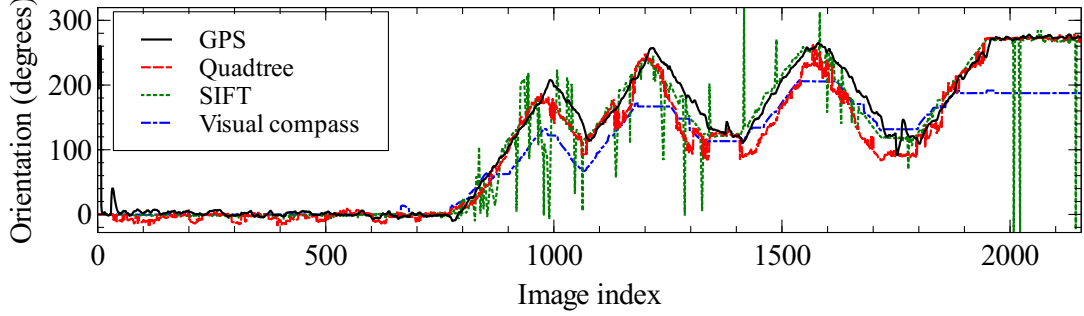


(a) Orientation estimation and ground truth

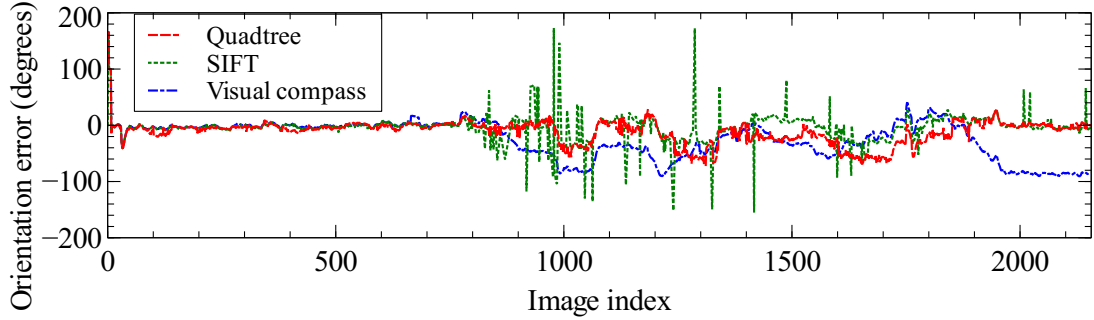


(b) Orientation error from ground truth

Figure 5.5: Experimental results for dataset FIELD, with a fixed reference image.



(a) Orientation estimation and ground truth



(b) Orientation error from ground truth

Figure 5.6: Experimental results for dataset TENERIFE, with a fixed reference image.

Two example results are given in Figures 5.7 and 5.8 after applying the SIFT method. In the first example, the number of detected keypoints from the image pair (Images 1399 and 1436) of the CARPARK dataset are 100 (Image 1399) and 90 (Image 1436), respectively, and the matches are 5. From the matching results shown in Figure 5.7 (bottom), we can see that the SIFT features can be matched correctly between two frames that are relatively closely spaced in the environment. In the second example, the number of detected keypoints from the image pair (Images 0 and 570) of the CARPARK dataset are 136 (Image 0) and 128 (Image 570), and the match is only 1: in fact, this match connects two points that do not actually correspond in the world. Therefore, when spatially distant reference frame is used, the SIFT method fails to find sufficient correct matches for orientation computing.



Figure 5.7: Top: an example image pair (Image 1399: upper, Image 1436: lower) from the CARPARK dataset. Bottom: SIFT matching result.

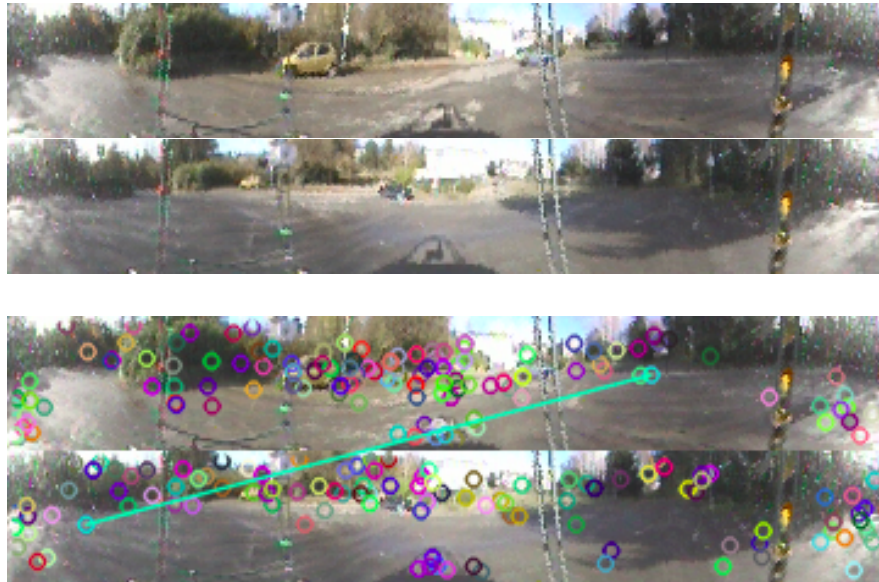
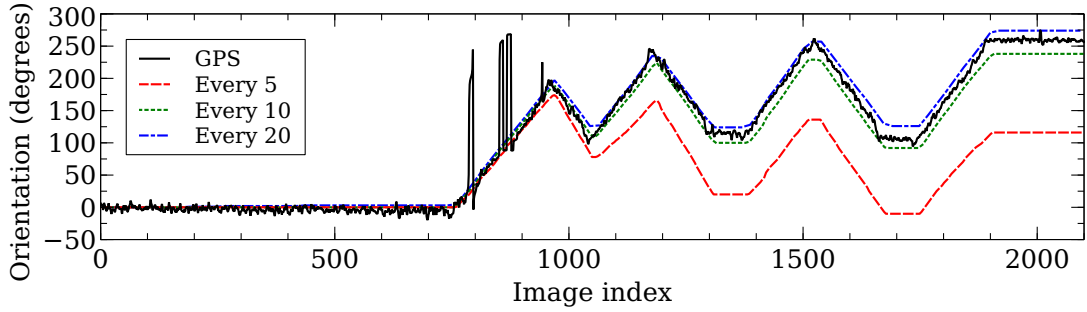


Figure 5.8: Top: an example image pair (Image 0: upper, Image 570: lower) from the CARPARK dataset. Bottom: SIFT matching result.

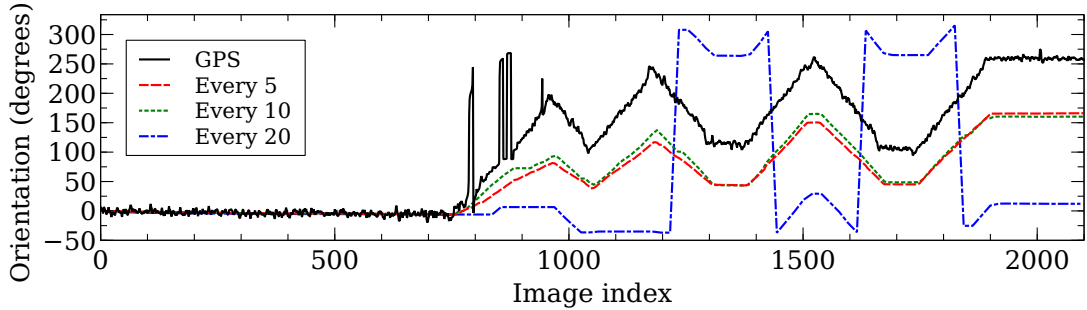
The SIFT method performs better than both appearance-based methods on the TENERIFE dataset. This is because the boundary between sky and land is very

strong, not visible all around the robot, and slanted with respect to the horizontal. Alignment of the images using pixel values will therefore tend to align the skyline, introducing a bias due to the slant.

Figures 5.9 to 5.11 show the results for the incremental QT and SIFT methods that use a moving reference. For both methods, pairs were created by skipping a fixed number of images, and results are given for different values of the number of images skipped (5, 10 and 20).



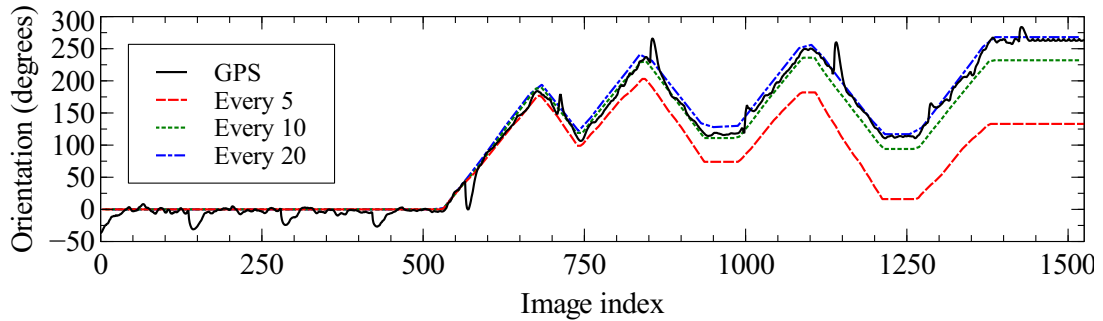
(a) Orientation from the QT method



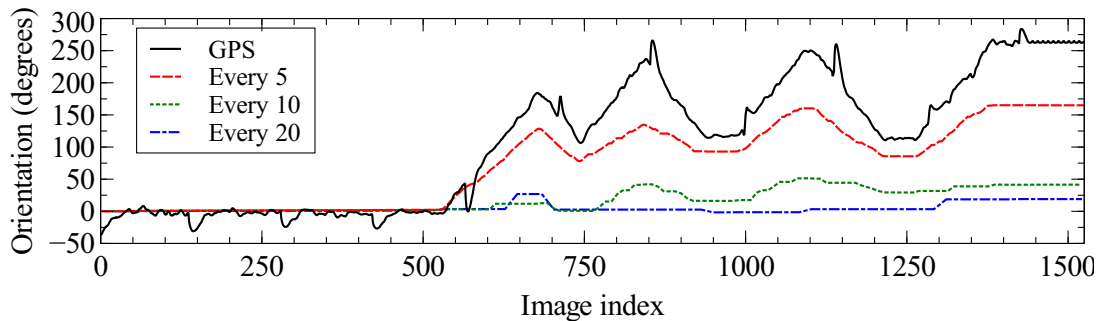
(b) Orientation from the SIFT method

Figure 5.9: Experimental results for dataset CARPARK, with a moving reference image.

These results clearly show that choosing the correct compromise between better short term rotation estimation and frequently-accumulating error is critical. In fact, none of these results are as good as that of the VC. This is due to the subpixel processing and the automatic, adaptive estimation of the best compromise performed in the VC. Nevertheless, the QT method performs similarly to



(a) Orientation from the QT method



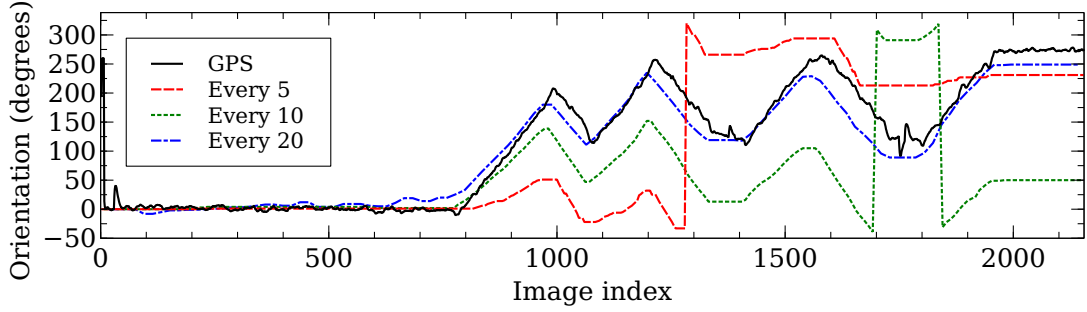
(b) Orientation from the SIFT method

Figure 5.10: Experimental results for dataset FIELD, with a moving reference image.

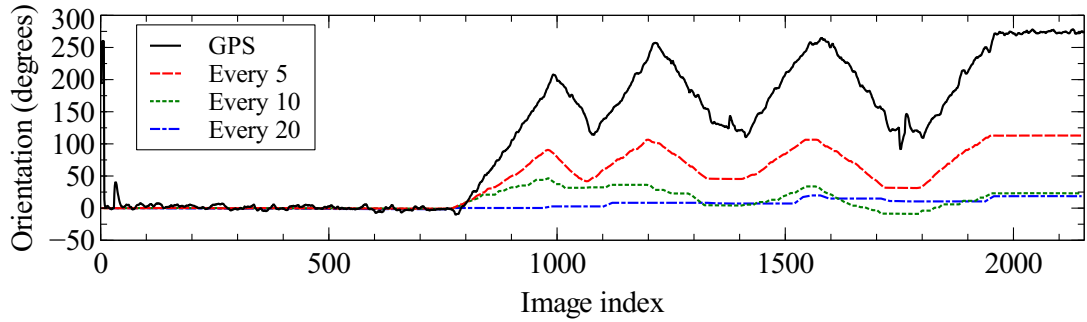
the SIFT method if orientation is accumulatively calculated by skipping more images. This is in line with the fact that the SIFT method performs better when the reference image is not too different from the processed images.

### 5.3.2 Indoor experimental results: ISL dataset

Both QT and SIFT methods were tested for performance in orientation estimation in the same manner used for the outdoor experiments: the current orientation of the robot is calculated using its previous orientation, and by accumulating changes in orientation (see Figure 5.3). The VC method uses exactly the same techniques as in the outdoor experiments. Ground truth was available through a VICON system.



(a) Orientation from the QT method



(b) Orientation from the SIFT method

Figure 5.11: Experimental results for dataset TENERIFE, with a moving reference image.

As mentioned in Chapter 3, the ISL dataset consists of four sub-datasets, each comprising three complete loops in dynamic environments that contain both static and moving obstacles. We use each sub-dataset as three independent datasets to evaluate the performance of orientation estimation, and test repeatability of all methods.

The quantitative results for four sub-datasets are summarized in Tables 5.2 to 5.5 that list the mean error, mean absolute error and standard deviation error. In the last column of each table, we present the maximum Mahalanobis distance of errors between each of two loops. This value reveals the difference in performance between each independent test, and allows us to evaluate the repeatability of orientation estimation accuracy for each method.

Table 5.2: Mean Error, Mean Absolute Error, Standard Deviation Error and Maximum Mahalanobis Distance for ISL 1 (\_3, \_5 and \_10 following SIFT and QT, respectively, indicate orientation estimation based on a moving reference with the corresponding skip value: these apply to the following tables).

Method	Mean Error			MAE			SD			MMD
	L1	L2	L3	L1	L2	L3	L1	L2	L3	
VC	5.80	8.63	8.09	7.30	11.35	<b>9.82</b>	6.59	9.07	<b>8.68</b>	0.36
SIFT_3	10.42	8.56	13.09	11.69	11.21	15.89	8.67	9.84	20.52	<b>0.34</b>
SIFT_5	11.27	6.7	8.74	12.59	9.77	12.55	9.10	9.24	13.32	0.50
SIFT_10	11.70	<b>5.73</b>	14.23	12.78	<b>8.80</b>	14.69	7.87	<b>8.66</b>	9.09	0.98
QT_3	-12.76	-30.63	-30.80	12.76	30.63	30.80	4.63	13.54	14.77	1.78
QT_5	<b>-2.69</b>	-17.61	-17.81	<b>3.49</b>	17.61	17.81	<b>3.41</b>	12.57	17.94	1.63
QT_10	15.67	-9.99	<b>1.08</b>	16.79	18.55	19.91	13.13	23.7	24.12	1.37

Table 5.3: Mean Error, Mean Absolute Error, Standard Deviation Error and Maximum Mahalanobis Distance for ISL 2.

Method	Mean Error			MAE			SD			MMD
	L1	L2	L3	L1	L2	L3	L1	L2	L3	
VC	3.92	5.02	8.13	5.49	<b>7.46</b>	8.74	5.10	<b>7.30</b>	<b>5.17</b>	0.82
SIFT_3	13.98	25.11	27.67	18.72	26.38	28.00	20.94	18.87	22.10	0.68
SIFT_5	17.16	18.96	34.37	18.38	20.42	34.51	13.83	14.26	21.75	0.95
SIFT_10	27.40	36.04	29.55	28.78	36.42	29.83	22.14	14.71	19.08	<b>0.47</b>
QT_3	<b>-1.26</b>	-12.15	-5.50	6.06	12.15	<b>7.51</b>	16.55	9.82	9.33	1.58
QT_5	3.03	-11.35	<b>2.19</b>	<b>4.46</b>	11.56	11.19	<b>4.29</b>	11.79	13.17	1.62
QT_10	15.06	<b>1.26</b>	-13.16	15.87	17.44	18.27	11.96	22.04	21.86	1.63

Table 5.4: Mean Error, Mean Absolute Error, Standard Deviation Error and Maximum Mahalanobis Distance for ISL 3.

Method	Mean Error			MAE			SD			MMD
	L1	L2	L3	L1	L2	L3	L1	L2	L3	
VC	8.84	13.69	8.87	10.15	14.56	10.42	<b>8.06</b>	<b>8.00</b>	<b>7.42</b>	0.62
SIFT_3	23.34	17.74	17.53	25.65	18.37	19.03	19.42	10.79	13.88	0.43
SIFT_5	23.30	17.98	16.45	24.08	18.62	17.89	17.00	10.25	13.38	0.45
SIFT_10	26.07	34.74	20.13	26.84	34.74	20.74	18.73	17.53	12.50	0.98
QT_3	<b>0.95</b>	<b>-2.19</b>	-7.25	<b>6.60</b>	<b>7.50</b>	<b>8.27</b>	9.54	9.56	9.39	1.10
QT_5	12.52	6.32	<b>5.23</b>	13.14	13.51	13.51	8.33	14.18	15.32	0.60
QT_10	-7.80	-9.02	-11.68	9.00	19.00	19.06	8.84	23.95	23.11	<b>0.23</b>

Table 5.5: Mean Error, Mean Absolute Error, Standard Deviation Error and Maximum Mahalanobis Distance for ISL 4.

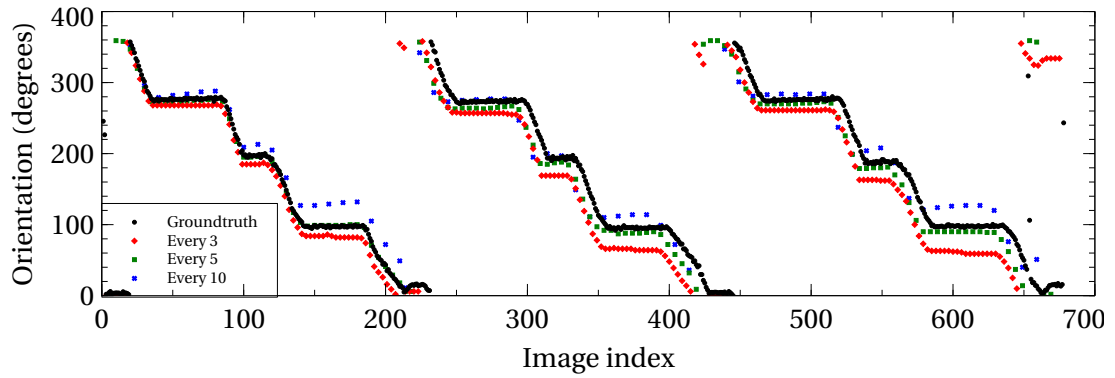
Method	Mean Error			MAE			SD			MMD
	L1	L2	L3	L1	L2	L3	L1	L2	L3	
VC	7.57	<b>7.49</b>	11.72	8.98	<b>8.63</b>	13.23	<b>6.96</b>	<b>5.75</b>	<b>8.09</b>	0.62
SIFT_3	22.15	19.60	16.56	23.12	20.38	17.92	17.08	11.89	12.45	0.37
SIFT_5	17.70	16.96	20.68	18.63	17.92	21.41	11.32	11.41	11.89	<b>0.32</b>
SIFT_10	15.70	15.03	24.20	16.37	15.87	24.84	10.31	9.96	13.71	0.80
QT_3	<b>4.04</b>	9.42	<b>-2.50</b>	8.66	12.73	<b>8.55</b>	11.23	11.61	10.72	1.07
QT_5	13.42	13.15	-4.09	14.70	17.05	13.08	12.73	14.69	16.11	1.24
QT_10	5.94	9.75	-23.08	<b>7.48</b>	21.93	25.72	9.41	27.19	26.75	1.55

Figure 5.12 and Table 5.2 show the results for the three methods on the dataset ISL 1. These show that there is a certain amount of drift for all methods. The QT method performs best when we choose the interval of five images to accumulate changes in orientation, as its results show little drift and is closest to the ground truth. Regarding repeatability, the VC and the SIFT methods both perform well and better than the QT method, and obtain consistent experimental results.

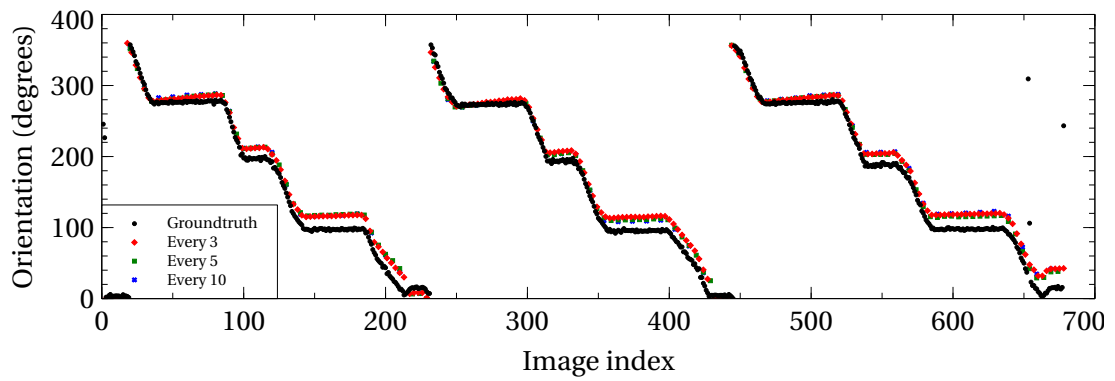
The results over dataset ISL 2 are shown in Figure. 5.13 and Table 5.3. The QT and the VC methods both perform well, while the SIFT method suffers from strong drift, although it has the best repeatability according to the maximum

---

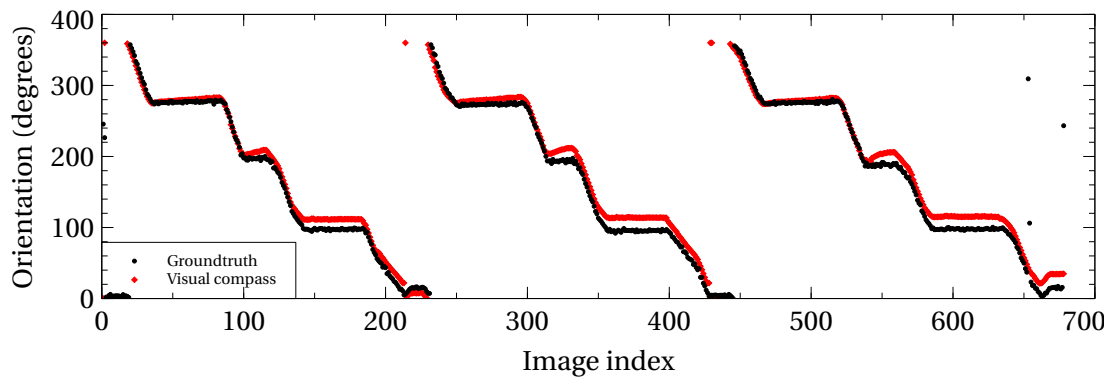
Mahalanobis distance value. Figure 5.14 and Table 5.4 show the results for the three methods on the dataset ISL 3. The performance of the quadtree method is surprisingly good, both in accuracy and repeatability, followed by that of the VC. Figure 5.15 and Table 5.5 show the results on the dataset ISL 4. We can see that the appearance-based method performs better than the SIFT method. The VC and the SIFT method both show consistency of experimental results.



(a) Orientation from the QT method

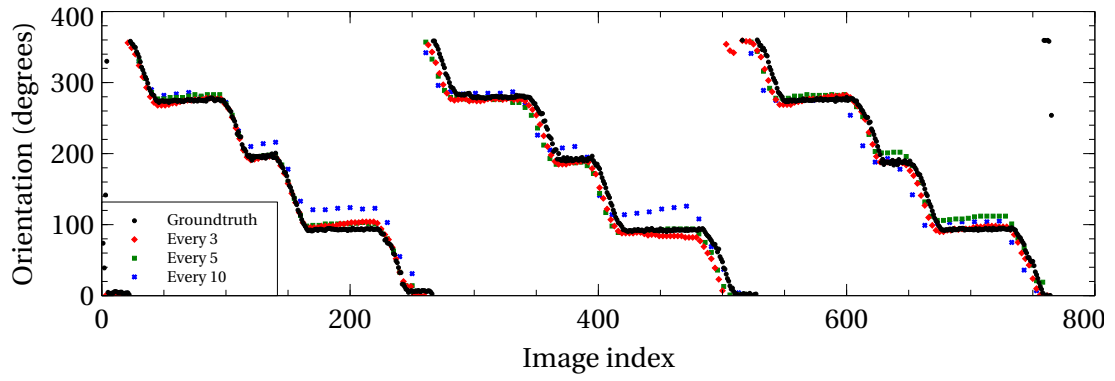


(b) Orientation from the SIFT method

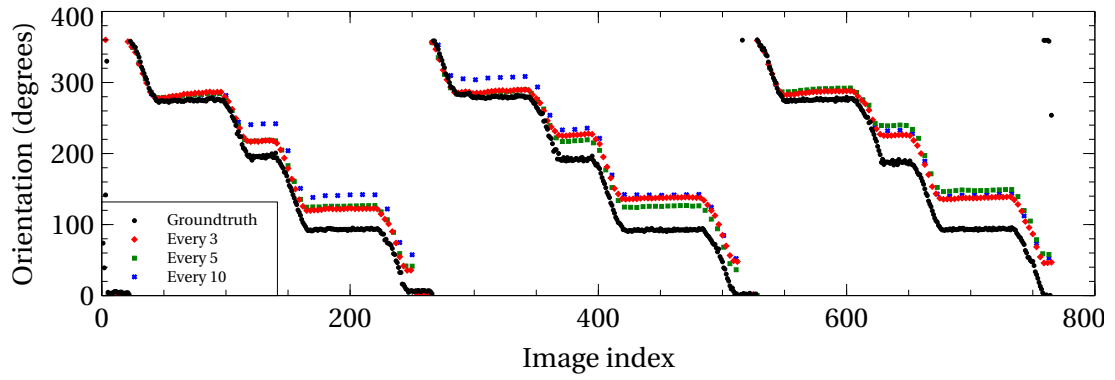


(c) Orientation from the VC method

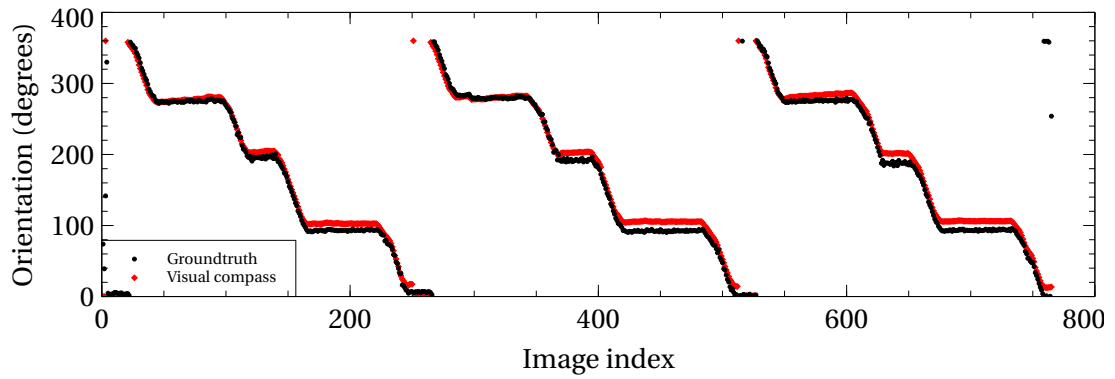
Figure 5.12: Experimental results for dataset ISL 1, with a moving reference image.



