VISUAL LOCALISATION IN DYNAMIC NON-UNIFORM LIGHTING

Stephen Thomas Nuske

A THESIS SUBMITTED TO THE UNIVERSITY OF QUEENSLAND
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
SCHOOL OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING
JULY 2009
Declaration by Author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis. I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award. I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the General Award Rules of The University of Queensland, immediately made available for research and study in accordance with the Copyright Act 1968. I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material.

________________________________________

Stephen Thomas Nuske
Contributions to Jointly Authored Works Contained in the Thesis

Stephen Nuske, Jonathan Roberts and Gordon Wyeth. 2009. Robust Outdoor Visual Localisation using a 3D-edge Map. International Journal of Field Robotics (accepted – in press) – Stephen Nuske conceived the work, designed the system, implemented the code, conducted the experiments and wrote up the paper. All authors contributed to the editing of the document.

Stephen Nuske, David Prasser, Jonathan Roberts and Gordon Wyeth. 2009. Experiments in Visual Localisation around Underwater Structures. Proceedings of International Conference on Field and Service Robotics, Cambridge, MA, U.S.A. – Stephen Nuske conceived the work, designed the system, implemented the code, conducted the experiments and wrote up the paper. David Prasser helped conduct the experiments in this paper. All authors contributed to the editing of the document.

Stephen Nuske, Jonathan Roberts and Gordon Wyeth. 2008. Visual Localisation in Outdoor Industrial Building Environments. Proceedings of International Conference on Robotics and Automation, Pasadena, California, U.S.A. – All authors contributed to the meetings which planned the research tasks and each contributed to proof reading. Stephen Nuske conceived the work, designed the system, implemented the code, conducted the experiments and wrote up the paper.

Jonathan Roberts, Ashley Tews and Stephen Nuske. 2008. Redundant Sensing for Localisation in Outdoor Industrial Environments. Proceedings of the 6th IARP/IEEE-RAS/EURON Workshop on Technical Challenges for Dependable Robots in Human Environments, Pasadena, California – Ashley Tews was responsible for designing and implementing the arbitration system, in charge of conducting the integration experiment and writing the related sections. Jonathan Roberts was contributed to the arbitration system design and writing the introduction and conclusion. Stephen Nuske was responsible for the design and implementation of the vision system, contributed to the integration and testing of the redundant localisation system, analysing the results and writing the literature review. All authors contributed to reviewing the paper.

Stephen Nuske and Manuel Yguel. 2007. Detecting Moving Pedestrians and Vehicles in Fluctuating Lighting Conditions. Proceedings of the 2007 Australasian Conference on Robotics and Automation, Brisbane, Queensland, Australia – Both authors contributed to the conception of the research. Stephen Nuske was responsible for the literature review, design and implementation of the system, experiments and writing the document.

Stephen Nuske, Jonathan Roberts and Gordon Wyeth. 2006. Extending the Dynamic Range of Robotic Vision Proceedings of the 2006 International Conference on Robotics and Automation, Orlando, Florida, U.S.A. – All authors contributed to the meetings which planned the research tasks and each
contributed to proof reading. Stephen Nuske conceived the work, designed the system, implemented the code, conducted the experiments and wrote up the paper.

**Contributions by Others to the Thesis as a Whole**

Jonathan Roberts and Gordon Wyeth contributed to meetings which planned the research tasks for this thesis. Ashley Tews and Jonathan Roberts were responsible for the design of the arbitration system and contributed to the related experiment and writing presented in Appendix A.

**Parts of the Thesis Submitted to Qualify for the Award of Another Degree**

None.

**Published Works by the Author Incorporated into the Thesis**


**Additional Published Works by the Author Relevant to the Thesis but not Forming Part of it**

None.
Acknowledgements

First and foremost, I would like to thank my brilliant family for the tremendous support they have given me throughout my studies and throughout my life in general.

Dr. Gordon Wyeth, my advisor, was the catalyst for my entry into the PhD program, who helped me right from my undergraduate years and offered sage advise at crucial stages of this PhD. My other advisor, Dr. Jonathan Roberts, has been a great supporter and promoter of my research, he is always positive, and has been absolutely fantastic in providing the necessary resources that made this project possible. Dr. Peter Corke must also be thanked for becoming my interim advisor while my two other advisors were on long service leave.

Many thanks have to go the CSIRO Autonomous Systems Laboratory, who along with the School of Information Technology and Electrical Engineering at the University of Queensland, funded my research (the Australian Government must also be thanked, somewhat indirectly, for its very valuable support of these research institutions). The Autonomous Systems Laboratory has been a great place to do my research and I am thankful for the great opportunity to be a part of this group. The engineering support available there is top notch and the staff always make time to help a research student such as myself. Special thanks goes to Paul Flick, Jon Whitam, Polly Alexander and Clinton Roy; who all have provided me much needed assistance along the way. The research staff at the lab have also been a great source of advice and help. Dr. Cedric Pradalier was especially encouraging and helpful in technical counsel. Dr. Ashley Tews, my team leader for the ground robotics research group, supported me with both resources and encouragement for my research. Thanks also to Dr. Matthew Dunbabin, who resourced the submarine project in this thesis and to Jock Cunningham who helped fund travel for me to present my research. There are also many others, too many to name, that contributed to make my time at CSIRO both pleasant and productive. Acknowledgements to my parents, Tony Roberts, Gordon Wyeth and Navid Nourani-Vatani; who provided the majority of the proof reading for this document. Thanks to the e-Motion laboratory at INRIA, Grenoble, France, who invited me to visit for a three month period. My stay there was fantastic and the people very friendly and fun. It was this trip, and the 2005 Cognitive Vision Summerschool at the University of Bonn, Germany, that sparked my imagination and love of this great field of research.

A big acknowledgement to the University of Queensland Football Club and also to the formidable Team NJ, Team Merino, Team Minto and the humble heroes of Team O’Neill all of whom have kept me fit and active and enabled me to meet many great friends. Finally, special thanks go to my several house-mates – Sunni, Jonno, Noogs, Cass, Marie, Reeds, Bethany, Sparticus, Bindi, Dukes, DK, Simone, Adam and Holly – and Chris and Natalie for looking after me in my final year, and all of my other friends, who all have made my life so very very enjoyable these last few years.
Abstract

Dynamic non-uniform lighting conditions, prevalent in many field robot applications, cause drastic changes in the visual information captured by camera images, resulting in major difficulties for mobile robots attempting to localise visually. Most current solutions to the problem rely on extracting visual information from images that is decoupled from the effects of lighting. This is not possible in many situations.

Chrominance information is often cited as having some invariance to lighting changes, which is confirmed by experiments in this thesis. However, in the bland application environments investigated, chrominance is not a pertinent metric, indicating that chrominance is not the complete solution to the lighting problem. Descriptions of the intensity gradient are also cited as having robustness to lighting changes. Many descriptions of image-point features are based on the intensity gradient and are commonly used as a basis for visual localisation. However, the non-uniform effects of lighting – shadows and shading – are tangled into the intensity gradient, making these descriptions sensitive to non-uniform lighting changes. Experiments are presented which reveal that image-point features recorded at one time of the day cannot be reliably matched with images captured only one or two hours later, after typical changes in sunlight.

It appears that autonomously building visual maps which permit geometric localisation in many lighting conditions remains an unsolved problem. Therefore, this thesis develops systems based on manually generated maps which are created a priori and ensure that only permanent, invariant, information is included within the map and allows reliable localisation to be achieved in many conditions.

The first proposed visual localisation system is for autonomous ground vehicles operating at outdoor industrial sites. The system avoids the problem of mapping from fluctuating visual information by using a professionally-surveyed 3D-edge map of the permanent buildings. The vehicles are fitted with fish-eye cameras that often have direct sunlight in the field of view, causing an issue of camera exposure – dealt with by using an intelligent exposure control algorithm. Results from the system show accurate localisation during the full range of lighting conditions experienced over a day.

The second visual localisation framework discussed is for submarines navigating underwater structures, where the only light source is a spotlight mounted on the vehicle. The moving vehicle and hence changing incident angle of the light source cause major variations in the appearance of the structure. This makes it difficult to employ traditional tracking techniques. The proposed localisation system uses the novel idea of incorporating a light source model. The light model is used to render synthetic images of scene – which accurately recreate the non-uniform lighting
effects. The synthetic images are compared with the real camera image to localise the vehicle. The idea of using a light model is partly motivated by the human visual system’s understanding of the light source within a scene and is also motivated by the limitations of the traditional approaches to factor-out lighting. Using a light model within a visual localisation system enables a more natural link between the internal environment representation and the image and is demonstrated to allow successful localisation in this difficult visual scenario.

The results of the two proposed localisation systems are distinguished, given the extremely challenging dynamic non-uniform lighting in each environment. Both systems have attracted the interest of industry partners and the projects continue to be developed into the future.

**Keywords**

visual localisation, robotics, computer vision, image processing, illumination changes

**Australian and New Zealand Standard Research Classifications (ANZSRC)**
080104 50%, 090602 50%
# Contents

<table>
<thead>
<tr>
<th>Glossary of terms</th>
<th>xiii</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1 Introduction</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1 Target Mobile Robot Applications</td>
<td>3</td>
</tr>
<tr>
<td>1.1.1 Autonomous Ground Vehicle in Outdoor Industrial Sites</td>
<td>3</td>
</tr>
<tr>
<td>1.1.2 Autonomous Submarine Navigating Underwater Structures</td>
<td>5</td>
</tr>
<tr>
<td>1.2 Research Questions</td>
<td>5</td>
</tr>
<tr>
<td>1.3 Thesis Layout and Methodology</td>
<td>6</td>
</tr>
<tr>
<td>1.4 Contributions</td>
<td>8</td>
</tr>
<tr>
<td><strong>2 Background</strong></td>
<td>11</td>
</tr>
<tr>
<td>2.1 Introduction to Visual Localisation</td>
<td>11</td>
</tr>
<tr>
<td>2.2 Historical Overview</td>
<td>13</td>
</tr>
<tr>
<td>2.3 Manual vs Autonomous Map Generation</td>
<td>14</td>
</tr>
<tr>
<td>2.4 3D-edge</td>
<td>15</td>
</tr>
<tr>
<td>2.5 Global Image Encodings</td>
<td>16</td>
</tr>
<tr>
<td>2.6 Image-point Features</td>
<td>19</td>
</tr>
<tr>
<td>2.7 Stereo Point Clouds</td>
<td>21</td>
</tr>
<tr>
<td>2.8 3D-surface</td>
<td>21</td>
</tr>
<tr>
<td>2.9 Discussion</td>
<td>23</td>
</tr>
<tr>
<td><strong>3 Extracting Visual Information in Dynamic Non-uniform Lighting</strong></td>
<td>27</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>28</td>
</tr>
<tr>
<td>3.1.1 Recognising Reflectance</td>
<td>29</td>
</tr>
<tr>
<td>3.2 Foreground Segmentation in Dynamic Non-uniform Lighting</td>
<td>30</td>
</tr>
<tr>
<td>3.2.1 Intensity Background Models</td>
<td>31</td>
</tr>
<tr>
<td>3.2.2 Gradient Background Models</td>
<td>31</td>
</tr>
<tr>
<td>3.2.3 Region Growing with Motion History</td>
<td>37</td>
</tr>
<tr>
<td>3.2.4 Results</td>
<td>41</td>
</tr>
<tr>
<td>3.2.5 Summary of Foreground Segmentation Study</td>
<td>41</td>
</tr>
<tr>
<td>3.3 Chrominance Spaces</td>
<td>42</td>
</tr>
<tr>
<td>3.3.1 Terminology</td>
<td>42</td>
</tr>
<tr>
<td>3.3.2 Background</td>
<td>42</td>
</tr>
<tr>
<td>3.3.3 Chrominance Spaces</td>
<td>43</td>
</tr>
<tr>
<td>3.3.4 Achromatic Variations in the Lighting</td>
<td>45</td>
</tr>
<tr>
<td>3.3.5 Chromatic Variations in the Lighting</td>
<td>53</td>
</tr>
</tbody>
</table>
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.5.1</td>
<td>Evaluation of Localisation</td>
</tr>
<tr>
<td>5.6</td>
<td>Initialisation Experiment</td>
</tr>
<tr>
<td>5.6.1</td>
<td>Comparison of Observation Functions at Initialisation</td>
</tr>
<tr>
<td>5.7</td>
<td>Operation Experiments</td>
</tr>
<tr>
<td>5.7.1</td>
<td>Bright Sunlight</td>
</tr>
<tr>
<td>5.7.2</td>
<td>All-day Experiment</td>
</tr>
<tr>
<td>5.7.3</td>
<td>Comparison of Observation Functions</td>
</tr>
<tr>
<td>5.7.4</td>
<td>Robustness of Image-point Feature Matching</td>
</tr>
<tr>
<td>5.7.5</td>
<td>Rain Experiment</td>
</tr>
<tr>
<td>5.8</td>
<td>Chapter Summary</td>
</tr>
</tbody>
</table>

## 6 Visual Localisation Incorporating A Light Source Model

<table>
<thead>
<tr>
<th>Subsection</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>System Design Motivations</td>
</tr>
<tr>
<td>6.1.1</td>
<td>Motivations from the Application Environment</td>
</tr>
<tr>
<td>6.1.2</td>
<td>Motivations from the Human Visual System</td>
</tr>
<tr>
<td>6.2</td>
<td>Related Work</td>
</tr>
<tr>
<td>6.3</td>
<td>Localisation Framework</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Submarine-setup</td>
</tr>
<tr>
<td>6.3.2</td>
<td>Oil-rig Scaffolding</td>
</tr>
<tr>
<td>6.3.3</td>
<td>Synthetic Image Generation and Light Source Model</td>
</tr>
<tr>
<td>6.3.4</td>
<td>Particle Filter Localisation</td>
</tr>
<tr>
<td>6.3.5</td>
<td>Gradient-domain Image Matching</td>
</tr>
<tr>
<td>6.3.6</td>
<td>GPU Implementation</td>
</tr>
<tr>
<td>6.4</td>
<td>Results</td>
</tr>
<tr>
<td>6.4.1</td>
<td>Implementation Notes</td>
</tr>
<tr>
<td>6.4.2</td>
<td>Initialisation</td>
</tr>
<tr>
<td>6.4.3</td>
<td>Experiment A</td>
</tr>
<tr>
<td>6.4.4</td>
<td>Experiment B</td>
</tr>
<tr>
<td>6.5</td>
<td>Discussion</td>
</tr>
<tr>
<td>6.5.1</td>
<td>Divergence</td>
</tr>
<tr>
<td>6.5.2</td>
<td>Efficiency</td>
</tr>
<tr>
<td>6.5.3</td>
<td>Accuracy</td>
</tr>
<tr>
<td>6.6</td>
<td>Chapter Summary</td>
</tr>
</tbody>
</table>

## 7 Thesis Conclusion

<table>
<thead>
<tr>
<th>Subsection</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1</td>
<td>Contributions</td>
</tr>
<tr>
<td>7.2</td>
<td>Research Questions and Answers</td>
</tr>
<tr>
<td>7.3</td>
<td>Discussion</td>
</tr>
</tbody>
</table>

## References

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>177</td>
</tr>
</tbody>
</table>

## A Redundant Localisation

<table>
<thead>
<tr>
<th>Subsection</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.1</td>
<td>Related work</td>
</tr>
<tr>
<td>A.2</td>
<td>Proposed System</td>
</tr>
<tr>
<td>A.2.1</td>
<td>Laser Localisation</td>
</tr>
<tr>
<td>A.2.2</td>
<td>Camera Localisation</td>
</tr>
</tbody>
</table>
Glossary of terms

**Localisation** - Estimating self position and self orientation to a map of an environment.

**Visual Localisation** - Localisation with a visible light sensor e.g. a camera.

**AUV** - Autonomous Underwater Vehicle.

**AGV** - Autonomous Ground Vehicle.

**Reflectance** - The properties of a surface relating to its reflection of light.

**Foreground segmentation** - Identifying the regions in an image that differ from a learnt background model.

**Edge/Gradient/Relative-intensity** - Terms all referring to the detection and description of spatially varying intensity values across an image.

**Chrominance** - the spectral signature of a surface or light source, otherwise known as colour. Although colour is also commonly used to mean greyscale intensity which is a different to chrominance.

**Image-point features** - descriptions of patches surrounding points of interest within an image.

**Dynamic range** - the range of measurements that a sensor can take simultaneously.

**Saturation** - when the sensory input exceeds the range of a sensor.

**Particle filter** - a nonparametric implementation of the Bayes filter representing a distribution as a set of samples.

**Odometry** - relative pose estimates.

**Propagation/Motion model** - a model of the motion/odometry of a vehicle used with a localisation filter.
**3D-edge map** - a three dimensional description of the boundaries of objects in the environment.

**3D-surface map** - a three dimensional description of an environment which often is a polygon mesh and also can have some associated reflectance information.

**Real image** - an image captured by a camera of the real world.

**Synthetic/Rendered/Generated image** - a simulated image created with a model of the objects in an environment and a light/reflectance model.

**CPU** - Central Processing unit, the core of most modern computers.

**GPU** - Graphics processing unit, a highly parallelised device used in most modern computers for fast graphics computations and displaying. GPU’s can also be used for fast parallel implementations of general processing tasks.
Introduction

Computer vision has existed as a field of research since the 1960’s. There have been many success stories of research methods progressing through into industrial applications, thus proving the value of work in this area. However, when placed in comparison with the sophistication and generalisability of many aspects of the human visual system, computer vision is still in its infancy. There are still many visual tasks that are easy for a human, but are beyond a robot’s capabilities. For vision to succeed as a perceptual mechanism for general robot applications, there are challenges that, for the moment, seem insurmountable.

Lighting that is non-uniform and dynamic is one such problem that disrupts many robotic vision systems. In comparison, the human visual system handles the lighting challenges with ease. The lighting issue is especially problematical in field robotics and this thesis investigates the task of visual localisation for two different field robot applications. The focus is on geometric visual localisation that is the task of deriving explicit self-position and self-orientation by aligning a 3D-map in a camera
image. Geometric localisation is different to the other visual localisation approach – image-based localisation (also known as appearance-based localisation), which is the task of place recognition. 

Localising in a 3D environment from a 2D camera image is by no means trivial. While it is true that camera images capture a large amount of rich information about the environment, fundamentally, they are just an array of light measurements. They do not make any direct observations of the properties of objects in the world. This is problematic for visual localisation that must link image intensity values with objects in the environment. Correlating intensity values with objects is difficult, even in the simplest indoor environments where the lighting is static and uniform. The problem becomes far more troublesome with dynamic and non-uniform lighting; here, the link between image values and objects in the world is even less direct. Furthermore, images cannot make direct measurements of distance as can a laser scanner, thus making geometric visual localisation difficult. Distances have to be inferred either with multiple cameras, over time via motion or by aligning an accurate 3D-map with the image; all of which becomes difficult in dynamic and non-uniform lighting conditions.

Previous research generally approaches the lighting issue by attempting to factor-out lighting when extracting information from images. DeSouza and Kak [1] in their review paper of robotic vision acknowledge the problem of lighting with a whole section dedicated to outdoor illumination conditions. The primary solution they suggest is to use chrominance information that has some lighting invariance. This thesis presents a detailed study of chrominance information to evaluate its properties through lighting changes. The outcome is that colour transforms, if used correctly, can factor-out the effects of lighting except for changes in the spectral composition of the light source. However, there are environments that are bland (such as the specific application environments investigated in this thesis) where chrominance information is not a distinguishing metric and monochrome intensity is far more pertinent. Thus, chrominance cannot be the full solution to the lighting problem.

Predominantly, the approaches to factor-out lighting from monochrome intensity are to process images in the intensity gradient domain. Jogan et al. [2] present an image-based localisation technique using a gradient filter that removes some of the lighting effects. Other localisation work uses image point features [3, 4] which are based on descriptions of intensity gradient orientations and also possess some lighting robustness. However, it is important to realise that the stability of the intensity gradient occurs with uniform lighting changes. Shadows and shading are inherently tangled into the patterns of the image and, therefore, non-uniform lighting changes will cause disruptions to the descriptions of intensity gradients. It is not surprising that many authors, such as Sim and Dudek [5], Lowe [6], Valgren and Lilienthal [7], have found sensitivities in such descriptions to lighting changes.
1.1 Target Mobile Robot Applications

Feature matching outliers caused by lighting or other means can be removed with robust constraints and robust comparison schemes, such as in the work of Valgren and Lilienthal in [8] and Cummins and Newman in [4]. Though it is not clear if robust localisation algorithms are enough to deal with the underlying instabilities through large non-uniform lighting changes, especially when only minimal inliers are found. Most authors test just one or two different lighting conditions. This thesis conducts a more detailed lighting change study in Section 3.4, testing image point feature stability with the full range of lighting changes experienced outdoors over a bright sunny day. The outcome of the study reveals that lighting disrupts the feature matching process sufficiently to question whether an autonomously built feature map can be used reliably for outdoor localisation over long periods and involving large lighting changes.

This thesis investigates two specific applications of a robot localising from known permanent structures. Maps of these structures were manually generated *a priori*, removing the difficulties of autonomously mapping from unstable image-level information. The initial expenditures to manually map an environment become insignificant if the map is to be used for many weeks or even for many years of productive operation. Furthermore, at the present moment, manual mapping is, apparently, necessary for applications that require reliable localisation over long periods in challenging lighting conditions, because autonomously generated maps have yet to be proven reliable.

### 1.1 Target Mobile Robot Applications

Visual localisation systems are developed for two field robot applications. Both application environments pose significant lighting challenges. However, each environment has unique problems for which the specialised systems are designed to solve.

#### 1.1.1 Autonomous Ground Vehicle in Outdoor Industrial Sites

The first application studied is for an industrial automated ground vehicle, shown in Figure 1.1(a), that transports large buckets of hot molten aluminium around an industrial site. The application dictates a high level of dependability. A localisation system would only be practical if it could function reliably for extended periods outdoors at any time of day. There are major difficulties for a vision system trying to achieve reliable localisation in the outdoor environment. The bright sunlight is often at an angle in the sky causing lens flares; parts of the scene tend to be in direct sunlight while other parts are in the shade. Under such conditions cameras are often not able to correctly expose the entire image
Figure 1.1: (a) An industrial automated ground vehicle (AGV) designed to carry large buckets of liquid aluminium around a smelter, (b) a camera image from the vehicle. Images often have the sun and the sky in view, causing lens flares. The buildings that are not in direct sunlight are dark and underexposed.

simultaneously. Another issue is the ever changing position of the sun throughout each day, causing dramatic changes in the information coming from the images. Literature surveys revealed that there are no visual localisation systems that have been proven to work over long periods in such difficult environments.
1.2 Research Questions

1.1.2 Autonomous Submarine Navigating Underwater Structures

The other robot application investigated in this thesis is for an autonomous submarine localising from underwater structures. The submarine’s role is to perform tasks, such as inspections, that are too expensive or dangerous with human divers, see Figure 1.2. The challenges of the lighting in this environment are far different from outdoor building environments. There is minimal or no natural lighting deep underwater and an artificial light source mounted on the submarine is required. Because the submarine is moving through the environment, the light source is moving with respect to the structure, causing the visual appearance of the structure to vary dramatically. This is quite different from regular environments, where the light source is far less dynamic and also where there is a significant level of ambient lighting.

An example camera image taken of an underwater structure can be seen in Figure 1.2(b). The visibility is quite poor; and the light mounted on the submarine illuminates only a part of the structure that is in view. Through literature surveys it would appear that existing visual localisation methods would struggle to operate reliably in such difficult conditions. The edges of the structure are not clearly visible; therefore a 3D-edge method is not applicable. The dominant visible features are specular highlights, the locations of which vary as the submarine moves; indicating that it is not feasible to employ a point-feature tracking technique to localise the vehicle.

1.2 Research Questions

The specific questions this thesis seeks to answer are;

- What are the properties of the various types of low level visual information in dynamic non-uniform lighting?
- How can this thesis improve the performance of robotic vision systems operating in highly-non-uniform outdoor lighting?
- How can visual localisation systems operate reliably in dynamic lighting conditions?

These research questions will be studied within the scope of the two specific robot applications; an outdoor industrial ground vehicle and an underwater submarine inspecting offshore structures.
1.3 Thesis Layout and Methodology

Before the visual localisation systems are developed for the two specific robot applications, a detailed study of existing visual localisation research is conducted in Chapter 2. The study outlines the strengths and weaknesses of different techniques before laying out the foundation for the rest of the thesis.
1.3 Thesis Layout and Methodology

It is apparent from the study of the existing research, that many authors have already identified lighting as a significant problem. A commonly-proposed solution is to factor-out lighting at the image level with chrominance information or with descriptions of the intensity gradient. Chapter 3 investigates if lighting can actually be factored-out at the image level and which types of visual information are useful for the visual localisation applications studied in this thesis. The chapter reveals that chrominance information, although a direct method to factor-out many of the effects of illumination, is not useful in the bland application environments studied in this thesis. The intensity gradient is more pertinent to the application environments and can remove the overall lighting level, but the effects of non-uniform lighting are not factored-out. Finally, Chapter 3 raises doubts on whether a map of image-point features can be reliably used for geometric localisation after significant changes occur in the lighting.

After the study of visual information in dynamic lighting conditions, Chapter 4 approaches the issue of the dynamic range limitations of cameras operating under highly non-uniform lighting conditions. Camera exposure is identified as a key method of improving the visual information available to a localisation system operating in environments with highly non-uniform lighting. A system is developed to rapidly collect multiple differently-exposed images to increase the amount of visual information extracted. The issue of camera exposure is revisited in the following chapter, Chapter 5, where it is demonstrated that a single exposure, intelligently controlled, is sufficient to capture the important information in an image. The developed exposure control algorithm adjusts the exposure according to the intensity of known pixels of importance, allowing a localisation system to operate even when a large portion of the field of view is consumed by the extremely bright sky and direct sunlight.

This intelligent exposure control algorithm presented within Chapter 5 is developed in the context of the main goal of the chapter, which is to design a system for localising a vehicle in an outdoor industrial environment. The application requires localisation with high levels of dependability. The studies in the initial chapters reveal that a map created autonomously from visual information may not provide dependable localisation during lighting changes. Therefore, a system is developed that uses a professionally surveyed 3D-edge map, ensuring that only permanent, accurate and explicitly lighting invariant information is included in the map. The experiments presented in Chapter 5 of successful localisation over extended periods and at different times of the day illustrate the robustness of the method and fulfillment of the application’s dependability requirements.

Chapter 6, develops a technique for autonomous submarines localising from underwater structures. After initial tests, it is apparent that existing methods struggle in this environment, because of the poor lighting provided by a spotlight mounted on the vehicle. This spotlight only partially lit the underwater structure causing the edges of the structure not to be detected in the image and, therefore,
are not possible to track. Also, the moving submarine and, hence, moving the light, resulted in a constantly changing appearance of the structure, making it difficult to employ a common image-point feature tracking technique. Instead of attempting to factor-out the lighting, which is the common approach, a novel localisation technique is developed that explicitly incorporates a light model within the system. A 3D-surface map of the structure and the light model are used together to render realistic synthetic images of the environment, see Figure 1.2(c), to compare with the real camera images. Visual localisation systems rarely if ever use a light model in this way. The framework is designed partly on the basis of motivations from the human visual system – a system which obviously makes use of the knowledge of the light source and its interaction with a scene. Results demonstrate that incorporating a light source model can provide a solution to this difficult visual localisation scenario that causes many problems for traditional methods that try to factor-out the effects of lighting.

Chapter 7 summarises the thesis with some concluding remarks.

1.4 Contributions

Section 3.2 develops an image segmentation algorithm to detect moving pedestrians and vehicles in fluctuating lighting conditions. The system uses a novel concept of edge-based motion history images to segment moving objects and enable the system to ignore appearing and disappearing shadows cast by static objects.

Section 3.3 studies discovered that the most suitable transform is hue from HSV. Certain capture and transmission scenarios were identified that should be avoided, because they result in unexpected chrominance variations.

Section 3.4 evaluates the performance of the SIFT technique in outdoor environments, revealing significant sensitivities of the SIFT features to outdoor lighting changes.

Chapter 4 develops a real-time, multiple-exposure technique for extending the dynamic range of any standard IEEE1394 digital camera. The system provides a solution for robots navigating from indoor environments to bright outdoor environments, where much of the visual information lies outside the range of a single exposure.

Chapter 5 develops a system to localise ground vehicles from a 3D-edge map in outdoor industrial sites. The technique is based on the work of Klein and Murray [9] with a number of modifications to improve the reliability of the localisation. The chapter proposes an intelligent exposure control algorithm to avoid the overcorrection problem of traditional exposure control algorithms that leave
buildings underexposed when the image is dominated by the sky and direct sunlight. Results show the system can operate over a wide range of difficult lighting conditions. In comparison, the SIFT technique, often used in visual localisation systems, performed poorly on the same data; this invites favourable review of the results of the proposed system.

Chapter 6, develops a novel localisation technique that explicitly incorporates a light model to enable an autonomous submarine to localise from underwater structures. Results from the system show that incorporating a light source model could provide a solution for the difficult task – which is problematic for other methods that alternatively try to factor-out the lighting conditions.
This chapter reviews the existing visual localisation techniques. It begins with an overview of visual localisation, a short chronological review of visual localisation techniques and a discussion on autonomous vs manual map creation, before moving into a detailed review of each category of visual representations. The categories are as follows; 3D-edges, global image encodings, image-point features, stereo point clouds and 3D-surfaces. The final section concludes the review leading into the direction and contributions of this thesis.

2.1 Introduction to Visual Localisation

Robots that do require knowledge of self-pose, derive pose information from a comparison between an internal representation of the environment (a map) and sensor measurements of the environment, see Figure 2.1(a). Not all robots need to localise; some robots such as certain vacuum cleaning robots,
can simply roam around avoiding obstacles. Other robots may need slightly more information of self-pose, such as a road-following robot, which need only know its location with respect to the centre of the road. The fully autonomous mobile robots investigated by this thesis need to know self-pose with respect to a map.

Visual localisation is performed using a camera as a sensor. The changing pose of the robot will cause the image to change, providing the necessary information to estimate the camera pose with respect to the map, as illustrated in Figure 2.1(b). Localising from camera images is extremely difficult. It requires locating 3D objects within the patterns of a 2D image or alternatively recognising an image that has been seen before. Images are a set of light measurements and do not make direct observations of the properties of objects in the world. An image cannot measure distance, distances have to be inferred either with multiple cameras, over time via motion or by aligning an accurate 3D-map with the image, making geometric localisation onerous. It is difficult even in the simplest indoor environments where the lighting is static and uniform. In static conditions, image variations can be assumed to be caused by the objects changing position within the image. However, in dynamic lighting, the link between image values and object reflectance is far less direct. Image variations are caused by lighting changes in addition to the changing camera pose, as indicated in Figure 2.1(b). A visual localisation system is primarily interested in positioning and orientating with respect to objects and is not concerned with the lighting conditions. Ideally, performance will be unaffected by the lighting, which must be derived from three areas of the system:

- the information extracted from the image
- the environment representation (the map)
• the comparison between map and image

The focus of most approaches is to extract invariant information from images, which is difficult, and arguably impossible for all situations. An image is fundamentally an array of light measurements. The effects of lighting, such as shadows and shading, are a major contributor to the patterns in the image, and there is no obvious way to factor these lighting effects out of an image.

If lighting cannot be factored-out directly at the image, a large importance is placed on having both a robust comparison technique and invariant information in the map. A map needs to allow comparisons with images captured in all conditions. This is why the review in this chapter groups techniques according to the type of information used in the map, but first the next section visits the history of the different visual representations used for localisation.

2.2 Historical Overview

The earliest works in the 1970’s and 1980’s proposed to use generalized high-level shape representations such as cylinder, cubes etc. An example of high-level shape representations is given by Brooks [10]. As pointed out by Keselman and Dickinson [11], matching these abstract models to an image is problematical, because there is no direct method of closing the representational gap. No functioning examples exist using these abstract representations and, as such, this chapter does not include this concept as a category of environment representation.

Progressing onwards from these early works there was a need for a more logical environment representation that could be more directly compared with an image. In the late 1980’s and early 1990’s, 3D-edge maps came into use [12, 13, 14] which could be aligned with edges extracted from images.

During the 1990’s, there was a focus on representations that were more of a direct image description, using image-point feature descriptions such as Harris corners [15] and global image encodings such as PCA [16]. These representations were popular because they were a more complex description than simple edges and could be autonomously or semi-autonomously learnt from images. A significant drawback was that they did not handle changes in scale or certain types of rotations.

A breakthrough at the end of the 1990’s was with Lowe’s SIFT descriptor [17], which offered a new level of robustness to changes in scale and rotation. The SIFT descriptor, and the more recent variant SURF [18], have become very popular and have dominated much of the visual localisation research in this century.
Another type of representation more recently proposed for visual localisation is stereo-point-clouds [19, 20]. These techniques, closely related to laser-point-cloud localisation techniques, build occupancy grids from point cloud data.

A less mainstream idea that has emerged in the 2000’s is the use of 3D-surface maps [21, 22]. The lack of attention that 3D-surface methods have received is most probably due to the inability of deriving the maps autonomously. The next section presents some arguments on the positive and negative aspects of both autonomous and manual map creation.

2.3 Manual vs Autonomous Map Generation

Maps used for localisation can be either given \textit{a priori} to the robot or the robot can autonomously build its own map. For navigating unknown environments, a robot would need to create its own map whilst simultaneously localising: a technique known as SLAM (Simultaneous Localization And Mapping) is required [23]. The majority of SLAM work has been restricted to laser-based, radar-based and other non-vision-based systems. These other sensors have the advantage of making direct measurements of distance, which is useful when mapping the environment.

Vision-based SLAM is far more challenging, because an image on its own does not provide direct measurements of distance. Distances have to be inferred from multiple images (either with stereo-pairs or over time with a single camera calculating depth from motion). Geometric visual SLAM techniques have been proposed [3][24] that triangulate image-point features and place them in a map. The map can then be refined with new observations of old features with an optimisation procedure minimising the reprojection error. Another approach to build geometric maps is based on grid maps built from stereo-point-clouds [19][20].

Appearance-based SLAM techniques do not have the problem of estimating distance, they are concerned only with recognising whether or not the current place has been previously visited by the robot. Newman et al. [25], Cummins and Newman [4] and Valgren and Lilienthal [7] collect sets of image-point features for each place then, at a later moment, try to identify if there are enough unique feature correspondences to confirm that a certain place has been previously visited.

Global image encodings can also be used for appearance-based SLAM, such as in the work of Jogan et al. [26], who learn an encoding for each pose of the robot in a learning phase. In comparison with image-point features Sim and Dudek [5] found that global image encodings methods at run-time were more efficient in computational load, but both performed poorly with lighting changes.
It is not conclusive that automatic techniques can build maps that can be used reliably over long periods of time owing to the disruptions from large changes in lighting. Image-point features do offer some robustness but have been noted to vary with the viewing angle and lighting conditions, especially when the planarity assumption is violated [27, 6]. Because of this, the map may be specific to the lighting condition or the view from which they were recorded. Thus, these maps may not be useful when trying to localise under another lighting condition or from another view. Section 3.4 presents experiments on the sensitivity of the popular SIFT technique, the aim being to evaluate if it is a suitable representation for maps used through large lighting changes.

There are applications where any susceptibility to failure is not acceptable. A visual localisation used for such applications would need a map which can be localised under all lighting conditions. A map of the 3D properties of an environment, such as a 3D-edge map [13, 14], is explicitly separate to lighting and viewing angles. Autonomously-creating these types of maps is an unsolved problem and is, therefore, only suitable in applications where the environment can be mapped a priori.

When evaluating a visual localisation technique, it is important to take the application into consideration. If the application environment permits an a priori map creation, then it may be beneficial to manually map the permanent landmarks with an invariant representation. Otherwise, if manual map creation is not feasible, an autonomous SLAM technique is the only valid option.

The following sections review the different categories of information used in maps for visual localisation, mentioning which representations can be autonomously mapped.

### 2.4 3D-edge

Maps of the structural edges in the environment can be used for localisation by aligning with edges extracted from an image. This was a big advance by the research community in the late 1980’s, because it was the first type of algorithm that closed the gap between map representation and the image in a realistic framework.

Pope [28] provides a survey of the 3D-edge work of the 1980’s and early 1990’s when processing power was an issue and real-time systems were not prevalent. Since then, there have been examples of indoor navigation systems [13, 14] which use a 3D-edge map of the doors and walls in the environment to match with edges extracted from images on a navigating robot’s camera. Towards the late 1990’s as processing power increased, Drummond and Cipolla [29] presented a fast technique to track 3D-edges that could operate at 30Hz. An implementation of this technique was later applied by Reitmayr and Drummond [30] outdoors. The 3D-edge based approach has a significant advantage
in that the representation is explicitly separate from the viewing conditions. This means that the map can be used from any viewing angle and also in any lighting condition.

However, 3D-edge localisation has yet to be tested over long periods through large lighting variations, perhaps because of the potential for the system to lose track of the map. Keselman and Dickinson [11] explain one reason why the 3D-edge framework suffers; it is because there is a discrepancy in the amount of information extracted from the image and the amount of information present in the 3D-edge representation. Much more information is extracted from the image, causing the potential for the map to be falsely aligned to the image.

The discrepancy between image and map is caused by a couple of factors. One problem is that it is practical to create only sparse 3D-edge-maps. Consequently, many objects that will be detected in the image are not included in the map. Lighting conditions is another factor. Edge-extraction uses relative-intensity that is robust to global changes in lighting, but is sensitive to non-uniform lighting changes. Edges will be extracted around shadows that were cast by non-uniform lighting, causing another discrepancy between the image and map.

Previous approaches to 3D-edge localisation are especially susceptible to failure from discrepancies between edge and map, because they maintain only a single pose estimate. A recent 3D-edge approach presented by Klein and Murray [9] tracks the map with a multiple pose hypothesis filter, which makes many comparisons between the image and the map each iteration. Chapter 5 develops a localisation system that is a modified version of Klein and Murray’s work and tested over long periods and through large lighting changes in an outdoor industrial environment.

### 2.5 Global Image Encodings

Representations were developed that formed global descriptions of images. These techniques have many advantages:

- complex objects can be described
- the descriptions can be learnt from images in an autonomous manner
- comparisons with the image are easy as the representations are just image encodings.

There are a group of localisation techniques which are variants of principal component analysis (PCA), a commonly used image encoding technique [31, 16, 32, 26]. The PCA technique transforms
the image of an object as a coordinate in a multidimensional space, by treating a 2D image as a decomposed single 1D vector. The training stage finds a multidimensional subspace whose basis eigenvectors correspond to the maximum-variance in the 1D vector [33]. Images from different poses correlate to different coordinates in the multidimensional space.