(a) Orientation from the QT method

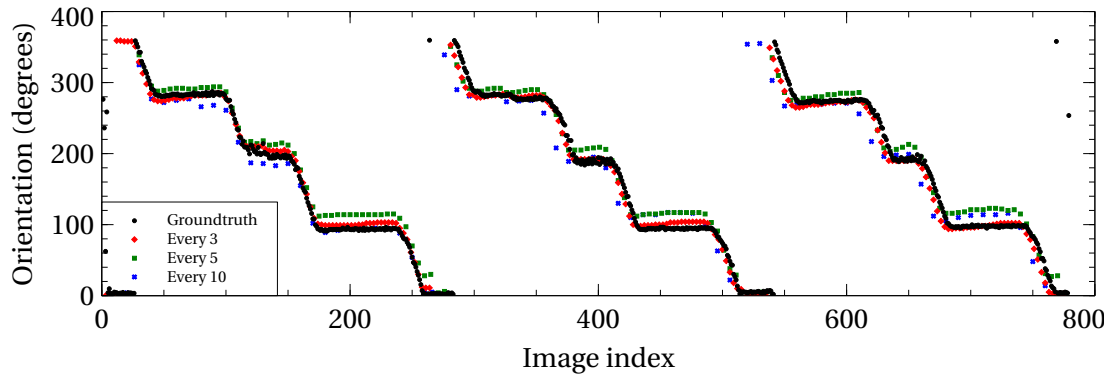


(b) Orientation from the SIFT method

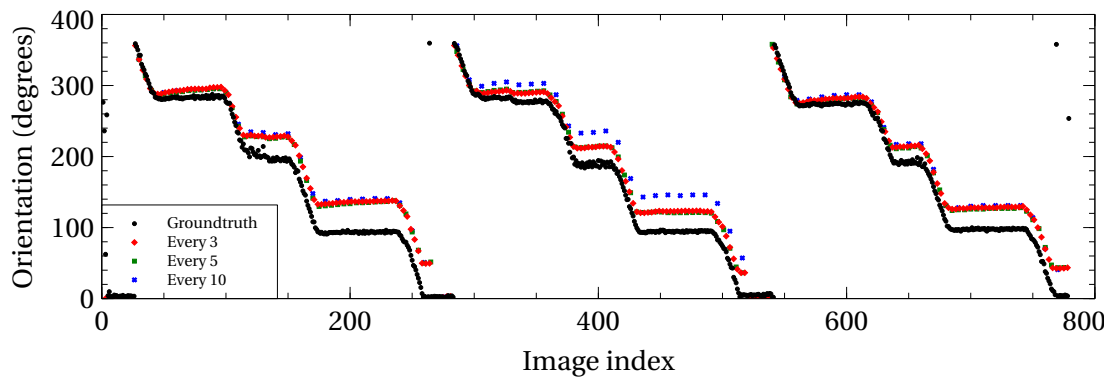


(c) Orientation from the VC method

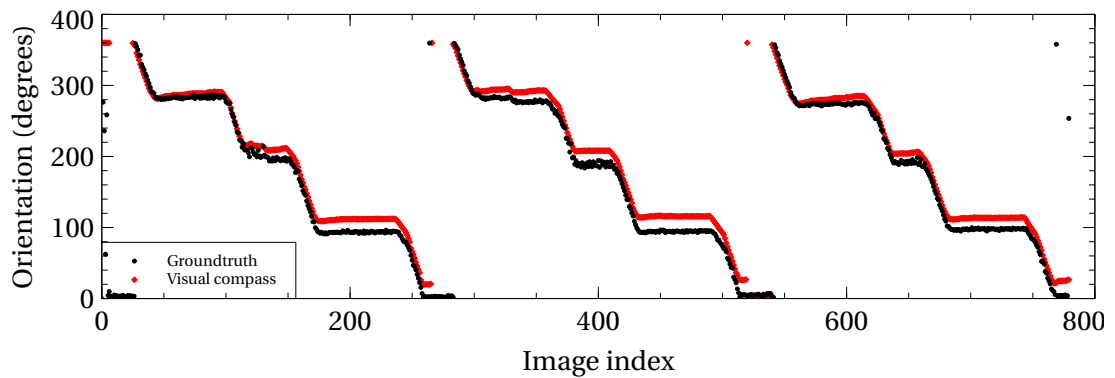
Figure 5.13: Experimental results for dataset ISL 2, with a moving reference image.



(a) Orientation from the QT method

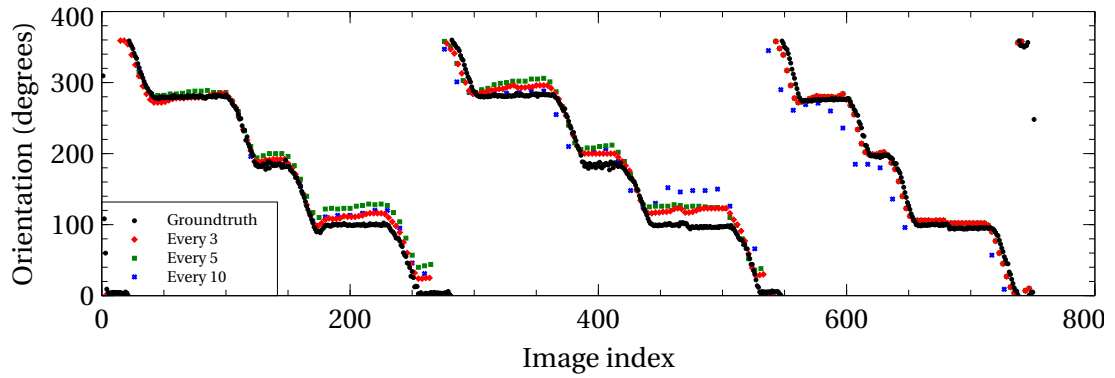


(b) Orientation from the SIFT method

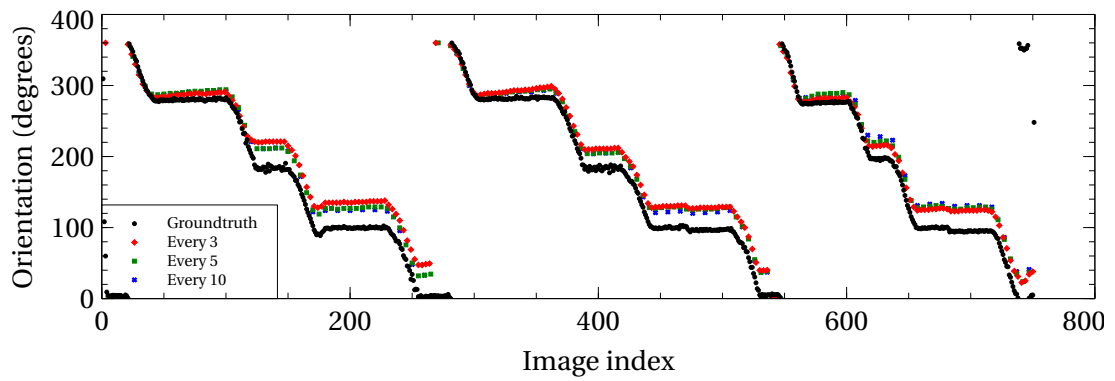


(c) Orientation from the VC method

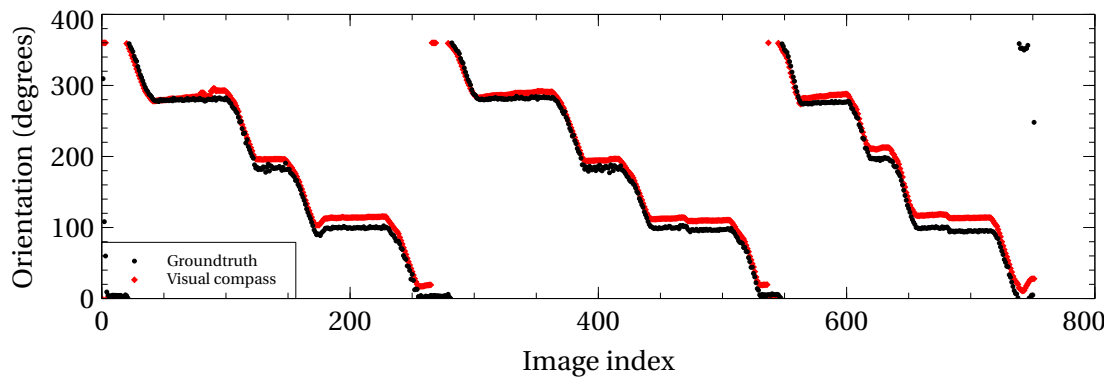
Figure 5.14: Experimental results for dataset ISL 3, with a moving reference image.



(a) Orientation from the QT method



(b) Orientation from the SIFT method



(c) Orientation from the VC method

Figure 5.15: Experimental results for dataset ISL 4, with a moving reference image.

---

In general, the different methods achieve similar performances on the ISL 1 dataset. For the ISL 2, ISL 3 and ISL 4 datasets, we found that using every third frame for estimating pairwise relative orientation worked best for the QT method, and its performance is slightly superior to that of the VC. It should be noted that the worst performance occurs with the SIFT method on the ISL 2, ISL 3 and ISL 4 datasets. It will be observed that the performance of the SIFT method decreased dramatically at the same point of different loops, where the robot starts to rotate and moves onto the other side of workspace. Examining the images taken around that location shows that a large red piece of furniture is present on the left of the image: as the robot moves, this piece of furniture is progressively absent on the left, and present on the right of the image. As described in Chapter 3, a large obstacle was introduced in the centre of the workspace for the ISL 2, ISL 3 and ISL 4 datasets: this is visible in the middle of the images. Due to the intervention of this obstacle, and discontinuous changes in the appearance caused by the furniture, the SIFT method fails to find the correct feature matches for orientation computing purposes. Moreover, since the orientation is calculated from the previous calculated orientation, the error is compounded over time. As a result of the factors mentioned above, the SIFT method was judged to give the least satisfactory performance when considering results on the ISL 2, ISL 3 and ISL 4 datasets.

### 5.3.3 Indoor experimental results: COLD dataset

The experiment presented here is inspired by (Payá et al. [2014]), which compares the performance of four types of global-appearance descriptors for map creation and localisation tasks. In (Payá et al. [2014]), the relative orientation between the current image and the images saved in the database is estimated to test the performance of the descriptors. We performed similar experiments using the QT, SIFT and VC methods on the same dataset (the Freiburg sub-dataset of the COLD dataset). As mentioned in Chapter 3, the raw omnidirectional images in the ISL and GummyBear dataset were unwrapped into  $360 \times 40$  and  $360 \times 55$  pixels panoramic images, respectively: this makes the angular resolution 1 pixel

---

per degree, such that the best shift between an image pair is directly equal to the relative rotation angle undertaken by the camera. It should be noted that the COLD dataset exploited in our research is a sequence of omnidirectional images with a resolution of  $640 \times 480$  pixels, which are then unwrapped for a  $360 \times 40$  pixels panoramic image for fair comparison with our collected datasets.

For the QT and SIFT methods, we compute a relative orientation between all image pairs of the dataset: each pair of images is chosen between the two consecutive images, as well as skipping one and two images. For the VC method, it is important to note that instead of using a reduced Field Of View (FOV) around the front and back of the omnidirectional camera, we used a wide FOV ( $100^\circ$ ) around the front of the camera only in order to remove visible intrusion by the support of the camera. We have shown the quantitative results for the three methods in Table 5.6.

As can be seen from Table 5.6, all methods have near-zero mean errors. This led to an investigation of the statistical difference of the mean errors away from zero. We tested the null hypothesis that the mean error is equal to zero, using one sample t-tests ( $\alpha = 0.05$ ). The statistical significance results are presented in parentheses, following the mean error values, in Table 5.6. It should be noted that the mean error of the SIFT\_1 method (which exploits consecutive images for relative rotation estimation) over the Night dataset is statistically different from zero, with a p-value of less than 0.05, while the others do not differ statistically significantly from zero. From the t-tests, we can conclude that all methods other than the SIFT\_1 method resulted in a similar average error (zero).

Table 5.6: Mean Error ( $p$ -Value), Mean Absolute Error, and Standard Deviation Error of relative orientation estimation, using the COLD dataset.

Method	Mean Error ( $p$ -Value)			MAE			SD		
	Sunny	Cloudy	Night	Sunny	Cloudy	Night	Sunny	Cloudy	Night
VC	0.03 (0.81)	0.07 (0.33)	0.10 (0.11)	2.41	1.17	0.66	5.29	2.68	<b>1.75</b>
SIFT_1	-0.04 (0.54)	-0.05 (0.45)	-0.15 ( <b>0.00</b> )	<b>1.22</b>	1.12	0.75	<b>2.76</b>	2.45	2.01
SIFT_2	<b>-0.01</b> (0.96)	-0.19 (0.25)	-0.24 (0.06)	2.30	2.01	1.49	5.06	4.38	3.95
SIFT_3	-0.18 (0.55)	-0.21 (0.45)	-0.39 (0.06)	3.19	2.65	2.03	6.92	6.18	5.27
QT_1	0.21 (0.36)	<b>0.02</b> (0.80)	-0.04 (0.32)	1.73	<b>1.03</b>	<b>0.60</b>	9.01	<b>2.25</b>	1.89
QT_2	-0.07 (0.66)	0.03 (0.81)	<b>0.01</b> (0.93)	2.40	1.89	1.13	4.72	3.57	2.91
QT_3	-0.08 (0.74)	-0.06 (0.78)	-0.02 (0.84)	3.15	2.55	1.61	5.54	4.43	2.87

From Table 5.6, we also can see that the SIFT\_1 method performs well on the Sunny dataset (with the lowest SD error), while the QT and VC methods perform well on the Cloudy and Night datasets. The major characteristic of images in the Sunny dataset is severe variations in illumination, which greatly affect the appearance of a room as a result of changes in highlights, shadows and reflectance. Moreover, all images were acquired with auto-exposure, leading to a decrease in contrast in the images. Appearance-based algorithms involving the direct comparison of pixels of images based purely on a Euclidean distance metric have difficulties in dealing with these images. This may explain why they perform less well than the SIFT method on the Sunny dataset.

The results from (Payá et al. [2014]) are presented in Table 5.7. From this we can see that the Rotational PCA descriptor-based method achieved marginally the best result among the four descriptors, providing a mean error of 0.75 and a standard deviation error of 1.3. This indicates that most of the estimated orientation errors fall within  $-0.41^\circ$  and  $2.41^\circ$ . It is noteworthy that only 200 images were used for the Rotational PCA-based method, while the whole dataset was utilized to evaluate the other three methods. Additionally, the Rotational PCA descriptor-based mapping scheme applied in (Payá et al. [2014]) cannot be created incrementally: in this case, a complete map of the environment must be available before the navigation, which limits the autonomy of the robot. Therefore, this method may be not appropriate for some realistic robotics tasks such as vSLAM. Another weakness of this method is the higher computational load involved. As

---

described in subsection (Subsection 5.2.4.2), in order to make localisation insensitive to in-plane orientation of the robot, a number of rotated panoramic images that all represent a single location are created, which encode the varying orientations of the robot. PCA is then applied on these spinning images to obtain the PCA subspace, and this incurs higher computational cost. However, the other three descriptors do not have the above-mentioned disadvantages. According to Table 5.7, similar results were obtained by the other three descriptors, and the the estimation mean errors can be limited to approximately one degree if the parameters are tuned properly.

Table 5.7: Mean and Standard Deviation Error of relative orientation estimation using different descriptors on the COLD dataset in (Payá et al. [2014]). Note that the mean presented here is the smallest mean: or mean with the smallest standard deviation, when the means are identical.

Descriptor	Mean	SD
Fourier Signature	1	1.41
Rotational PCA	0.75	1.32
HOG	1	1.73
Gist	1	1.41

Comparing the results in Table 5.7 with the results of Table 5.6, it may be seen that, when working on the Cloudy and Night datasets, the SIFT\_1 and QT\_1 methods and the VC method when applied to the Night dataset alone, result in the smaller mean error, tending to zero, and a slightly larger standard deviation error. However, the small difference between the mean error and the standard deviation makes it very difficult to draw conclusions concerning the difference in accuracy between the various approaches. Further studies are necessary to investigate whether there are statistically significant differences between these approaches.

---

### 5.3.4 Indoor experimental results: COLD dataset (based on HS colour space and log transformation)

In this subsection, we evaluate the performance of the proposed method (QT) for estimating the pairwise relative orientation between frames, based on the COLD dataset, but in the HSV (Hue, Saturation, and Value) colour space rather than RGB colour space. In addition, the logarithmic transformation of the COLD dataset has been used for evaluation of the QT method. The experiments were conducted in exactly the same way as in the previous subsection (Subsection 5.3.3). Orientation estimates are obtained between each pair of images created by choosing two consecutive frames, skipping one frame, and skipping two frames.

#### 5.3.4.1 Experimental results: HS colour space

HSV is a perceptual colour space, designed by Smith [1978]. It is defined in a way that is similar to human perception, which separates luminance from colour information. This is very useful in many applications, such as tracking, human detection, and medical image processing. In this experiment, the value component of HSV is discarded, which determines the image brightness. The conversion from RGB to HSV colour space is defined by the following equations mentioned by Ososinski and Labrosse [2013]. For  $0 \leq R, G, B \leq 1$ ,

$$\begin{aligned} M &= \max(R, G, B), \\ m &= \min(R, G, B), \\ C &= M - m. \end{aligned} \tag{5.1}$$

---

With this, HSV is defined as

$$H = 60^\circ \times \begin{cases} \text{undefined} & \text{if } C = 0, \\ \frac{G-B}{C} \bmod 6 & \text{if } M = R, \\ \frac{B-R}{C} + 2 & \text{if } M = G, \\ \frac{R-G}{C} + 4 & \text{if } M = B, \end{cases} \quad (5.2)$$

$$S = \begin{cases} 0 & \text{if } C = 0, \\ \frac{C}{V} & \text{otherwise,} \end{cases}$$

$$V = M.$$

Table 5.8: Mean Error ( $p$ -Value), Mean Absolute Error, and Standard Deviation Error of relative orientation estimation, using the COLD dataset in HS colour space.

Method	Mean Error ( $p$ -Value)			MAE			SD		
	Sunny	Cloudy	Night	Sunny	Cloudy	Night	Sunny	Cloudy	Night
QT_1	0.21 (0.99/0.42)	0.10 (0.83/0.79)	-0.19 ( <b>0.03/0.00</b> )	2.17	2.90	0.77	10.22	14.07	2.32
QT_2	0.13 (0.67/0.79)	0.63 (0.42/0.39)	-0.41 ( <b>0.02/0.00</b> )	4.41	5.20	1.64	14.07	19.75	4.52
QT_3	0.88 (0.27/0.30)	0.47 (0.54/0.57)	-0.61 ( <b>0.04/0.01</b> )	6.68	5.94	2.53	19.36	18.15	6.59

Table 5.8 summarises the estimation accuracy with respect to different frame rates. We also made a statistical comparison of the QT method based on HS colour space and RGB colour space. We evaluate the difference of mean error between the HS model and RGB model using a two-sample t-test, with significance level of 0.05. The statistical significance results (p-value) are presented in parentheses (the left-hand number), following the mean error values, in Table 5.8. In addition, we tested the null hypothesis that the mean error of estimation based on the HS model is equal to zero, using one sample t-test, with significance level of 0.05. The statistical significance results (p-value) are presented in parentheses (the right-hand number), following the mean error values, in Table 5.8. The estimation results of the QT method, based on the RGB colour space are shown in Table 5.6 (on page 133).

---

For the Sunny dataset, the mean errors, based on the HS model, are 0.21, 0.13, and 0.88, with a standard deviation of 10.22, 14.07, and 19.36, and the corresponding mean errors, based on RGB model, are 0.21, -0.07, and -0.08, with a standard deviation of 9.01, 4.72, and 5.54. The p-values for the paired t-tests are 0.99, 0.67, and 0.27. It can be seen that the estimation accuracy of the HS model based method is statistically comparable with that of the RGB model based method for the Sunny dataset, when looking only at the mean error. However, when comparing the MAE (Mean absolute error) and SD error values of these two colour space based QT methods, we observe that the HS model based method yielded a less precise estimate than the RGB model based method, as the HS model based method produced higher values of MAE and SD.

For the cloudy dataset, the mean errors, based on the HS model, are 0.1, 0.63, and 0.47, with a standard deviation of 14.07, 19.75, and 18.15, and the corresponding mean errors, based on the RGB model, are 0.02, 0.03, and -0.06, with a standard deviation of 2.25, 3.57, and 4.43. The p-values for the paired t-tests are 0.83, 0.42, and 0.54. As can be seen, there was no statistically significant difference between the mean errors of the compared methods. However, the HS model based method is less precise than the RGB model based method, giving a higher SD value.

For the Night dataset, the mean errors, based on the HS model, are -0.19, -0.41, and -0.61, with a standard deviation of 2.32, 4.52, and 6.59, and the corresponding mean errors, based on the RGB model, are -0.04, 0.01, and -0.02, with a standard deviation of 1.89, 2.91, and 2.87. The p-values for the paired t-tests are 0.03, 0.02, and 0.04. It can be seen that the estimation mean errors of the two methods are statistically significantly different. The p-values for the one sample t-test of the HS model based method on the Night dataset are all approximately equal to zero. This indicates that the mean error of the HS model based method is significantly different from zero.

In general, we found that ignoring the value (V) component of the HSV colour space does not improve the accuracy of orientation estimation, or the robustness of the QT method against illumination variations. Moreover, the HS model based QT method performed slightly worse than the RGB model based QT method on the Night dataset. This seems to be a consequence of the characteristics of the

---

Night dataset, which exhibits low luminance, drab colour, and low contrast. In addition, the projection of the higher dimensional RGB onto the lower dimensional HS model leads to loss of information. This might be another reason for the inferior performance of the HS model based QT method. From the above experimental validation, we may conclude that the QT method is more suited to the orientation estimation in an RGB colour space than in an HS colour space.

#### 5.3.4.2 Experimental results: log transformation

In this subsection, we apply the QT method to the logarithmic transformed COLD data, and make a statistical comparison between its performance and that of the RGB model based QT method on the orientation estimation task. The experiments were conducted in exactly the same way as in Subsection 5.3.3. Orientation estimates are obtained between each pair of images created by choosing two consecutive frames, skipping one frame, and skipping two frames, respectively.

The logarithmic transformation of an RGB image can be mathematically expressed as:

$$s = c * \log(1 + r) \quad (5.3)$$

where  $c$  is a constant,  $r$  is the original pixel value, and  $s$  is the resulting pixel value. The log transformation is a nonlinear transformation: it maps a narrow range of low pixel values in the input image into a wider range of output levels (Jain et al. [1995]). In our experiment  $c$  is set to 1. We apply the logarithmic transformation to each of the three colour components (R, G and B) separately. Figure 5.16 shows three images selected from COLD dataset. The left column shows the original images and the right column is the resulting log-transformed images.

Table 5.9 illustrates the mean error, mean absolute error, and standard deviation error of the relative orientation estimation, using the log-transformed COLD dataset. In addition, a paired t-test, with significance level of 0.05, has been utilised to evaluate the difference of mean error between the QT methods based

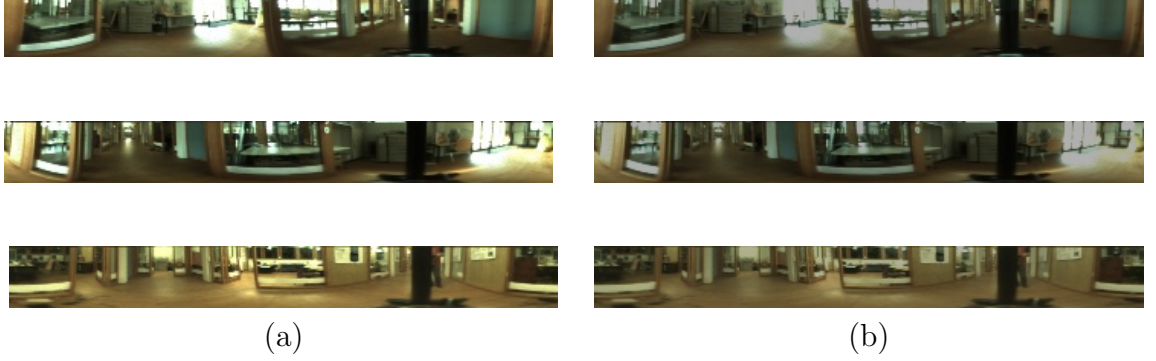


Figure 5.16: Example images from Sunny (top row), Cloudy (middle row), and Night (bottom row) dataset of COLD datasets, (a) original images, and (b) corresponding log-transformed images.

on the log-transformed data and the RGB model. The statistical significance results (p-values) are presented in parentheses (the left-hand number), following the mean error values, in Table 5.9. A one sample t-test, with significance level of 0.05, has also been used to test the null hypothesis that the mean error of estimation based on log-transformed data is equal to zero. The statistical significance results (p-values) are presented in parentheses (the right-hand number), following the mean error values, in Table 5.9.

Table 5.9: Mean Error ( $p$ -Value), Mean Absolute Error, and Standard Deviation Error of relative orientation estimation, using the COLD dataset after logarithmic transformation.

Method	Mean Error ( $p$ -Value)			MAE			SD		
	Sunny	Cloudy	Night	Sunny	Cloudy	Night	Sunny	Cloudy	Night
QT_1	-0.23 (0.34/0.29)	-0.21 ( <b>0.00/0.00</b> )	-0.34 (8.27/1.32)	1.20	0.96	0.65	5.71	2.27	2.15
QT_2	-0.15 (0.74/0.37)	-0.21 (0.20/0.12)	-0.32 ( <b>0.02/0.00</b> )	2.49	1.82	1.17	4.77	3.56	3.14
QT_3	-0.18 (0.77/0.44)	-0.10 (0.86/0.59)	-0.19 (0.37/0.20)	3.06	2.56	1.75	5.38	4.34	3.77

As we can see from the results in Table 5.9 and in Table 5.6 (on page 133), the QT\_1 method using the log-transformed Sunny dataset achieves slightly better performance than using the same images in RGB colour space, since it yields a lower SD value (5.71), while the two methods have statistically similar mean errors (p-value for the paired t-test is 0.34, i.e., greater than 0.05). For the Cloudy dataset, the QT\_1 method using log-transformed images achieved a worse

---

performance than when using images in an RGB colour space: the difference in mean error between these two methods is statistically significantly different (the p-value for the paired t-test is 0.007, i.e., smaller than 0.05). Similarly, worse results were obtained by the QT \_2 method on the log-transformed Night dataset. We can see clearly that with this method there are some overexposed areas and shadows in the images from the Sunny dataset. As we know, the log-transformation can improve contrast in a poor quality image, because low pixel values are mapped over a wider range, while the higher values are compressed. This might explain why the QT \_1 method produced better results on the log-transformed Sunny dataset. However, directly using logarithmic mapping on the individual R, G, and B channels has the effect of reducing the contrast between the colour channels, resulting in desaturation. This can be seen from the bottom column of Figure 5.16. It should be emphasised that the worst performances of the QT \_1 and QT \_2 methods were achieved on the Cloudy and Night datasets, respectively.

## 5.4 Conclusions

In this chapter we have evaluated three methods for robot orientation estimation with panoramic images, in order to determine what image-based techniques are suitable for this task. The outdoor experimental results show that the QT method performs better than the SIFT method when the distance between pairs of images becomes high, while the SIFT based method does well over short image separation distances. This implies that the appearance-based method is likely to work better at lower frame rates, and so be more appropriate for loop closure tasks, at least for orientation estimation. Moreover, the appearance-based methods (QT and VC) perform better than the feature-based method when the environment is visually variable, but not too contrasty. The experimental results on the ISL dataset show that drift is unavoidable for all methods over the course of a whole experiment. By comparison, the QT and VC methods suffer less from drift than the SIFT method. The results over the COLD dataset show that the QT and VC methods work well under stable illumination conditions, while the SIFT method works

---

better when there are significant illumination changes in the environments. The results from (Payá et al. [2014]) show that Rotational PCA achieved the best rotation estimates, based on the reduced COLD dataset. However, when new images are incorporated into the created map, the projection results of PCA have to be recalculated for all images. This does not meet the requirement of many real world problems, such as vSLAM. The DFT, HOG, and Gist tend to produce similar results if the parameters are tuned properly. Further studies are needed to test for statistically significant differences between the VC, SIFT and QT methods, and those deployed in (Payá et al. [2014]). Finally, we investigated the performance of QT on the COLD dataset in an HS colour space and after logarithmic transformation. The experimental results show that an RGB colour space is more suitable for our proposed QT method than an HS colour space. The experimental results on the log-transformed COLD dataset show that, to some extent, the logarithmic transformation could improve the performance of the QT method when confronted with changes in illumination, with a standard deviation error of roughly  $5^\circ$  on Sunny dataset. Therefore, further work will also be necessary with a view to enhancing the robustness of the proposed method to illumination variations.

## Chapter 6

# A quadtree-based method for loop closure detection

### 6.1 Introduction

A mobile robot should be able to determine when it has returned to a visited place after some time: this is known as loop closure which is an essential part of vSLAM system. Loop closures technique based on visual information has received much attention in recent years, as cameras have become more easily available and the opportunity has been grasped to exploit the rich visual detail embedded in images to match images collected along robot routes. Some examples are the local feature-based methods of Angeli et al. [2008b]; Cummins and Newman [2008a]; Labbe and Michaud [2013]; Scaramuzza et al. [2010]: and the global descriptor-based methods of Badino et al. [2012]; Liu and Zhang [2012]; Sunderhauf and Protzel [2011]; and Wu et al. [2014]. More recently, with the great boost in performance of Convolutional Neural Networks (CNNs) on image classification and object recognition tasks, deep features from various layers of CNNs can be applied to describe the image and implement the task of loop closure detection (Hou et al. [2015]; Sunderhauf et al. [2015]). However, in natural environments, repetitive structures and dynamic objects continue to pose severe challenges for any place recognition system.

---

In our work, we use an omnidirectional camera as the only sensor modality. A quadtree-based image comparison method incorporating Euclidean distance and Pearson Correlation coefficient metric is used to evaluate loop closure through very simple decision rules. The decision will be made by comparing the similarity score between two scenes returned by the quadtree-based method with the selected loop closure threshold. The overall procedures of our method will be described in the next section. The quadtree decomposition process employed in our method is concerned with the spatial structure property of an image, rather than detailed textural information: this focus renders our method robust against dynamic changes in scenes, such as the movement of objects within a scene. The detail of our quadtree method was described in Chapter 4.

## 6.2 Methodology

Measuring the distance or similarity between the current observation and the view of a location in the built map is a fundamental problem in a visual loop closure system. We approach this by simply calculating the similarity of two views after removing areas that are marked as too different, using our quadtree decomposition method.

Our method recursively compares quadrants of two images to be compared using the Euclidean distance metric or Pearson’s correlation coefficient until either the two quadrants are judged similar, or the quadrants become too small. When two quadrants are not judged sufficiently similar, they are each separated into four quadrants of the same size, and the process is repeated. Through this process, the locations of visual changes between image pairs will be indicated: exploiting this information, we then can obtain the Euclidean distance or Pearson Correlation coefficient of similar areas of the two images. We then apply this scoring ( $Sc$ ) to determine loop closure acceptance or rejection. There are two thresholds ( $T_{quadtree}$  and  $T_{loop}$ ) that are the main factors affecting the quadtree construction process and the performance of loop closure detection:

$T_{quadtree}$ : This is the threshold for quadtree decomposition. During the compari-

---

son, the two images or sub-regions of each image pair are considered as similar if their Euclidean distance (resp. Pearson's correlation coefficient) is below (resp. above) this threshold. Through all our experiments,  $T_{quadtree}$  for the Euclidean distance metric is set to 42, while 0.6 is used for the Pearson's correlation coefficient, since they appeared to give the best results in general.

$T_{loop}$ : This is the threshold for determining loop-closure acceptance or rejection. If the score  $Sc$  returned by the quadtree-based method is higher than this threshold, we accept that the observation comes from the same place as the reference image. This parameter is difficult to set in order to obtain both high precision and high recall rate loop closures. We have explored different ways to determine this threshold, and will discuss this in more detail below in subsection (Subsection 6.2). The relationship between the precision rate, the recall rate, and this threshold are discussed in discussion section (Section 6.3), where precision-recall statistics are generated by varying threshold  $T_{loop}$ .

It will be apparent that the depth of the quadtree depends on the value for threshold  $T_{quadtree}$  and the smallest region size predefined. We stop splitting a quadrant when it is sufficiently small. The performance difference of our algorithm with different smallest sizes ( $20 \times 20$ ,  $10 \times 10$  and  $5 \times 5$  of pixels) of quadrant have been compared: as there were no significant differences between them, we chose  $20 \times 20$  pixels as the smallest quadrant size, in order to increase the comparison speed of the algorithm.

## How to choose $T_{loop}$ ?

We selected the widely-used Euclidean distance (represented as E) or Pearson Correlation coefficient (represented as C) metrics for quadtree decomposition. We then choose one of them to yield the final distance between two given images, using our algorithm. Consequently, a total of four cases will be adduced, labeled CC, CE, EE and EC, respectively.

To determine an appropriate threshold ( $T_{loop}$ ) for loop closure detection, the following three procedures are performed.

---

First, we plot the histograms of the scores between the first image and all images of each dataset using different distance metrics. This provide evidence that the scores from all datasets can be modeled as an asymmetrical distribution with a long tail on one side (see Figure A.1 - A.4 on page 178 - 181).

Secondly, in terms of the shape of the histograms we acquired, we fit Lognormal, Weibull, Gamma, Normal and Logistic distributions, and use PPCC (Probability Plot Correlation Coefficient) to discover the distribution family most appropriate for our data. The PPCC test is known to be a powerful, but easy-to-use suitability-of-fit test that indicates whether or not it is reasonable to assume that the observed data comes from a specific distribution. It should be noted that Normal and Logistic distributions are symmetrical, and apparently not a good fit to our data. We calculated the PPCC values in an attempt to provide an intuitive comparison with other distributions and a reference for the reader.

Table A.1 - A.4 in Appendix A (page 178 - 181) show the maximum PPCC values for different distributions. Almost all tests support the conclusion that our data give a reasonably good fit to the log-normal distribution, except in two cases that utilize the CE method and the CC method on the ISL 1 dataset. In these two cases, the Weibull distribution is a slightly better fit than the others. Overall, the Weibull, Gamma and Log-normal densities are similar in shape for the same coefficient of variation.

Finally, after choosing a Log-normal distribution to fit the data, we estimate parameters (meanlog  $\mu$  and stdlog  $\sigma$ ) of this model. We then obtain the mean and standard deviation of the data, using the following equations:

$$mean = e^{(\mu + \sigma^2/2)}, \quad (6.1)$$

$$SD = (e^{\sigma^2} - 1)(e^{2\mu + \sigma^2}); \quad (6.2)$$

Finally, the threshold  $T_{loop}$  is set as  $mean + i \times SD$ , where  $i = 1, 1.5, 2, 2.5, 3, 3.5$ , and 4.

---

## 6.3 Experimental results and discussion

In this section, we carry out some experiments to evaluate which distance measure is the most appropriate for our application. We then perform the comparison between the proposed method and the four state-of-the-art descriptor-based schemes in the loop closure detection task.

A series of experiments were performed based on the ISL dataset, which was collected by ourselves within an indoor environment and contains four sub-datasets. ISL 1 and ISL 2 are characterized by high perceptual aliasing in a static environment: while ISL 3 and ISL 4 were recorded to validate the matching capability in the presence of scene changes. More detail on this dataset has been provided in Chapter 3. We chose one particular position (marked in red in Figure 3.2 ) which interested us to determine whether the algorithm would be able to detect a loop closure at this particular place.

### 6.3.1 Evaluation loop closure accuracy

A precision-recall metric is used for performance evaluation in the following experiments. Precision is defined as the number of correct loop closure detections divided by the total number of detections, and recall as the number of correct loop closure detections divided by the number of ground truth loop closures. Expected correct detections are defined as previously visited VICON locations within a given distance  $dist$  (e.g.,  $dist = 1m$ ) of the current location. This parameter can be designed according to the requirements (coarse, or more accurate) of the application in question. To compare different precision-recall curves, we calculate the average precision, which can be estimated geometrically by the area under the precision-recall curve: a high score represents both high recall, and high precision. In addition, the best recall rates at 100% precision for all methods are compared. The experimental results are presented in the following subsections.

---

### 6.3.2 Evaluation the proposed method

We present our results on the detection accuracy, and evaluate how the accuracy depends on the chosen distance metrics.

Precision-Recall (PR) curves for each sub-dataset are illustrated in Figures 6.1, 6.2, 6.3 and 6.4, respectively. Each figure contains four plots showing the results of different distance metrics used to determine the thresholds for loop closure detection, and each precision-recall statistic of the plot is calculated for varying  $dist$ , which was carried out with  $dist = 0.1m, 0.2m, 0.3m \dots, 1m$  for a total of 10 values. It is apparent that the parameter  $dist$  accommodates varying levels of detection quality at ranges from  $0.1m$  to  $1m$ . When  $dist$  increases, precision increases and recall decreases. Each curve is produced by applying a specific parameter  $T_{loop}$  ( $mean + i \times SD$ ). A range of values (1, 1.5, 2, 2.5, 3, 3.5, and 4) for  $i$  were tested. Furthermore, each curve is summarised by average precision (AP), a measure that corresponds to the area under the precision-recall curve. Tables 6.1- 6.4 show the AP for each sub-dataset using different distance metrics.

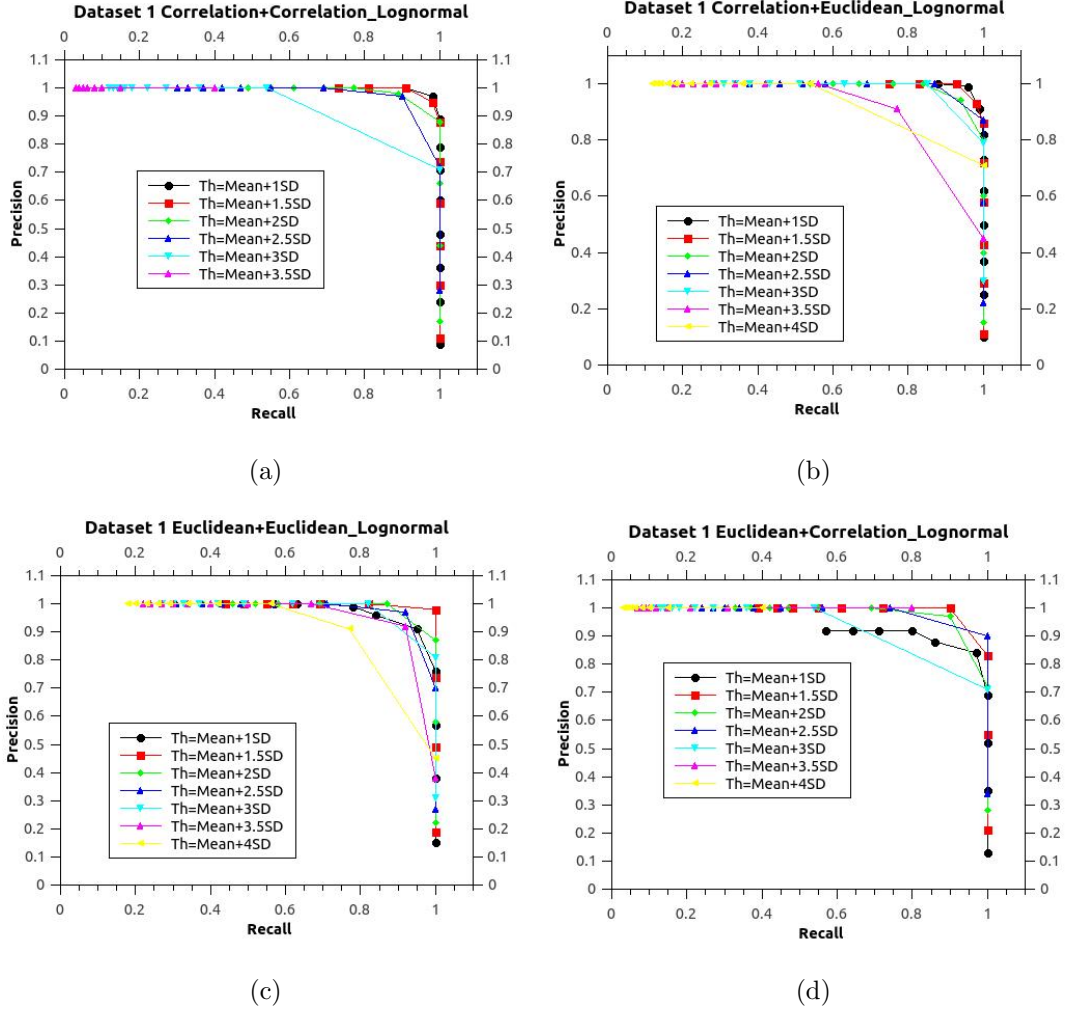


Figure 6.1: Recall and precision curves, depending on the parameter of correct detection criteria, for ISL 1. The first distance metric in the caption is used for quadtree decomposition: the second, for calculating the distance between similar areas of two images applies to the following figures.

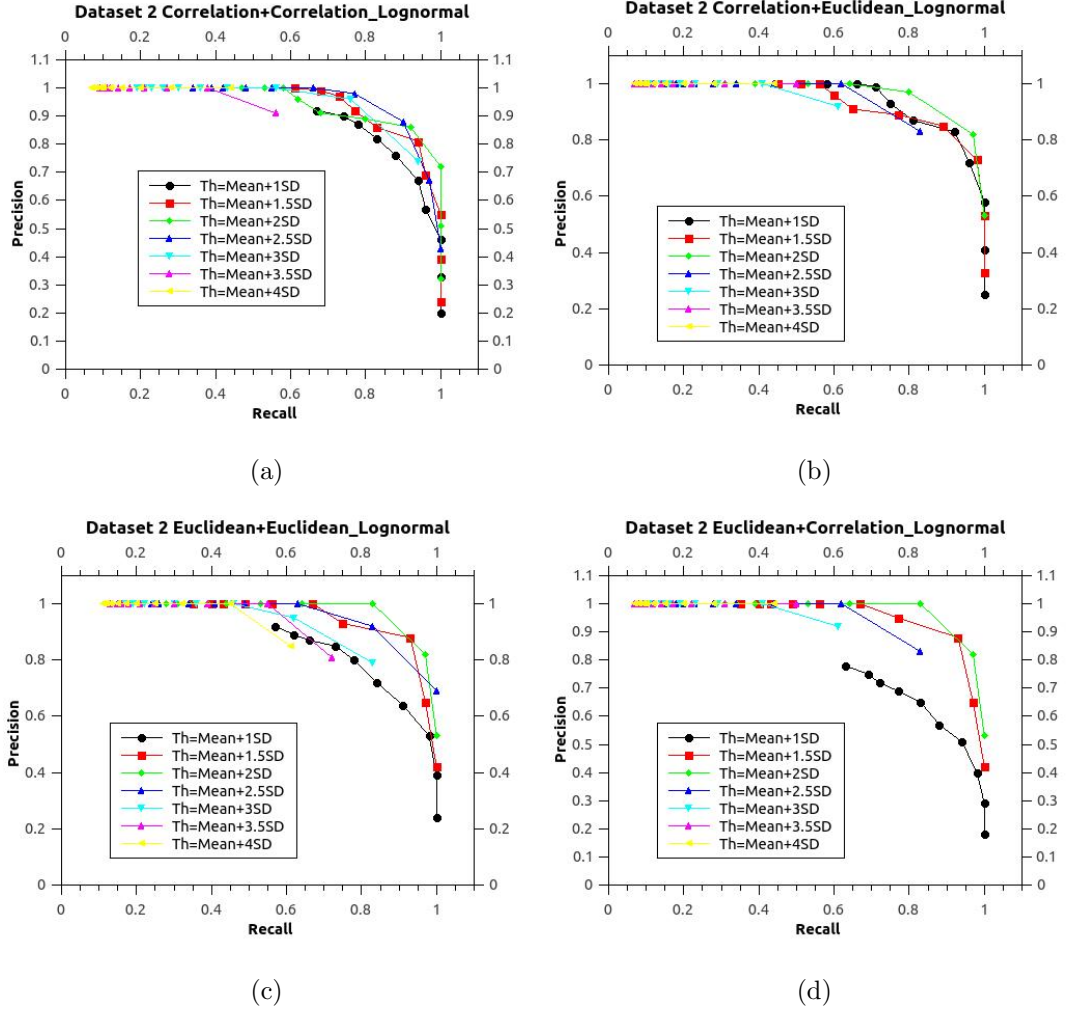


Figure 6.2: Recall and precision curves, depending on the parameter of correct detection criteria, for ISL 2.

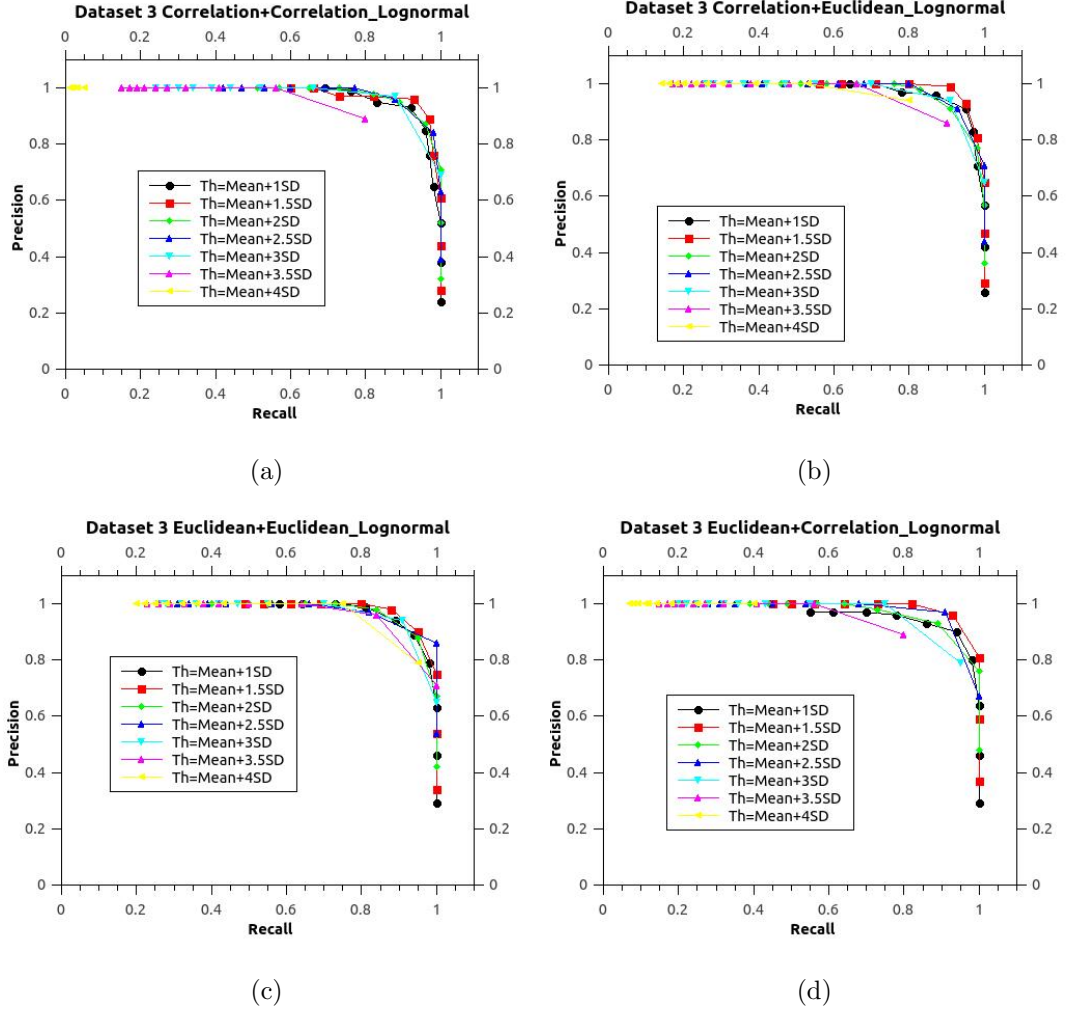


Figure 6.3: Recall and precision curves, depending on the parameter of correct detection criteria, for ISL 3.

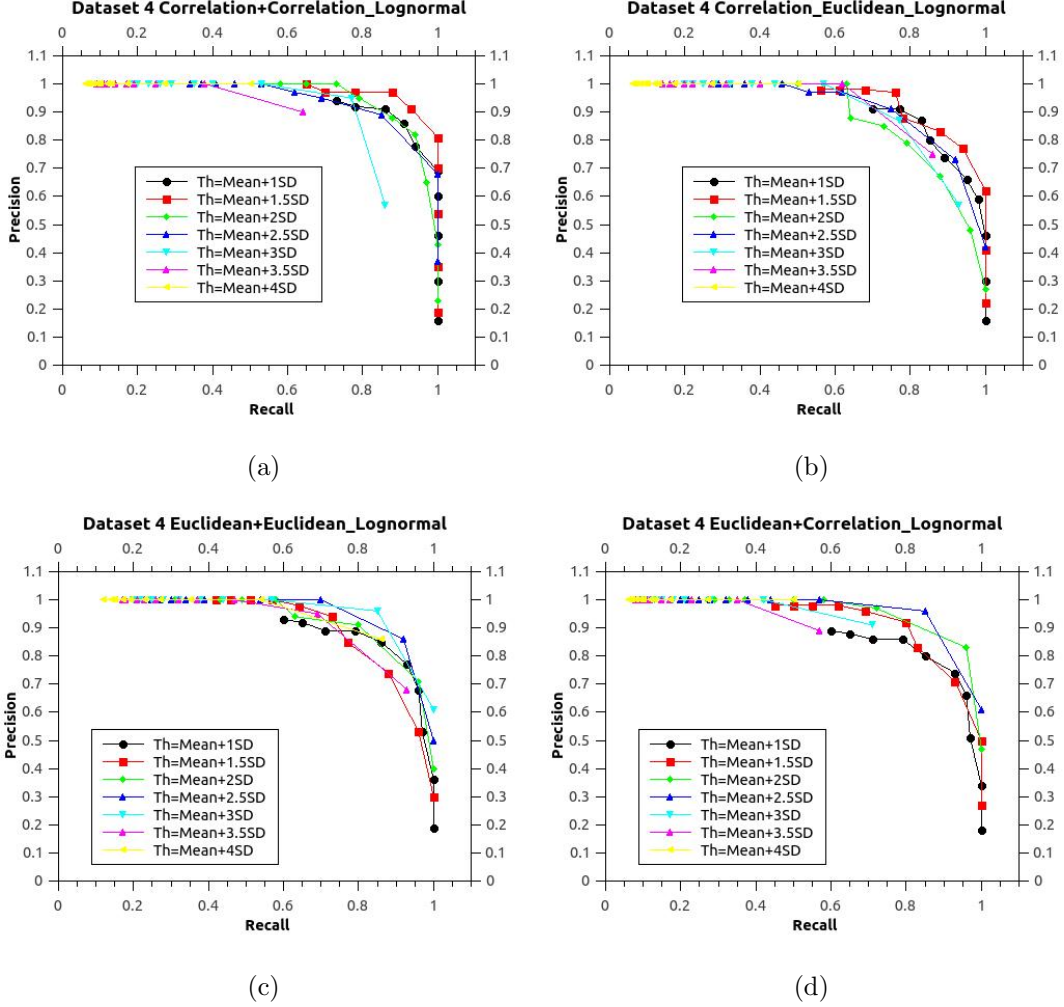


Figure 6.4: Recall and precision curves, depending on the parameter of correct detection criteria, for ISL 4.

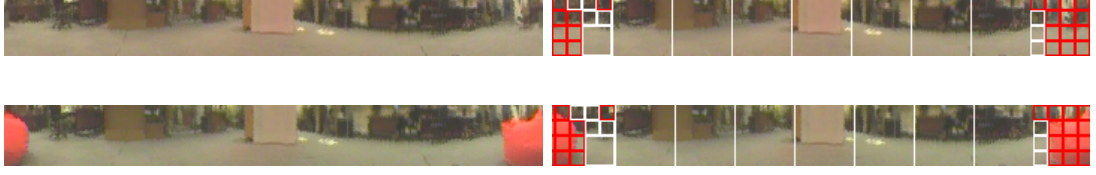
From Figure 6.1 we can see that the detection quality on the dataset ISL 1 is good, and that full precision can be obtained if we tune the parameters  $d$  and  $T_{loop}$ , but at the cost of sacrificing recall. Except in the case shown in Figure 6.1(d), lower precision is obtained when  $T_{loop}$  is set to  $mean + SD$ . It should be noted that loop closure cannot be detected when  $T_{loop}$  is set to  $mean + 4SD$ , as in Figure 6.1.

Figure 6.2 shows the results on the dataset ISL 2. We can observe that they are slightly inferior to the results on ISL 1, especially, when  $T_{loop}$  is  $mean + SD$ , which

---

is the hardest requirement for a true positive.

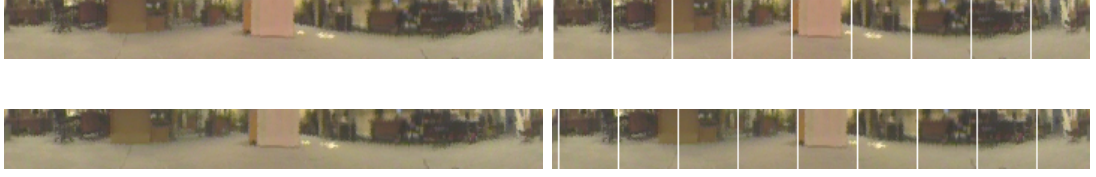
Figures 6.3 and 6.4 show the precision-recall curves resulting from the datasets ISL 3 and ISL 4: these two datasets are challenging due to the appearance of obstacles. However, the proposed method is able to reliably identify the loops, and this is attributable to the success of the method in identifying and removing the part of the compared image pair that genuinely corresponds to the areas of the environment affected by variabilities. Several examples can be found in Figures 6.5 and 6.6, where the left column shows the original three loop closure image pairs detected by our method, the right column shows the corresponding visualisation of image pair comparisons: red rectangles indicate the patch matches exceeding the similarity threshold  $T_{quadtree}$ , while the bottom shows the distance produced by our method between the Image 0 and all images, based on ISL 3 and ISL 4, respectively.