Pourraz and Crowley in [32] explain the localisation procedure. It begins with a training phase where several images from different poses in an environment are recorded then projected as coordinates in an eigenspace. Each coordinate corresponds to the pose from which the image was taken. By grouping coordinates, a continuous parametric manifold can be formed in the eigenspace, where the parameters of the manifold are the camera pose. Afterwards, in the localisation phase, an image is taken from the moving robot and encoded as coordinate in the eigenspace. The point is then projected to the nearest coordinate in the manifold, which corresponds to an approximate location of where the image was taken. The manifold is specific to one orientation of the camera and a new manifold must be learnt for each rotation.

Georghiades et al. [34] explain that the eigenspace recognition of an image under a particular lighting and pose can be performed reliably, provided an image had been previously seen under similar circumstances. Therefore, to recognise environments under all poses, the training phase must include images from the full range of poses. This is a big undertaking when merely three degrees of freedom are considered. The training phase becomes even more difficult with more degrees of freedom in the pose and would be logistically impossible if the learning permutations included lighting changes.

PCA descriptions are taken directly from the image and there is no direct relationship between image intensity and object reflectance – an issue discussed in Section 3.1.1 – and, therefore, the PCA descriptions will be describing both the object and the current lighting condition. There is research in building representations that account for variations in lighting conditions and viewing pose [35, 36, 37]. This research is designed for face tracking and may not easily be applied to robot localisation for a number of reasons, such as, not accounting for scale changes, using restrictive assumptions about the lighting and requiring access to specific different lighting conditions in the a priori learning phase.

Jogan et al. [2] show that rather than learning different lighting conditions, some of the effects of lighting can be removed from the eigenspace using gradient filtering. Gradient filtering will remove the overall lighting level from the eigenspace; however non-uniform lighting effects will not be removed. They present successful localisation results in a number of different indoor lighting conditions, though it is unclear how well such methods perform in the highly non-uniform lighting conditions in outdoor environments.
There are many alternative global image descriptions proposed for visual localisation in contrast to these eigenspace encodings. Ulrich and Nourbaksh [38] use chrominance information for their localisation system, instead of it being based on intensity information. Chrominance information is an attractive option, because it provides some resilience to lighting changes and is often cited as a solution to the lighting problem, such as suggested by DeSouza and Kak [1] and also shown through experimentation by this thesis in Section 3.3. However, spectral changes in the lighting cannot be factored out by chrominance information, which is demonstrated by Austin and Barnes [39]. Furthermore, chrominance information is not useful in greyish environments where monochrome intensity is the pertinent visual cue.

The RatSLAM framework has proposed some different global image encodings to couple with a computational model of the mapping process in the rodent brain. In [40], an image is encoded as an array of the average intensity of the image columns. In [41], chrominance histograms are used. Gabor filters are used in [42]. Any of these different encodings can become a local view cell in the RatSLAM system and inject activity into the competitive attractor neural network of pose cells.

Zhang and Kleeman [43] store Fourier transforms of images recorded along a route. They restrict the localisation to a specific route which enables large scale outdoor environments to be navigated. Zhang and Kleeman present a patch normalisation process, which attempts to cancel the affects of the lighting. The normalisation cancels the overall lighting level; however, the normalisation will not remove the effects of shadows and shading caused by non-uniform lighting, which can be noticed in the figures in [43].

In summary, the different image encoding techniques share the common trait of learning the environment as an encoding recorded at each pose in the environment. This learning process can be considered an advantage, because the representation can be generated in an autonomous or semi-autonomous manner. However, this is also a disadvantage, because the encoding will be specific for the lighting condition under which it was recorded. Gradient filters have been proposed to factor out the overall lighting level, though non-uniform lighting is still a problem. Depending on how specific the encoding is to a lighting condition and how drastic the lighting changes are, the visual localisation system may fail in another lighting condition. The only obvious solution to this problem is that images need to be learnt for the many permutations of viewing pose and lighting in an logistically intensive learning phase.
2.6 Image-point Features

The image-encoding methods reviewed in the previous section encode entire images, where the encoding from each camera pose provides a unique description. The goal of the point-feature-based techniques reviewed in this section is to reliably select and describe points in a representation that remains constant through many viewing conditions. If successful, this process reduces the intensiveness of the learning phase, because features learnt under a specific viewing condition can be used for localisation from another viewing angle or lighting condition.

One of the first examples of robot localisation is based on Harris point features [15]. Many recent techniques use the more robust scale invariant feature transform (SIFT) developed by Lowe [17, 6]. The SIFT features are distinctive, which allows a single feature to be matched uniquely against a large database of features, thus providing a good basis for pose estimation. The main advantage of SIFT is robustness to scale and viewing rotations, making it one of the most popular techniques. A more recent variant that is more efficiently computed is the SURF technique [18].

Point-feature techniques can be applied to visual localisation in different ways. They can be used in a geometric manner such as in the work of Harris [15], and Se and Lowe [3], positioning the features into a 3D map using stereo cameras or over time using the motion of the vehicle. The approach is to match features in consecutive images, then using a least squares or alternative minimisation approach to find the movement of the vehicle that brings the image positions of the features together such that the discrepancy is minimised. Integrating these motion measurements over time allows the positioning of feature landmarks into a global map. The uncertainty when estimating the locations of the features is often modeled by an individual Kalman filter (or variant, such as the Extended Kalman Filter or Unscented Kalman Filter) for each feature location. The uncertainty grows as the robot explores the environment and when the robot completes a loop and recognises old features, an optimisation procedure can be triggered to minimise the reprojection error across all features observations. Limitations with this type of approach include computation complexity as the size of the map increases, difficulty maintaining maps over long periods of time and a significant problem is the growing uncertainty in the feature positions. Uncertainty in the locations of features can get so large that loop closure events become difficult to detect and, also, can be too large to reliably perform the map optimisation procedure resulting from loop closures.

Cummins and Newman [4] and Valgren and Lilienthal [7] present alternative localisation approaches that avoid accumulating geometric errors. Their methods use features in another non-geometric manner, instead of creating geometric maps with the features, the sets of features detected at each place are recorded. Places which have been seen previously can be recognised as having significant correspondence with the learned feature set. Cummins and Newman’s robust algorithm can deal
with the situation where two different places have high correlation by inherently recognising the distinctiveness of each place and, therefore, only recognising places that have a high distinctiveness. These types of localisation approaches, whilst avoiding the accumulation of errors and providing recognition of distinct places, cannot, on their own, provide accurate geometric self-pose estimates. Newman et al. [25] show how to use these place recognition techniques to trigger a loop closure event in another geometric localisation system.

The different approaches and algorithms for localisation from image-point features will, no doubt, each have different robustness properties. However, the fundamental properties of the features themselves can be studied to promote an understanding of the performance of this group of techniques.

The SIFT description has been noted to be sensitive to viewing pose changes when the local-planarity assumption is violated, as revealed by Vedaldi in [27] and also mentioned by Lowe [6]. The problem is that the SIFT features, and other feature descriptors, are not a 3D description, they are just transforms of a 2D image and, therefore, do not correctly represent 3D objects in the environment.

Instabilities to lighting changes have been noted by many authors. Lowe [6] found light disrupted his descriptor especially when the planarity assumption was violated. Sim and Dudek [5], discovered the sensitivities of features with experiments through light changes. Valgren and Lilienthal [7] find difficulties localising with features recorded under different lighting conditions to the reference dataset, but produced better results in later work on the same datasets with improvements to their matching scheme [8]. They attribute their improvements to three separate factors:

- processing the images in full resolution
- introducing an epipolar constraint
- using RANSAC to remove outliers.

It is not clear if robust techniques to remove outliers are sufficient to enable reliable localisation when the number of inliers is low after large changes in lighting. Moreover, the experiments performed previously have only incorporated a few separate lighting conditions. This thesis conducts a more detailed lighting change study in Section 3.4, testing SIFT features with the full range of lighting changes experienced outdoors over a complete day.

In summary, image-point features have the major advantage of enabling localisation and autonomous mapping – no doubt one of the primary appeals of these techniques. However, the features are not entirely robust, which is of concern for applications that require operation over long periods.
2.7 Stereo Point Clouds

Stereo point clouds are another type of visual information that can be used for SLAM. There is much localisation work using laser scanners to localise and build maps from scanned 3D point clouds. Similar methodologies have been adopted in vision work, using dense stereo point clouds instead of laser-scanned point clouds.

Elinas et al. [19] present a particle filter approach in indoor environments, using sparse features for visual odometry, then using dense stereo to build a 2D occupancy grid. Marks et al. [20] present a similar approach, although using a variance grid representation instead of an occupancy grid, and show successful results in outdoor off-road terrain. Marks et al. do not remember all previous individual observations and, therefore, do not explicitly deal with loop closure events, by optimising the error in each observation. Loop closures are accommodated intrinsically with their particle filter approach. They merge all observations into a map for each particle and, if the odometry of the system is sufficiently accurate, there will be enough particles at or near the correct vehicle pose when there is a loop closure. These particles will be intrinsically highly weighted because their current observations should be similar to their map and, therefore, particles are likely to be resampled at the correct pose.

Herath et al. [44] provide a study of stereo vision for localisation and mapping, identifying and characterising errors. They state that stereo vision techniques have not yet proven they can perform better then laser based systems for extended periods of time. This is due to a number of issues with stereo vision, including the limited field of view, spurious measurements and structured environments not providing suitably textured surfaces for generating reliable depth measurements. Although generating accurate point clouds in structured environments is problematic, stereo vision is more suited to unstructured environments and has been applied successfully on the surface of Mars [45].

Specular reflections are also known to cause errors with stereo correspondences, because the appearance of an object can be different in each camera.

2.8 3D-surface

Gerard and Gagalowicz [21], Noyer et al. [46], Gupta and Brennan [47], and Ho and Jarvis [22] present visual localisation work with the use of 3D surface models. Like the 3D-edge representation, 3D-surface representations have the advantage of being explicitly separate from viewing conditions. This is an advantage, because the model can be used from any viewing angle with an explicit 3D-to-2D transform. But unlike the 3D-edge representations, 3D-surface maps do not need to be sparse and can
represent complex objects with textured surfaces. The image-to-model comparison can be performed at the image-level, between real images and synthetic images rendered of the environment. At the beginning of the 1990’s, Gagalowicz [48] predicted that, in the future, computer graphics technology will be used within robotic vision to solve hard visual problems; now, through the advancement of graphics processing hardware, this concept is beginning to be possible.

Gupta and Brennan [47] present an orientation-only localisation method for ground vehicles in large outdoor environments, using a surveyed 3D-terrain map. Gupta and Brennan project points from the map onto the virtual image plane, selecting the highest points to form an expected horizon line. This virtual horizon line is matched to the real horizon line that is extracted from the camera image, giving an estimate of the current roll, pitch and yaw angles of the vehicle. This technique provides an accurate means to estimate the orientation of a vehicle in open terrain where the horizon is in clear view.

Noyer et al. [46] and Ho and Jarvis [22] present orientation and position estimation employing a particle filter that uses a rendered synthetic image of the 3D-model for each particle. This requires a large amount of processing and Ho and Jarvis perform all the rendering off-line to build an image encoding database of Haar wavelets. However, it is now becoming feasible, with the progression of graphics processors, to quickly generate and process a large number of synthetic images in real-time.

One issue with the existing localisation approaches using 3D-surface models, is that they only consider textured models and do not properly consider the surface reflectance properties and the interactions with lighting.

Work in the domain of face identification [34, 49, 50, 51] introduces the idea of using a 3D reflectance model together with a light model. They show how to perform face identification in unknown lighting conditions, by first estimating the current lighting source then rendering the face models using the estimated light model.

This is a new way of approaching robotic vision and has the potential to solve many difficult visual scenarios. Whereas, most previous techniques have focused on factoring out lighting at the image-level, which is explained in [50] and also shown in Chapter 3 to be problematic, if not impossible. This new work in face identification accounts for lighting with a new approach by explicitly incorporating light information within the visual process. This framework has the potential to lead the way for a higher level of competency in robotic vision systems, through the understanding and use of lighting information, rather than trying to factor it out.
2.9 Discussion

Techniques using a 3D-reflectance-map in conjunction with a lighting model have only been applied for face recognition, but have not yet been applied to localisation. This thesis will look at applying this type of framework to localisation in Chapter 6.

2.9 Discussion

This chapter has reviewed the existing visual localisation research, categorising the techniques by their type of environment representation. The categories were 3D-edge, global image encodings, point-features, point-clouds and 3D-surface. Table 2.1 provides a summary of the advantages and disadvantages.

Of all the techniques reviewed, those that can semi- or fully-autonomously, learn maps have an obvious advantage for exploring robots. For applications that do not need to explore unknown environments and that require high dependability, autonomously mapping an environment is less of an advantage, while there is the major disadvantage if the map contains non-permanent information, because these will be susceptible to failure. Representations that are mapped manually, such as 3D-edge and 3D-surface maps, ensure that the information in the map is permanent and explicitly separate from the lighting conditions. Manually-generated maps are expected to be more reliable over long periods. However, such maps are not applicable for environments that are inaccessible a priori.

The review highlighted that each method has its niche application environments and does not perform as well in other environments. Stereo point clouds, for example, can be used for autonomous map generation in unstructured terrain. However, stereo data is not as reliable in structured, texture-less environments where dense correlations are difficult to acquire. 3D-edge maps are good at describing simple structures where the edges are pertinent visual features, though they are not good at describing complex environments with a lot of clutter. Global image encodings and image-point features are better at describing complex cluttered scenes. However, because the representations are learned from images, large non-uniform lighting changes may cause disruptions.

3D-surface maps are a less common type of representation proposed for visual localisation systems. Previous techniques that use 3D-surface maps do not properly deal with lighting variations. They do not consider the reflectance properties of the surfaces and, therefore, cannot generalise the maps to new lighting conditions. Authors have presented 3D-surface model work for face identification that includes a light model within the process to relight the surface model before comparisons. This idea to use a light source model has the potential to solve visual problems that cannot be solved by the old factor-out lighting approaches. A light source model has been rarely, if ever, applied to visual
Table 2.1: Advantages and disadvantages of different visual localisation techniques.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Advantages (+) and Disadvantages (-)</th>
</tr>
</thead>
</table>
| 3D Edge                | +Explicitly separate from lighting and viewing pose.  
                           +Real-time.  
                           -Manual map creation.  
                           -Sparse and cannot describe complex objects.  
                           -Discrepancy between information in image and map. |
| Global Image Encodings | +Can describe complex objects.  
                           +Autonomous map creation.  
                           -Specific to lighting (non-uniform) and viewing pose.  
                           -Must have recorded and described images for each lighting condition and viewing pose. |
| Image-Point Feature    | +Can describe complex objects.  
                           +Autonomous map creation.  
                           -Not completely robust to lighting conditions (how bad is this problem?).  
                           -Difficult to deal with accumulating errors and accumulating computation complexity.  
                           -Description is not 3D and therefore not stable when local planarity assumption violated.  
                           -Map maintenance is difficult. |
| Stereo Point Clouds    | +Can manually map the environment.  
                           +Proven to be useful for unstructured terrain.  
                           +Describes the explicit geometry in the environment.  
                           -Stereo data is erroneous in structured texture-less environments.  
                           -Specular reflections cause errors. |
| 3D Surface             | +Can describe complex objects.  
                           +Explicitly separate from lighting and viewing pose.  
                           +Can potentially include lighting model within the process.  
                           -Manual map creation.  
                           -Large amount of processing required (modern GPU’s can provide this?). |

localisation. This thesis will attempt to develop a visual localisation framework in Chapter 6 that incorporates a light model.

Chapter 5 looks to develop another visual localisation system to localise a ground vehicle in an outdoor industrial site. This application requires reliable real-time localisation over long periods of time. Therefore a map is required that comprises permanent information that can be used for localisation in many lighting conditions. 3D-edge maps are identified as the most probable technique to provide this type of performance. 3D-edge maps have rarely been applied to outdoor localisation,
perhaps because they are sparse and do not include much information that is extracted from images. Recent work by Klein and Murray [9], propose a multiple hypothesis technique, providing robustness to this superfluous edge information. However, it is yet to be applied to outdoor vehicle localisation. Chapter 5 develops a multiple hypothesis 3D-edge localisation system and evaluates its reliability in the challenging conditions in outdoor industrial sites.

Before these visual localisation systems are developed, the following chapter, Chapter 3, evaluates the various classes of visual information in dynamic lighting conditions to develop a greater understanding of the lighting problem. The chapter will reiterate the difficulties acknowledged by other authors of autonomous mapping in fluctuating lighting, and specifically investigates the sensitivity of SIFT, one of the most commonly used image-point features in visual SLAM.
Any visual localisation system's robustness to dynamic non-uniform lighting is owed largely to the robustness of the visual information upon which the system is based. Even though a system can use robust algorithms to deal with some variance in the incoming information, ultimately these algorithms will fail if the visual information does not have a level of resilience to lighting changes. SLAM systems are even more susceptible to lighting invariance, because these systems must not only localise from unstable information, they must also build and maintain a map which itself must be permanent, and permit localisation under different lighting conditions.

When designing a visual localisation system, it is important to have an understanding of the stability of the type of information upon which the system is based. This chapter conducts an initial study of visual information to lighting changes and performs experiments in dynamic lighting to evaluate the robustness of different classes of visual information. The chapter begins with an introduction to image
formation and the perception of object reflectance in dynamic lighting conditions in Section 3.1. Section 3.2 continues the discussion of object reflectance recognition in the context of the basic visual task of foreground segmentation. Section 3.3 presents experiments on the stability of chrominance information during lighting changes. Section 3.4 looks at the performance of the commonly used SIFT image point feature technique during outdoor lighting changes. Concluding remarks from the study and experiments are presented in Section 3.5.

3.1 Introduction

At this stage, it is prudent to consider the theory of image formation, because it is the basis of any visual system. The light that reaches the camera depends on the sources of light, $L$, and the reflecting surfaces, $O$, of objects in view. These two factors have complicated interactions and cannot be modelled a linear relationship. The relationship is a complicated combination of a number of variables, including surface normal, incident angle, albedo, distances, inter-reflections, wavelengths, among others. The measured light intensity, $I$, of each pixel, $p$, varies according to the amount of light reaching the camera, and can be expressed as a function of light $L$, the object surface reflectance $S$, and the sensitivity of the photoreceptor, $P$:

$$I(p) = f(L, S, P)$$  \hspace{1cm} (3.1)

In real-world environments, the behaviour of $L$ can vary dramatically. The lighting can be dynamic, in that the light source is moving or changing in its properties over time, and the lighting can also be non-uniform, in that objects in the camera’s view range from being directly facing the light, to being on an angle to the light source and other objects in the shade not directly lit by the light source at all.

Outdoors, on clear days, the sun behaves like a point light source and the angle of the sun in the sky is the significant variable, causing large non-uniformities between those surfaces receiving direct light and others in shadows. From sunrise through to sunset, the intensity of light will change and is accompanied by shifts in the spectral composition of the light. On overcast days, the cloud cover (or pollution) disperses the light more uniformly, removing the large shadows and the angle of the sun in the sky does not have such a large impact. However, even on overcast days the light is not completely uniform, there is still a difference in shading on the underside of objects compared to the surfaces facing upwards. The lighting on clear days is far more challenging, because the direct sunlight causes large non-uniformities in the lighting of the objects.
In Australia, where this thesis tests its visual localisation systems, there are many challenging days where the weather is clear. These clear days are, perhaps, clearer then many other locations owing to a lack of pollution. Also, in the northern regions of Australia which are closer to the Equator, the lighting is even more challenging because the sun is brighter. All of these reasons result in the lighting conditions in Australia being more challenging then the lighting conditions often present in much of the robotic vision experiments in the literature. Experiments presented in the literature are often conducted in cloudier climates, with more pollution and farther away from the Equator. These differences in conditions must be taken into account when designing and evaluating systems and experiments.