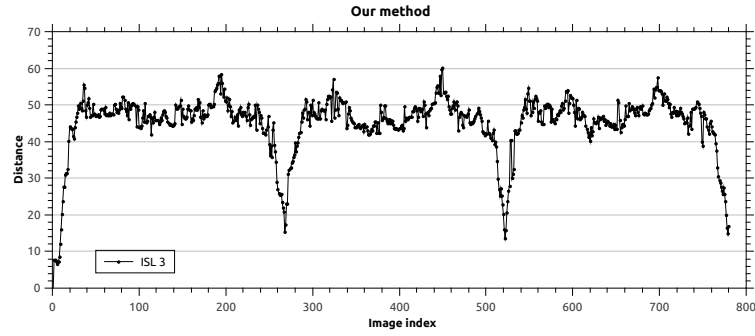
(a) Left: Image 268 (Bottom) closes a loop with Image 0 (Top). Right: visualisation of left image pair comparison using our proposed loop closure detection method, the distance is **15.38** computed after neglecting the parts of image pair which represent changes in the environment labeled by red rectangles.



(b) Left: Image 523 (Bottom) closes a loop with Image 0 (Top). Right: visualisation of left image pair comparison using our proposed loop closure detection method: the distance is **15.75**.

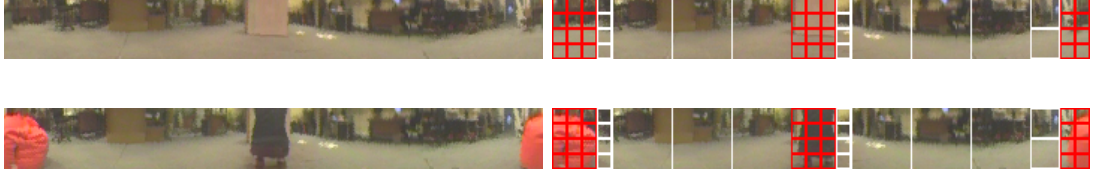


(c) Left: Image 779 (Bottom) closes a loop with Image 0 (Top). Right: visualisation of left image pair comparison using our proposed loop closure detection method: the distance is **14.83**.

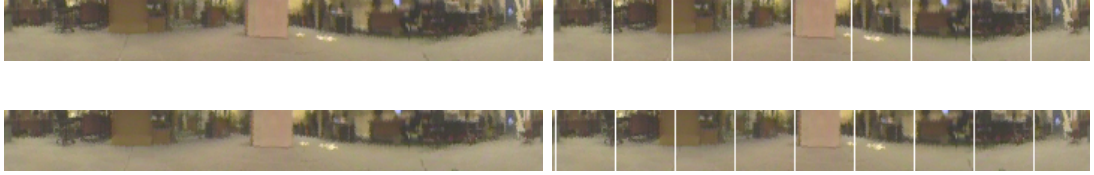


(d) Distance between Image 0 and all Images.

Figure 6.5: Examples of loop closure detection based on the CE scheme. Images 0, 268, 523 and 779 from the ISL 3 dataset were captured at nearly the same location, but with slight offset or rotation of the camera viewpoint (see Figure 3.5(c)).



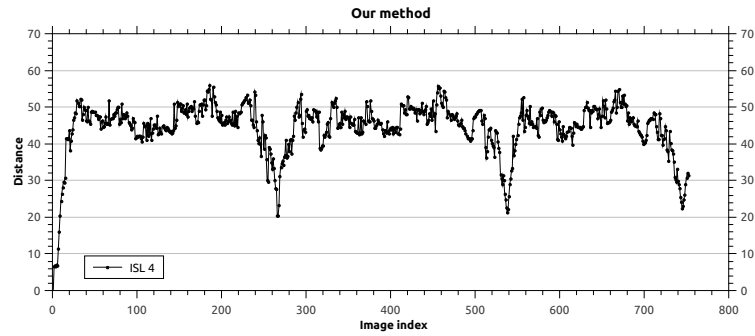
(a) Left: Image 266 (Bottom) closes a loop with Image 0 (Top). Right: visualisation of left image pair comparison using our proposed loop closure detection method: the distance is **20.42**, computed after neglecting the parts of the image pair that represents changes in the environment, as labeled by red rectangles.



(b) Left: Image 538 (Bottom) closes a loop with Image 0 (Top). Right: visualisation of left image pair comparison using our proposed loop closure detection method: the distance is **21.34**.



(c) Left: Image 745 (Bottom) closes a loop with Image 0 (Top). Right: visualisation of left image pair comparison using our proposed loop closure detection method: the distance is **22.37**.



(d) Distance between Image 0 and all Images.

Figure 6.6: Examples of loop closure detection based on the CE scheme. Images 0, 266, 538 and 745 from the ISL 4 dataset were captured at nearly the same location, but with slight offset or rotation of camera viewpoint (see Figure 3.6(c)).

---

We can see from Figure 3.5(c) that Images 0, 268, 523 and 779 from the ISL 3 dataset are taken from nearly the same position. The distances between Image 0 and Image 268, Image 0 and Image 523, and Image 0 and Image 779 computed by our method are 15.38, 15.75 and 14.83, respectively. As can be seen in Figure 6.5, a bean bag is present in Image 268 that is not seen in the previous visit (Image 0), and disappears in the two subsequent visits (Images 523 and 779). For the three ISL 4 examples, we can see from Figure 3.6(c) that Images 0, 266, 528 and 745 from the ISL 4 dataset are also collected from nearly the same location. the distance between Images 0 and 266, Images 0 and 528, and Images 0 and 745 are 20.42, 21.34 and 22.37, respectively. In these cases, there are more changes in the appearance, which are present in Image 268 and absent in Images 0, 538 and 745. As we expected, our method is able to find these loop closures in spite of the changes in the environment.

Table 6.1: Average precision for the ISL 1 dataset.

Threshold	CC	CE	EE	EC
$Mean + 1SD$	<b>99.76%</b>	<b>99.68%</b>	98.27%	90.76%
$Mean + 1.5SD$	99.66%	99.62%	<b>99.82%</b>	<b>99.15%</b>
$Mean + 2SD$	99.11%	98.92%	99.16%	98.14%
$Mean + 2.5SD$	98.14%	99.16%	98.37%	98.70%
$Mean + 3SD$	93.33%	98.43%	98.29%	93.33%
$Mean + 3.5SD$	40.00%	91.70%	96.20%	80.00%
$Mean + 4SD$	—	93.33%	91.70%	40.00%

Table 6.2: Average precision for the ISL 2 dataset.

Threshold	CC	CE	EE	EC
$Mean + 1SD$	87.32%	95.27%	84.97%	72.28%
$Mean + 1.5SD$	95.15%	94.21%	95.68%	96.06%
$Mean + 2SD$	95.15%	<b>97.00%</b>	<b>97.77%</b>	<b>97.77%</b>
$Mean + 2.5SD$	<b>96.06%</b>	81.22%	95.89%	81.22%
$Mean + 3SD$	90.90%	60.20%	79.87%	60.20%
$Mean + 3.5SD$	55.19%	50.00%	70.39%	50.00%
$Mean + 4SD$	44.00%	44.00%	59.80%	44.00%

---

Table 6.3: Average precision for the ISL 3 dataset.

Threshold	CC	CE	EE	EC
$Mean + 1SD$	97.46%	97.85%	98.04%	95.34%
$Mean + 1.5SD$	98.14%	<b>98.86%</b>	<b>98.63%</b>	<b>98.98%</b>
$Mean + 2SD$	98.20%	97.71%	97.99%	97.49%
$Mean + 2.5SD$	<b>98.25%</b>	98.09%	98.23%	98.04%
$Mean + 3SD$	97.63%	97.53%	97.53%	92.90%
$Mean + 3.5SD$	78.68%	88.32%	96.96%	78.68%
$Mean + 4SD$	5.00%	79.19%	92.90%	40.00%

Table 6.4: Average precision for the ISL 4 dataset.

Threshold	CC	CE	EE	EC
$Mean + 1SD$	91.89%	87.29%	88.85%	85.03%
$Mean + 1.5SD$	<b>98.11%</b>	<b>96.81%</b>	91.64%	92.45%
$Mean + 2SD$	96.01%	89.32%	93.76%	95.99%
$Mean + 2.5SD$	95.08%	92.39%	95.90%	<b>96.22%</b>
$Mean + 3SD$	83.24%	87.22%	<b>96.22%</b>	69.70%
$Mean + 3.5SD$	62.70%	83.00%	88.01%	55.79%
$Mean + 4SD$	50.00%	50.00%	83.76%	50.00%

As indicated in Table 6.1-Table 6.4, the best detection performance was achieved by EE when the threshold value was set at  $mean + 1.5SD$ , providing an average precision (AP) of 99.82% on the ISL 1 dataset. This was closely followed by the CC, CE and EC, which obtained 99.76%, 99.68%, and 99.15% AP, respectively. On the ISL 2 dataset, the EE and EC both produced equal best performances: the obtained largest AP was 97.77%, closely followed by the CE, with 97% AP. The CC yielded the worst results, with the obtained largest AP at 96.06%. On the ISL 3 dataset, performance from best to worst was: EC, CE, EE, and CC, with the AP of 98.98%, 98.86%, 98.63%, and 98.25%, respectively. On the ISL 4 dataset, the CC obtained a performance of 98.11% AP, followed by CE with 96.81%: EE and EC achieved the lowest performance, at 96.22% AP.

In general, the different metrics perform similarly on loop closure detection for each sub-dataset. In our subsequent comparative study, we select Pearson Correlation coefficient and Euclidean distance metrics as the representative metrics

---

for quadtree decomposition, and final distance measure between images (CE), respectively.

### 6.3.3 Comparison with other methods

In this section, we compare the accuracy of the proposed method against the BRIEF-Gist (Sunderhauf and Protzel [2011]), LDB-based method, WI-SIFT and WI-SURF (Badino et al. [2012]) for loop closure detection, using the ISL dataset. We selected the same position chosen in the previous experiment to evaluate the performance of all methods at this particular place (marked in red in Figure 3.2).

In the case of BRIEF-Gist, we used the same implementation used in Chapter 4 for testing. However, we split each panoramic image vertically into six blocks and resized each block to  $60 \times 60$  pixels in order to obtain a suitable patch size for descriptor generation. The center of each image block is defined as the keypoint without rotation or scale: the BRIEF descriptor is computed for each block, so a complete image will contain six descriptors. The final distance (in appearance space) between two scenes is the average Hamming distance of six pairs of descriptors.

The same algorithm is also applied to the LDB, SIFT and SURF descriptors: in this way we obtain the global binary (LDB) or floating-point (SIFT and SURF) descriptions of the image. For convenience, the LDB-based method is named LDB-Gist. The Hamming distance is used for computing distances in LDB-Gist, while the Euclidean distance is used in WI-SIFT and WI-SURF. The implementation of the LDB descriptor is taken from (Yang and Cheng [2014a]), and the SIFT and SURF descriptors from OpenCV (Bradski [2000]). The default feature dimensions of LDB, SIFT and SURF were chosen in our experiments, and were 32, 128 and 64 bytes, respectively. Figures 6.8 - 6.11 present the distance (in appearance space) produced by all methods between the image of interest (Image 0) and all the images of the ISL dataset.

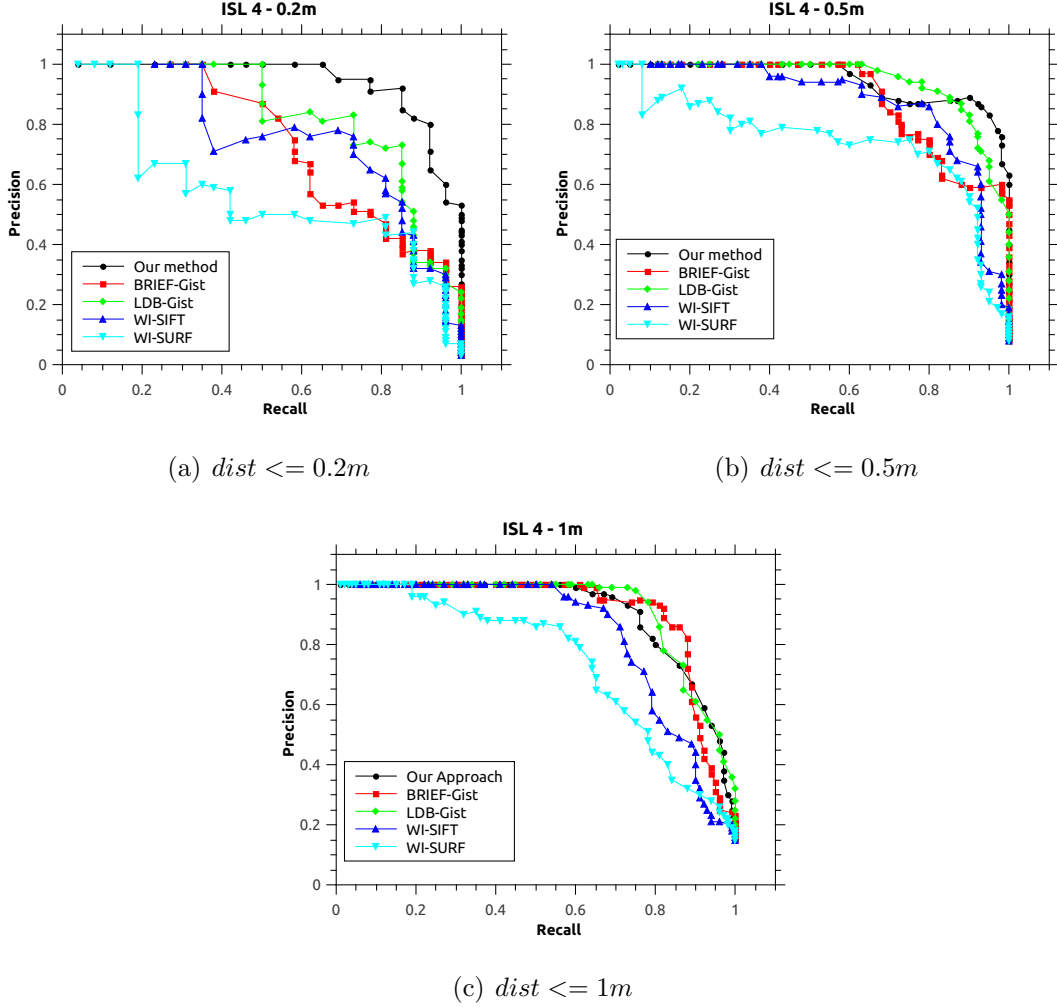


Figure 6.7: Precision and recall curves for the ISL 4 dataset.

Algorithm	$dist \leq 0.2m$		$dist \leq 0.5m$		$dist \leq 1m$	
	AP	R	AP	R	AP	R
Our method	94.21%	65.00%	94.75%	57.00%	90.63%	58.00%
BRIEF-Gist	74.38%	35.00%	89.42%	62.00%	90.89%	61.00%
LDB-Gist	82.93%	50.00%	94.62%	63.00%	91.81%	64.00%
WI-SIFT	75.75%	35.00%	87.91%	38.00%	83.67%	54.00%
WI-SURF	57.82%	19.00%	74.47%	8.00%	74.06%	19.00%

Table 6.5: Average precision (AP), and best recall (R) at 100% precision, for the ISL 4 dataset.

For purposes of comparison, the max-normalization scheme was used to bring

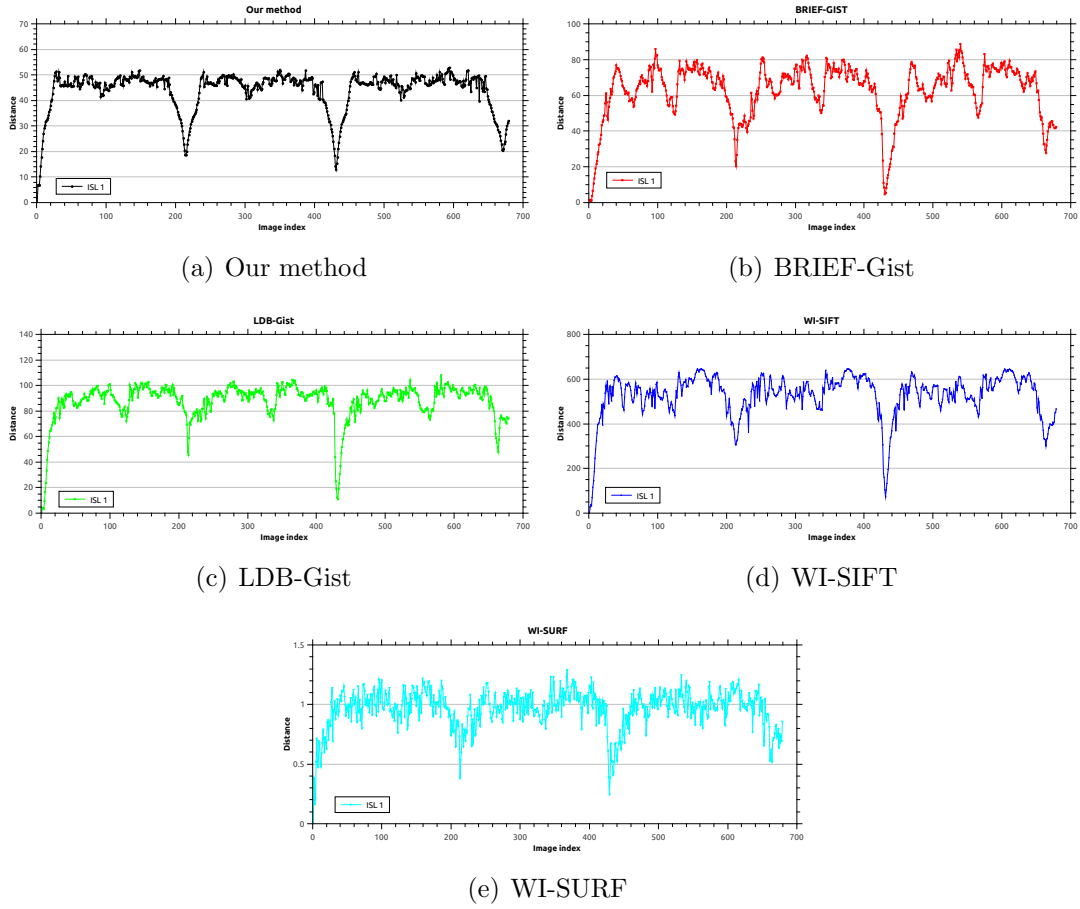
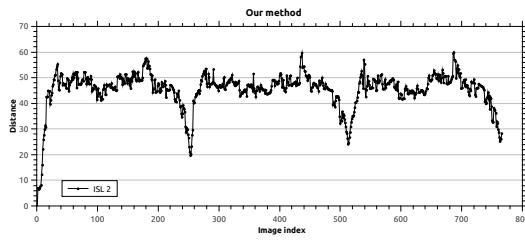
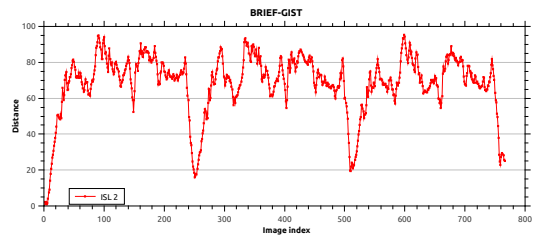


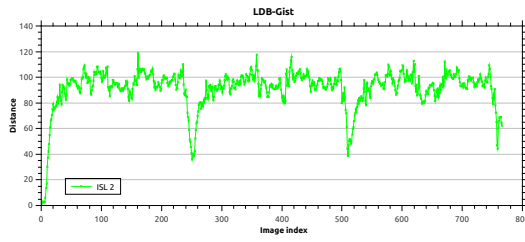
Figure 6.8: Distance (in appearance space) between Image 0 and all images of the ISL 1 dataset: (a) Euclidean distance of our method; (b) Hamming distance of BRIEF-Gist; (c) Hamming distance of LDB-Gist; (d) Euclidean distance of WI-SIFT; and (e) Euclidean distance of WI-SURF.



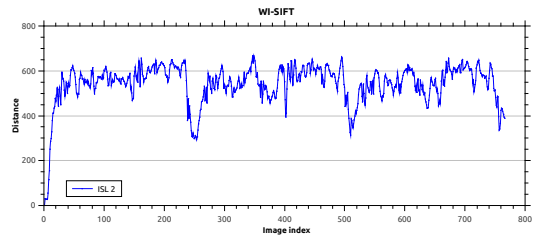
(a) Our method



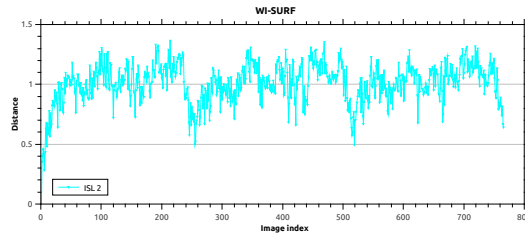
(b) BRIEF-Gist



(c) LDB-Gist



(d) WI-SIFT



(e) WI-SURF

Figure 6.9: Distance between Image 0 and all images of the ISL 2 dataset: (a) Euclidean distance of our method; (b) Hamming distance of BRIEF-Gist; (c) Hamming distance of LDB-Gist; (d) Euclidean distance of WI-SIFT; and (e) Euclidean distance of WI-SURF.

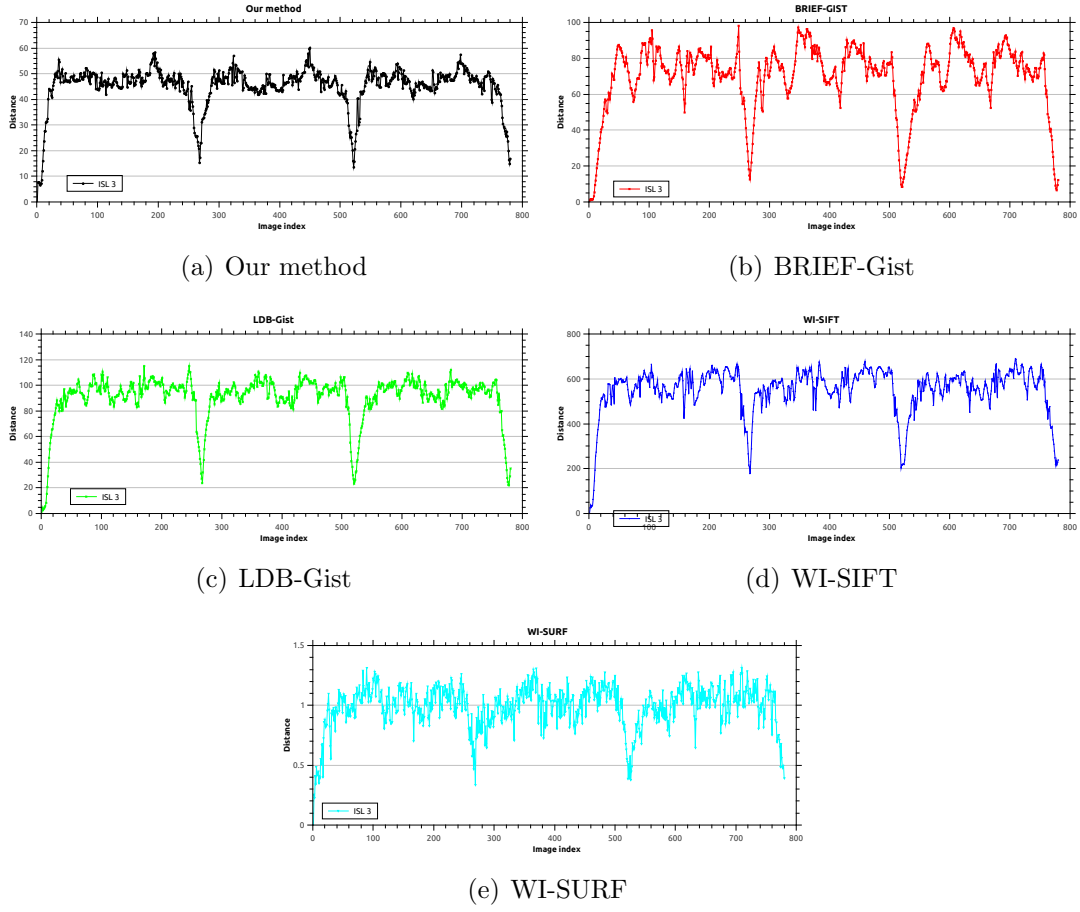


Figure 6.10: Distance between Image 0 and all images of the ISL 3 dataset: (a) Euclidean distance of our method; (b) Hamming distance of BRIEF-Gist; (c) Hamming distance of LDB-Gist; (d) Euclidean distance of WI-SIFT; and (e) Euclidean distance of WI-SURF.

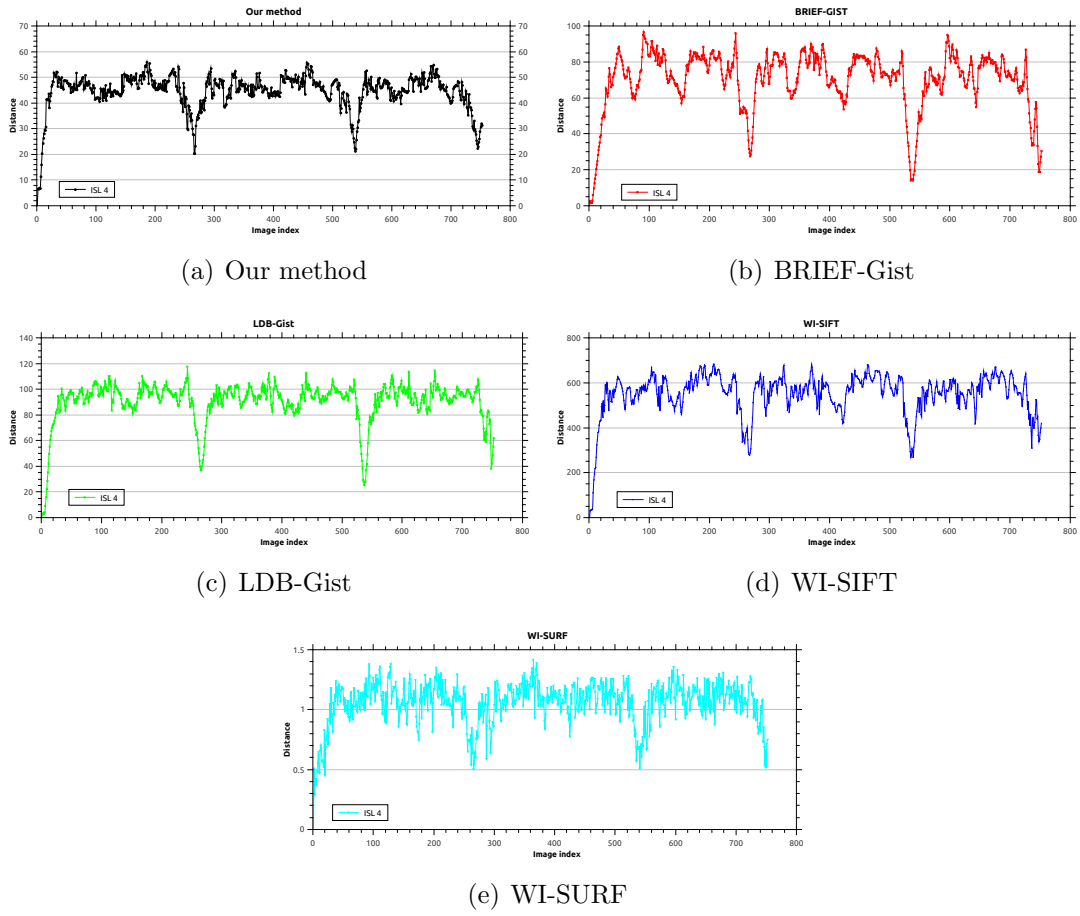


Figure 6.11: Distance between Image 0 and all images of the ISL 4 dataset: (a) Euclidean distance of our method; (b) Hamming distance of BRIEF-Gist; (c) Hamming distance of LDB-Gist; (d) Euclidean distance of WI-SIFT; and (e) Euclidean distance of WI-SURF.