Regardless of the location in the world, the problem remains that lighting changes translate into complicated variations in the intensity values in the image, even if the object in view remain static in the scene. It is not as simple as factoring out a linear increase in the power of the light source, pixel values will change and will change non-linearly with respect to each other.

3.1.1 Recognising Reflectance

Robotic vision depends on detecting, locating and interacting with objects in the environment. Therefore, the factor of interest in Equation 3.1 is the object reflectance, because it is the only factor that holds the visual information about the objects in the environment. Unfortunately, there is no simple relationship between pixel intensity and the object reflectance, unless the lighting is uniform and static. In non-uniform dynamic lighting conditions, there is no obvious way to factor-out the complex lighting factor.

It is documented that human vision has the ability to perceive object reflectance throughout lighting changes. Cornsweet [52] discusses this ability and explains that humans achieve it by estimating the lighting conditions present in the scene and adjusting their perception accordingly. Cornsweet describes an experiment to prove this. The experiment is conducted in a room where there are a variety of different surfaces fixed to a wall. A viewer is asked to identify the surfaces using a reference chart placed on the desk in front of them. When an additional light source is focused on certain surfaces, and the viewer is aware of this light, the viewer adjusts their perception and can still correctly identify the surfaces. However, when the light source is focused on certain surfaces in such a way that the viewer does not realise that some surfaces are lit by the new light source while the other surfaces are still lit just by the ambient light, the viewer then loses the ability to identify surfaces. Cornsweet’s study illustrates the importance of understanding the present light source when evaluating surface reflectance properties.
Robotic vision has yet to progress to the level of perception that is displayed by the human visual system. Robotic systems developed to recognise object surface reflectance generally do not estimate or include a model of the light source. Primarily, lighting conditions are dealt with by robotic vision systems by attempting to factor-out lighting or by constantly adapting models of surface reflectance.

3.2 Foreground Segmentation in Dynamic Non-uniform Lighting

This section studies the low-level task of foreground segmentation, and although it is slightly off-topic from visual localisation, it serves as a good example of the problem of attempting to recognise the reflectance properties of objects during lighting changes. The foreground segmentation task is to learn a model of the reflectance properties of the background in a scene with a static camera. Then, when areas in the image conflict with the background model, these are detected as belonging to foreground objects that have different surface reflectance. Ideally, the process should be unaffected by fluctuating lighting and attempts to achieve robustness usually do so by constantly adapting the background model.

Traditional background models are based on raw intensity measurements. There is no direct relationship between intensity and object reflectance, as discussed earlier. Changes in pixel intensity can either be from changes in lighting or changes in object reflectance, and it is not feasible to competently differentiate between the two causes of intensity changes at the pixel level. These intensity-based approaches must adjust their background model over time to account for lighting changes. However, these approaches are susceptible to significant errors while the lighting is fluctuating.

This section reiterates the problem of modelling reflectance as raw pixel intensity in dynamic lighting, and shows that it is more appropriate to model reflectance in the intensity gradient domain as proposed in recent works [53],[54],[55]. Intensity gradient has the advantage of being intrinsically robust to uniform lighting changes, though non-uniform lighting changes must be addressed. For example, if a moving cloud quickly reveals the sun, the sharp shadows cast by buildings and other stationary objects will cause large changes in the intensity gradient and false object detections. A novel algorithm is proposed to deal with these types of appearing shadow edges. The algorithm has been published in [56].
3.2 Foreground Segmentation in Dynamic Non-uniform Lighting

3.2.1 Intensity Background Models

Background models are traditionally based on intensity measurements; Piccardi [57] provides a survey of the predominant intensity-based techniques. An image intensity measurement, $I$, at pixel $p$, is a fusion of illumination source $L$, object surface reflectance $S$ and camera’s photoreceptor sensitivity $P$, over light wavelength $\lambda$:

$$I(p) = \int_{\lambda} f(p, \lambda, L, S, P) d\lambda$$  \hspace{1cm} (3.2)

Ignoring the sensitivity of the photo-receptors and the wavelength distribution the intensity formation equation simplifies to:

$$I(p) = f(p, L, S)$$  \hspace{1cm} (3.3)

Assuming a static camera, a change in intensity will be due to either a change in the object surface reflectance $S$ or a change in the illumination source $L$. The factor of interest is $S$, considering the task is detecting foreground objects. However, obtaining direct observations of $S$ from $I$ during fluctuating lighting $L$ is problematic. This can be seen in Figure 3.1, showing a typical change in cloud cover, where the two camera images were 40 seconds apart in the video sequence. However, the plot of the intensity of pixel intensity from the road through the 40 second sequence shows a drastic change in intensity of 30 percent. Understandably, any intensity-based foreground extraction will create false detections due to such lighting changes. This can be seen in the output of an intensity-based algorithm where both sunshine and a person are detected. To avoid this problem, more and more complex adaptive background models have been introduced; but, fundamentally, any intensity-based approach will be susceptible to errors such as seen in Figure 3.1.

3.2.2 Gradient Background Models

The three recent papers [53][54][55] avoid such dramatic errors during fluctuating lighting, because they model discontinuities (edges) in the intensity gradient, thus providing intrinsic robustness to global lighting levels. Most edge-filters consider neighbourhoods of pixels and for the purpose of a simple demonstration, consider only the relative intensity between two pixels, $a$ and $b$, that neighbour either side of pixel $p$, to illustrate the robustness to the lighting levels of an edge image, $E$:

$$E(p) = |I(a) - I(b)|$$  \hspace{1cm} (3.4)
Assuming for this demonstration that \( L \) and \( S \) are linear factors, the equation can be rewritten as:

\[
E(p) = |L_a S_a - L_b S_b| \tag{3.5}
\]

Assuming uniform lighting; \( L_a = L_b \), the edge image will have the lighting invariant property of being zero in areas of homogeneous object reflectance:

\[
E(p) = \begin{cases} 
0, & \text{if } S_a = S_b \\
> 0, & \text{if } S_a \neq S_b 
\end{cases} \tag{3.6}
\]
3.2 Foreground Segmentation in Dynamic Non-uniform Lighting

![Edge Image with Cloud Cover](image1)

(a) Edge Image with Cloud Cover  
(b) Edge Image with Direct sunlight

![Output of gradient based foreground segmentation](image2)

(c) Output of gradient based foreground segmentation

**Figure 3.2:** Gradient based (edge based) foreground segmentation example from the same image sequence from Figure 3.1. Here, the person is detected along with the edges of the shadows that appeared when clouds moved revealing the sun, thus illustrating that the intensity gradient is not stable to non-uniform lighting changes.

Changes in $E$, from zero to non-zero values, can therefore be attributed to a change in object reflectance, with certainty that it is not due to a uniform lighting change. Therefore, an edge based foreground segmentation will be stable during uniform lighting changes.

However, this stability does not hold for the case of non-uniform lighting changes. For example, the lighting change presented in Figure 3.1 reveals a sharp shadow cast by the building. An edge-based foreground segmentation would not be robust to this shadow. Figure 3.2 shows the shadow being detected by a basic edge-based foreground segmentation.

The edge-based techniques of Javed *et al.* [54] and Davis and Sharma [53] do not provide adequate solutions to avoid detecting static shadow edges. Yokoyama and Poggio [55] present a better solution by only detecting only the edges at pixels with optical flow vectors and, therefore, do not detect static shadow edges. The edge-based method of Smith *et al.* [58] will also be robust to appearing static shadows by estimating motion models for each edge, then cluster and segmented edges considering the similarity in the estimated motion model.
All the edge-based segmentation approaches discussed above attempt to cluster edge pixels into distinct objects, which is very difficult at the pixel-level. Furthermore, foreground areas are created without directly-available information in the image, indicating the presence of a foreground object. The method proposed in this section proposes a two-fold approach to detect internal foreground areas with more direct information, whilst also being robust to appearing static shadow edges. The initial stage increases the internal foreground area detected by using a multi-resolution pyramid background model, and with the second stage presenting the novel idea of using edge-based motion history images to segment whole foreground areas of moving objects. The details of the algorithm are explained in the following two sections.

**Temporal Background Filter**

In this work, a temporal median filter is used to form the background model, which stores a set of previous edge images. Traditional median filters store every image in the sequence, which is expensive. Cucchiara et al. [59] propose to using a single temporal scale of every 10 frames to reduce the storage and processing cost. However, with a single temporal scale, frame $t - 100$ holds as much weight as frame $t - 10$. This section introduces a background model with increasing time scales. The background model $B$, at pixel $p$, at time $t$ is calculated as:

$$B(x, y) = \frac{\sum_{l=0}^{m} E_{t-2^l}(x, y)}{m}$$  \hfill (3.7)

Where $E_k$ is the input edge image at time $k$. $m$ is the number of distinct images to include in the

<table>
<thead>
<tr>
<th>Traditional Median</th>
<th>Cucchiara et al.</th>
<th>Multiple Temporal Scales</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_{t-1}$</td>
<td>$I_{t-10}$</td>
<td>$I_{t-1}$</td>
</tr>
<tr>
<td>$I_{t-2}$</td>
<td>$I_{t-20}$</td>
<td>$I_{t-2}$</td>
</tr>
<tr>
<td>$I_{t-3}$</td>
<td>$I_{t-30}$</td>
<td>$I_{t-4}$</td>
</tr>
<tr>
<td>$I_{t-4}$</td>
<td>$I_{t-40}$</td>
<td>$I_{t-8}$</td>
</tr>
<tr>
<td>$I_{t-5}$</td>
<td>$I_{t-50}$</td>
<td>$I_{t-16}$</td>
</tr>
<tr>
<td>$I_{t-6}$</td>
<td>$I_{t-60}$</td>
<td>$I_{t-32}$</td>
</tr>
<tr>
<td>$I_{t-7}$</td>
<td>$I_{t-70}$</td>
<td>$I_{t-64}$</td>
</tr>
<tr>
<td>$I_{t-8}$</td>
<td>$I_{t-80}$</td>
<td>$I_{t-128}$</td>
</tr>
</tbody>
</table>

**Figure 3.3**: Comparison of temporal median filters.
model and in this work $m = 7$ and therefore frame $t - 128$ is the furthest image into the past that is incorporated. In the background, model the frames $t - 1$ to $t - 8$ will have the same weight as frames $t - 16$ to $t - 128$. This model places greater importance of events nearer in the past, whilst still incorporating events over a long period. Figure 3.3 illustrates this technique. To avoid the memory cost of storing 128 full images, only $m$ images are stored at any time. The edge images, $E_{2^k}$, used to calculate $B$ are only temporally approximate and updated only every $2^{(k-1)}$ frames, as follows:

$$E_{t-2^{(k-1)}} \rightarrow E_{t-2^k}$$

(3.8)

This reduces the memory overhead of image storage. The filter does require per-pixel processing of each of the $m$ background images; however the processing is not as intensive in comparison with the updating steps of alternative background models, such as the Gaussian mixture model [60].

A set of binary images $T$ are calculated from thresholded subtractions performed between the current edge image and the edge-images in the background model. The subtraction detects those pixels with an edge value a threshold, $\psi$, above the background image value:

$$T_m(x, y) = \begin{cases} 
true, & \text{if } E_t(x, y) - B_m(x, y) > \psi \\
false, & \text{otherwise} 
\end{cases}$$

(3.9)

Through empirical experimenting to find an appropriate threshold value, $\psi = 10$ provided the best results. The next stage is to calculate the final foreground image $F$ by selecting pixels with a majority of positives from the set $T$:

$$F(x, y) = \begin{cases} 
true, & \text{if } \sum_{i=0}^m T_i(x, y) > \frac{m}{2} \\
false, & \text{otherwise} 
\end{cases}$$

(3.10)

If only the highest resolution is used, there is a lot of internal foreground area that is not detected, as seen in the middle column of Figure 3.4. Therefore, multiple-scales are used to increase the area of foreground segmented. A standard pyramid approach is adopted, by recursively down-sampling each input edge image into a multiple spatial scale image pyramid. The image-pyramid is passed through the edge filter to form an edge-pyramid. A background-model is recorded as a set of edge-pyramids using the temporal filter. Foreground pixels are detected from each level in the pyramid and merged into one high-resolution output image. Examples of the final merged output are shown in the right hand column of Figure 3.4.
Figure 3.4: Images showing the results of the edge-based background modeling. Top: A moving van. Bottom: Person walking around a loading bay. Left: Raw video. Middle: Foreground at the finest scale. Right: Multiple-spatial-scale foreground extraction merged together. Even with multiple-scales not all of the object area is detected.
3.2 Foreground Segmentation in Dynamic Non-uniform Lighting

3.2.3 Region Growing with Motion History

Even with the use of multiple-scales, there are still situations where there is not a complete detection of objects. This can be seen in Figure 3.4, where there are internal areas of the van that are not detected. Other relative-intensity approaches attempts to detect full object areas by first clustering edges into an object, which is problematic, as stated by Yokoyama and Poggio [55]. Instead of clustering, this paper proposes to use edge-motion-history-images introduced as a basis for segmenting entire foreground regions and also being robust to appearing static shadows.

Motion-History-Images (MHI) were developed by Davis and Bobick [61] and are traditionally intensity-based images. This paper presents the novel use of edge information with MHI (Edge-MHI). Edge-MHI enables complete foreground areas to be segmented while also being robust to global lighting conditions.

**Edge-MHI**

The Edge-MHI, $F^H_t(x, y)$, describes the history of the edge foreground extraction calculated as follows:

$$F^H_t(x, y) = \begin{cases} 
\text{MAX}(0, F^H_{t-1}(x, y) - 1), & \text{if } F_t(x, y) = \text{false} \\
\kappa, & \text{if } F_t(x, y) = \text{true}
\end{cases}$$

(3.11)

where $\kappa$ is the number of frames to record the motion history. $\kappa$ is dependant on a number of factors, the frame rate, the image resolution, the speed of the moving objects and the distance the moving objects are from the camera. In this work, $\kappa = 15$ and was selected through experimentation. Examples of Edge-MHI can be seen in the left column of Figure 3.5 (videos attachments – labeled mhiWalk.mpg and mhiVan.mpg – also show the Edge-MHI). The detected object area, with the edge-MHI, is noticeably more complete when compared with the straight edge extraction in Figure 3.4.

**Region Growing Algorithm**

The area detected by the Edge-MHI is more complete, but there are trails in the Edge-MHI left behind the objects, which need to be removed. Furthermore, there are areas surrounding the object, caused by the use of larger scales, that also need to be removed to give an accurate high-resolution segmentation.
Examples of both these problems can be seen in Figure 3.5. A seeded region growing algorithm is developed to discard the unwanted areas and select the actual foreground areas.

A related edge-restricted growing method is proposed in [58]. The algorithm presented in this paper restricts the growing at the fine-scale foreground edges and is summarized as follows:

1. Distance transform on the Edge-MHI to find central pixels.
2. Breadth-first region growing in the Edge-MHI, seeded from pixels discovered in the previous step. Stop growing when restricted by fine-scale foreground pixels or the boundary of the Edge-MHI. The middle column of Figure 3.5 shows examples of the regions grown.
3. Regions are selected as foreground when they have a significant count of boundary pixels that are either fine-scale foreground pixels or neighbour previously selected regions, as defined in Equation 3.12.

The grown regions, $F^R$, pass through a selection process based on its boundary pixels, $\beta$, to decide whether this region is an internal foreground region or a motion history trail. The decision metric is the ratio between the count of boundary pixels that are deemed valid and those that are deemed invalid. A valid boundary pixel, $\beta_v$, is one that is a fine-scale foreground pixel or is a pixel belonging to a region that has previously passed the selection process. An invalid boundary pixel, $\beta_i$, is one which is not a fine-scale foreground pixel or is a pixel belonging to a region that has previously failed the selection process. The decision is made by a threshold $\mu$:

$$F^R = \begin{cases} 
\text{foreground}, & \text{if } \frac{\beta_v}{\beta_i} > \mu \\
\text{background}, & \text{otherwise}
\end{cases}$$  

Equation 3.12

During the experiments the threshold $\mu = 1$ proved to select correct regions in most of the cases. The middle column of Figure 3.5 shows the decisions made by the selection process and the right column shows regions that passed the process. Examples of the final foreground extraction are in Figure 3.6 (videos attachments – labeled outputWalk.mpg and outputVan.mpg – also show the final segmented foreground).
Figure 3.5: Examples of edge-MHI and the region growing process. Top: A moving van. Bottom: Person walking around a loading bay. Left: Edge-MHI, trails are left behind objects. Middle: Region growing. Right: Final output, trails are removed.
Figure 3.6: Top: Raw video. Bottom: Final output. There is no false-positive detection, showing an improvement of the algorithm in this section over the intensity-based method shown in Figure 3.1.
3.2 Foreground Segmentation in Dynamic Non-uniform Lighting

3.2.4 Results

The algorithm is tested on both moving pedestrians and vehicles during quick changes in lighting. When processing $640 \times 480$ images at half-resolution, the entire algorithm presented in this system functions at 20Hz on a 3.2GHz CPU. The quick changes were a result of moving cloud cover, when the sun goes from being totally blocked by clouds to being completely uncovered within a space of around 30 seconds.

In the third row of the examples in the Figure 3.6, the pedestrian is detected without any false positives from the shadow. Shadows cast by buildings are static and will not have motion history and therefore will not be segmented with the algorithm presented in this section. This shows the improvement of this algorithm over a traditional intensity-based technique as shown in Figure 3.1, and also an improvement over a basic edge-based segmentation as shown in Figure 3.2. Only shadows that are cast by moving objects will result in false detections from the method in this section. Some techniques to remove cast shadows have been proposed, such as in [62], however, it is difficult to remove all the effects of non-uniform lighting for all situations directly at the pixel level.

3.2.5 Summary of Foreground Segmentation Study

This section presented a study on foreground segmentation in dynamic non-uniform lighting. Although the task is unrelated to visual localisation, the study provided an understanding of the difficulties of recognising object reflectance in changing lighting conditions. Traditional foreground extraction techniques are based on raw-intensity measurements, which are sensitive to changes in lighting. Edge information is derived from the intensity-gradient which is intrinsically robust to global lighting levels. However, non-uniform lighting will cause disruptions such as at the edge of appearing shadows. A technique has been developed that proposes the use of edge motion history images to ignore static shadow edges and to segment whole foreground areas. Moving cast by moving objects will be detected, and it appears that pixel level techniques are not able to remove all the effects of non-uniform lighting.
3.3 Chrominance Spaces

Colour is a powerful and direct cue in computer vision and is the basis of some visual localisation systems, such as the work of Ulrich and Nourbaksh [38]. Colour is often considered as being intrinsically separate from illumination levels. The recommendation in DeSouza and Kak’s review paper [1] is that chrominance on its own is a solution to the problems with outdoor lighting changes. Most colour spaces have separate channels for luminance; however, experiments presented in this section reveal that some of the colour channels are not explicitly separate to overall light levels. The experiments compare commonly used colour spaces and evaluate stability in changing lighting conditions.

In addition to revealing transforms which are unstable to the most basic of light changes, the experiments in this section highlight a number of other factors that cause chrominance instability, low-light sensor noise, sensor saturation, inconsistent video encoding and decoding standards and camera brightness parameters.

Importantly, this section also identifies that chrominance is not useful in bland environments, such as the application environments investigated in this environment.

3.3.1 Terminology

The term chrominance is adopted for the remainder of this section to avoid the ambiguity with using the term colour. In fields such as painting; black, grey and white are considered individual colors. Whereas from a computer vision perspective, they essentially share the same colour, and are distinguished by intensity, not chrominance. Chrominance refers to the spectral signature and this section will discover the most suitable light invariant transform.