---

all distances obtained from all methods into the range  $[0, 1]$ . The decisions for closing loops were based on thresholding of the normalized distances ( $T_{loop}$ ).

We first examined the effect of  $T_{loop}$  on precision and recall. We varied the  $T_{loop}$  within the range from 0 to 1 with a step of 0.01, in order to generate well-defined curves. We also tested the performance of all methods under different detection quality constraints (closeness to ground truth). As described in the discussion of our previous experiments, we applied a distance threshold  $dist$  (in metric space), so that all the possible pairs taken within, for example, 1m were considered a true positive. In this test, the distance threshold  $dist$  was set to 0.2, 0.5 and 1m, successively, and precision and recall results for each method at these different thresholds are generated: these are shown in Figures 6.12 - 6.15. Tables 6.6- 6.9 summarize the average precision, and the best recall rates at precision of 100%, on each sub-dataset respectively.

The ISL 1 dataset is characterized by perceptual aliasing, which proves challenging for all vision-based place recognition frameworks. The experimental results depicted in Figure 6.12 and Table 6.6 show that surprisingly good results are obtained using the proposed method. The average precision varies between 96.96% and 99.85%, and the best recall varies between 65% and 93%, depending on the value set for threshold  $dist$ . Recall results drop severely for the other methods: the BRIEF-Gist and WI-SIFT methods perform similarly on this dataset in terms of average precision and best recall.

In the case of ISL 2, the presence of objects within the experiment area helps to relieve the perceptual aliasing problem. The accuracy of BRIEF-Gist, LDB-Gist and WI-SIFT were significantly increased with  $dist \leq 0.5m$ , as is seen from the experimental results presented in Figure 6.13 and Table 6.7. Our proposed method presents the best performance with  $dist \leq 0.2m$  and  $dist \leq 1m$ , achieving 96.36% and 91.16% recall respectively, with no false positives. WI-SURF again records the worst performance on this dataset.

Figure 6.14 shows the precision-recall curve for ISL 3, where all methods except WI-SURF exhibit the same behaviour with the stricter threshold  $dist$  (smaller value): our method performs best with  $dist \leq 1m$ . It can be seen that the

---

presence of an object in this dataset (Images 159 to Image 384) does not affect the accuracy of any of the methods.

ISL 4 was more challenging, as the environment contained more changes, which were present from Image 156 to Image 393. It can be observed from Figure 6.15 and Table 6.9 that the average precision of our method still reached 94.21% and recall reached 65% at 100% precision, under the strictest situation ( $dist \leq 0.2m$ ). However, the average precision and the recall of BRIEF-Gist, LDB-Gist and WI-SIFT show substantially lower scores for this dataset than for ISL 3. Our method, BRIEF-Gist and LDB-Gist recorded similar performances with  $dist \leq 0.5m$  and  $dist \leq 1m$ , followed by WI-SIFT.

It can be seen from Table 6.6- 6.9 that the performance of each method does not always improve when we increase the distance threshold  $dist$  (from 0.2m to 1m). This threshold determines the range within which all possible pairs are considered a true positive loop closure event. From Figures 6.8 - 6.11, we can see that the distance curves produced by all methods between the image of interest (Image 0) and all the images of the ISL dataset is jagged rather than smooth. In this case, the algorithm has probably yielded a number of error matches under the restrictive detection quality constraints applied in our experiment. Moreover, the ground truth for loop closure evaluation contains all frames in the sequence of each ISL dataset whose distance falls within the  $dist$  relative to the current frame, excluding the most closely adjacent 50 frames. This means that the number of successful ground truth for loop closures will rise with the increase of  $dist$ , possibly resulting in reduced recall rates. This may explain why not all results improve for all values of  $dist$  when  $dist$  is increased from 0.2m to 1m. For instance, the best recall of BRIEF-Gist at the 100% precision level presented in Table 6.7 increases from 41% to 81% when increasing the  $dist$  value from 0.2m to 0.5m, but falls from 81% to 64% when increasing the  $dist$  value from 0.5m to 1m. Similarly, the AP increases from 84.54% to 98.9% when increasing the  $dist$  value from 0.2m to 0.5m, but falls from 98.9% to 86.59% when increasing the  $dist$  value from 0.5m to 1m.

Overall, from Figures 6.12 - 6.15 and Table 6.6 - 6.9 we can conclude that our method achieved a higher recall rate, while maintaining 100% precision with

---

a restrictive value of  $dist$ , for all datasets. Moreover, satisfactory results were obtained by our method using ISL 1, which implies that our method is suitable for more accurate loop closure detection under conditions of strong perceptual aliasing. We believe that one reason for this is that the colour images used in our method contain richer information than the grayscale images of other methods.

Two binary descriptors, BRIEF-Gist and LDB-Gist, exhibited comparable discriminative power with SIFT: that is, when converted to a global descriptor; in addition, they were less expensive to compute. It is notable that the drop in recall results was most significant for WI-SURF over the whole dataset, showing this descriptor to be insufficiently discriminating for the loop closure detection task. As may be observed that, the LDB-Gist obtains a slightly better result compared to BRIEF-Gist on ISL 4, showing greater robustness to perceptual changes. One possible reason is that the LDB descriptor exploits intensity and gradient pairwise comparison: this provides superior discrimination capability, as the BRIEF descriptor makes use of the intensity comparison only.

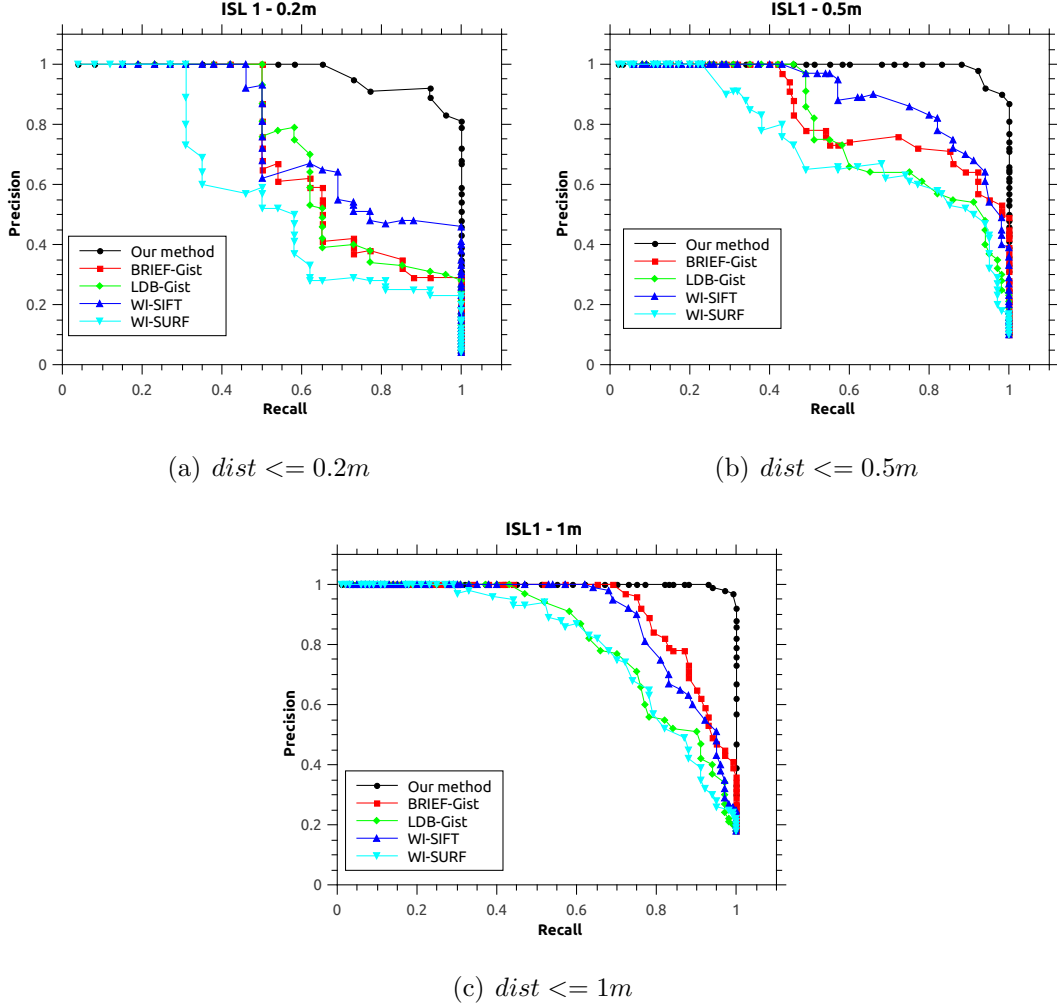


Figure 6.12: Precision and recall curves for the ISL 1 dataset.

Algorithm	$dist \leq 0.2m$		$dist \leq 0.5m$		$dist \leq 1m$	
	AP	R	AP	R	AP	R
Our method	96.96%	65.00%	99.27%	88.00%	99.85%	93.00%
BRIEF-Gist	71.47%	50.00%	83.80%	42.00%	91.92%	69.00%
LDB-Gist	72.72%	50.00%	79.52%	46.00%	82.52%	43.00%
WI-SIFT	77.10%	46.00%	89.91%	43.00%	89.51%	62.00%
WI-SURF	58.09%	31.00%	73.76%	23.00%	80.63%	29.00%

Table 6.6: Average precision (AP), and best recall (R) at 100% precision, for the ISL 1 dataset.

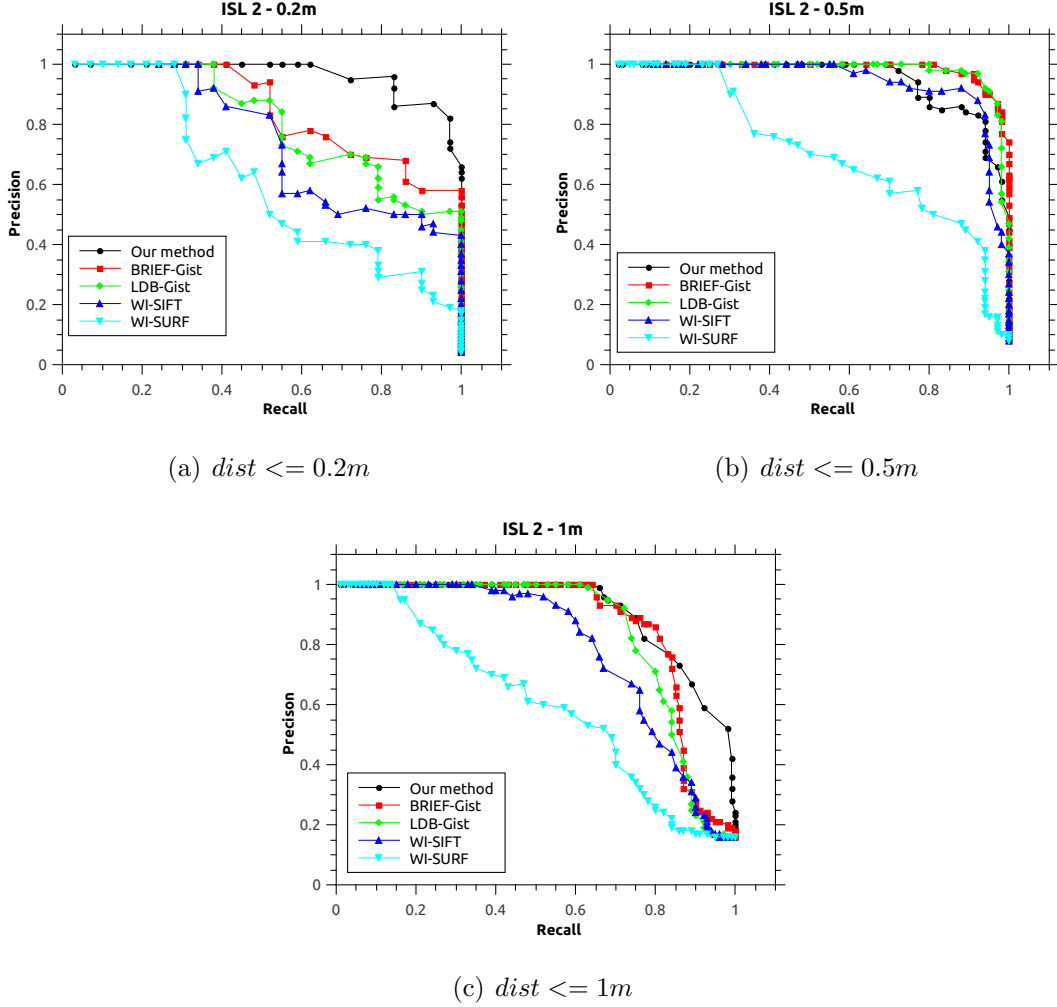


Figure 6.13: Precision and recall curves for the ISL 2 dataset.

Algorithm	$dist \leq 0.2m$		$dist \leq 0.5m$		$dist \leq 1m$	
	AP	R	AP	R	AP	R
Our method	96.36%	62.00%	94.99%	70.00%	91.16%	63.00%
BRIEF-Gist	84.54%	41.00%	98.60%	81.00%	86.59%	64.00%
LDB-Gist	80.58%	38.00%	98.14%	80.00%	84.47%	61.00%
WI-SIFT	74.84%	34.00%	94.45%	56.00%	78.48%	34.00%
WI-SURF	61.35%	28.00%	71.39%	27.00%	60.03%	14.00%

Table 6.7: Average precision (AP), and best recall (R) at 100% precision, for the ISL 2 dataset.

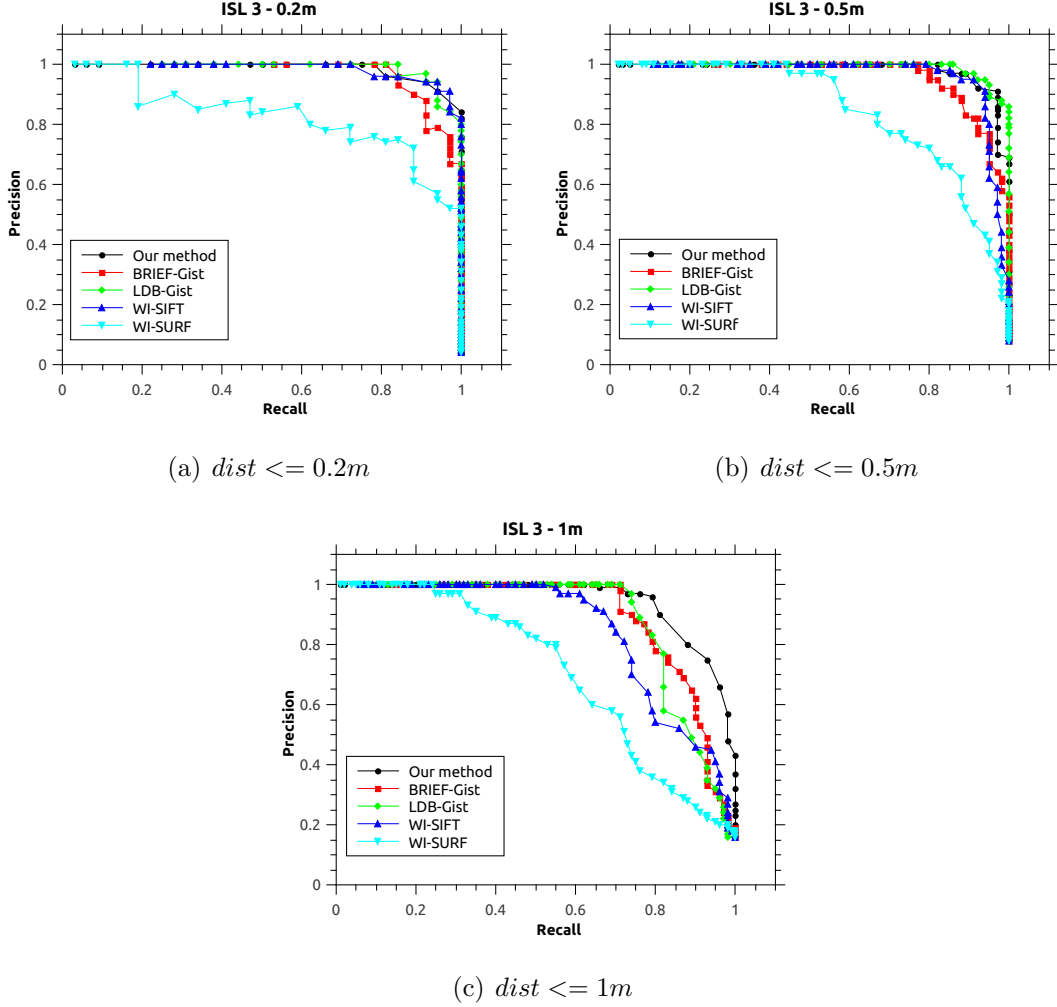


Figure 6.14: Precision and recall curves for the ISL 3 dataset.

Algorithm	$dist \leq 0.2m$		$dist \leq 0.5m$		$dist \leq 1m$	
	AP	R	AP	R	AP	R
Our method	98.47%	78.00%	98.30%	82.00%	94.65%	64.00%
BRIEF-Gist	96.96%	81.00%	96.04%	77.00%	89.39%	71.00%
LDB-Gist	98.59%	84.00%	99.08%	86.00%	88.17%	71.00%
WI-SIFT	98.33%	72.00%	96.55%	79.00%	84.85%	52.00%
WI-SURF	82.59%	19.00%	84.91%	44.00%	70.85%	24.00%

Table 6.8: Average precision (AP), and best recall (R) at 100% precision, for the ISL 3 dataset.

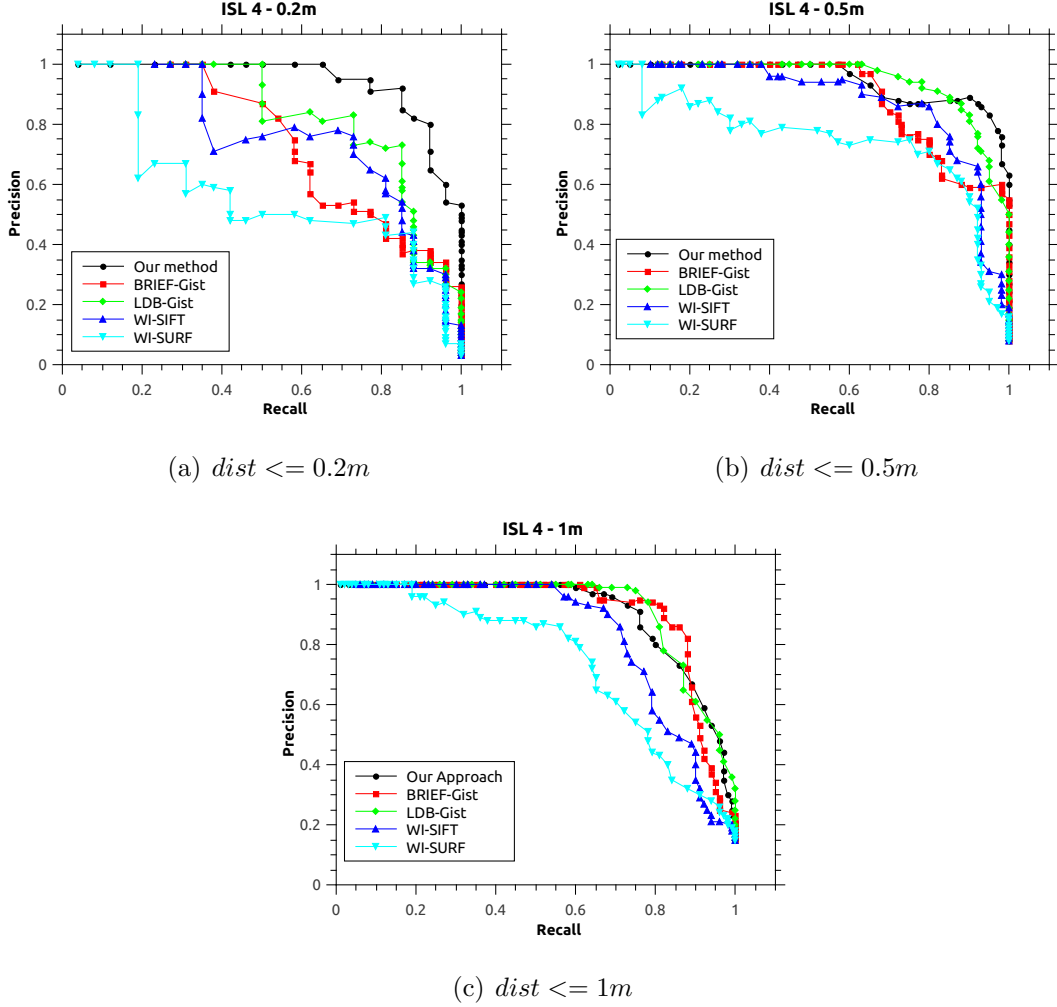


Figure 6.15: Precision and recall curves for the ISL 4 dataset.

Algorithm	$dist \leq 0.2m$		$dist \leq 0.5m$		$dist \leq 1m$	
	AP	R	AP	R	AP	R
Our method	94.21%	65.00%	94.75%	57.00%	90.63%	58.00%
BRIEF-Gist	74.38%	35.00%	89.42%	62.00%	90.89%	61.00%
LDB-Gist	82.93%	50.00%	94.62%	63.00%	91.81%	64.00%
WI-SIFT	75.75%	35.00%	87.91%	38.00%	83.67%	54.00%
WI-SURF	57.82%	19.00%	74.47%	8.00%	74.06%	19.00%

Table 6.9: Average precision (AP), and best recall (R) at 100% precision, for the ISL 4 dataset.

---

## 6.4 Conclusions

In this chapter, we have presented a new method to solve the loop closure problem. Our method is a straightforward matching between the observation and the reference view, no prior knowledge being required. This approach is based on a quadtree decomposition process, which allows us to ignore any dynamic changes within the scene that may have occurred between multiple visits. This capability provides benefits when dealing with a changing environment. The final decision of loop closure acceptance or rejection is simply achieved by comparing the similarity measurement between a pair of images and the threshold value. To our knowledge, the proposed method is a first attempt to model visual loop closure based on a quadtree concept. We have evaluated the algorithm using our own collected dataset (ISL). Experimental results show that our algorithm is effective, in most cases: the algorithm can attain 100% precision, with higher recall when the parameters are well tuned. Moreover, we performed a comparison between the proposed method and other, state-of-the-art descriptor-based methods in the visual loop closure detection application. The experimental results show that the best performance was achieved by the proposed method in strong perceptual aliasing scenarios, and more precise detection results were obtained in terms of closeness to the ground truth.

# Chapter 7

## Conclusions

This chapter summarises the work performed in this thesis, discusses the research outcomes, points out the limitations of the current work, and outlines possible directions for future research.

### 7.1 Summary of thesis

To build a fully autonomous mobile robot that is capable of operating for long periods in real environments, we must develop place recognition strategies that can handle unknown or changing environments. A visual sensor can provide such a robot with an incredible amount of information required to perceive its environment. This thesis has described the development of appearance-based place recognition strategies that aim to yield robustness to perceptual aliasing and dynamic changes of the environment.

We started with a brief background discussion of the selected topics of this thesis. Following on from this, we carried out a literature review of the main solutions to the SLAM problem with a camera as the only sensor, covering image detectors, descriptors, approaches based on BoWs schemes and a few dimensionality-reduction techniques for image descriptors. We then discussed the studies that overlap with our research, dividing these into four subgroups: place recognition,

---

visual odometry, loop closure and quadtree structure. Although place recognition has been widely researched, and demonstrated successfully in many real world implementations, significant challenges remain, because a robot is likely to encounter perceptual variability and perceptual aliasing problems. Based on this review, we concluded that it is still highly appropriate to design and implement place recognition methods in order to increase the robustness of vSLAM solutions. This provided the main motivation for the original works that follows (Chapter 4-6). The main outcome of this investigation was the decision to focus effort on combining the quadtree decomposition concept with the omnidirectional vision sensors in order to tackle changes in the environment.

Four datasets were utilised to evaluate the proposed methods, that were chosen to span a variety of environments. For the outdoor scenes, the GummyBear dataset was captured in field-like, car park and Mars-like surroundings, including ground truth provided by an RTK GPS, which allows centimetre-level accuracy of positioning. A publicly-available dataset, New College 1, was also utilized: this consists of sequences recorded within the New College Campus in a dynamic environment comprising multiple unidirectional and bidirectional loops. For the indoor scenes, the ISL dataset was collected from a laboratory environment. A VICON motion capture system provided the ground truth. Moreover, the COLD dataset (Ullah et al. [2007]), commonly utilised as a benchmark for place recognition, was adopted: this was obtained under various weather and illumination conditions (sunny, cloudy and night). Consequently, the results reported in the evaluation in this thesis can be compared with the results of other researchers, or reproduced by others. The detailed description of above four datasets was given in an individual Chapter (Chapter 3).

Many image matching techniques have shown good results in the place recognition task in recent years. However, their reliance on local features limits their ability to handle a relatively featureless environment. Moreover, these techniques ignore the spatial information contained in the image. The techniques based on global descriptors cannot deal with severe viewpoint changes and partial occlusions. Therefore, a new image comparison method based on a recently developed quadtree decomposition algorithm (Cao et al. [2012]) was proposed (Chapter 4).

---

The process of comparison corresponds to a top-down quadtree construction procedure, which gives the robot the ability to capture the variation between image pairs. In addition, to compensate for unknown rotation of the robot between visits, alignment is carried out to find the maximum similarity of the compared image pair, this alignment consisting of a simple column-by-column shift. Our collected dataset, which contains the sequence of images captured along a carefully designed “Gummy Bear” path, has been used to evaluate the algorithm. The qualitative results of the experiment have demonstrated that this method is effective at dealing with self-similar environments, and is robust to viewpoint changes of the robot. Moreover, the proposed method has been compared with three alternative existing methods (FAB-MAP(Cummins and Newman [2008a]), BRIEF-Gist(Sunderhauf and Protzel [2011]), and ABLE-P(Arroyo et al. [2014])) on the same dataset (New College 1) in the loop closure detection task without additive noise. The experimental results have shown that the proposed method is close to the two comparative methods (BRIEF-Gist and ABLE-P) if only unidirectional loop closure detection is needed, while it achieves much better recall rate at 100% precision than other methods if both unidirectional and bidirectional loop closure detection are considered. The worst results were obtained by FAB-MAP in both cases: the reasons for its poor performance might be a shortage of training data and the severe challenge of a self-similar environment.

Another experiment was carried out to analyse the performance of the proposed method on noisy images. The experimental results demonstrate that the proposed method is robust to noise. A quantitative comparison between the proposed method and the BRIEF-Gist and ABLE-P methods was conducted, and the experimental results indicate that both the proposed method and the BRIEF-Gist method achieve similar performance when Gaussian noise is introduced, while the ABLE-P method yields inferior results.

Taking into account long-term navigation, an important prerequisite for a mobile robot is to successfully determine its own orientation. In order to investigate what image-based techniques can give us a good and reliable estimation result, we evaluated the performance of three methods (quadtree-based, visual compass and SIFT-based methods) on three datasets (GummyBear, ISL, COLD) for the

---

purpose of robot orientation estimation (Chapter 5) in indoor and outdoor environments.

We introduced two ways to test the performance of the quadtree and SIFT methods on GummyBear and ISL datasets. First, a fixed image is used as a reference image and the current orientation is given relative to the reference image. This process has the drawback that the estimation becomes less reliable as the images become more different. Second, a moving reference image is used with specified skips, and the current orientation is calculated by accumulating the changes in orientation. It is apparent that the estimation becomes less and less accurate due to accumulated errors.

The evaluation of the quadtree and SIFT methods on the COLD dataset were carried out in a similar way to that of (Payá et al. [2014]). The relative orientation between all image pairs from the dataset was calculated, each pairs of images is chosen between the two consecutive images, as well as skipping one and two images. The visual compass method was validated on all datasets using a moving reference with automatically adjusted skips.

The experimental results on the GummyBear dataset revealed that the two appearance-based methods were superior to the SIFT method at lower frame rates. The results on the ISL dataset showed that less drift occurred using the appearance-based methods, while better repeatability was presented using the SIFT method in most cases. The COLD dataset results showed that appearance-based methods performed better under stable illumination, while the SIFT method performed well when illumination variations were large (the Sunny dataset). The experimental results on the COLD dataset in HS colour space and after logarithmic transformation indicated that RGB colour space is more suitable for QT than HS colour space, and that logarithmic transformation could be useful to some degree for increasing robustness against illumination changes. Compared with the result of Payá et al. [2014], all three methods achieved smaller mean errors, but with larger standard deviation. It is obviously difficult to draw conclusions about the differences in performance between different methods. Moreover, the different sizes of datasets used in the experiments makes direct comparisons difficult.

---

Recently, loop closure techniques using visual have received a great deal of attention. Visual cues can play an important role in correcting for accumulated errors, and in obtaining an overall consistent map, especially when a robot is operating over a large area and is in motion over long periods. However, these techniques have to tackle changing environments, and even a single erroneous loop closure incorporated into the map can lead to system failure. Consequently, much work has focused on pushing false positive rates closer to zero, while maintaining a high percentage of correctly-recognized loop closures.

Following the current trend, we developed a new loop closure detection method based on the quadtree decomposition algorithm (Chapter 6). The task of deciding whether a robot has returned to a previously visited area or not is formulated as a binary classification problem. Based on a similarity value calculated by using our quadtree-based method and a predefined threshold, a decision is made whether the two scenes are sufficiently similar to meet the identity criteria. The experimental results have shown that the proposed method provides a very effective performance of loop closure detection on the ISL dataset. We have achieved 100% precision with higher recall when the parameters are tuned properly. In addition, the proposed method has been compared to other state-of-the-art descriptor-based algorithms: BRIEF-Gist, LDB-based method, WI-SIFT and WI-SURF, using the ISL dataset. The results have shown that the proposed method is capable of detecting the loop closure with accuracy in strong perceptual aliasing scenarios, and under the stricter ground truth criteria.

The main drawback of our quadtree-based algorithms is lengthy computation times caused by the necessity of exhaustively matching between image pairs. The typical cost of matching is about three seconds on a computer with an Intel Core i3 1.7GHz processor and 4GB RAM. Therefore, these algorithms are currently not capable of meeting real-time constraints. In the next Section, we suggest several possible strategies to speed up computation.

In summary, we conclude that the developed quadtree-based image comparison algorithm has been utilised effectively for the task of place recognition, in that we have shown that it can handle ambiguous data and adapt to changing environments. This indicates that the methods presented in this thesis have the

---

potential to become an essential component of a full vSLAM system that relies on omnidirectional images alone.

## 7.2 Future work

In this section, we list a number of possible research directions that might be investigated in future research on the basis of the work presented in this thesis.

- Improvements on the complexity of the quadtree decomposition algorithm

As our quadtree decomposition algorithm uses an exhaustive search strategy to find the best match (maximum similarity) between two panoramic images, the computational cost increases with the dimensions of the images. Seeking a fast and simple strategy would undoubtedly be beneficial in improving the efficiency of methods based on this algorithm, including the proposed orientation estimation and loop closure detection approaches. One interesting option would be to find a local minimum, thereby avoiding exhaustive searching: one such strategy is deployed in Labrosse [2006]. Another possibility might be to exploit a parallel solution for quadtree decomposition, or to make use of specialist hardware units such as a graphics processing unit (GPU) in order to allow real-time operation.

- Investigation of different tree structures, and partition or segmentation approaches.

In this thesis, we used the fixed-partition quadtree matching model. In our future study, the effects of different tree structures and partition or segmentation approaches will be investigated. Specifically, instead of using the fixed-partition quadtree representation, variable size of patches containing more semantic content might be studied. An interesting work in this context was presented recently in (Milford et al. [2014]), where sub-image patch matching processes with high-tolerance properties were conducted for place recognition tasks under dynamic conditions.

- 
- Improvements on robustness against illumination variations

In order to make mobile robot operation feasible for real-world applications, the algorithm must be able to adapt to illumination changes. In fact, this is a very challenging problem in robot vision, and far from being solved. As seen in Chapter 5, the three sub-datasets of the COLD dataset (Sunny, Cloudy and Night) contain different lighting conditions, from brightly sunlit, shadowed and reflective areas to artificial fluorescent light. Our method involving the direct comparison of pixels of images is more sensitive to changing illumination conditions than the SIFT method. An alternative approach could use chrominance colour spaces that separate luminance and chromatic components, as suggested by (Ososinski and Labrosse [2013]), such as the log-chromaticity colour space (LCCS), in which an illumination-invariant representation of images can be obtained. Another strategy, presented by (Maddern et al. [2014]), might be a way to mitigate this issue: this uses an illumination-invariant transformation to reduce the problems associated with illumination changes due to sunlight and shadow.

- Combination of appearance-based and feature-based methods into a single scheme

The comparative experiments in Chapter 5 show that a major advantage of appearance-based methods over feature-based methods is realized in situations in which features cannot be extracted easily due to a featureless environment, or feasible matches cannot be found when the distance between two images becomes high. However, appearance-based methods perform poorly when the environment becomes too contrasty (as in the TENERIFE dataset). Combining them would be a possible direction for future research. A suitable binary descriptor, the local difference binary (LDB) descriptor, described in (Yang and Cheng [2014a]) demonstrated a good performance on image matching in our comparative study, and is computationally inexpensive.

- Dynamically adjusting configuration parameters of loop closure detection

The algorithm proposed in Chapter 6 has two key parameters that allow it to

---

carry out quadtree decomposition and loop closure validation. Currently, the performance of the proposed method is heavily dependent on tuning the parameters manually. Therefore, providing loosely estimated ranges for these two parameters that can be learned autonomously by the robot, thus allowing the system to adapt to different accuracy requirements and environments, is of great potential interest.

- Improvement on orientation estimation accuracy

The experimental results in Chapter 5 show that all the surveyed methods, as with odometry technique, inevitably exhibit drift because of compounding of small errors. Using a more sophisticated method for orientation estimation that incorporates the quadtree method will be explored. One possible extension will be employing an advanced probabilistic framework for making decisions on loop closures.

- Development of a novel visual SLAM system in the topological paradigm

Our next step of research is to generate topological maps and implement a novel visual SLAM system that integrates the proposed loop closure detection method and benefits from visual odometry measurements. The ideas developed in Clipp et al. [2010], which introduces a vSLAM system utilizing the parallelism strategy to perform visual odometry and loop closure, may be of assistance here. However, this is only a weak intuition and it would need to be thoroughly examined.

# Appendix A

## A.1 Histograms for the ISL dataset

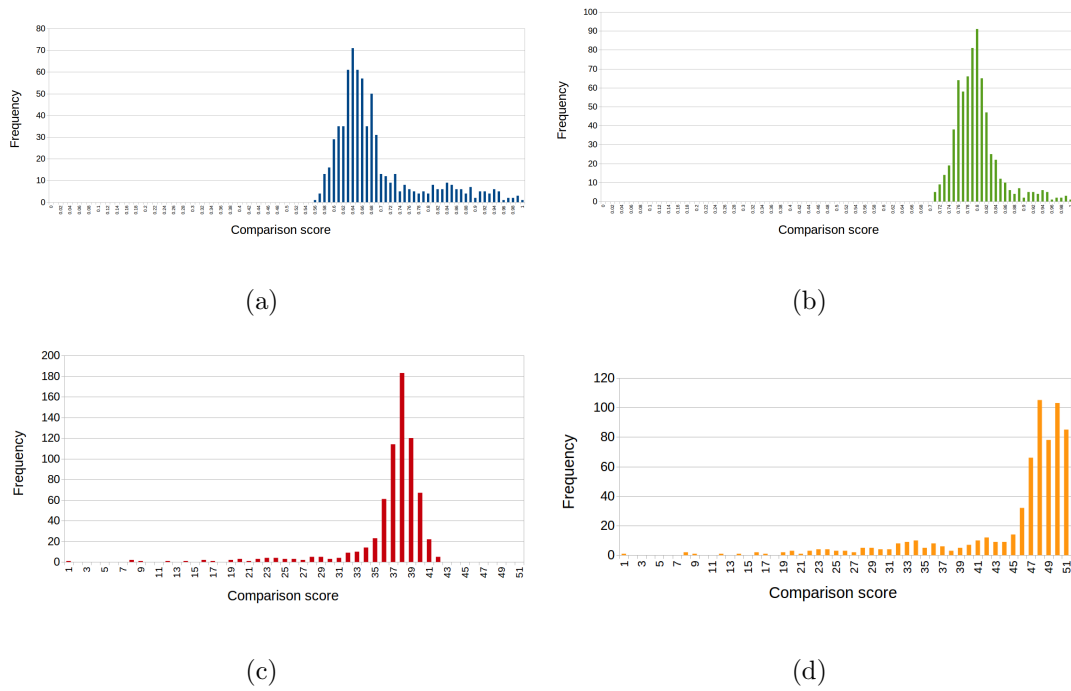


Figure A.1: Histograms of the comparison scores between the image of interest (Image 0) and all the images of the ISL 1 dataset, based on (a) CC; (b) EC; (c) EE; and (d) CE methods.

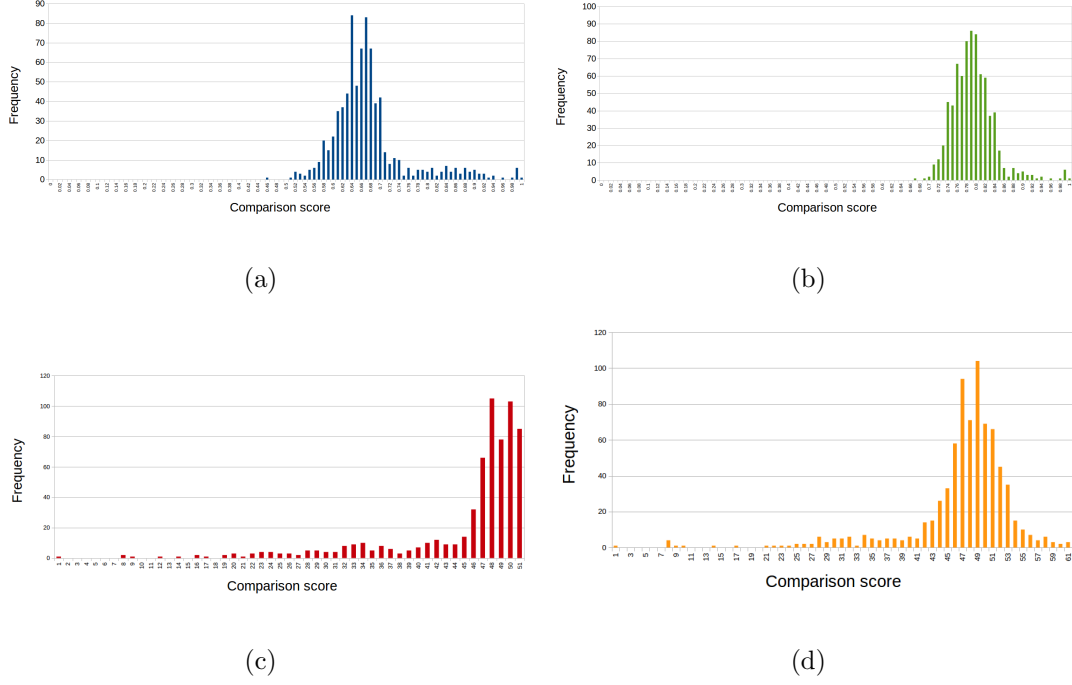


Figure A.2: Histograms of the comparison scores between the image of interest (Image 0) and all the images of the ISL 2 dataset, based on (a) CC; (b) EC; (c) EE; and (d) CE methods.

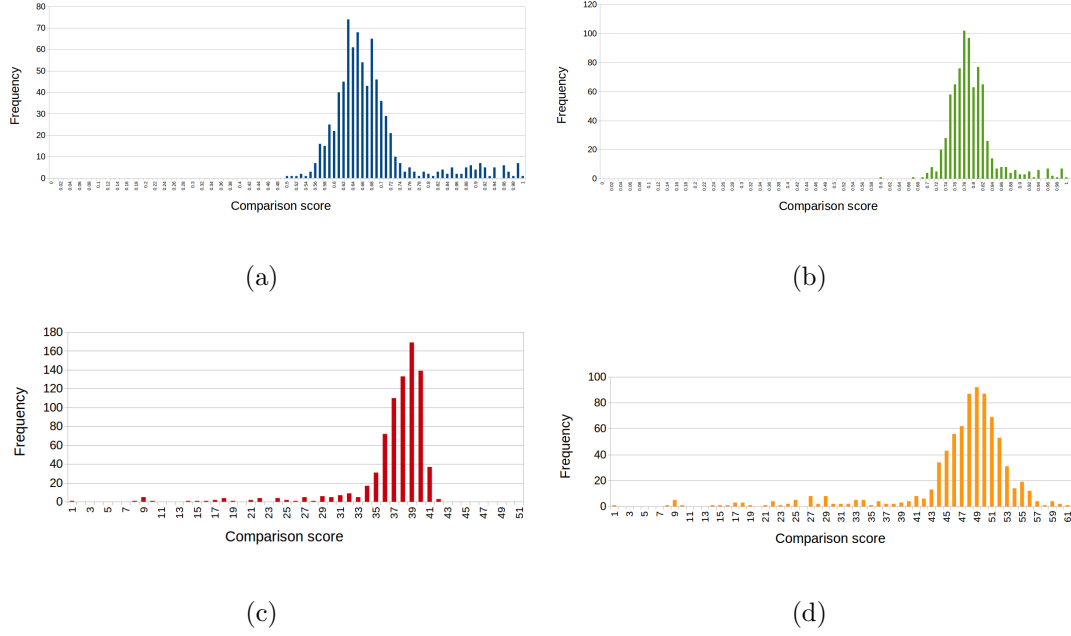


Figure A.3: Histograms of the comparison scores between the image of interest (Image 0) and all the images of the ISL 3 dataset, based on (a) CC; (b) EC; (c) EE; and (d) CE methods.

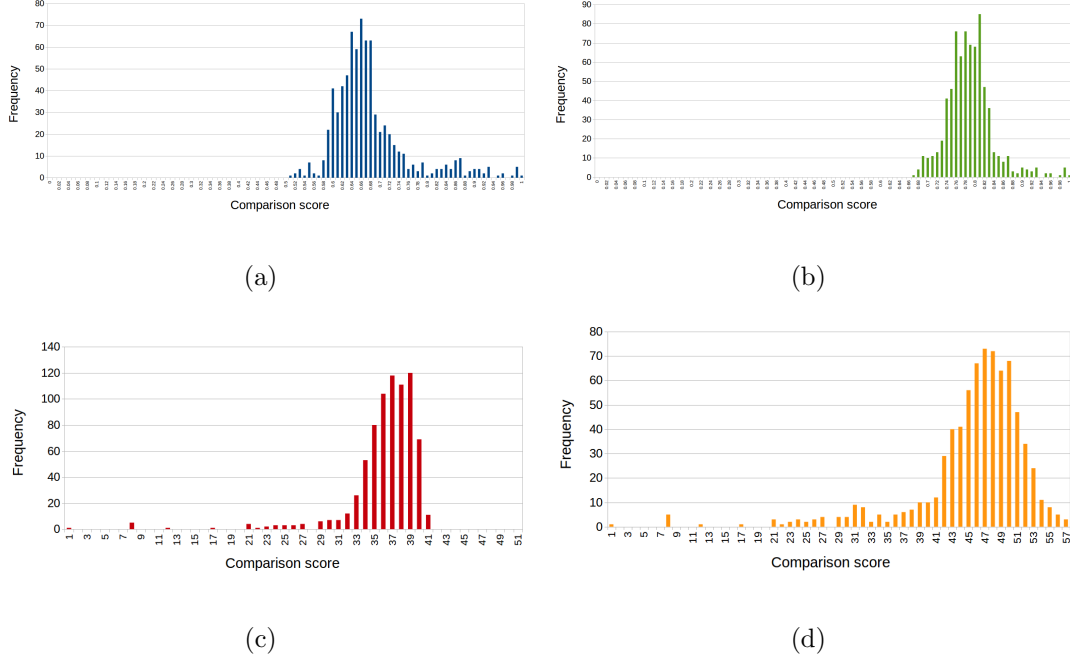


Figure A.4: Histograms of the comparison scores between the image of interest (Image 0) and all the images of the ISL 4 dataset, based on (a) CC; (b) EC; (c) EE; and (d) CE methods.

## A.2 PPCC coefficients

Table A.1: PPCC coefficients for the histograms shown as Figures A.1(a) - A.4(a)

PPCC_Correlation+Correlation					
	Lognormal	Weibull	Gamma	Normal	Logistic
<b>Dataset 1</b>	0.9689229411	0.977073821	0.9761692924	0.9055135176	0.9043657409
<b>Dataset 2</b>	0.9739792096	0.9700923616	0.9716783767	0.9217004098	0.9328743971
<b>Dataset 3</b>	0.9667442036	0.9678026507	0.968114015	0.902265964	0.9105684899
<b>Dataset 4</b>	0.979823357	0.9790401456	0.9798161287	0.9228542565	0.9309041085

Table A.2: PPCC coefficients for the histograms shown as Figures A.1(b) - A.4(b)

PPCC_Correlation+Euclidean					
	Lognormal	Weibull	Gamma	Normal	Logistic
<b>Dataset 1</b>	0.975212627	0.980127937	0.980127937	0.8466752474	0.8561546343
<b>Dataset 2</b>	0.9796200224	0.9696720903	0.9696720903	0.8773302345	0.8974234873
<b>Dataset 3</b>	0.9706217127	0.9655686403	0.9655686403	0.8582467033	0.876328254
<b>Dataset 4</b>	0.9879187903	0.9838271258	0.9838271258	0.8993611044	0.9127139509

Table A.3: PPCC coefficients for the histograms shown as Figures A.1(c) - A.4(c)

PPCC_Euclidean+Correlation					
	Lognormal	Weibull	Gamma	Normal	Logistic
<b>Dataset 1</b>	0.9856892722	0.9807424123	0.9832184175	0.9423689004	0.9495512008
<b>Dataset 2</b>	0.9856892722	0.9783347547	0.9833856761	0.9626279286	0.970204955
<b>Dataset 3</b>	0.9746428723	0.9671987264	0.9708461519	0.9348337956	0.9475945034
<b>Dataset 4</b>	0.983731783	0.9754622127	0.9815156935	0.9665888421	0.9750520421

Table A.4: PPCC coefficients for the histograms shown as Figures A.1(d) - A.4(d)

PPCC_Euclidean+Euclidean					
	Lognormal	Weibull	Gamma	Normal	Logistic
<b>Dataset 1</b>	0.9796472521	0.9473764055	0.9473764055	0.7878638068	0.8122517511
<b>Dataset 2</b>	0.9735673268	0.9380902614	0.9380902614	0.8129470282	0.8358561496
<b>Dataset 3</b>	0.9729729171	0.9401481371	0.9401481371	0.7746009186	0.7981593895
<b>Dataset 4</b>	0.9817191156	0.9508582257	0.9508582257	0.8212775985	0.8426485947

# References

- Abdel-Hakim, A. and Farag, A. CSIFT: A SIFT descriptor with color invariant characteristics. In *Int. Conf. Computer Vision and Pattern Recognition*, volume 2, pages 1978–1983, 2006. 19
- Agrawal, M., Konolige, K., and Blas, M. CenSurE: Centersurround extremas for realtime feature detection and matching. In *the 10th European Conference on Computer Vision*, volume 5305, pages 102–115, 2008. 17
- Alvarez-Mozos, J., Lopez, A., and Baldrich, R. Illuminant-invariant model-based road segmentation. In *IEEE Intelligent Vehicles Symposium*, pages 1175–1180, 2008. 44
- Anati, R. and Daniilidis, K. Constructing Topological Maps using Markov Random Fields and Loop-Closure Detection. In *Int. Conf. Advances in Neural Information Processing Systems*, pages 37–45. 2009. 4, 50
- Angeli, A., Doncieux, S., Meyer, J.-A., and Filliat, D. Real-time visual loop-closure detection. In *Int. Conf. Robotics and Automation*, pages 1842–1847, 2008a. 9, 51
- Angeli, A., Filliat, D., Doncieux, S., and Meyer, J.-A. A fast and incremental method for loop-closure detection using bags of visual words. *IEEE Transactions On Robotics, Special Issue on Visual SLAM*, 2008b. 4, 24, 57, 141
- Angeli, A., Doncieux, S., Meyer, J.-A., and Filliat, D. Visual topological SLAM and global localization. In *Int. Conf. Robotics and Automation*, pages 4300–4305, 2009. 49, 58

## REFERENCES

---

- Arroyo, R., Alcantarilla, P., Bergasa, L., Yebes, J., and Gamez, S. Bidirectional loop closure detection on panoramas for visual navigation. In *IEEE Intelligent Vehicles Symposium Proceedings*, pages 1378–1383, 2014. 4, 51, 81, 99, 172
- Artac, M., Jogan, M., and Leonardis, A. Mobile robot localization using an incremental eigenspace model. In *IEEE Conference of Robotics and Automation*, pages 1025–1030, 2002. 26, 28, 43
- Asensio, J., Montiel, J., and Montano, L. Goal directed reactive robot navigation with relocation using laser and vision. In *Proc. Int. Conf. Robotics and Automation*, pages 2905–2910, 1999. 33
- Asmar, D. *Vision-Inertial SLAM using Natural Features in Outdoor Environments*. PhD thesis, University of Waterloo, Canada, 2006. 33, 35
- Astua, C., Barber, R., Crespo, J., and Jardon, A. Object detection techniques applied on mobile robot semantic navigation. *Sensors*, 14(4):6734–6757, 2014. 11
- Bacca, B., Salvi, J., and Cufi, X. Appearance-based mapping and localization for mobile robots using a feature stability histogram. *Robotics and Autonomous Systems*, 59(10):840–857, 2011. 43, 46
- Badino, H., Huber, D., and Kanade, T. Real-time topometric localization. In *Proc. Int. Conf. Robotics and Automation*, pages 1635–1642, 2012. 4, 51, 141, 156
- Bay, H., Ess, A., Tuytelaars, T., and Gool, L. SURF: Speeded up robust features. *Computer Vision and Image Understanding*, 110(3):346–359, 2008. 17, 19
- Beaudet, P. Rotationally invariant image operators. In *Proc. Int. Joint Conf. on Pattern Recognition*, pages 579–583, 1978. 16
- Beeson, P., Jong, N. K., and Kuipers, B. Towards autonomous topological place detection using the extended Voronoi graph. In *Proc. Int. Conf. Robotics and Automation*, pages 4373–4379, 2005. 9

## REFERENCES

---

- Beeson, P., Modayil, J., and Kuipers, B. Factoring the mapping problem: Mobile robot map-building in the hybrid spatial semantic hierarchy. *Robotics Research*, 29(4):428–459, 2010. 11
- Bellotto, N., Burn, K., Fletcher, E., and Wermter, S. Appearance-based localization for mobile robots using digital zoom and visual compass. *Robotics and Autonomous Systems*, 56(2):143–156, 2008. 3, 28, 39, 46, 57
- Betke, M. and Gurvits, L. Mobile robot localization using landmarks. *Robotics and Automation*, 13(2):251–263, 1997. 36
- Blaer, P. and Allen, P. Topological mobile robot localization using fast vision techniques. In *Proc. Int. Conf. Robotics and Automation*, pages 1031–1036, 2002. 17, 28, 30, 43
- Blaer, P. and Allen, P. A Hybrid Approach to Topological Mobile Robot Localization. Technical report, Department of Computer Science, Columbia University, 2005. 31
- Blanco, J.-L., Fernández-madrigal, J.-A., and Gonzalez, J. Towards a unified Bayesian approach to hybrid metric-topological SLAM. *IEEE Transactions on Robotics*, 24:259–270, 2008. 10
- Bonin-Font, F., Ortiz, A., and Oliver, G. Visual navigation for mobile robots: A survey. *Journal of Intelligent and Robotic Systems*, 53(3):263–296, 2008. 9
- Booiij, O., Terwijn, B., Zivkovic, Z., and Krose, B. Navigation using an appearance based topological map. In *Proc. Int. Conf. Robotics and Automation*, pages 3927–3932, 2007. 9, 29, 57
- Bosse, M., Neal, P., Leonard, J., and Teller, S. SLAM in large-scale cyclic environments using the atlas framework. *International Journal of Robotics Research*, pages 1113–1139, 2004. 10
- Botterill, T., Mills, S., and Green, R. Bag-of-words-driven, single-camera simultaneous localization and mapping. *Journal of Field Robotics*, 28(2):204226, 2011. 9, 58

## REFERENCES

---

- Bradley, D., Patel, R. N. V., and Thayer, S. M. Real-time image-based topological localization in large outdoor environments. In *Proc. Int. Conf. on Intelligent Robots and Systems*, pages 3062–3069, 2005. 17, 56
- Bradski, G. The opencv library. *Doctor Dobbs Journal*, 25(11):384–386, 2000. 98, 156
- Briggs, A., Scharstein, D., Braziunas, D., Dima, C., and Wall, P. Mobile robot navigation using self-similar landmarks. In *Proc. Int. Conf. Robotics and Automation*, pages 1428–1434, 2000. 33, 36
- Bulow, H. and Birk, A. Fast and robust photomapping with an unmanned aerial vehicle (UAV). In *Proc. Int. Conf. Intelligent Robots and Systems, 2009*, pages 3368–3373, 2009. 22, 47, 48, 57
- Burgard, W., Brock, O., and Stachniss, C. Online learning for offroad robots: spatial label propagation to learn long-range traversability. In *Proc. of Robotics: Science and Systems*, pages 17 – 23, 2007. 52
- Burt, P. and Adelson, E. The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, 1983. 52
- Cadena, C., Gálvez, D., Ramos, F., Tardós, J., and Neira, J. Robust place recognition with stereo cameras. In *Proc. Int. Conf. Intelligent Robots and Systems*, 2010. 15, 24, 58
- Cadena, C., Galvez-Lopez, D., Tardos, J., and Neira, J. Robust place recognition with stereo sequences. *IEEE Transactions on Robotics*, 28(4):871–885, 2012. ISSN 1552-3098. 24, 59
- Calonder, M., Lepetit, V., Strecha, C., and Fua, P. Brief: Binary robust independent elementary features. In *Proceedings of the 11th European Conference on Computer Vision*, pages 778–792, 2010. 17, 20, 41
- Campos, F., Correia, L., and Calado, J. Global localization with non-quantized local image features. *Robotics and Autonomous Systems*, 60(8):1011 – 1020, 2012. 35, 70