3.3.2 Background

There is a variety of different chrominance transforms, and some evaluations of these are available in the literature. In their qualitative experiments, Gevers and Smeulders [63] discovered that HSV and Normalized RGB are invariant to achromatic light changes. Finlayson and Schaefer [64] consider the affects of gamma for the HSV space. Zarit et al. [65] compare different transforms for skin classification.
3.3 Chrominance Spaces

Figure 3.7: Intensity coded histograms of the 16 million RGB values in the various chrominance spaces.

These previous studies give predominantly qualitative comparisons of the different transforms. There appears to be no quantitative study of the actual values produced by these transforms during lighting changes. This section performs rigorous experiments to provide quantitative results.

3.3.3 Chrominance Spaces

The first analysis in this section investigates the spread of RGB values throughout the well known chrominance spaces, YCrCb, HSV and Normalized RGB. The spread of RGB values is an important characteristic of a chrominance space, because it ascertains the uniformity across the space and indicates the ability to distinguish objects.

Figure 3.7 presents histograms of the 16 million RGB values in the different chrominance spaces. Following is an introduction to each chrominance spaces and a discussion of the uniformity of values:

YCrCb

The $Y C_r C_b$ chrominance space was designed for television sets to enable backwards compatibility with black-and-white systems. It is sometimes thought that cameras can sense in $Y C_r C_b$ format. This is a misconception brought about by the fact that many cameras transmit in $Y C_r C_b$ image format. In-fact, these cameras actually have red, green and blue photo-receptors and apply a transform into YCrCb. The standard digital television transform [66], calculated from digitized 8-bit RGB camera
images is:

\[ Y = \frac{77}{256} R + \frac{150}{256} G + \frac{29}{256} B \]
\[ C_r = \frac{131}{256} R - \frac{110}{256} G - \frac{21}{256} B + 128 \]
\[ C_b = -\frac{44}{256} R - \frac{87}{256} G + \frac{131}{256} B + 128 \]  \hspace{1cm} (3.13)

This digital encoding is actually a part of a family of different television standards that vary between countries and also vary between digital and analog systems. If the encoding of one format is decoded using another format, this can cause instability to the lighting in chrominance spaces computed from the incorrectly decoded \(RGB\) data.

Figure 3.7(a) shows the spread of values in the \(C_r\) and \(C_b\) planes; it can be seen that many \(RGB\) values map near the white-point located in the centre of the space and the density of values decreases moving away from the white-point. This is not ideal for computer vision purposes, because the many values near the white-point will be difficult to distinguish.

**Normalized RGB**

Hunt presents the \textit{Color Triangle} chrominance transform in [67]. The space is also known as Normalised \(RGB\), because the chrominance values, \(N_r, N_g, N_b\), are derived from each RGB value divided by the sum of the RGB triplet. Any pair of the Normalised RGB triplet can be used to describe a unique chrominance value (because \(N_r + N_g + N_b = 1\)).

\[ N_r = \frac{R}{R+G+B} \]
\[ N_g = \frac{G}{R+G+B} \]
\[ N_b = \frac{B}{R+G+B} \]  \hspace{1cm} (3.14)

The spread of Normalised RGB in the \(N_r\) and \(N_g\) planes is shown in Figure 3.7(c). A large portion of the space is not utilised and there is a large clump of values at the white-point of the Normalised RGB space. Moving away from the white-point the density of the values decreases quickly. The small dense clump of values in the Normalised RGB space is even more pronounced than the \(YCrCb\) space, and would indicate potential difficulties distinguishing chrominance values.
3.3 Chrominance Spaces

HSV

HSV, introduced by Alvy Ray Smith in [68], was designed to mimic human color perception, with three components, hue \( H \), saturation \( S \) and intensity \( V \).

\[
H = \frac{\theta_1 - \theta_2}{M - m} \\
S = \frac{M - m}{M} \\
V = \frac{R + G + B}{3}
\]

(3.15)

Where \( M \) and \( m \) are the maximum and minimum values of the RGB triplet, respectively, and \( \theta_1 \) and \( \theta_2 \) can represent any of the RGB triplet depending on which are the maximum and minimum values. This transform is defined in detail in [68].

The spread of the RGB values in the HSV space is presented in Figure 3.7(b) as a polar coordinate, with a white-point located in the center. The polar angle is hue and the radial axis is the saturation component. There is an almost perfect even spread of RGB values and there is no large gathering of values at the white-point, which the other spaces do have. This spread of values in the HSV space appears to be more suitable for computer vision purposes because chrominance values should be more easily distinguished from each other.

The outcome of the spread analysis is that HSV space has the most suitable chrominance spread for computer vision, because of its even and encompassing representation of the set of RGB combinations.

3.3.4 Achromatic Variations in the Lighting

This section presents both a theoretical and an empirical analysis of the ability of each chrominance space to maintain stable values through spectrally-uniform lighting changes. Gevers and Smeulders provided a comparison in [63] of Normalized RGB and HSV, and they found both to be invariant to light levels. Their analysis is revisited here with the inclusion of a YCrCb analysis. Also presented here are quantitative results and an examination of a range of image capture and transmission issues.

To perform a theoretical study, a lighting change is simulated for a RGB sensor and then translated to each chrominance space. The three different types of pixels in an RGB camera measure the intensity of light at the red, green and blue sections of the visible wavelength spectrum. Because these measurements are of intensity, on their own they will be sensitive to light changes. Though, when the RGB values are transformed into a chrominance space, ideally, the values in the chrominance space
will be stable to light changes. A linear light increase, $\kappa$, the simplest type of light change, is applied equally to each RGB channel. The equations are assumed to be linear and the light change can be simply written as:

$$
R(p)' = \kappa(R(p)) \\
G(p)' = \kappa(G(p)) \\
B(p)' = \kappa(B(p))
$$

(3.16)

This light change can now be propagated into each chrominance space. Unchanged chrominance values are the desired scenario, because this indicates stability to spectrally-uniform light changes.

**Normalized RGB and HSV**

Gevers and Smeulders in [63] show that the spectrally-uniform light change, ($\kappa$), is factored out of the Normalised RGB and HSV transforms.

**YCrCb**

Gevers and Smeulders do not provide a similar analysis of the YCrCb space, so this will be provided here. Considering the spectrally-uniform increase in light source intensity, $\kappa$, Equation 3.13 can be rewritten as:

$$
Y' = \frac{77}{256} \kappa R + \frac{150}{256} \kappa G + \frac{29}{256} \kappa B \neq Y \\
C'_r = \frac{131}{256} \kappa R - \frac{110}{256} \kappa G - \frac{21}{256} \kappa B + 128 \neq C_r \\
C'_b = -\frac{44}{256} \kappa R - \frac{87}{256} \kappa G + \frac{131}{256} \kappa B + 128 \neq C_b
$$

(3.17)

$Y'$, the intensity channel, will have an increased value as expected. Looking at $C'_r$ and $C'_b$, when the original RGB values are equal, the $C_r$ and $C_b$ values will remain the same after the light change. However, when the RGB values are equal to the surface is either grey white or blank and there is essentially no chrominance information to distinguish the object. The other combinations of RGB values, that are not equal, alter the $C_r$ and $C_b$ values. This indicates that even though there are separate chrominance and luminance channels in the YCrCb space, it is a misconception that the chrominance channels are invariant to light intensity. Work in computer vision, which uses $YCrCb$ as a basis, such as [69] which attempts to adapt YCrCb-based-models to lighting conditions, would
be better served using HSV or Normalised HSV, because they would actually factor out lighting and avoid the need for adapting the models.

### Experiment Setup

Experiments are now conducted to evaluate and confirm the earlier theory. The experiment aims at creating spectrally uniform light changes. Looking at the colour formation equations:

\[
R = f(L, S, P_R) \\
G = f(L, S, P_G) \\
B = f(L, S, P_B) 
\] (3.18)

These are identical to the intensity Equation 3.1, though with three different types of photo-receptors, \(P_R, P_G, P_B\), each have different sensitivities to different parts of the light wavelength spectrum. An ideal experimental setup is where the camera is stationary so that object reflectance, \(S\), and photoreceptor sensitivity, \(P\), are constant and the power of the light source is increased. This is possible, but it is difficult to verify that the light is increasing uniformly across the spectrum. It would also be possible to affect \(P\) through the camera exposure or through analog to digital gain. Here, a different approach is proposed to ensure the spectrum of the light is not altered as the light level reaching the camera is increased, by adjusting the aperture of the lens iris. Because the iris is just a physical opening, it affects only the quantity of light reaching the image sensor, the spectral make-up of the light source reaching the image sensor will remain constant.

The RGB intensities increase according to the following relationship with the iris diameter \(\phi\):

\[ I' \propto \phi^2 I \] (3.19)

\(\phi^2\) in the experiment, is the light level factor, \(\kappa\), used in the theory presented earlier.

A motorised lens is used to adjust the iris diameter, where a positive charge is applied to open the iris and a negative charge to close. It is difficult to estimate the actual aperture of the iris in diameter; however, it is sufficient for this experiment to simply ensure an increasing amount of light reaches the sensor. This is achieved by totally closing the iris and supplying a positive charge until the iris is entirely open. While the iris is being opened an image stream is recorded from the camera. The camera’s automatic exposure and color-balancing features are disabled to maintain a constant photoreceptor sensitivity.
A colour-chart is fixed in front of the stationary camera, see Figure 3.8. The chart has a variety of different patches including red, green, blue, yellow, cyan and magenta, representing the different sections of the visible light spectrum. The RGB values are sampled from a rectangle of pixels in each patch. The sampled RGB values are converted into the different chrominance spaces and plotted on graphs. The values of an ideal chrominance space will remain stable through the experiment.

**Experimental Results**

Chrominance values from the cyan patch chart are presented in Figure 3.9. For brevity, the values of other patches are not presented. The cyan patch was not chosen for any particular reason, it is representative of the behavior exhibited by the other patches.

The raw RGB values from the experiment are presented in Figure 3.9(a). As expected the RGB intensity values increase as the iris is opened. The non-linear shape of the graph is indicative of the $\phi^2$ iris diameter relationship.

The YCrCb values are presented in Figure 3.9(b) are coordinate pair with the white point at [0.5, 0.5]. It is surprising how sensitive YCrCb chrominance is to the light level given that it is such a commonly used chrominance space. This experiment affirms the earlier theory that YCrCb chrominance is not stable to light levels and that it is unreliable for computer vision purposes.

Normalized RGB chrominance is shown in Figure 3.9(c) and is also a coordinate pair but with a white point at [0.33, 0.33]. For the most part, Normalized RGB does exhibit stability to light increases,
3.3 Chrominance Spaces

![Graphs showing RGB, YCrCb, Normalized RGB, and HSV chrominance values against iris diameter.]

**Figure 3.9:** The chrominance values of various spaces as the iris is progressively opened. Sampled as the average of the pixels from the cyan patch of the chart seen in Figure 3.8. The relationship between iris diameter $\phi$ and intensity is defined in Equation 3.19.

which is in accordance with theoretic predictions. The periods of instability at the beginning and end of the experiment are caused by sensor issues which are discussed later.

On the HSV graph in Figure 3.9(d), hue is presented as a value from 0 to 1, this represents the hue angles from $0^\circ$ through to $360^\circ$. As the theory predicts, HSV also shows stability to light levels, with only short moments of instability that are discussed in the following sections.

The brief periods of instability in Normalized RGB and HSV values at the beginning and end of this experiment are due to two factors, sensor noise and sensor saturation.
**Figure 3.10:** Graphs illustrating the sensor related problems. 3.10(a) is the HSV and RGB values as the iris is progressively opened. The left vertical line indicates when the sensor noise stops dominating the signal. The right line indicates when the photo-receptors begin to saturate. 3.10(b) is the min, max and mean hue values from the cyan patch, illustrating problems with sensor noise.

**Figure 3.11:** 3.11(a) Problems with using different video encoding and decoding standards. Normalized RGB and saturation from HSV are no longer robust to light levels. However, hue from HSV does remain stable. The 3.11(b) Graph using images from a camera set with the brightness parameter set to 5% of the intensity range ($\omega$).
3.3 Chrominance Spaces

Sensor Saturation

Cameras have limited dynamic range and the output intensity values represent only a portion of the absolute intensity scale. Figure 3.10(a) shows that the hue values become unstable after the iris is opened too far, when the RGB photo-receptors begin to saturate. The point when the first channel saturates is indicated by the rightmost vertical line on the graph. When the final channel saturates, there is no longer any valid information coming from the photo-receptors and, therefore, the calculation of hue is essentially meaningless.

Sensor Noise

A CCD camera as used in these experiments produces measurements with associated electrical noise. The noise has a negative effect on the calculation of chrominance, illustrated in Figure 3.10(b). At the beginning, the range of hue values from the rectangular patch (the difference between the maximum and minimum values) is large because of a small signal-to-noise ratio. The point when the signal-to-noise ratio ceases to be a significant disruption is signified by the leftmost vertical line in Figure 3.10(a), after this point, the hue deviation across the patch is small.

Encoding/Decoding Inconsistency

As discussed in Section 3.3.3, there are a number of video encoding/decoding standards that vary from country to country and between analog and digital transmissions. If the RGB values recorded by the camera are encoded and transmitted in one format, and decoded using another format, the result will be slightly different from the original RGB values.

Figure 3.11(a) has an example of the instability in chrominance values caused by incorrectly decoded images. In this particular example the incorrect YCrCb-to-RGB decode transform from [70] is used, while the images were originally encoded from RGB-to-YCrCb according to the transform in [66]. While the decoded images appear correct to the human eye, the chrominance values can be disturbed by this issue. The Normalized RGB values, as seen in Figure 3.11(a), have lost the property of being stable to light level changes. The saturation channel from HSV, is also no longer stable to light changes, though hue from HSV is robust.
Inspecting the decoding transform in [70], the decoded values, $R'G'B'$, are:

$$
R' = 1.164(Y - 16) + 1.596(C_r - 128) \\
G' = 1.164(Y - 16) - 0.813(C_r - 128) - 0.391(C_b - 128) \\
B' = 1.164(Y - 16) + 2.018(C_b - 128)
$$

(3.20)

then inserting the encoding transform from [66] gives:

$$
R' = 1.164R - 19 \\
G' = 1.164G - 19 \\
B' = 1.164B - 19
$$

(3.21)

These inconsistencies between the raw RGB and the decoded $R'G'B'$ values cause the variations seen in the chrominance values in Figure 3.11(a). It is a simple mistake to use the incorrect video decoding format because of the variety of slightly different standards in use. In addition to ensuring the correct standards are used, it is prudent to use HSV hue, because it is not sensitive to this issue.

**Camera Brightness Parameter**

It is common for cameras to have three different methods for adjusting the output image, gain, gamma and brightness or $\nu$, $\gamma$ and $\omega$ respectively. These affect the intensity values as follows:

$$
I' = \nu I^\gamma + \omega
$$

(3.22)

$\nu$ is a linear factor that behaves the same as the light level changes that were discussed earlier. Gamma, $\gamma$, is a non-linear variable and Finlayson and Schaefer in [64] present a new HSV hue transform to negate its effect. Brightness, $\omega$, is an additive variable which can disturb chrominance transforms, as shown in Figure 3.11(b).

A theoretical spectrally-uniform light change, $(\kappa)$, with a fixed $\omega$, is applied to each RGB value. Then, the Normalised RGB transform (from Equation 3.14) becomes:

$$
N'_r = \frac{\kappa R' + \omega}{\kappa R + \kappa G + \kappa B + 3\omega} \neq N_r \\
N'_g = \frac{\kappa G' + \omega}{\kappa R + \kappa G + \kappa B + 3\omega} \neq N_g
$$

(3.23)
3.3 Chrominance Spaces

With the presence of a non-zero $\omega$, Normalised RGB will no longer be robust to light level changes, because $\kappa$ cannot be canceled out of Equation 3.23. This is consistent with Figure 3.11(b).

A light change and a fixed $\omega$ can be applied to the HSV transform from Equation 3.15, and this becomes:

$$H' = \frac{\kappa(\theta_1 + \omega) - \kappa(\theta_2 + \omega)}{\kappa M + \omega} = \frac{\theta_1 - \theta_2}{M - m} = H$$

$$S' = \frac{\kappa M + \omega - (nm + \omega)}{\kappa M + \omega} \neq S$$

(3.24)

where $\theta$ is taken to be any of the RGB values according to the HSV transform described in [68], $\omega$ is subtracted out of the $H'$ equation, meaning that hue is robust to camera brightness. However, because of the $\omega$ in the denominator of the ratio in the $S'$ equation, the $\kappa$ cannot be canceled out. This is consistent with the graph in Figure 3.11(b).

Normalized RGB and HSV saturation are robust only with multiplicative variables. The hue from $HSV$ is more robust than the other chrominance spaces, it is robust to both multiplicative and additive disruptions, because of its subtractive terms.

3.3.5 Chromatic Variations in the Lighting

The previous section studied the situation where a light source changed uniformly across all wavelengths. The hue channel from HSV was found to be the most robust transform to these uniform lighting changes. However, if the light source changes non-uniformly across the wavelength spectrum, HSV values, and all other chrominance transform values, will vary. Austin and Barnes [39] presented experiments showing instability in the HSV space with standard sunlight changes.

Cornsweet [52] explained that the human visual system can estimate chromatic light changes and accordingly adjust the perception of objects. Similar functionality has yet to be replicated in robot vision, and colour constancy is still an active area of research. Certain cameras have white-balancing functions, which can perform basic adjustments to correct for chromatic light changes. However, white-balancing features use a naive grey world assumption, which can be fooled when the field of view of the camera is dominated by an object of a certain chrominance. Chromatic changes in the light source are a significant problem for a chrominance-based visual localisation system.
3.3.6 Bland Environments

In addition to environments where there are chromatic light changes, chrominance is not a useful cue in environments which are bland. Sural et al. [71] define a threshold to differentiate whether the hue or the intensity of a pixel is more pertinent. The threshold is on the saturation value in the HSV space, where, if the saturation is below 20%, the pixel’s hue should not be considered. If the saturation is too low, chrominance is not a pertinent metric in distinguishing the bland surface.

Figure 3.12 presents an example of this threshold on the RGB images recorded in the two application environments investigated by the thesis. The pixels in the image that have saturation values below 20% are drawn as grey, to indicate they have no useful chrominance information. The building faces in the industrial environment are uniformly coloured and bland. The blue sky is the most colourful.
area and some bright yellow is detected on the vehicle. It is obvious that chrominance information would not be a useful cue to perform localisation in this environment.

In the underwater environment there are only a few randomly-scattered pixels which pass the hue saturation threshold. Again, this indicates that chrominance would not be useful in this underwater environment.

### 3.3.7 Summary of Chrominance Study

This study has compared the properties of different chrominance space with a theoretical and experimental study. The experiments were conducted in a variety of different scenarios and a number of key issues have been identified.

Results from experiments show that YCrCb is sensitive to light levels and is not suitable for vision systems operating in dynamic lighting conditions. Normalized RGB and HSV are more stable, though there were four issues raised by the results which cause instability, sensor saturation, sensor noise, inconsistent video encoding/decoding formats and also the camera brightness parameter.

 Unexpectedly, Normalized RGB and the saturation component of HSV were found to become variant with light levels when the camera’s brightness parameter was non-zero and also when there were discrepancies with encoding/decoding formats for video transmission. Hue from HSV performs most reliably in the situations tested during the experiments. It also has the most uniform and encompassing spread of the RGB values. Therefore HSV, especially the hue channel, is found to have properties most useful in dynamic lighting conditions.