## REFERENCES

---

- Cao, J., Labrosse, F., and Dee, H. A novel image similarity measure for place recognition in visual robotic navigation. In *Conf. of Towards Autonomous Robotic Systems*, pages 414–415, 2012. 6, 108, 171
- Cao, J., Labrosse, F., and Dee, H. An evaluation of image-based robot orientation estimation. In *Conf. of Towards Autonomous Robotic Systems*, pages 135–147, 2013. 6
- Casasent, D. and Psaltis, D. Position, rotation, and scale invariant optical correlation. *Applied Optics*, 15(7):1795–1799, 1976. 22
- Case, C., Suresh, B., Coates, A., and Ng, A. Y. Autonomous sign reading for semantic mapping. In *Proc. Int. Conf. on Robotics and Automation*, pages 3297–3303, 2011. 33, 36
- Cassandra, A., Kaelbling, L., and Kurien, J. Acting under uncertainty: discrete Bayesian models for mobile-robot navigation. In *Proc. Int. Conf. on Intelligent Robots and Systems*, volume 2, pages 963–972, 1996. 11
- Castle, R. O., Gawley, D. J., Klein, G., and Murray, D. W. Towards simultaneous recognition, localization and mapping for hand-held and wearable cameras. In *Proc. Int. Conf. on Robotics and Automation*, pages 4102–4107, 2007. 42
- Chandrasekhar, V., Takacs, G., Chen, D., Tsai, S., Grzeszczuk, R., and Girod, B. CHoG: Compressed histogram of gradients A low bit-rate feature descriptor. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pages 2504–2511, 2009a. 27
- Chandrasekhar, V., Takacs, G., Chen, D., Tsai, S., Singh, J., and Girod, B. Transform coding of image feature descriptors. In *Society of Photo-Optical Instrumentation Engineers Conference Series*, volume 7257, page 9, 2009b. 28
- Chang, C.-K., Siagian, C., and Itti, L. Mobile robot vision navigation & localization using Gist and Saliency. In *Proc. Int. Conf. Intelligent Robots and Systems*, pages 4147–4154, 2010. 41, 58

## REFERENCES

---

- Chang, C.-K., Siagian, C., and Itti, L. Mobile robot vision navigation and obstacle avoidance based on gist and saliency algorithms. *Journal of Vision*, 11(11):927, 2011. 41
- Chapoulie, A., Rives, P., and Filliat, D. A spherical representation for efficient visual loop closing. In *Int. Conf. Computer Vision Workshops*, pages 335–342, 2011. 9
- Chatila, R. and Laumond, J. Position referencing and consistent world modeling for mobile robots. In *Proc. Int. Conf. Robotics and Automation*, volume 2, pages 138–145, 1985. 1, 10
- Cheng, Y., Maimone, M., and Matthies, L. Visual odometry on the Mars exploration rovers - a tool to ensure accurate driving and science imaging. *Robotics Automation Magazine*, 13(2):54–62, 2006. 3
- Chong, K. S. and Kleeman, L. Feature-based mapping in real, large scale environments using an ultrasonic array. *The International Journal of Robotics Research*, 18(1):3–19, 1999. 2
- Choset, H. and Nagatani, K. Topological simultaneous localization and mapping (SLAM): toward exact localization without explicit localization. *IEEE Transactions on Robotics and Automation*, 17(2):125–137, 2001. 9, 11
- Clemente, L., Davison, A., Reid, I., Neira, J., and Tardós, J. D. Mapping large loops with a single hand-held camera. In *Proc. Robotics: Sci. Syst*, 2007. 50
- Clipp, B., Lim, J., Frahm, J.-M., and Pollefeys, M. Parallel, real-time visual SLAM. In *Proc. Int. Conf. on Intelligent Robots and Systems*, pages 3961–3968, 2010. 3, 177
- Comport, A. I., Malis, E., and Rives, P. Real-time quadrifocal visual odometry. *International Journal of Robotics Research*, 29(2-3):245–266, 2010. 58
- Corke, P., Paul, R., Churchill, W., and Newman, P. Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localisation. In *Proc. Int. Conf. Intelligent Robots and Systems*, pages 2085–2092, 2013. 44

## REFERENCES

---

- Crowley, J. World modeling and position estimation for a mobile robot using ultrasonic ranging. In *Proc. Int. Conf. Robotics and Automation*, volume 2, pages 674–680, 1989. 2
- Cummins, M. and Newman, P. FAB-MAP: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research*, 27(6):647–665, 2008a. 9, 15, 23, 49, 50, 57, 141, 172
- Cummins, M. and Newman, P. Accelerated appearance-only SLAM. *Int. Conf. on In Robotics and Automation*, pages 1828–1833, 2008b. 81
- Cummins, M. and Newman, P. Appearance-only SLAM at large scale with FAB-MAP 2.0. *The International Journal of Robotics Research*, 2010. 3, 4, 58
- Dalal, N. and Triggs, B. Histograms of oriented gradients for human detection. In Schmid, C., Soatto, S., and Tomasi, C., editors, *Int. Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 886–893, 2005. 113
- Davison, A. Real-time simultaneous localisation and mapping with a single camera. In *Proc. Int. Conf. on Computer Vision*, volume 2, pages 1403–1410, 2003. 1, 9, 13, 14, 56
- Duda, R. O., Hart, P. E., and Stork, D. G. Pattern classification. 2nd. *Edition*. New York, 2001. 82, 83
- Durrant-Whyte, H. and Bailey, T. Simultaneous localisation and mapping (SLAM): Part I the essential algorithms. *IEEE Robotics and Automation Magazine*, 2:2006, 2006. 3
- Durrant-Whyte, H., Rye, D., and Nebot, E. Localization of autonomous guided vehicles. In *Robotics Research*, pages 613–625. 1996. 1
- Eade, E. and Drummond, T. Scalable monocular SLAM. In *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 469–476, 2006. 13, 56
- Eade, E. and Drummond, T. Unified loop closing and recovery for real time monocular SLAM. In *British Machine Vision Conference*, 2008. 57

## REFERENCES

---

- Ekvall, S., Jensfelt, P., and Kragic, D. Integrating active mobile robot object recognition and SLAM in natural environments. In *Proc. Int. Conf. Intelligent Robots and Systems*, 2006. 37
- Elfes, A. Using occupancy grids for mobile robot perception and navigation. *Computer*, 22(6):46–57, 1989. 9
- Estrada, C., Neira, J., and Tardos, J. Hierarchical SLAM: Real-time accurate mapping of large environments. *IEEE Transactions on Robotics*, 21(4):588–596, 2005. 10
- Eze, L. and Benosman, R. Visual localization using an optimal sampling of bags-of-features with entropy. *Proc. Int. Conf. Intelligent Robots and Systems*, pages 1332 – 1338, 2007. 52
- Fairfield, N. and Maxwell, B. Mobile robot localization with sparse landmarks. In *Proc. SPIE*, 2001. 33, 36
- Fazi-Ersi, E. and Tsotsos, J. K. Histogram of oriented uniform patterns for robust place recognition and categorization. *International Journal of Robotics Research*, 31(4):468–483, 2012. 17, 32, 59
- Fei-Fei, L. and Perona, P. A Bayesian hierarchical model for learning natural scene categories. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, volume 2, pages 524–531, 2005. 29
- Ferdaus, S., Vardy, A., Mann, G., and Gosine, R. Comparing global measures of image similarity for use in topological localization of mobile robots. In *Canadian Conference on Electrical and Computer Engineering*, 2008. 25, 43
- Fernández, L., Payá, L., Reinoso, Ó., and Amorós, F. Appearance-based visual odometry with omnidirectional images — A practical application to topological mapping. In *Proc. of ICINCO*, pages 205–210, 2011. 47, 48
- Filliat, D. A visual bag of words method for interactive qualitative localization and mapping. In *Proc. Int. Conf. on Robotics and Automation*, 2007. 24, 28, 57

## REFERENCES

---

- Fuentes-Pacheco, J., Ruiz-Ascencio, J., and Rendon-Mancha, J. Visual simultaneous localization and mapping: a survey. *Artificial Intelligence Review*, pages 1–27, 2012. 9
- Gallegos, G. and Rives, P. Indoor SLAM based on composite sensor mixing laser scans and omnidirectional images. In *Proc. Int. Conf. Robotics and Automation*, pages 3519–3524, 2010. 45
- García, D., Rojo, L., Aparicio, A., Castelló, L., and García, O. Visual odometry through appearance- and feature-based method with omnidirectional images. *Journal of Robotics*, pages 1–13, 2012. 47, 48, 59
- Garcia-Fidalgo, E. and Ortiz, A. Probabilistic appearance-based mapping and localization using visual features. In *The 6th Iberian Conference on Pattern Recognition and Image Analysis*, volume 7887, pages 277–285, 2013. 4
- Gaspar, J., Winters, N., and Santos-Victor, J. Vision-based navigation and environmental representations with an omni-directional camera. *IEEE Transactions on Robotics and Automation*, 16:890–898, 2000. 26, 28, 43
- Gelfand, I. and Yaglom, A. Calculation of the amount of information about a random function contained in another such function. *American Mathematical Society*, 2(12):199–246, 1959. 86
- Glover, A., Maddern, W., Warren, M., Reid, S., Milford, M., and Wyeth, G. OpenFABMAP: An open source toolbox for appearance-based loop closure detection. In *Proc. Int. Conf. Robotics and Automation*, pages 4730–4735, 2012. 98
- Glover, A., Maddern, W., Milford, M., and Wyeth, G. FAB-MAP + RatSLAM: Appearance-based SLAM for multiple times of day. In *Proc. Int. Conf. Robotics and Automation*, pages 3507–3512, 2010. 15, 42, 58
- Goecke, R., Asthana, A., Pettersson, N., and Petersson, L. Visual vehicle ego-motion estimation using the Fourier-Mellin transform. In *Intelligent Vehicles Symposium*, pages 450–455, 2007. 22, 47

## REFERENCES

---

- Goedemé, T., Nuttin, M., Tuytelaars, T., and Van Gool, L. Omnidirectional vision based topological navigation. *International Journal of Computer Vision*, 74(3):219–236, 2007. 4, 46, 49, 57
- Goedemé, T., Tuytelaars, T., and Gool, L. *Visual Topological Map Building in Self-similar Environments*, volume 15. Springer Berlin Heidelberg, 2008. 9, 12
- Gonzalez-Barbosa, J.-J. and Lacroix, S. Rover localization in natural environments by indexing panoramic images. In *Proc. Int. Conf. Robotics and Automation*, volume 2, pages 1365–1370, 2002. 43
- Goshtasby, A. *Image Registration - Principles, Tools and Methods*. Advances in Computer Vision and Pattern Recognition. Springer, 2012. 82
- Grisetti, G., Stachniss, C., and Burgard, W. Improved techniques for grid mapping with Rao-Blackwellized particle filters. *IEEE Transactions on Robotics*, 23(1):34–46, 2007. 9
- Gross, H.-M., Koenig, A., Schroeter, C., and Boehme, H.-J. Omnivision-based probabilistic self-localization for a mobile shopping assistant continued. In *Proc. Int. Conf. Intelligent Robots and Systems*, volume 2, pages 1505–1511, 2003. 29, 56
- Guivant, J. and Nebot, E. Optimization of the simultaneous localization and map building algorithm for real time implementation. *IEEE Transactions on Robotics and Automation*, 17:242–257, 2001. 1
- Guivant, J., Nebot, E., and Baiker, S. Localization and map building using laser range sensors in outdoor applications. *Journal of Robotic Systems*, 17(10): 565–583, 2000. 2
- Guivant, J. E., Nebot, E. M., Nieto, J. I., and Masson, F. R. Navigation and mapping in large unstructured environments. *International Journal of Robotics Research*, 23:449–472, 2004. 52
- Gutierrez-Osuna, R. and Luo, R. C. LOLA probabilistic navigation for topological maps. *AI Magazine*, 17(1):55–62, 1996. 11

## REFERENCES

---

- Harris, C. and Stephens, M. A combined corner and edge detector. In *Proc. of Fourth Alvey Vision Conference*, pages 147–151, 1988. 16
- Hashem, T. and Andreas, Z. Vision based localization of mobile robots using kernel approaches. In *Proc. Int. Conf. Intelligent Robots and Systems*, 2004. 39, 56
- Hayet, J., Devy, M., and Lerasle, F. Visual landmarks detection and recognition for mobile robot navigation. In *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, pages 313–318, 2003. 33, 34, 56
- Ho, K. and Newman, P. Detecting loop closure with scene sequences. *International Journal of Computer Vision*, 74(3):261–286, 2007. 9, 50, 57
- Hofmann-Wellenhof, B., Lichtenegger, H., and Collins, J. GPS - Global positioning system. Theory and practice. *GPS - Global Positioning System. Theory and practice.*, 1997. 73
- Hou, Y., Zhang, H., and Zhou, S. Convolutional neural network-based image representation for visual loop closure detection. In *Proc. Int. Conf. on Information and Automation*, pages 2238–2245, 2015. 141
- Howard, A. and Kitchen, L. Navigation using natural landmarks. *Robotics and Automous Systems*, 26:348–355, 1999. 33
- Hua, G., Brown, M., and Winder, S. Discriminant embedding for local image descriptors. In *Int. Conf. on Computer Vision*, 2007. 28
- Huh, J., Lee, K., Chung, W. K., Jeong, W. S., and Kim, K. K. Mobile robot exploration in indoor environment using topological structure with invisible barcode. In *Proc. Int. Conf. Intelligent Robots and Systems*, pages 5265–5272, 2006. 33, 36
- Huntington, E. V. Mathematics and statistics, with an elementary account of the correlation coefficient and the correlation ratio. *American Mathematical Monthly*, 26(10):421–435, 1919. 84

## REFERENCES

---

- Huynh, D., Saini, A., and Liu, W. Evaluation of three local descriptors on low resolution images for robot navigation. In *Int. Conf. Image and Vision Computing New Zealand*, pages 113–118, 2009. 18
- Iocchi, L. and Nardi, D. Self-localization in the robocup environment. In *RoboCup-99: Robot Soccer World Cup III*, volume 1856 of *Lecture Notes in Computer Science*, pages 318–330. Springer Berlin Heidelberg, 2000. 36
- Ishiguro, H. and Tsuji, S. Image-based memory of environment. In *Proc. Int. Conf. Intelligent Robots and Systems*, page 634639, 1996. 25
- Ishizuka, D., Yamashita, A., Kawanishi, R., Kaneko, T., and Asama, H. Self-localization of mobile robot equipped with omnidirectional camera using image matching and 3d-2d edge matching. In *Computational Methods for the Innovative Design of Electrical Devices’11*, pages 272–279, 2011. 27
- Itti, L., Koch, C., and Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998. 17
- Jain, R., Kasturi, R., and Schunck, B. *Machine Vision*. McGraw-Hill, Inc., 1995. 137
- Jang, G., Lee, S., and Kweon, I. Color landmark based self-localization for indoor mobile robots. In *Proc. Int. Conf. Robotics and Automation*, volume 1, pages 1037–1042, 2002. 36
- Jennings, C., Murray, D., and Little, J. Cooperative robot localization with vision-based mapping. *Proc. Int. Conf. on Robotics and Automation*, pages 2659–2665, 1999. 33
- Jensfelt, P., Kragic, D., Folkesson, J., and Björkman, M. A framework for vision based bearing only 3D SLAM. In *Proc. Int. Conf. on Robotics and Automation*, 2006. 18, 56
- Jogan, M. and Leonardis, A. Panoramic eigenimages for spatial localisation. In *8th CAIP*, pages 558–567. Springer Verlag, 1999. 26

## REFERENCES

---

- Jogan, M. and Leonardis, A. Robust localization using eigenspace of spinning-images. In *IEEE Workshop on Omnidirectional Vision*, 2000. 112
- Jones, C. B., Abdelmoty, A. I., Finch, D., Fu, G., and Vaid, S. The SPIRIT spatial search engine: Architecture, ontologies and spatial indexing. *Lecture Notes in Computer Science*, 3234:125–139, 2004. 52
- Kaess, M. and Dellaert, F. Probabilistic structure matching for visual SLAM with a multi-camera rig. *Computer Vision and Image Understanding*, 114:286–296, Feb 2010. 9, 58
- Kalman, R. E. A new approach to linear filtering and prediction problems. *ASME. J. Basic Eng.*, 82(1):35–45, 1960. 74
- Kalman, R. E. and Bucy, R. S. New results in linear filtering and prediction theory. In *Trans. ASME, Ser. D, J. Basic Eng.*, page 109, 1961. 74
- Karlsson, N., Di Bernardo, E., Ostrowski, J., Goncalves, L., Pirjanian, P., and Munich, M. The vSLAM algorithm for robust localization and mapping. In *Proc. Int. Conf. Robotics and Automation*, pages 24–29, 2005. 14
- Kawewong, A., Tongprasit, N., Tangruamsub, S., and Hasegawa, O. Online and incremental appearance-based SLAM in highly dynamic environments. *I. J. Robotic Res.*, 30(1):33–55, 2011. 15, 58
- Kazik, T. and Goktogan, A. Visual odometry based on the Fourier-Mellin transform for a rover using a monocular ground-facing camera. In *Proc. Int. Conf. Mechatronics*, pages 469–474, 2011. 22
- Ke, Y. and Sukthankar, R. PCA-SIFT: A more distinctive representation for local image descriptors. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pages 506–513, 2004. 19, 26
- Klasing, K., Lidoris, G., Bauer, A., Rohrmüller, F., Wollherr, D., and Buss, M. The autonomous city explorer: Towards semantic navigation in urban environments. In *Proc. Int. Workosop on Cognition for Technical Systems*, 2008. 11

## REFERENCES

---

- Klein, G. and Murray, D. Parallel tracking and mapping for small AR workspaces. In *Proc. Int. Symposium on Mixed and Augmented Reality*, 2007. 14
- Knappek, M., Oropeza, R. S., Swain, R., David, O., and Kriegman, D. J. Selecting promising landmarks. In *Proc. Int. Conf. on Robotics and Automation*, pages 3771–3777, 2000. 34
- Koch, O., Walter, M., Huang, A. S., and Teller, S. Ground robot navigation using uncalibrated cameras. In *Proc. Int. Conf. on Robotics and Automation*, Anchorage, AL, USA, May 2010. 50, 58
- Konolige, K. and Agrawal, M. FrameSLAM: From bundle adjustment to real-time visual mapping. *IEEE Transactions on Robotics*, 24(5):1066–1077, 2008. 14
- Konolige, K., Bowman, J., Chen, J., Mihelich, P., Calonder, M., Lepetit, V., and Fua, P. View-based maps. *Int. J. Rob. Res.*, 29(8):941–957, 2010. 9, 14, 58
- Konolige, K., Marder-Eppstein, E., and Marthi, B. Navigation in hybrid metric-topological maps. In *Proc. Int. Conf. Robotics and Automation*, pages 3041–3047, 2011. 10
- Korrapati, H. and Mezouar, Y. Vision-based sparse topological mapping. *Robotics and Autonomous Systems*, 62(9):1259 – 1270, 2014. Intelligent Autonomous Systems. 9
- Kosecka, J. and Li, F. Vision based topological Markov localization. In *Proc. Int. Conf. Robotics and Automation*, pages 1481–1486, 2004. 11, 28, 29
- Kouzoubov, K. and Austin, D. Hybrid topological/metric approach to SLAM. In *Proc. Int. Conf. Robotics and Automation.*, volume 1, pages 872–877, 2004. 10
- Košecká, J., L., Z., Barbara, P., and Duric, Z. Qualitative image based localization in indoors environments. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pages 3–10, 2003. 31, 56

## REFERENCES

---

- Kreucher, C. and Lakshmanan, S. LANA: a lane extraction algorithm that uses frequency domain features. *IEEE Transactions on Robotics & Automation*, 15(2):343–350, 1999. 52
- Kröse, B., Vlassis, N., Bunschoten, R., and Motomura, Y. A probabilistic model for appearance-based robot localization. In *First European Symposium on Ambience Intelligence*, pages 264–274, 2000. 26, 43
- Kriechbaumer, T., Blackburn, K., Breckon, T., Hamilton, O., and Riva-Casado, M. Quantitative evaluation of stereo visual odometry for autonomous vessel localisation in inland waterway sensing applications. *Sensors*, 15(12):31869–31887, 2015. 46, 48
- Kristopher, R. B. and Wesley, H. H. Loop closing in topological maps. In *Proc. Int. Conf. Robotics and Automation*, 2005. 11
- Kuipers, B. and Byun, Y.-T. A robot exploration and mapping strategy based on a semantic hierarchy of spatial representations. *Journal of Robotics and Autonomous Systems*, 8:47–63, 1991. 9
- Kuipers, B., Modayil, J., Beeson, P., MacMahon, M., and Savelli, F. Local metrical and global topological maps in the hybrid spatial semantic hierarchy. In *Int. Conf. on Robotics & Automation*, pages 4845–4851, 2004. 10
- Kunttu, I., Lepisto, L., Rauhamaa, J., and Visa, A. Multiscale fourier descriptor for shape-based image retrieval. In *Proc. Int. Conf. Pattern Recognition*, volume 2, pages 765–768, 2004. 17
- Labbani-Igbida, O., Charron, C., and Mouaddib, E. M. Haar invariant signatures and spatial recognition using omnidirectional visual information only. *Autonomous Robots*, 30(3):333–349, 2011. 43, 45
- Labbe, M. and Michaud, F. Appearance-based loop closure detection for online large-scale and long-term operation. *IEEE Transactions on Robotics*, 29(3):734–745, 2013. 51, 59, 141

## REFERENCES

---

- Labrosse, F. The visual compass: Performance and limitations of an appearance-based method. *Journal of Field Robotics*, 23(10):913–941, 2006. 47, 108, 109, 175
- Labrosse, F. Short and long-range visual navigation using warped panoramic images. *Robotics and Autonomous Systems*, 55(9):675 – 684, 2007. 3
- Lamon, P., Nourbakhsh, I., Jensen, B., and Siegwart, R. Deriving and matching image fingerprint sequences for mobile robot localization. In *Proc. Int. Conf. Robotics and Automation*, 2001. 34
- Lamon, P., Tapus, A., Glauser, E., and Tomatis, N. Environmental modeling with fingerprint sequences for topological global localization. In *Proc. Int. Conf. Intelligent Robots and Systems*, pages 3781–3786, 2003. 11
- Laurent Kneip, M. C. and Siegwart, R. Robust real-time visual odometry with a single camera and an IMU. In *Proc. of the British Machine Vision Conference*, pages 16.1–16.11, 2011. 4
- Lazebnik, S., Schmid, C., and Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2:2169–2178, 2006. 52
- Leonard, J. and Durrant-Whyte, H. Mobile robot localization by tracking geometric beacons. *IEEE Transactions on Robotics and Automation*, 7(3):376–382, 1991. 12
- Leutenegger, S., Chli, M., and Siegwart, R. BRISK: Binary robust invariant scalable keypoints. *Proc. Int. Conf. Computer Vision*, 0:2548–2555, 2011. 20, 21
- Levin, A. and Szeliski, R. Visual odometry and map correlation. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pages 611–618, 2004. 49
- Li, F. Probabilistic location recognition using reduced feature set. In *Proc. Int. Conf. on Robotics and Automation*, 2006. 35

## REFERENCES

---

- Li, J., Wang, J., and Wiederhold, G. IRM: integrated region matching for image retrieval. In *Proc. Int. Conf. Multimedia*, pages 147–156, 2000. 29
- Li, L., Su, H., Li, F., and Xing, E. P. Object Bank: A high-level image representation for scene classification & semantic feature sparsification. *Advances in Neural Information Processing Systems*, pages 1378–1386, 2010. 52
- Lin, H.-Y., Lin, Y.-H., and Yao, J.-W. Scene change detection and topological map construction using omnidirectional image sequences. In *MVA*, pages 57–60, 2013. 9, 59
- Linde, O. and Lindeberg, T. Object recognition using composed receptive field histograms of higher dimensionality. In *Proc. Int. Conf. Pattern Recognition*, 2004. 19
- Linåker, F. and Ishikawa, M. Real-time appearance-based Monte Carlo localization. *Robotics and Autonomous Systems*, 54(3):205–220, 2006. 43
- Liu, M. and Siegwart, R. Topological mapping and scene recognition with lightweight color descriptors for an omnidirectional camera. *IEEE Transactions on Robotics*, 30(2):310–324, 2014. 28, 35, 70
- Liu, Y. and Zhang, H. Visual loop closure detection with a compact image descriptor. In *Proc. Int. Conf. Intelligent Robots and Systems*, pages 1051–1056, 2012. 51, 141
- Lowe, D. Object recognition from local scale-invariant features. In *Proc. Int. Conf. Computer Vision*, pages 1150–1157, 1999. 17, 18
- Lowe, D. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 17, 108, 109
- Maddern, W., Milford, M., and Wyeth, G. CAT-SLAM: Probabilistic localisation and mapping using a continuous appearance-based trajectory. *Int. J. Rob. Res.*, 31(4):429–451, 2012. 3, 15, 59
- Maddern, W., Stewart, A., McManus, C., Upcroft, B., Churchill, W., and Newman, P. Illumination invariant imaging: Applications in robust vision-based

## REFERENCES

---

- localisation, mapping and classification for autonomous vehicles. In *Proc. Int. Conf. Robotics and Automation*, 2014. 3, 44, 176
- Maddern, W. and Vidas, S. Towards robust night and day place recognition using visible and thermal imaging. *Rss Beyond Laser & Vision Alternative Sensing Techniques for Robotic Perception*, 2012. 16
- Maddern, W., Milford, M., and Wyeth, G. Continuous appearance-based trajectory SLAM. In *Proc. Int. Conf. Robotics and Automation*, pages 3595–3600, 2011. 15, 58
- Magnabosco, M. and Breckon, T. P. Cross-spectral visual simultaneous localization and mapping (SLAM) with sensor handover. *Robotics and Autonomous Systems*, 61(2):195 – 208, 2013. 3, 16, 59
- Maimone, M., Cheng, Y., and Matthies, L. Two years of visual odometry on the Mars Exploration Rovers. *Journal of Field Robotics*, 24(3):169–186, 2007. 46, 57
- Majdik, A., Albers-Schoenberg, Y., and Scaramuzza, D. MAV urban localization from Google street view data. In *Proc. Int. Conf. Intelligent Robots and Systems*, pages 3979–3986, 2013. 42
- Mariottini, G. and Roumeliotis, S. I. Active vision-based robot localization and navigation in a visual memory. In *Proc. Int. Conf. Robotics and Automation*, pages 6192–6198, 2011. 24, 58
- Mata, M., Armingol, J. M., Escalera, A. D. L., and Salichs, M. A. Using learned visual landmarks for intelligent topological navigation of mobile robots. In *Proc. Int. Conf. Robotics and Automation*, pages 1324–1329, 2003. 36
- Matas, J., Chum, O., Urban, M., and Pajdla, T. Robust wide-baseline stereo from maximally stable extremal regions. *Image Vision Comput.*, 22(10):761–767, 2004. 17
- Matsumoto, Y., Inaba, M., and Inoue, H. Visual navigation using view-sequenced route representation. *Proc. Int. Conf. Robotics and Automation*, pages 83–88, 1996. 38