Important aspects of using chrominance information that have been identified in this section can bring an improvement for future chrominance-based vision systems. However, for the bland application environments presented in this thesis, chrominance information was found not to be pertinent. Furthermore, there is no complete solution to the problem of chromatic variations in the light source.
3.4 Image Point Features

There are a group of related techniques that locate, describe and match points in an image. The process, in general, is to discover points of interest in an image, form a description of the neighbourhood of pixels around the points and then match points, at a later stage, by comparing descriptions. Many different variants of this process exist, such as the work of Harris and Stephens [72], Schmid and Mohr [73], Mikolajczyk and Schmid [74], Lowe [6] and Bay et al. [18]. This section reviews only Lowe’s SIFT transform [6] because it has become one of the most commonly-used techniques.

SIFT has become popular because, unlike many other techniques, it offers robustness to changes in viewing scale. Robustness to scale is achieved in both the keypoint selection process, using multiple image resolutions, and also in the neighbourhood description which uses histograms of orientations in the intensity gradient. The result is a technique that can match feature points through large changes in viewing angle and position. Results presented in [6] show the impressive ability of SIFT to match features through large changes in viewing pose.

The SIFT process, however, can be expected to vary from the typical lighting changes that occur outdoors. Consider that the keypoint selection process looks for points of high contrast. The moving sun causes shadows to move and, therefore, points of high contrast in the image also change. Thus, some variations in the keypoint selection process can be expected. Variations can also be expected from the SIFT descriptor which is formed from histograms of gradient orientation that are essentially measurements of relative-intensity. This chapter has already discussed that relative-intensity will be robust to changes of global light level, but non-uniform lighting changes will cause disruptions. However, it is not understood how significantly the SIFT process will be disrupted by non-uniform lighting changes and whether a robust localisation algorithm will be able to cope with the level of insensitivity.

Many authors have noticed instability in image point features. Lowe [6] found light disrupted his descriptor especially when the planarity assumption was violated. Sim and Dudek [5] discovered sensitivities of the features with light changes. Valgren and Lilienthal [7] find difficulties localising with different lighting conditions to the reference dataset, but produce better results in later work on the same datasets with improvements to their matching scheme [8]. The improvements are attributed to processing the images in full resolution, introducing an epipolar constraint and using RANSAC to remove outliers.

Robust techniques may remove outliers, but still require a minimum number of inliers which may not be present if the feature matching process is unstable to lighting conditions. Previous experiments
with robust algorithms generally only use a few distinct lighting conditions. The experiments of Mikolajczyk and Schmid [75] only adjust the camera aperture to test the lighting sensitivity, which is only a global adjustment to the light in the image. Valgren and Lilienthal [7] test six separate outdoor conditions; however, their tests look more at variations in snow cover and foliage changes than lighting. Only two of their tested conditions seem to have direct bright sunlight with shadows, the other conditions have similar soft uniform lighting conditions. Furthermore, the reference dataset is taken from soft uniform lighting conditions, which would be easier to easier to match. This chapter conducts a more detailed lighting change study, testing SIFT features with the full range of non-uniform lighting changes experienced outdoors over a day. Several different reference images are used to compare with all other images recorded on the day.

3.4.1 Dawn to Dusk Experiment

To ensure a rigorous test is conducted, four different stationary cameras were setup in four different locations. The cameras employed for the experiment were Canon VB-C50ir which are designed for outdoor usage. Images were recorded from just after sunrise to just before sunset on a clear sunny day, so that the full effect of the sun’s angle was present. A sample of images from the experiment is presented in Figure 3.13. An image from each camera was recorded every two minutes from sunrise at 6:25 am right through to dusk at 5:10 pm, producing a total of 300 images recorded from each camera.

The experimental work presented here was aimed at evaluating SIFT’s robustness to outdoor illumination changes. It was not looking at the performance of SIFT through changes in viewing pose and, therefore, the viewing pose of the camera was kept constant while the illumination naturally varied with the moving sun. This would determine if SIFT could reliably match features while there was only one variable, the lighting.

In this experiment the SIFT technique was processed using Matlab’s version of SIFT available freely online for evaluation. It was used in its original configuration, that is, no parameters were adjusted.
Figure 3.13: Left to right: Images from different locations around the site. The cameras remained stationary while capturing images throughout a day. Example images from different times of the day are top to bottom: 6:25 am, 9:00 am, noon, 2:30 pm, 5:10 pm.

3.4.2 Keypoint Selection

The first set of data collected from the experiment is simply the number of keypoints detected in each image. The count of keypoints will give an initial evaluation of the constancy of the keypoint selection process through lighting variations. Figure 3.14 presents a graph of the keypoint counts from the four different locations over the day.
The number of keypoints detected is not entirely stable at any of the camera locations; however, the variations are generally minor and there are several thousand keypoints found at almost all times. Towards the end of the day, there was approximately a 50% drop in the number of keypoints detected at locations B and D, this could only be explained by the lighting changes. There were still several hundred keypoints detected at this stage, which is sufficient for visual localisation if the keypoints can be correctly matched.

3.4.3 Number of Matches

A better indication of the robustness of the whole SIFT process is the number of feature matches. Five reference images are selected from each camera at 2.5 hour increments over the day. The reference images are shown in Figure 3.13. For the 300 images recorded from each camera, SIFT features are extracted and are matched to the reference images. The results of the number of detected feature matches from the five different reference images are presented in Figure 3.15. The number of features matched is presented as a percentage of the total features extracted from the corresponding reference image. The graph shows clearly that the most matches are recorded at the time when the reference image was recorded, which is predictable. What is surprising is that the number of matches drops so suddenly before and after the reference image time. For the majority of the day, the number of features matched with the reference image is less than 20 percent. This is consistent at all locations and all reference images. The reduced number of matches indicated that the SIFT process struggles to maintain matches through lighting changes. However, even if the low number of matches are correct matches, there still may be enough information to perform visual localisation. Therefore, it is required to find out if these are correct matches.
Figure 3.15: SIFT match counts taken over a day. Graphs of the four different locations each with five different reference images. The y-axis is the number of SIFT matches as a percentage of total number of keypoints found in the corresponding reference image.

3.4.4 Match Location Disparity

The camera in the experiment was stationary and therefore a simple discrepancy between the locations of two matched features indicates the validity of the match. The stationary cameras mean that the discrepancy in matched feature locations should be zero.

Figure 3.16 is a 2D histogram on location disparity of matched SIFT features. The histogram shows that most SIFT matches are recorded within 10 pixels of each other, giving an indication of the accuracy of the keypoint selection process. However, this is a stationary test and does not necessarily translate to the same accuracy with viewing pose changes. Furthermore, although most matches are with 1 or 2 pixels discrepancy, there are matches recorded with a greater discrepancy, which would translate to a significant error when triangulating into geometric positions.
3.4 Image Point Features

2D histogram on disparity in matched SIFT locations

Figure 3.16: 2D histogram on location discrepancy between matched SIFT features. The histogram is taken from all the SIFT matches recorded over the day long experiment.

The histogram also provides a means to differentiate correct matches from incorrect matches. A threshold on disparity can be set, and all matches with disparity within the threshold can be considered correct matches and all matches with a disparity greater than 10 pixels are deemed incorrect. A threshold of 10 pixels is chosen from the graph in Figure 3.16. This threshold will detect no false negatives, because, with a stationary camera, all correct matches must have a disparity close to zero. There may be a small fraction of false positives, because incorrect matches will have arbitrary disparity and by chance might be within the 10 pixel radius.

3.4.5 Incorrect Matches

It is now possible to determine which matched features are correct using the 10 pixel disparity threshold, chosen in the previous section. The graphs in Figure 3.17 show the number of incorrect SIFT matches recorded at each time of the day when matched against the five different reference images. The number of incorrect matches is presented as a percentage of the total number of matches.
Figure 3.17: Incorrect SIFT matches over a day. Five reference images are compared at four different locations. An incorrect match is defined as a greater than 10 pixel discrepancy in the two keypoint locations. The y-axis is the number of incorrect matches as a percentage of the total number of matches.
3.4 Image Point Features

(a) Location A
(b) Location B

Figure 3.18: Example image pairs with matched features shown with lines. Green lines are correct matches and magenta lines are incorrect. Left images are all from 9:00 and the right are from different times of the day, from top to bottom: 6:25 am, 9:00 am, noon, 2:30 pm, 5:10 pm. Image pairs in (a) are from location A and (b) are from location B.
Figure 3.19: Example image pairs with matched features shown with lines. Green lines are correct matches and magenta lines are incorrect. Left images are all from 9:00 and the right are from different times of the day, from top to bottom: 6:25 am, 9:00 am, noon, 2:30 pm, 5:10 pm. Image pairs in (b) are from location C and (b) are from location D.
3.4 Image Point Features

Table 3.1: Aggregated SIFT results. Columns left to right: mean outlier rate, the mean number of inliers, the percentage of images where there are more outliers than inliers and the percentage of images with less than eight inliers.

(a) Aggregated by location

<table>
<thead>
<tr>
<th>Location</th>
<th>Error</th>
<th>Mean inliers</th>
<th>Majority outliers</th>
<th>&lt;8 inliers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location A</td>
<td>11.21%</td>
<td>134</td>
<td>5.11%</td>
<td>14.55%</td>
</tr>
<tr>
<td>Location B</td>
<td>13.06%</td>
<td>144</td>
<td>10.87%</td>
<td>22.47%</td>
</tr>
<tr>
<td>Location C</td>
<td>25.54%</td>
<td>71</td>
<td>16.43%</td>
<td>30.06%</td>
</tr>
<tr>
<td>Location D</td>
<td>22.41%</td>
<td>107</td>
<td>16.12%</td>
<td>31.04%</td>
</tr>
<tr>
<td>Overall</td>
<td>18.14%</td>
<td>114</td>
<td>12.15%</td>
<td>24.56%</td>
</tr>
</tbody>
</table>

(b) Aggregated by reference image

<table>
<thead>
<tr>
<th>Ref. time</th>
<th>Error</th>
<th>Mean inliers</th>
<th>Majority outliers</th>
<th>&lt;8 inliers</th>
</tr>
</thead>
<tbody>
<tr>
<td>06:25</td>
<td>10.64%</td>
<td>86</td>
<td>4.07%</td>
<td>16.27%</td>
</tr>
<tr>
<td>09:00</td>
<td>17.16%</td>
<td>137</td>
<td>11.65%</td>
<td>28.25%</td>
</tr>
<tr>
<td>Noon</td>
<td>18.26%</td>
<td>133</td>
<td>13.28%</td>
<td>22.26%</td>
</tr>
<tr>
<td>14:30</td>
<td>23.42%</td>
<td>123</td>
<td>18.71%</td>
<td>24.21%</td>
</tr>
<tr>
<td>17:10</td>
<td>21.57%</td>
<td>88</td>
<td>13.14%</td>
<td>32.62%</td>
</tr>
<tr>
<td>Overall</td>
<td>18.15%</td>
<td>114</td>
<td>12.15%</td>
<td>24.56%</td>
</tr>
</tbody>
</table>

The graphs show that there are minimal incorrect matches around the time of the reference image, which is as expected. However there is an alarming reduction of correct matches at other times, and an alarming increase in the number of incorrect matches.

Table 3.1(a) shows the errors recorded at each location. The results are presented in four different columns, the average error rate, the average number of inliers recorded in each image, the percentage of images where there are more outliers than inliers and the percentage of images with less than eight correct matches. If only the first two columns are considered, it would appear that these figures suggest suitable results on which to base a localisation system. On average, there are 70% inliers in the matches recorded in every image, and the average number of inliers is sufficient to localise a camera. However, the final two result columns reveal that these first two columns are skewed by the periods of good results around the reference image time. In actual fact, for 10% of the time, there are more outliers recorded than inliers, which, perhaps, is not a large issue if robust constraints or a RANSAC algorithm are employed. However, the final column indicates that there are often not enough inliers recorded to use robust localisation algorithms. In general, eight correct feature matches are required to estimate the relative camera pose between two different images [76, 77], a necessary task in geometric SLAM systems such as [3]. The experiments reveal that a map of SIFT features will not be useful for geometric localisation for 24% of the time in the average case across the four locations and five reference times.
Locations C and D have more incorrect matches in comparison to the locations A and B. Potentially, the better matching performance at locations A and B is due to more of the tree-line silhouette on the horizon being visible, which will not be affected by moving shadows. Figure 3.18 illustrates this, where it is clear that more correct matches are recorded on the tree-line in the distance with only a few features matched on the buildings. This is not ideal for visual localisation, because distant features are only suitable for accurately estimating self-orientation, not self-position.

Table 3.1(b) shows the errors recorded with each reference image time. The 6:25 am reference image time produces the least errors, which is probably due to the mild nature of the dawn light, giving ideal conditions to form feature descriptions.

3.4.6 Summary of Image Feature Study

This section has presented a study of the SIFT transform, a predominant image feature matching technique. Experiments were presented over the duration of a clear sunny day when the changing angle of the sun in the sky produce lighting variations. This provided a good test of SIFT’s stability in outdoor conditions.

The results show some instability to lighting in the keypoint selection process, but, more importantly, were the instabilities discovered in the feature matching process. The SIFT matching process not only had trouble maintaining enough correct matches through lighting changes, but also produced significant levels of false matches. The performance was better at some locations, which is thought to be because of the tree-line on the horizon being visible. The performance was also better when using an early morning reference image, which is probably because of the soft nature of the dawn lighting. However, the general consensus of the results is that SIFT features cannot be reliably matched through lighting changes. Although the experiment did test some difficult lighting conditions, other aspects of the experiment provided favourable conditions for feature matching. A stationary camera was used, so changing the viewing pose may have further disrupt the results. The cameras were setup with a large portion of the environment in view, including trees and other clutter, meaning that there was much for the SIFT process to describe. A moving vehicle would often be in positions where there is far less within the field of view, meaning far less visual features to describe.

The conclusions drawn from this study are that a system attempting to geometrically localise based on SIFT feature maps recorded under a varying lighting conditions will be susceptible to failures.
3.5 Chapter Summary

This chapter investigated different classes of visual information giving quantitative results from experiments through significant lighting changes.

The first study was conducted on the difficulties of recognising object reflectance from intensity values in fluctuating lighting conditions. The study was framed in the context of foreground segmentation and reiterated the fact that there is no direct relationship between single image intensity values and object reflectance. Relative-intensity values are stable to uniform lighting variations and can reliably detect the edge of objects; however, non-uniform light changes will cause disruptions. A foreground segmentation method was designed to only detect objects that have a history of motion and, therefore, the method is robust to appearing static shadows.

The next study was on chrominance information. Various chrominance transforms were studied and the hue channel from HSV was found to be the most stable to achromatic light changes, though spectral changes in the light source cannot be factored-out. Importantly, chrominance information was also deemed to be not pertinent in the dull application environments studied in this thesis.

Finally, a study on the stability of image-point features during outdoor lighting changes raised doubts on whether maps created from image-point features can be reliably used for localisation after significant changes in the lighting.
The previous chapter evaluated the second phase of the visual process; the extraction of low-level visual information from an image. The first phase of the visual process is the image acquisition by an image sensor. The image sensor of a robot vision system is a camera, which is limited in the range of lighting it can perceive correctly; which is a problem for operation in highly-non-uniform lighting.

The human eye can capture approximately five orders of magnitude of dynamic range in radiance [78]. Conventional cameras (both digital and analog), by comparison, have dynamic ranges of approximately three orders of magnitude. As a result of the limited dynamic range of conventional cameras, information is lost at the sensor level in non-uniformly lit scenes. For example, the scene shown in Figure 4.1 is made up of some areas in direct sunlight and some lit by standard artificial lights. It is clear that information in this scene lies outside the range of a conventional camera. This causes a problem for robotic vision systems operating in wide radiance range scenes because they simply cannot see parts of the environment. An initial study of this thesis is to investigate methods for
This chapter presents a technique for an extended dynamic range vision system, that was published in [56]. It is designed for a conventional, unmodified digital camera, thus contrasts to the techniques reviewed in Section 4.1.1 that require complex optics/hardware modifications [79, 80] or result in reduced spatial resolution [81]. The method is a sequential exposure changing technique and is distinct from the other sequential exposure changing techniques, such as [82], because it allows operations in a non-static, wide radiance range scene in real-time. Kang et al. [83] have presented a method of performing real-time sequential exposure changing, using a camera with modified firmware and with the off-line combination of the sequence. The technique developed in this project performs operating image sensors in highly non-uniform lighting conditions and how to improve the low-level visual information available to the higher levels of a robotic vision system.

Figure 4.1: Example showing the limitations of conventional cameras. Top: Image taken with a regular exposure; areas outside are completely over-exposed. Bottom: Image of the same scene taken with a short exposure; areas inside are completely under-exposed.
4.1 Existing HDR Technologies

This section reviews the existing high dynamic range (HDR) technologies. The review first covers the different sensor technologies designed to create high dynamic range images.

With the emergence of technologies that can create HDR images, there also comes a problem of what to do with these HDR images. In many situations, high-bit-depth images need to be compressed into a more compact form; there is a growing area of research investigating this challenge. The second part of the chapter reviews this research in HDR reduction technologies.

4.1.1 HDR Sensor Technologies

The high radiance range problem is well documented, with much literature on the attempts to increase the dynamic range of sensors. This section will review the existing HDR techniques.

*Sequential Exposure Changing* – sequentially grabbing images of a scene on a single camera, changing the exposure settings (gain, shutter time, aperture) between captures to produce multiple images that can then be merged into a single HDR image [82].

*Multiple Sensor Arrays* – a technique detailed in [80] employs complex optics to split the light beam onto multiple sensors. Each detector has a different sensitivity and will produce different images of the same scene. The benefit of this method is that it is not temporal; each detector can capture images simultaneously and thus not being affected by the problem of dynamic scenes. To accurately replicate the light onto multiple sensors requires precision optics.

*Multiple Sensor Elements Per Pixel* – this technique uses one sensor array that has multiple sensor elements for each pixel [81]. Neighboring sensor elements with different, fixed sensitivities, are combined into a HDR representation. This method captures measurements simultaneously and is suitable for dynamic scenes avoiding the need for complex optics. However, it compromises the resolution of the outgoing image because multiple sensor elements are combined to form a single HDR pixel.
*Sequential Illumination Changing* – this is a method proposed in [84] for grabbing multiple images with the same exposure settings but with different illumination conditions. It is performed by using a camera flash with different radiant intensities to obtain images of the same scene with different lighting conditions. This method requires that the scene remains static while the images are being captured. Also, the camera flash has to be of significant radiance with respect to the natural lighting conditions of the scene being captured; thus, the differences in the flash will appear as significant differences in the captured images.

*Adaptive Dynamic Range Pixels* – the technique detailed in [79] shows that the dynamic range of each pixel can be increased by individually controlling the amount of light reaching each sensor element. A liquid crystal light modulator that has a separate attenuator element for each sensor element is placed in front of the sensor. The modulator works by attenuating the light reaching saturated sensor elements. Although the each element provides conventional low-bit depth readings, the modulator allows HDR scenes to be recorded.

### 4.1.2 HDR Image Reduction Techniques

HDR images cannot be displayed on conventional display media, such as monitors, printers or photographs, because the images span a far greater range of intensities than that can be reproduced (except for recent developments in HDR display technology [85, 78]). Furthermore computer vision systems when presented with high detailed and high contrasting images cannot efficiently and effectively process the floating point HDR information. Thus in many situations HDR images need to be compressed into a low-bit-depth form. This concept does appear bizarre at first, for why spend all this effort to produce a HDR image, only to throw all the information away? HDR imagers give us more information about a scene than is generally required, for it is unlikely that natural brains encode perceived sensations with more than 6 bits of dynamic range [86].

HDR sensors are so important because they recover information about the scene that cannot be recorded with conventional sensors, but there is still the problem of preserving all necessary information from a HDR image in a low-bit-depth representation. This section details the existing HDR reduction methods.

### 4.1.3 Tone Mapping

A common technique of dynamic range reduction is tone mapping (commonly referred to as tone reproduction curves), where incoming HDR data are directly mapped to a low-bit-depth. This is a
4.1 Existing HDR Technologies

spatially invariant, global method, that is generally efficient and avoids the halo (reverse gradient) problem. However, it can suffer from a loss of local detail.

Linear: a single linear mapping applied globally to a HDR image cannot preserve contrast or leave an accurate impression of brightness [87] and, hence, cannot comply with either of the criteria stated by [88].

Non-Linear: more sophisticated functions than a simple linear mapping have been developed. One popular method is first, to take the logarithm of the incoming HDR data before applying a linear tone map. These types of global tone curve functions can provide more realistic low-bit-depth images [89], but user interaction is generally needed to tweak the parameters. Moreover global mappings have difficulties preserving local contrasts where the intensities of objects in the image span large areas of the dynamic range [90].

4.1.4 Tone Reproduction Operators

Tone curves can provide suitable results for photography, because user interaction is acceptable; however, other applications need to perform high dynamic range reduction automatically and also retain local detail. Tone operators, which have spatial context, attempt to preserve local detail but can lead to halo artifacts, caused by reverse gradients. The concept of local tone reproduction operators is an attempt to match the human visual system by responding to local intensity changes rather than absolute global intensity levels [91].

Gradient: by measuring the intensity gradients, tone reproduction can be adapted in a local manner to preserve detail. Many operators work at multiple resolutions for the purpose of represent gradients that occur at different scales. In [92], Pattanaik explores how the human visual system responds to light intensity in a spatial context and this was used as the basis for the multi-scale tone reduction operator.

As stated in [90], multi-scale gradient operators will typically produce halo artifacts if the different resolution levels are manipulated separately. Thus, a method is proposed in [90] to avoid this where a gradient map is created from a HDR image that, in turn, is attenuated and reversed back into intensity space in a low-bit-depth form. A HDR scene gives rise to significant large-scale intensity gradients, whereas fine detail and texture corresponds to gradients of much smaller magnitude. The method is then to attenuate significant large-scale gradients whilst retaining smaller fine-scale gradients. As a result, drastic changes are compressed whilst preserving fine detail.
Segmentation: a different method of local adaptation shown in [93] is developed to avoid halos. First regions are segmented based on luminance levels and then tone mapping is applied individually to each segment.

### 4.1.5 Summary of HDR Technologies

The HDR sensor techniques reviewed in the first section of this chapter all had pitfalls. They either have reduced spatial resolution [81], require complex optics/hardware [79, 80] or are unable to cope with motion [82]. There is work on true real-time HDR sensors [94] and in the future these may become more widely available.

The literature published on HDR image reduction presented in the second section of this chapter is generally aimed at display purposes. Either designed for photography or for conventional monitors, with the principal criteria of making the image \textit{appear} real. The techniques are designed to work off-line with the process generally taking several seconds to compute, some methods even need user parameter tweaking to achieve acceptable results. These methods are not designed for robotics which requires real-time functioning and are more interested in information content rather than how real the end result appears.

### 4.2 Introduction to the System

The system captures a sequence of three images. Each image is exposed differently according to exposure control algorithms that aim to maximise the information coming from the images. The sequence of images is taken on a moving robot and, therefore, spatial discontinuity in the image sequence needs to be accounted for before the sequence can be merged. Registering images with different exposures is difficult because they hold different information. Kang et al. [83] pixel boost images of short exposures to match image intensities of the reference image before registration. Conversely, the technique developed in this project registers the inter-sequence error and then interpolates the error within the sequence. Real-time image registration is performed by conducting a coarse-to-fine registration process and then performing fine-resolution registration only at points of interest.

After registration, the next stage of the system is to merge all necessary information from the sequence. In [82], high-bit-depth images are constructed from exposure sequences however, high-bit-depth images contain too much information to be used effectively and often bit-depth compression
is required. Previous work published on high-bit-depth image compression was reviewed in Section 4.1.2. Most work is generally aimed at display purposes, either for photography or for conventional monitors, with the principal criteria of making the image appear real. The techniques are designed to work off-line with the process generally taking several seconds to compute; some methods require manual parameter tweaking to achieve acceptable results, thus making them unsuitable for real-time robotics.

The technique developed in this chapter applies a sequential exposure technique to robotics. The sequence is merged into a representation not designed for human viewing, but in a form suitable for robotic vision processing, with edge and chrominance information being retained, as opposed to absolute RGB intensity values. The key challenge is to develop algorithms that achieve this functionality on a conventional, unmodified digital video camera in real-time.

### 4.3 Real-time Exposure Changing

The exposure level of the camera is changed between frames to form a sequence of differently exposed images. This section explains how this is performed without hardware modification. To achieve real-time operation, the number of exposures in the sequence is limited. This restricts the range of radiance covered by the sequence and introduces the need for individual exposure control of each image to maximize the information across the sequence.

Kang et al. [83] show a technique for modifying a camera’s firmware to perform real-time exposure changing on moving scenes. This chapter shows that the exposure changing can be performed without camera modification, with any IIDC [95] compliant IEEE1394 digital camera. IIDC compliant digital cameras have the advantage of enabling both video data transmission and the control of the camera over a single connection. Exposure settings can be controlled directly from a PC, see Figure 4.4. However, if the camera is placed in the asynchronous (continuous shot) mode from the IIDC camera specification, it is unknown exactly when, in the capture-transmit cycle, the new exposure setting command will be received by the camera. This is because the camera’s capture-transmit cycle is decoupled from the PC control software. To avoid this problem the alternative synchronous (one shot) capture mode can be used. The PC sends an exposure setting (analog-to-digital gain and exposure time) to the camera before sending a one shot command, which ensures the next image received is captured with that exposure setting. The PC waits for the image to be transmitted, before sending the next exposure setting followed by another one shot command, see Figure 4.2. The disadvantage of the one shot mode is that the maximum frame rate of the system is reduced, because the camera cannot capture a frame while the previous frame’s data is being transmitted.
Figure 4.2: Timing diagrams for the IIDC digital camera specification [95]. To perform sequential exposure changing the PC needs to know when the camera is capturing images. Left: Asynchronous mode. The camera’s capture-transmit cycle is decoupled from the PC; therefore, an individual exposure level for each image cannot be reliably set. Right: Synchronous one shot mode. In this mode, an exposure setting can be defined for each image.

4.4 Automatic Multiple Exposure Control

Using the sequential exposure changing process, any number of different exposures can be employed in the sequence; however, the more exposures that are used, the lower the cycle rate of the system. Three different exposures (low, regular and high exposures, see Figure 4.3) were deemed appropriate to span a wide enough range of radiance, whilst enabling a sufficient cycle rate. This section describes the control algorithms for the exposure levels.

Generic automatic exposure control (AEC) algorithms work through a grey-world assumption that the surface reflectance of the objects in the field of view will always average out to a constant value. The AEC algorithm will adjust the exposure to return the average image intensity to the desired value, which cancels the effects of uniform lighting changes. This generic algorithm fails in non-uniform lighting conditions, an example being when the field of view is a combination of a bright outdoor environment and a darker indoor environment.

To maximize the correctly exposed areas of a non-uniformly lit scene, the exposure level of each image in the sequence must be controlled. For the regular exposure, a generic AEC algorithm is
4.4 Automatic Multiple Exposure Control

Figure 4.3: Example illustrating natural radiance levels compared with the simultaneous radiance range of the human eye and that of a camera. Also shows the extended range of the sequential exposure technique. Rough estimates of range taken from [96, 78].

employed in an attempt to keep the mean intensity of the whole image at 50%. It would then be possible to fix the low and high exposures to a predefined offset either side of the regular exposure level, which would give a fixed simultaneous dynamic range. However, to cover the full range of radiance in a real world scene, more than three exposure levels are needed, see Figure 4.3. To maximize the viewable area across the three exposures and to account for non-uniform lighting changes, individual (de-coupled) control strategies for the low and high exposures are employed, aimed at viewing the under-exposed and over-exposed areas of the regular image. The exposure level is set as follows:

$$\xi_t = \xi_{t-1} + (\xi_{t-1} \times \rho \times e) \quad (4.1)$$

where $\rho$ is a tuning constant and $e$ is the error calculated for each of the three exposures, see Equations 4.2, 4.3 and 4.4. The IEEE1394 IIIDC [95] digital camera used in experimentation has two exposure parameters available, an analog-to-digital gain and a shutter time. There is no mention in the IIIDC standard of how a camera must implement the exposure settings. To make this technique portable between different cameras these two parameters are scaled between 0 and 1, then combined as a single parameter $\xi_t$ at time $t$. The scaling can be calculated at run-time by accessing the minimum and maximum exposure values in the inquiry registers of the camera. Note in Equation 4.1 that the exposure level adjustments are proportional to the current exposure level.

The control equation Equation 4.1 is essentially the same for the three exposures; however, the respective error measurement is different for each exposure. As stated earlier, the regular exposure mean intensity of the whole image is kept at 50%, so the error for the regular exposure control is:

$$e_r = \frac{\sum_{p=0}^{n} I(p)}{n} - I_d \quad (4.2)$$
where \( n \) is the number of pixels in the image, \( I(p) \) is the intensity of pixel \( p \) in the regular exposure and \( I_d \) is the desired intensity which is set to 50% of an 8-bit intensity image.

The error for the \textit{high} exposure, \( e_h \), is the mean intensity of the \textit{darkest} third of the image, defined as follows:

\[
e_h = \frac{\sum_{k=0}^{D} \text{hist}_h(k) \times k}{n/3} - I_d
\]  

(4.3)

where \( D \) is the intensity value which is brighter than the darkest third of the pixels in the \textit{high} exposure and where \( \text{hist}_h(k) \) is the number of pixels in the \textit{high} exposure that have intensity value \( k \).

The error for the \textit{low} exposure, \( e_l \), is the mean intensity of the \textit{brightest} third of the image, defined as follows:

\[
e_l = \frac{\sum_{k=B}^{255} \text{hist}_l(k) \times k}{n/3} - I_d
\]  

(4.4)

Where \( B \) is the intensity value which is darker than the brightest third of the pixels in the \textit{low} exposure and where \( \text{hist}_l(k) \) returns the number of pixels in the \textit{low} exposure that have intensity value \( k \).

### 4.5 Iris Control

Altering the aperture of the iris improves the range covered by the exposure sequence. A motorized-iris lens is controlled through the PC’s parallel port, see Figure 4.4. The iris is opened and closed; the objective being to keep the \textit{regular} exposure level in the middle of the exposure setting range. Note that iris control does not directly increase the simultaneous dynamic range of the camera, as iris adjustments occur over longer periods of time. However, it does allow a greater simultaneous dynamic range by enabling the exposure settings of the sequence to be more spread out.
Figure 4.4: Diagram of the overall system. Showing automatic control of the three sequential exposures and the lens aperture followed by image registration and merging.
4.6 Fusion of Edge and Chrominance Information

It is possible to create an absolute radiance map with a large bit-depth, as shown in [82]. High-bit-depth image construction has been generally conducted in an off-line process, and the constructed images contain too much information to be efficiently and effectively processed by robotic vision systems, hence the need for compression. Section 4.1.2 discusses the various approaches of high-bit-depth image compression. The reduction techniques are designed to work off-line and some methods even need parameter tweaking to achieve acceptable results. It is accepted that the human eye is far more sensitive to local intensity changes than absolute intensity levels [91]. Therefore, instead of working with absolute intensity levels, this system extracts and merges chrominance and relative-intensity information from each of the RGB images.
4.6 Fusion of Edge and Chrominance Information

Figure 4.6: Left: Three example RGB images from the exposure sequence. Middle: Edge images of exposure sequence, with chrominance overlaid. Darker regions correspond to larger edges. Top-Bottom: low exposure, regular exposure, high exposure. Right: Merged image from the sequence.

4.6.1 Multi-Scale Edge Representation

A standard multi-scale edge operator is employed to preserve visual information across the sequence of exposures. The operator works on multiple scales (similar to the method in [90]) to handle image noise and to emphasize important information. A pyramid of sub-resolution images is created by starting at the bottom with the original image and creating lower-resolution images with the mean intensity of the corresponding pixels from the higher-resolution image. Then the pyramid of images is converted into a pyramid of edge maps using a $3 \times 3$ edge-magnitude operator which takes the mean
difference in intensity between the center pixel and the eight neighboring pixels. This operation is performed by starting at the top of the pyramid and working down by taking half the edge magnitude at the current level and half the corresponding edge from the level above. The resulting edge image at the bottom of the pyramid has a softer edge for fine texture/detail and strong edges on the boundaries that occur at larger scales.

Now that a sequence of edge images is available, each with differing information, the missing details in the regular image, caused by over-exposure or under-exposure, can be recovered in the low and high images, as seen in Figures 4.5 and 4.6. Because the edge images are simply local intensity differences and are independent of the radiance ranges, corresponding edge values between the images can be directly combined without absolute radiance map construction. To merge the information, at each point the maximum edge magnitude from the sequence is taken with the concept that a stronger edge has more information, and more information infers more reliable information.

### 4.6.2 Chrominance Fusion

In addition to the fusion of edge information, chrominance information from the sequence is also combined. The concept that more information infers better information is also applied to chrominance where more information is a higher saturation of chrominance. This can be calculated using the HSV color space, which is comprised of three channels of information, hue, saturation and intensity. To do this, each image in the sequence is transformed from the native RGB space into the HSV space. At each point, the hue and saturation is taken from the pixel with the highest saturation.

### 4.7 Image Registration

There will be spatial discontinuity between images when there is motion between the camera and the scene which results in artifacts when merging. This system has been tested with a forward-facing camera on a ground robot, and artifacts do appear with the robot turning. The image motion is predominantly through horizontal translation. A real-time image registration process is presented that finds the most likely translation discrepancies between images.

Kang et al. [83] present an off-line technique for boosting images of short exposures to match the intensities of the reference image before pair-wise registering. The technique requires off-line processing and a larger set of images with relatively similar exposures. Pixel boosting will not work with the three images, because they have significantly different exposure levels and thus significantly
4.7 Image Registration

Figure 4.7: Top: Merged sequence of exposures without registration. Notice the artifacts at the fire hose (top-left), traffic cones (center) and drawers (right). Bottom: Merged sequence of exposures with automatic real-time registration and interpolation.

different visual information that cannot be registered with each other. A new technique has been developed to avoid this problem by registering two consecutive sequences of exposures and then interpolating the error to images within the sequence.

The method of Kang et al. [83] takes a total of 32 seconds to compute. The system presented in this chapter is concerned with real-time operation and, therefore, a more efficient technique is developed that only takes into account image motion. It must take into account the full effects of 3D camera motions along with the 3D scene structure properly remove all registration artifacts. The simple method presented here does not take all these aspects into account, but can be computed in real-time and does improve the image motion problem.

To enable real-time performance, a coarse to fine registration refinement process is used similar to the method in [97]. The edge image from consecutive images of the regular exposure are first registered in the coarsest resolution. As stated in [97], coarse registration is beneficial not only for computational efficiency but also to avoid fine scale aliasing problems. The registration is an energy minimization
process for finding the motion \( M \) which produces the most correlated gradients, \( G([i, j]) \). A set of motions \( M \) is searched that constitute the range of expected motions defined by the characteristics of the platform at hand. Each motion \( M \) produces a 2D pixel motion \([u, v]\) for each location in the image \([i, j]\);

\[
[u, v] = M([i, j])
\]  
(4.5)

The energy minimization equation to find the motion at time \( t \) is as follows;

\[
M_t = \arg\min_{(M)} \sum_{i=0}^{w} \sum_{j=0}^{h} \frac{|G_{t-1}([i, j]) - G_t([i, j] + M([i, j]))|}{w \times h}
\]  
(4.6)

where \( w \) and \( h \) are the width and height of the image respectively.

Points of interest, \( P \), which are points with the strongest gradient magnitude, are found at the coarse resolution, that further accelerate the fine registration process. Registration at the fine-resolution is performed only at the points of interest using the estimated coarse motion to refine/restrict the motion set to \( M' \);

\[
M_t = \arg\min_{(M')} \sum_{P} \frac{|G_{t-1}([i, j]) - G_t([i, j] + M([i, j]))|}{N}
\]  
(4.7)

where \( N \) is the number of interest points in \( P \).

The motion between the other images in the sequence is calculated by interpolating the inter-sequence translation. By assuming constant velocity, the interpolation is linear based upon the timestamps of each image. Once the motion is found the images can be corrected to the first image in the sequence. An example of the registration results is shown in Figure 4.7.

## 4.8 Results and Discussion

To demonstrate the system, a forward-facing camera was mounted on an experimental four-wheeled robot and driven from an indoor environment lit by artificial light to a bright sunny outdoor environment (The video attachment named regular.mpg shows the limitations of using a single exposure in this situation and the videos named merge-not-registered.mpg and merge-registered.mpg show the output of the proposed system with-, and without-, image registration).
4.8 Results and Discussion

4.8.1 Multiple Exposure Control Results

Chrominance and edge images from the experiment can be seen in Figures 4.5 and 4.6 that clearly show the information gained from areas both indoors and outdoors. Figure 4.8 illustrates the problem of using a single exposure with a generic exposure control algorithm. At times, the correctly exposed area is limited to around 20%, where a properly exposed pixel for an 8-bit image is deemed to have an intensity value above decimal 50 and below 255. Figure 4.9 shows the increased correctly exposed area provided by the proposed system. Figure 4.9 also shows that all of the information about a wide radiance range scene cannot be captured in a sequence of three exposures. There is missing information outside the range of the three exposures and intra-range information missing, because the radiance ranges do not always overlap (caused by the decoupled control algorithm, see section 4.4).

![Generic AEC Algorithm Graph](image)

**Figure 4.8:** Graph showing the limitations of a single exposure with a generic exposure control algorithm. At times, there is only 20% correctly exposed area in the image. Graph of: over, under and properly exposed areas of the regular exposure.

4.8.2 Image Registration Results

The spatial error between images depends on both the motion between the camera and scene, and the frame-rate of the camera. To achieve suitable results, the camera was configured to capture 500 × 400 pixels images enabling frame-rates in excess of 30 Hz. The exposure time also affects the frame-rate; upper limits on exposure time were calculated so that in the worst case scenario, frame-rates would drop to 25 Hz. If dark areas need to be viewed, the limits could be extended but at the expense of the frame-rate. Similarly, the resolution could also be increased.
Figure 4.9: Graph showing the information gained from the low and high exposures. Also shows that not all information can be captured with only three exposures. Graph of the combined correctly exposed areas, over-exposed area in the low image, under-exposed area in the high image, and intra-range information missed because intensity ranges of the sequence do not always overlap.

The robot turned corners during the maneuver, which created significant motion between an evaluation of the image registration. Ground-truth image motion was recorded by manually aligning images recorded during the experiment. The automatic image registration output was compared with this ground-truth motion to give an error measurement. From Figures 4.10 and 4.7, it is clear that the image registration technique copes with translational image motion with a reasonable degree of accuracy. Figure 4.10 also shows that the registration errors are worse towards the end of the experiment. It is during this period that the robot is traveling through the doorway, creating a looming effect, which translational correction cannot resolve. Camera pitch and roll are also not accounted for; however, the effects of these motions are minor with a four wheeled robot on flat ground.