## REFERENCES

---

- Matsumoto, Y., Sakai, K., Inaba, M., and Inoue, H. Visual navigation using omnidirectional view sequence. In *Proc. Int. Conf. Intelligent Robots and Systems*, pages 317–322, 1999. 38
- Matsumoto, Y., Sakai, K., Inaba, M., and Inoue, H. View-based approach to robot navigation. In *Proc. Int. Conf. Intelligent Robots and Systems*, pages 1702–1708, 2000. 38
- Matthies, L. and Shafer, S. Error modeling in stereo navigation. *IEEE Journal of Robotics and Automation*, 3(3):239–250, 1987. 46, 47
- Mei, C., Sibley, G., Cummins, M., Newman, P., and Reid, I. A constant time efficient stereo SLAM system. In *British Machine Vision Conference*, 2009. 3, 9, 14, 53, 58
- Mei, C., Sibley, G., Cummins, M., Newman, P., and Reid, I. RSLAM: A system for large-scale mapping in constant-time using stereo. *International Journal of Computer Vision*, pages 1–17, 2010. 14, 58
- Menegatti, E., Zoccarato, M., Pagello, E., and Ishiguro, H. Hierarchical image-based localisation for mobile robots with Monte-Carlo localisation. In *Proc. of European Conference on Mobile Robots*, pages 13–20, 2003. 25, 28, 45
- Menegatti, E., Maeda, T., and Ishiguro, H. Image-based memory for robot navigation using properties of omnidirectional images. *Robotics and Autonomous Systems*, 47(4):251–267, 2004a. 17, 25, 56
- Menegatti, E., Zoccarato, M., Pagello, E., and Ishiguro, H. Image-based Monte-Carlo localisation without a map. *Robotics and Autonomous Systems*, 48:17–30, 2004b. 25, 43, 56
- Mikolajczyk, K. and Schmid, C. Indexing based on scale invariant interest points. In *Proc. Int. Conf. Computer Vision*, volume 1, pages 525–531, 2001. 16
- Mikolajczyk, K. and Schmid, C. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004. 17

## REFERENCES

---

- Mikolajczyk, K. and Schmid, C. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005. 17, 18, 19
- Milford, M. Vision-based place recognition: how low can you go? *I. J. Robotic Res.*, 32(7):766–789, 2013. 16, 42, 59
- Milford, M., Scheirer, W., Vig, E., Glover, A., Baumann, O., Mattingley, J., and Cox, D. Condition-invariant, top-down visual place recognition. In *Proc. Int. Conf. Robotics and Automation*, pages 5571–5577, 2014. 175
- Milford, M. and Wyeth, G. Single camera vision-only SLAM on a suburban road network. In *Proc. Int. Conf. Robotics and Automation*, pages 3684–3689, 2008a. 47, 48, 57
- Milford, M. and Wyeth, G. Mapping a suburb with a single camera using a biologically inspired SLAM system. *IEEE Transactions on Robotics*, 24(5):1038–1053, 2008b. 15
- Milford, M. and Wyeth, G. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In *Proc. Int. Conf. Robotics and Automation*, pages 1643–1649, 2012. 15
- Milford, M., Wyeth, G., and Prasser, D. RatSLAM: A hippocampal model for simultaneous localization and mapping. In *Proc. Int. Conf. Robotics and Automation*, volume 1, pages 403–408, 2004. 15, 56
- Möller, R., Horst, M., and Fleer, D. Illumination tolerance for visual navigation with the holistic min-warping method. *Robotics*, 3(1):22–67, 2014. 43, 44
- Montemerlo, M., Thrun, S., Koller, D., and Wegbreit, B. FastSLAM: A factored solution to the simultaneous localization and mapping problem. In *Proc. Int. Conf. on Artificial Intelligence*, 2002. 9, 13, 56
- Moravec, H. Obstacle avoidance and navigation in the real world by a seeing robot rover. Technical report, Robotics Institute, Carnegie Mellon University, 1980. 46

## REFERENCES

---

- Moravec, H. Sensor fusion in certainty grids for mobile robots. *AI Mag.*, 9(2): 61–74, 1988. 9
- Morel, J.-M. and Yu, G. ASIFT: A new framework for fully affine invariant image comparison. *Siam Journal on Imaging Sciences*, 2(2):438–469, 2009. 19, 42
- Mortensen, E., Deng, H., and Shapiro, L. A SIFT descriptor with global context. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, volume 1, pages 184–190, 2005. 19
- Mouragnon, E., Lhuillier, M., Dhome, M., Dekeyser, F., and Sayd, P. Real time localization and 3D reconstruction. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pages 363–370, 2006. 14
- Murillo, A. C., Guerrero, J. J., and Sags, C. SURF features for efficient robot localization with omnidirectional images. In *Proc. Int. Conf. Robotics and Automation*, pages 3901–3907, 2007. 28
- Murillo, A. and Kosecka, J. Experiments in place recognition using gist panoramas. In *Int. Conf. on Computer Vision Workshops*, pages 2196–2203, 2009. 41, 58
- Murphy, K. Bayesian Map Learning in Dynamic Environments. In *Proc. Int. Conf. Neural Information Processing Systems*, volume 12, pages 1015–1021, 1999. 13
- Neal, M. and Labrosse, F. Rotation-invariant appearance based maps for robot navigation using an artificial immune network algorithm. In *Congress on Evolutionary Computation*, volume 1, pages 863–870, 2004. 9
- Neubert, P., Sunderhauf, N., and Protzel, P. Appearance change prediction for long-term navigation across seasons. *Robotics and Autonomous Systems*, 69(1): 198–203, 2013. 16
- Newman, P., Sibley, G., Smith, M., Cummins, M., and Harrison, A. Navigating, recognizing and describing urban spaces with vision and lasers. *International Journal of Robotics Research*, 28:1406–1433, 2009. 101

## REFERENCES

---

- Nistér, D. and Stewénus, H. Scalable recognition with a vocabulary tree. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pages 2161–2168, 2006. 23, 46
- Nistr, D., Naroditsky, O., and Bergen, J. Visual odometry. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2004. 9, 14, 56
- Oliva, A. and Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42: 145–175, 2001. 22, 39, 114
- Oliva, A. and Torralba, A. Building the gist of a scene: the role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006. 39, 114
- Olson, C., Matthies, L., Schoppers, M., and Maimone, M. Rover navigation using stereo ego-motion. *Robotics and Autonomous Systems*, 43(4):215–229, 2003. 46, 47
- Ortiz, R. FREAK: Fast Retina Keypoint. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pages 510–517, 2012. 20, 21
- Ososinski, M. and Labrosse, F. Automatic driving on ill-defined roads: An adaptive, shape-constrained, color-based method. *Journal of Field Robotics*, pages 504–533, 2013. 134, 176
- Payá, L., Amors, F., Fernández, L., and Reinoso, O. Performance of global-appearance descriptors in map building and localization using omnidirectional vision. *Sensors*, 14(2):3033–3064, 2014. 108, 109, 111, 113, 114, 130, 132, 133, 140, 173
- Pearson, K. Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Machine learning*, 187:253–318, 1896. 84
- Pearson, K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901. 26

## REFERENCES

---

- Pinies, P. and Tardos, J. Large-scale SLAM building conditionally independent local maps: application to monocular vision. *IEEE Transactions on Robotics*, 24(5):1094–1106, 2008. 9, 57
- Pirker, K. Histogram of oriented cameras - a new descriptor for visual SLAM in dynamic environments. In *British Machine Vision Conference*, pages 1–12, 2010. 52
- Prasser, D., Milford, M., and Wyeth, G. Outdoor simultaneous localisation and mapping using RatSLAM. In *Proc. Int. Conf. Field and Service Robots*, pages 143–154, 2005. 15
- Pretto, A., Menegatti, E., Jitsukawa, Y., Ueda, R., and Arai, T. Image similarity based on discrete wavelet transform for robots with low-computational resources. *Robotics and Autonomous Systems*, 58(7):879 – 888, 2010. 29
- Pronobis, A. and Caputo, B. Confidence-based cue integration for visual place recognition. In *Proc. Int. Conf. Intelligent Robots and Systems*, 2007. 40, 57
- Pronobis, A., Caputo, B., Jensfelt, P., and Christensen, H. A discriminative approach to robust visual place recognition. In *Proc. Int. Conf. Intelligent Robots and Systems*, pages 3829–3836, 2006. 32, 56
- Qamar, S., Khawaja, F., Qureshi, A., Muhammad, N., Ayaz, Y., and Abbasi, A. A solution to perceptual aliasing through probabilistic fuzzy logic and SIFT. In *Proc. Int. Conf. Advanced Intelligent Mechatronics*, pages 1393–1398, 2013. 46
- Ramos, F., Upcroft, B., Kumar, S., and Durrant-Whyte, H. A Bayesian approach for place recognition. *Robotics and Autonomous Systems*, 60(4):487–497, 2012. 27, 59
- Ranganathan, A. and Dellaert, F. Semantic modeling of places using objects. In *Robotics: Science and Systems*, pages 939–940, 2007a. 37
- Ranganathan, A. and Dellaert, F. Probabilistic topological mapping for mobile robots using urn models. Technical report, GVU, 2007b. 45

## REFERENCES

---

- Ranganathan, A., Menegatti, E., and Dellaert, F. Bayesian inference in the space of topological maps. *IEEE Transactions on Robotics*, 22(1):92–107, 2006. 9, 11
- Remolina, E. and Kuipers, B. Towards a general theory of topological maps. *Artificial Intelligence*, 152:47–104, 2002. 9
- Rencken, W. Concurrent localisation and map building for mobile robots using ultrasonic sensors. In *Proc. Int. Conf. Intelligent Robots and Systems*, volume 3, pages 2192–2197, 1993. 2
- Ribas, D., Ridao, P., Tards, J. D., and Neira, J. Underwater SLAM in man-made structured environments. *Journal of Field Robotics*, 25(11-12):898–921, 2008. 2
- Rostami, V., Ramli, A., and Sojodishijani, O. Integration of global and local salient features for scene modeling in mobile robot applications. *Journal of Intelligent and Robotic Systems*, 2013. 29, 59
- Rosten, E., Porter, R., and Drummond, T. Faster and better: a machine learning approach to corner detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):105–119, 2010. 21
- Rosten, E. and Drummond, T. Machine learning for high-speed corner detection. In *Proc. Int. Conf. Computer Vision*, pages 430–443, 2006. 17, 21
- Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. ORB: An Efficient Alternative to SIFT or SURF. In *Proc. Int. Conf. Computer Vision*, 2011. 17, 20, 21
- Rubner, Y., Guibas, L., and Tomasi, C. The earth movers distance, multi-dimensional scaling, and color-based image retrieval. In *Proc. of the ARPA Image Understanding Workshop*, pages 661–668, 1997. 17
- Rubner, Y., Tomasi, C., and Guibas, L. The earth mover’s distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000. 85

## REFERENCES

---

- Salakhutdinov, R. and Hinton, G. E. Learning a nonlinear embedding by preserving class neighbourhood structure. In *Proc. Int. Conf. Artificial Intelligence and Statistics*, volume 2, pages 412–419, 2007. 28
- Samet, H. The quadtree and related hierarchical data structures. *ACM Comput. Surv.*, 16(2):187–260, 1984. 52
- Saudabayev, A., Kungozhin, F., Nurseitov, D., and Varol, H. A. Locomotion strategy selection for a hybrid mobile robot using time of flight depth sensor. *Journal of Sensors*, 2015:1–14, 2015. 53
- Scaramuzza, D. and Fraundorfer, F. Visual odometry [tutorial]. *Robotics Automation Magazine*, 18(4):80–92, 2011. 4, 46
- Scaramuzza, D. and Siegwart, R. Appearance-guided monocular omnidirectional visual odometry for outdoor ground vehicles. *IEEE Transactions on Robotics*, 2008. 9, 57
- Scaramuzza, D., Fraundorfer, F., and Pollefeys, M. Closing the loop in appearance-guided omnidirectional visual odometry by using vocabulary trees. *Robot. Auton. Syst.*, 58(6):820–827, 2010. 4, 50, 58, 141
- Schaffalitzky, F. and Zisserman, A. Automated location matching in movies. *Computer Vision and Image Understanding*, 92:236–264, 2003. 45
- Schindler, G., Brown, M., and Szeliski, R. City-scale location recognition. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2007. 23
- Schmidt, A., Kraft, M., and Kasiski, A. An evaluation of image feature detectors and descriptors for robot navigation. In *Computer Vision and Graphics*, volume 6375, pages 251–259. Springer Berlin Heidelberg, 2010. 18
- Schmidt, A., Kraft, M., Fularz, M., and Domagaa, Z. Comparative assessment of point feature detectors and descriptors in the context of robot navigation. *Journal of Automation, Mobile Robotics & Intelligent Systems*, 7:11–20, 2013. 21

## REFERENCES

---

- Se, S., Lowe, D., and Little, J. Local and global localization for mobile robots using visual landmarks. In *Proc. Int. Conf. Intelligent Robots and Systems*, volume 1, pages 414–420, 2001a. 18
- Se, S., Lowe, D., and Little, J. Vision-based mobile robot localization and mapping using scale-invariant features. In *Proc. Int. Conf. Robotics and Automation*, pages 2051–2058, 2001b. 18
- Se, S., Lowe, D., and Little, J. Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *International Journal of Robotics Research*, 21:735–758, 2002. 14, 18, 42
- Se, S., Lowe, D., and Little, J. Vision-based global localization and mapping for mobile robots. *IEEE Transactions on Robotics*, 21:364–375, 2005. 18, 56
- Segvic, S. and Ribaric, S. Determining the absolute orientation in a corridor using projective geometry and active vision. *IEEE Transactions on Industrial Electronics*, 48(3):696–710, 2001. 33
- Shannon, C. *The mathematical theory of communication*. University of Illinois Press, 1949. 86
- Shatkay, H. and Kaelbling, L. P. Learning topological maps with weak local odometric information. In *Int. Joint Conference on Artificial Intelligence*, pages 920–929, 1997. 11
- Shlomo, A.-E. Using image signatures for place recognition. *Pattern Recognition Letters*, 19(10):941–951, 1998. 38
- Shojaeipour, S., Haris, S., Khalili, K., and Shojaeipour, A. Motion planning for mobile robot navigation using combine quad-tree decomposition and voronoi diagrams. In *Int. Conf. Computer and Automation Engineering*, volume 1, pages 90–93, 2010. 52
- Siagian, C. and Itti, L. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(2):300–312, 2007. 40

## REFERENCES

---

- Siagian, C. and Itti, L. Biologically inspired mobile robot vision localization. *IEEE Transactions on Robotics*, 2009. 9, 40, 58
- Siagian, C., Chang, C.-K., and Itti, L. Autonomous mobile robot localization and navigation using hierarchical map representation primarily guided by vision. *Journal of Field Robotics*, 2014. 10, 41, 59
- Silpa-Anan, C. and Hartley, R. Visual localization and loop-back detection with a high resolution omnidirectional camera. In *Proc. Int. Conf. Computer Vision*, pages 1–8, 2005. 49
- Sim, R., Elinas, P., and Griffin, M. Vision-based SLAM using the rao-blackwellised particle filter. In *IJCAI Workshop on Reasoning with Uncertainty in Robotics*, 2005. 13, 56
- Singh, G. Visual loop closing using gist descriptors in Manhattan World. In *Omnidirectional Robot Vision workshop, held with IEEE ICRA*, 2010. 41, 58
- Sinha, S., Frahm, J.-M., Pollefeys, M., and Genc, Y. GPU-based Video Feature Tracking and Matching. Technical report, In *Workshop on Edge Computing Using New Commodity Architectures*, 2006. 19
- Sivic, J. and Zisserman, A. Video Google: A text retrieval approach to object matching in videos. In *Proc. Int. Conf. Computer Vision*, volume 2, pages 1470–1477, 2003. 23
- Smith, A. R. Color gamut transform pairs. *SIGGRAPH Computer Graphics*, 12 (3):12–19, 1978. 134
- Smith, M., Baldwin, I., Churchill, W., Paul, R., and Newman, P. The New College vision and laser data set. *The International Journal of Robotics Research*, 28 (5):595–599, 2009. 60, 77, 97, 100
- Smith, R., Self, M., and Cheeseman, P. Estimating uncertain spatial relationships in robotics. In *Proc. Int. Conf. Robotics and Automation.*, volume 4, pages 850–850, 1987. 1, 12

## REFERENCES

---

- Sogo, T., Ishiguro, H., and Ishida, T. Acquisition and propagation of spatial constraints based on qualitative information. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23:2001, 2001. 9
- Sousa, A., Santiago, C., Malheiros, P., Costa, P., and Moreira, A. P. Using barcodes for robotic landmarks. *Fourteenth Portuguese Conference on Artificial Intelligence*, 2009. 33, 36
- Sousa, A. J., Costa, J., Moreira, A., and Carvalho, A. Self localization of an autonomous robot: using an EKF to merge odometry and vision based landmarks. In *Proc. Int. Conf. Emerging Technologies and Factory Automation*, 2005. 36
- Strasdat, H., Montiel, J. M. M., and Davison, A. Scale drift-aware large scale monocular SLAM. In *Proceedings of Robotics: Science and Systems*, 2010a. 3, 14, 53, 58
- Strasdat, H., Montiel, J. M. M., and Davison, A. Real-time monocular SLAM: Why filter? In *Proc. Int. Conf. Robotics and Automation*, pages 2657–2664, 2010b. 14, 58
- Strecha, C., Bronstein, A. M., Bronstein, M. M., and Fua, P. LDAHash: Improved matching with smaller descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1), 2012. 28
- Stürzl, W. and Zeil, J. Depth, contrast and view-based homing in outdoor scenes. *Biological Cybernetics*, 96(5):519–531, 2007. 43
- Sujan, V., Meggiolaro, M., and Belo, F. Information based indoor environment robotic exploration and modeling using 2-d images and graphs. *Autonomous Robots*, 21(1):15–28, 2006. 52
- Sunderhauf, N. and Protzel, P. BRIEF-Gist - closing the loop by simple means. In *Proc. Int. Conf. Intelligent Robots and Systems*, pages 1234–1241, 2011. 4, 40, 51, 81, 98, 141, 156, 172

## REFERENCES

---

- Sünderhauf, N., Dayoub, F., Shirazi, S., Upcroft, B., and Milford, M. On the performance of ConvNet features for place recognition. In *Proc. Int. Conf. Intell. Robot. Syst.*, pages 4297–4304, 2015. 141
- Swain, M. and Ballard, D. H. Color indexing. *International Journal of Computer Vision*, 7:11–32, 1991. 30, 84
- Takacs, G., Chandrasekhar, V., Gelfand, N., Xiong, Y., Chen, W.-C., Bismpiannis, T., Grzeszczuk, R., Pulli, K., and Girod, B. Outdoors augmented reality on mobile phone using loxel-based visual feature organization. In *Proc. Int. Conf. Multimedia Information Retrieval*, pages 427–434, 2008. 28, 57
- Tenenbaum, J., de Silva, V., and Langford, J. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. 27
- Thompson, S., Matsui, T., Zelinsky, A., and Zelinsky, A. Localisation using automatically selected landmarks from panoramic images. In *Proc. of Australian Conference on Robotics and Automation*, 2000. 28, 34
- Thorpe, C., Carlson, J., Duggins, D., Gowdy, J., Maclachlan, R., Mertz, C., Suppe, A., and Wang, B. *Safe Robot Driving in Cluttered Environments*. Springer Berlin Heidelberg, 2005. 52
- Thrun, S. Finding landmarks for mobile robot navigation. In *Proc. Int. Conf. Robotics and Automation*, pages 958–963, 1998. 33
- Thrun, S., Montemerlo, M., Dahlkamp, H., Stavens, D., Aron, A., Diebel, J., Fong, P., Gale, J., Halpenny, M., Lau, K., Oakley, C., Palatucci, M., Pratt, V., Stang, P., Strohb, S., Dupont, C., erik Jendrossek, L., Koelen, C., Markey, C., Rummel, C., Niekerk, J. V., Jensen, E., Bradski, G., Davies, B., Ettinger, S., Kaehler, A., Nefian, A., and Mahoney, P. The robot that won the DARPA Grand Challenge. *Journal of Field Robotics*, 23:661–692, 2006. 2
- Tomasi, C. and Shi, J. Direction of heading from image deformations. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pages 422–427, 1993. 46

## REFERENCES

---

- Tomatis, N., Nourbakhsh, I., and Siagian, R. Hybrid simultaneous localization and map building: a natural integration of topological and metric. *Robotics and Autonomous Systems*, pages 3–14, 2003. 10
- Torralba, A., Murphy, K., Freeman, W., and Rubin, M. Context-based vision system for place and object recognition. In *Proc. Int. Conf. Computer Vision*, pages 273–280, 2003. 28, 40
- Torralba, A., Fergus, R., and Weiss, Y. Small codes and large image databases for recognition. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2008. 28
- Tuytelaars, T. and Schmid, C. Vector quantizing feature space with a regular lattice. In *Proc. Int. Conf. Computer Vision*, pages 1–8, 2007. 27
- Ullah, M. M., Pronobis, A., Caputo, B., Luo, J., and Jensfelt, P. The COLD database. Technical report, Kungliga Tekniska Högskolan, CVAP/CAS, 2007. 60, 70, 171
- Ulrich, I. and Nourbakhsh, I. Appearance-based place recognition for topological localization. In *Proc. Int. Conf. Robotics and Automation*, 2000. 17, 28, 30, 43
- Vale, A. and Ribeiro, M. I. Environment mapping as a topological representation. In *Proc. Int. Conf. Advanced Robotics*, 2003. 11
- Valenzuela, R., Schwartz, W., and Pedrini, H. Dimensionality reduction through PCA over SIFT and SURF descriptors. In *Proc. Int. Conf. Cybernetic Intelligent Systems*, page 5863, 2012. 26
- Valgren, C. and Lilienthal, A. Incremental spectral clustering and seasons: Appearance-based localization in outdoor environments. In *Proc. Int. Conf. Robotics and Automation*, 2008. 19, 28, 42, 57
- Valgren, C., Lilienthal, A., and Duckett, T. Incremental topological mapping using omnidirectional vision. In *Proc. Int. Conf. Intelligent Robots and Systems*, pages 3441–3447, 2006. 49, 56

## REFERENCES

---

- Valgren, C., Duckett, T., and Lilienthal, A. Incremental spectral clustering and its application to topological mapping. In *Proc. Int. Conf. Robotics and Automation*, pages 4283–4288, 2007. 49
- Vasudevan, S. and Siegwart, R. Bayesian space conceptualization and place classification for semantic maps in mobile robotics. *Robot. Auton. Syst.*, 56(6): 522–537, 2008. 11
- Vasudevan, S., Gehler, S., Berger, M., and Siegwart, R. Cognitive maps for mobile robots – an object based approach. *Robotics & Autonomous Systems*, 55(5):359–371, 2007. 37, 57
- Viola, P. *Alignment by Maximization of Mutual Information*. PhD thesis, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1995. 86
- Wang, J. and Yagi, Y. Robust location recognition based on efficient feature integration. In *Proc. Int. Conf. Robotics and Biomimetics*, pages 97–101, 2012. 9
- Wang, J. and Yagi, Y. Efficient topological localization using global and local feature matching. *International Journal of Advanced Robotic Systems*, 10(1): 1–9, 2013. 10, 32, 59
- Wang, J., Cipolla, R., and Zha, H. Vision-based global localization using a visual vocabulary. In *Proc. Int. Conf. Robotics and Automation*, pages 4230–4235, 2005. 23, 56
- Wang, M.-L. and Lin, H.-Y. An extended-HCT semantic description for visual place recognition. *The International Journal of Robotics Research*, 30(11): 1403–1420, 2011. 28, 70
- Weiss, C., Masselli, A., Tamimi, H., and Zell, A. Fast outdoor robot localization using integral invariants. In *Proc. Int. Conf. Computer Vision Systems*, page 24, 2007a. 17, 57

## REFERENCES

---

- Weiss, C., Tamimi, H., Masselli, A., and Zell, A. A hybrid approach for vision-based outdoor robot localization using global and local image features. In *Proc. Int. Conf. Intelligent Robots and Systems*, pages 1047–1052, 2007b. 10
- Werner, F., Sitte, J., and Maire, F. Visual topological mapping and localisation using colour histograms. In *Proc. Int. Conf. Control, Automation, Robotics and Vision*, pages 341–346, 2008. 57
- Werner, F., Maire, F., and Sitte, J. Topological SLAM using fast vision techniques. In *FIRA RoboWorld Congress*, volume 5744 of *Lecture Notes in Computer Science*, pages 187–196, 2009. 45
- Williams, B. and Reid, I. On combining visual SLAM and visual odometry. In *Proc. Int. Conf. Robotics and Automation*, pages 3494–3500, 2010. 3
- Williams, B., Cummins, M., Neira, J., Newman, P., Reid, I., and Tardos, J. An image-to-map loop closing method for monocular SLAM. In *Proc. Int. Conf. Intelligent Robots and Systems*, pages 2053–2059, 2008. 50
- Williams, B., Cummins, M., Neira, J., Newman, P., Reid, I., and Tardós, J. A comparison of loop closing techniques in monocular SLAM. *Robotics and Autonomous Systems*, 2009. 50
- Williams, S., Newman, P., Dissanayake, G., and Durrant-Whyte, H. Autonomous underwater simultaneous localisation and map building. In *Proc. Int. Conf. Robotics and Automation*, volume 2, pages 1793–1798, 2000. 1
- Willsky, A. Multiresolution Markov models for signal and image processing. *Proceedings of the IEEE*, 90(8):1396–1458, 2002. 52
- Winder, S. and Brown, M. Learning local image descriptors. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2007. 18
- Winder, S., Hua, G., and Brown, M. Picking the best daisy. In *Computer Vision and Pattern Recognition*. IEEE Computer Society, 2009. 27

## REFERENCES

---

- Wolf, J., Burgard, W., and Burkhardt, H. Robust vision-based localization by combining an image-retrieval system with Monte Carlo localization. *Trans. Rob.*, 21(2):208–216, 2005. 35, 56
- Wu, J. and Rehg, J. Where am I: Place instance and category recognition using spatial PACT. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pages 1–8, 2008. 32
- Wu, J., Zhang, H., and Guan, Y. An efficient visual loop closure detection method in a map of 20 million key locations. In *Proc. Int. Conf. Robotics and Automation*, pages 861–866, 2014. 4, 51, 141
- Yagi, Y., Fujimura, S., and Yachida, M. Route representation for mobile robot navigation by omnidirectional route panorama Fourier transformation. In *Proc. Int. Conf. Robotics and Automation*, volume 2, pages 1250–1255, 1998. 25
- Yang, X. and Cheng, K. T. Local difference binary for ultrafast and distinctive feature description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):188–194, 2014a. 20, 21, 51, 99, 156, 176
- Yang, X. and Cheng, K. T. Learning optimized local difference binaries for scalable augmented reality on mobile devices. *IEEE Transactions on Visualization & Computer Graphics*, 20(6):852–865, 2014b. 21
- Yeo, C., Ahammad, P., and Ramchandran, K. Rate-efficient visual correspondences using random projections. In *Proc. Int. Conf. Image Processing*, pages 217–220, 2008. 28
- Yoon, K. and Kweon, I.-S. Landmark design and real-time landmark tracking for mobile robot localization. In *Proc. Int. Society for Optical Engineering*, volume 4573, pages 219–226, 2002. 33, 36
- Zhou, C., Wei, Y., and Tan, T. Mobile robot self-localization based on global visual appearance features. In *Proc. Int. Conf. on Robotics and Automation*, pages 1271–1276, 2003. 17, 31

## REFERENCES

---

- Zingaretti, P. and Frontoni, E. Vision and sonar sensor fusion for mobile robot localization in aliased environments. In *Proc. Int. Conf. Mechatronic and Embedded Systems and Applications*, pages 1–6, 2006. 45
- Zivkovic, Z., Bakker, B., and Krose, B. Hierarchical map building using visual landmarks and geometric constraints. In *Proc. Int. Conf. Intelligent Robots and Systems*, pages 2480–2485, 2005. 11