Linear image motion is assumed for the interpolation of translational error. If there are significant accelerations in image motion, the interpolation will be incorrect. However, if the cycle rate is kept high, the image motion is low.

When operating the system on a 3.6 GHz Pentium 4 machine, image acquisition, sequential exposure changing, exposure control and gradient map construction, are successfully operated at 30 Hz. Image
registration and interpolation with edge and color merging of the three image sequence is processed at 10 Hz.

4.9 Chapter Summary

This chapter presented a system that improves robotic vision operations in highly non-uniform lighting conditions. When operating in non-uniform scenes, such as those comprising of indoor and sunny outdoor areas, much information is lost at the sensor level because of the limited dynamic range of conventional cameras.

The work in this chapter was published in [56]. The system enables conventional digital cameras to be operated in wide radiance range scenes. The approach is to extend the effective simultaneous dynamic range of a camera, by changing the exposure level of the camera in real-time, to form a sequence of images that together cover a wide range of radiance. The individual control algorithms for each image in the sequence provide the maximal viewable range across the sequence.

Visual information was preserved in a suitable format without having to going through the process of high-bit-depth image compression. The sequence was combined by merging color and edge
information, which avoids the time expensive operations of constructing absolute radiance maps and image reduction, whilst retaining information useful for a robotic vision system. Spatial discrepancies between images were improved with a real-time image registration method. All these techniques were integrated into one system showing that it is possible for a robotic vision system to be operated in wide radiance range scenes.
The previous two chapters looked at the early phases of a visual system; the image acquisition and the extraction of low-level visual information. This chapter looks at taking the extracted visual information and localising the camera via a comparison with a map. The specific application this chapter focuses on is a large forklift type robotic vehicle, as seen in Figure 5.1.

The vehicle’s application dictates that the localisation system must operate reliably over long periods. This is a challenge for a vision-based solution because of the extremities of the outdoor lighting conditions. It is not certain that existing visual localisation techniques can provide the required dependability in outdoor environments. Image-point features are probably the most popular basis for modern visual localisation research, but systems using these features have not been proven over the wide range of lighting conditions experienced outdoors. Section 3.4 presented experiments with the commonly-used image-point feature technique which cast doubts about whether an image-point
feature map generated under one lighting condition could be used for geometric localisation under a different lighting condition.

To ensure reliability the system presented here is based on a surveyed 3D-edge map of the buildings, which includes only permanent features and is explicitly separate from lighting effects. Therefore, the map is usable for localisation over long periods of time.

The edge-based localisation system is inspired by the work in [9]. This algorithm uses a particle filter which maintains multiple pose hypotheses where each particle represents a possible pose of the vehicle. This is different to traditional edge-model localisation frameworks which maintain a single hypothesis. Multi-hypothesis particle filters have some notable advantages:

- the ability to commence operation with large uncertainty in the initial pose of the vehicle, and,
- the ability to deal with local minima caused by spurious image-edges by forming multi-modal distributions.

The contribution of this chapter lies in the area of the observation function, with a new function that further improves the initialisation process and robustness to local minima in the particle filter. The new observation function conducts, for each particle, a search in the image-plane for image-edges, whereas the previous observation functions only consider image-edges which directly align with the projected 3D-edge map. The implicit downside of these previous functions is that a small change in pose causes a large drop in probability, requiring a tight distribution of many particles. This chapter shows situations illustrating the benefits of the new observation function where the filter can converge reliably even when given sparse particle distributions.

An additional contribution of this chapter is the development of an intelligent exposure control algorithm to deal with the issues of non-uniformity of lighting across the scene (in particular shadows) and the problem of direct sunlight in the camera’s field-of-view.

A variety of experiments are conducted with a vehicle operating in an industrial environment, demonstrating the performance and successful operation outdoors.

The remainder of the chapter is structured as follows. The next section will introduce the application and motivations for the design of the visual localisation system. In Section 5.3, an edge-based localisation technique is described that can be used to estimate the position and heading of the mobile vehicle given a sparse 3D-edge map of the doors in the environment. Section 5.4 discusses the issue of camera exposure and shows how knowledge of the scene can be used to intelligently adjust the camera exposure to improve the quality of information in the image. The experimental setup is described in
Section 5.5. Initialisation results are in Section 5.6. Extended operation results are in Section 5.7. Results from an entire sunny day are presented in Section 5.7.2. Results from rainy weather are presented in Section 5.7.5. Finally, conclusions are drawn in Section 5.8.

## 5.1 System Design Motivations

In recent years, heavy industry has begun to investigate the use of automated mobile equipment to resolve productivity and personnel safety issues. The work reported here is a vision-based method of vehicle localisation and is a part of a larger plan to fully automate large ground vehicles operating in outdoor environments. The technique has been developed and tested on a large forklift type robotic vehicle [98] which operates in an outdoor setting, handling large loads in the steel and aluminium industry. These vehicles operate for long periods outdoors during dynamic and non-uniform lighting conditions. The operating conditions of the vehicle requires a vision system that is real-time, accurate and robust to lighting changes.

The ultimate aim is to create a dependable fully-autonomous vehicle. A robust visual localisation system is one element required to achieve dependability.

*Figure 5.1:* The large forklift vehicle in its operating environment
5.1.1 Motivations from Redundant Localisation

The robust visual localisation system described in this chapter is one aspect of developing an autonomous vehicle which can operate dependably. High levels of dependability can be achieved with multiple independent localisation systems, based on entirely different sensor types, allowing the vehicle’s navigation system to have redundancy in localisation. It is only in very recent times that field roboticists have had the ability to compare pose estimates from independent localisation systems because, until now, it has been difficult to deploy more than one working localisation system on a field robot. The redundant localisation systems can be cross-checked and compared to automatically detect failure or degradation in one of the systems.

To date, a scanning laser localisation system has been developed for the vehicle [99], which has proven to be reliable and accurate. Throughout the many hours of operation in the field, there have been isolated incidents where the laser scanners have failed. The failure is thought to be caused by large magnetic fields in the environment causing disturbances in the spinning metal parts within the laser scanner. The vision-based localisation system presented in this chapter acts as an independent...
5.1 System Design Motivations

A system based on a completely different sensor. The extra localisation system provides redundancy and, hence, increases the dependability of the autonomous vehicle.

The vehicle’s navigation system is developed to take input from the multiple localisation systems and compare and arbitrate the input information to decide upon appropriate navigation actions. The proposed navigation system is presented in Appendix A with results from the autonomous experiments.

The use of multiple sensors for localisation has been well researched and has been widely applied in the area of field robotics. For the most part, multiple-sensor information is fused to form a single localisation system. This approach can improve the situation where the sensors cannot individually provide enough information for continuous and/or reliable localisation. In multi-sensor data fusion (Figure 5.2(a)), the aim is to provide a single localisation system with a more complete set of input sensor data by fusing all available sensor information. However, when sensors fail, provide erroneous readings or have a limited view of the world, the accuracy and confidence of the localisation estimates degrade. Hence, the data fusion process is not focused on providing redundancy. Examples of sensor fusion in the literature are:

- Majumder et al. [100] fuse sonar and camera information for an underwater vehicle
- Miura et al. [101] fuse laser and stereo camera data into an obstacle map
- Arras and Tomatis [102] fuse tracked features from laser and camera data.

In this study, lasers and cameras are used in outdoor environments; however, most previous laser and camera systems were developed for indoor environments. Newman et al. [25] is one of the only examples of outdoor localisation using both laser and a camera. They use these sensors in a single localisation system, where the vision system performs place recognition to trigger a loop closure event in a geometric laser localisation system. In contrast, two individual and unrelated geometric localisation systems are used.

This thesis presents a system of using multiple sensors in an alternative and more dependable manner. Two independent localisation systems, based on unrelated sensors, provide redundancy to the navigation system. It is thought that the use of multiple sensors for multiple-independent-localisation systems has rarely been investigated in the area of field robotics research. Figure 5.2 shows the fundamental difference of this approach. A system using independent localisation systems (Figure 5.2(b)), employs an additional process – an arbitrator or comparator – to monitor the pose estimates from the multiple localisation systems and cross-checks them for consistency.
5.1.2 Motivations from the Lighting Conditions

The outdoor environment presents major difficulties for visual localisation, and provides motivations for the design of the visual localisation system in this chapter.

Difficulties include the bright sun that is often at an angle in the sky, causing parts of the scene to be in direct sunlight while other parts are in shadows. Also, the sun itself can actually be within the field of view. In such conditions, cameras often cannot correctly expose the entire image simultaneously. The other key issue is the ever changing position of the sun throughout each day and the varying cloud cover, causing dramatic changes in the information within the images. There are no methods to completely factor out the effects of lighting when extracting information from the image. This is a problem for autonomously building a visual map, because the map will be specific to the lighting condition under which it was recorded. Section 3.4 presented experiments on the SIFT technique, which is commonly used in the autonomous construction of visual maps. The experiments indicated that failures are expected when trying to localise from an autonomously generated visual map after the light changes.

The problems discussed above lead to the conclusion that current approaches to autonomously build visual maps are not suitable for the high robustness required for the targeted application area, which is why this chapter develops a system based on a manually surveyed map. This is a logical choice to meet the reliability requirement imposed by the industrial application.

The application environment contains large buildings and sheds that are to be repetitively navigated. A map of the environment is surveyed in just one day, a comparatively short period for a map that will be useful for countless hours – if not decades – of productive operation. Manual map surveying ensures that only permanent objects are included in the map and that the description of the environment is not specific to a lighting condition, all of which allow for reliable long term operations under many conditions.

5.2 Related Work

Most visual localisation work has been performed by aligning the 3D world locations of features to their corresponding 2D locations in the image plane, the work of [103] and [3] being notable examples.

For many years was restricted to indoor environments, but recently there are more outdoor experiments being performed where the lighting conditions present a significant challenge. The work
5.2 Related Work

of [4] and [8] both present recent results that improve operations even when lighting changes cause disruptions in the image-point feature descriptions. However, these authors are solving a different problem to ours – that of appearance-based localisation (recognising a previously visited place) rather than geometric localisation (deriving an explicit orientation and pose of a vehicle with enough accuracy to allow autonomous navigation).

The monocular visual odometry system of [104] and the stereo visual odometry system of [45] are examples of geometric localisation in outdoor environments. Both odometry systems present very impressive and useful results, although for long term autonomous operations the frame-to-frame error (however small) will accumulate and must be removed using a map of the environment.

Outdoor visual localisation examples that do perform geometric localisation with a map can be seen in the dense 3D-point cloud stereo vision work of [20] in unstructured natural environments and [105] show a stereo system operating in an urban environment. Stereo is well known to be suited to environments with a lot of small-scale texture and not suited to environments with large homogeneous texture-less walls, where the dominant visual features are at the edges of the doors, walls, windows and roof-lines of the buildings.

There are a class of localisation algorithms that compare edges extracted from images with 3D-edge maps. One of the initial edge-based localisation techniques was developed in [106] for navigating indoor hallways using a 3D-edge map of the doors and walls. This type of technique has, for the most part, been applied in indoor environments. A more recent real-time technique developed in [29] has been applied outdoors in [30]. A GPU accelerated version of this system was presented in [107] and another outdoor example of this type of system is presented in [108]. The main limitation of these 3D-edge localisation approaches listed above are that they need to be initialised with an accurate prior estimate of the pose, and must remain with an accurate estimate at all times. This is because they maintain only a single pose hypothesis at a time which makes them susceptible to the edges of scene clutter and shadows extracted in the camera image which are not in the 3D-edge map.

Whereas the multiple pose hypothesis particle filter has the ability to deal with spurious edges detected by the camera by forming multi-modal distributions which correctly account for ambiguities in the observation function. A 3D-edge particle filter was presented in [9] which uses an observation function for each particle that measures the quality of alignment of the 3D-edge map with the camera image. Their system is tested and evaluated against the proposed system of this chapter.

The first instantiation of this chapter’s system computes the observation function with image edge weighted equally [109], rather than weighting each pixel equally as in [9]. Computing the observation function on a local-basis gives equal weighting to each edge, avoiding the observation function being dominated by large edges, so that smaller edges are not ignored in situations where they are in
fact providing useful localisation information. The results of [109] showed that the filter using an observation function that weights edges equally was more accurate at estimating the orientation of the vehicle.

Even with an observation function that is computed on a per-edge basis, small changes of pose produce large changes of likelihood in the observation function. This is because the alignment between 3D-edge map and camera image is evaluated only at the locations of the projected 3D-edge map. This chapter proposes a new observation function that searches outwards from the projected edges for the nearest edge in the camera image. The design of this new observation function is aimed at allowing the filter to converge even when given sparse particle distributions. Convergence is more likely because small changes of pose will not produce large changes in the likelihood measured by the observation function. Particles that are moderate distances or orientations away from the correct pose will still generate moderate alignment probabilities.

5.3 3D-edge Map Localisation

Our goal of a robust outdoor visual localisation system led us to explore the use of a 3D map of the edges of the buildings in the environment. Edges are found on the buildings themselves (e.g. door ways, windows, ventilation openings, etc.) and at the visual boundary of the buildings with the background (e.g. the roof line against the sky, and at the sides of buildings with the general background). The attractive property of the building edges in this environment is that they are static and their precise location can be measured in three dimensions.

5.3.1 A Sparse Map

The 3D-edge map of industrial buildings can be extremely sparse. The edge features stored in the 3D map include the building roof lines (typically, the gutter lines), the door frames (the edge of the large industrial roller doors) and on some of the buildings, some of the lines that define the edges of some other significant features in the buildings shape (such as overhangs). For a building with three roller doors and a single visible roof line, this translates into just 14 3D data points in the model (4 for each door and 2 for the roof line). This sparsity of data means that it is both feasible (in terms of cost and effort) and desirable (in terms of confidence in accuracy) to manually survey the model data points. Such a survey ensures that only permanent parts of the buildings are included — which is difficult to guarantee in an autonomous map-building system. An example of the map is shown in Figure 5.3(a).
5.3 3D-edge Map Localisation

Figure 5.3: Examples of the surveyed 3D-edge map, fish-eye camera-setup and calibration. Two fish-eye cameras are placed at the front of the vehicle facing sideways. The hemispheres represent the field of view of the cameras.

To give the reader an idea of scale, the roofline in the figure is 9 m high and the doors are 4 m wide by 6 m tall.

The cost of the survey was approximately USD$4000, which is an insignificant cost considering the large value of the raw minerals transported by the vehicle, and also can be considered minor when compared with the price of a standard laser scanner, which can be more than USD$4000. The map took a single day to survey, which is a one off delay that is not a concern for an application where the vehicles operate day after day for several years of productive operation. The survey was of 19 industrial buildings ranging in height from 6 m to 17 m and together have a footprint on the ground of approximately 290 m by 100 m. The experiments in this chapter are conducted in the industrial compound portion of the site which is surrounded by seven buildings that have a footprint
of around 70 m by 45 m. Figure 5.4 presents an overhead view of the site and survey, showing both the dimensions of the overall footprint and the compound footprint.

The survey was taken using a total station and the professional surveyor conducting the survey quoted the accuracy of the measurements to be within 50 mm. The accuracy depends on the surface properties of the point being measured and the viewing angle and the distance to the point.

**Figure 5.4:** Top; Aerial image of the site. Bottom; Overhead view of the survey of the buildings.

### 5.3.2 Wide FOV Imaging

In the target application, the distance between the vehicle (and hence the cameras) and the buildings can vary from 1m to 50m. This geometry indicates that an imaging system is needed that can cover as much of the environment as possible at all times in order to guarantee that the edges in the model can
be seen. Two fish-eye cameras (with 185 degree FOV) are mounted on the vehicle (Figure 5.3(b)). A typical image from one of these cameras is shown in Figure 5.3(c).

A specialised lens model is adopted for calibration from Geyer and Daniilidis [110], that was shown to be applicable to fish-eye cameras by Ying and Hu [111]. The model assumes that all pixels in the fish-eye image map onto a sphere located in front of the image plane. The model consists of four parameters, \( \frac{m}{l} \), the distance from the center of the sphere to the image plane, \( C_x \) and \( C_y \), the center of projection on the image plane, and, \( l \), the distance from the center of the sphere to the intersecting focus point of the light rays and the center of projection line.

These calibrated parameters enable the fish-eye image to be transformed into an undistorted image. Examples of distorted/undistorted images can be seen in Figure 5.3. The corrected image is generated as follows; for each pixel coordinate \([U_u, U_v]\) in the corrected image (Figure 5.3(d)), with centre \([C_x, C_y]\), the corresponding distorted coordinate \([D_u, D_v]\) in the fish-eye image (Figure 5.3(c)) is calculated by:

\[
D_u = R \cos(\arctan\left(\frac{U_v}{U_u}\right)) + C_x \tag{5.1}
\]
\[
D_v = R \sin(\arctan\left(\frac{U_v}{U_u}\right)) + C_y \tag{5.2}
\]

where

\[
R = \frac{\sin(\theta)(m + l)}{\cos(\theta) + l} \tag{5.3}
\]

and

\[
\theta = \arctan\left(\frac{\sqrt{U_u^2 + U_v^2}}{f}\right) \tag{5.4}
\]

where \( f \) is the effective focal length of the undistorted projective image, measured in pixels from the center of the sphere. \( f \) is also used in the projection model to project the 3D-edge-map onto the image plane and can be selected according to the effective FOV that is required. The resulting undistorted image is converted into an edge-image using Canny’s algorithm [112] with a \( 3 \times 3 \) kernel.

### 5.3.3 Particle Filter Localisation

Most previous 3D-edge based techniques calculate only a single pose estimate for each iteration, which requires a precise initial estimate of the pose and is susceptible to failure. Recently in [9] a particle filter method was presented that maintains many pose estimates per frame. The technique was applied to the tracking of a single object, such as a printer, from a range of just a few metres in a
regular indoor environment. In this chapter a particle filter is also presented, but for tracking outdoor industrial buildings. The comparison between the map and the camera images is calculated for each pose hypothesis in a particle filter and provides a likelihood measure, discussed in Section 5.3.4.

In [113] the use of a particle filter for pose estimation is described in detail. In brief, the filter is a set of $N$ pose hypotheses (particles) $X_t = x^{(1)}_t, x^{(2)}_t, x^{(3)}_t \ldots, x^{(N)}_t$. The pose is a six degree-of-freedom translation and rotation ($t_x, t_y, t_z, r_x, r_y, r_z$). The dominant degrees-of-freedom are 2D horizontal translations and a rotation around the vertical axis ($t_x, t_y, r_z$). The additional degrees of freedom are used to deal with any deviations in the ground plane causing slight rolls and pitches in the vehicle and slight changes in the vertical displacement. The coordinate system is defined from the centre point of the axle joining the two rear wheels and positive rotations in the vehicle’s heading are defined by anti-clockwise rotations around the vertical axis. Vehicle odometry from wheel and steering encoders is used to estimate changes in horizontal translation and heading angle. The other degrees of freedom are not measured by additional sensors, but are included as small perturbations in the filter, more details on the propagation model are in Section 5.3.6.

The set of poses is sampled from the previous set $X_{t-1}$ using a propagation model $m_t$ and a corresponding set of weights (probabilities), $W$. The weight of particle $n$ is calculated at time $k$ as follows:

$$W_k^{(n)} = p(y_k|x_k^{(n)}) \quad (5.5)$$

where $y$ is the comparison between the edge-image extracted from the camera image and the 3D-edge map projected to the image plane from the pose of each particle $x$. The comparison between the map and camera image provides a likelihood measurement based on observation functions described in the next section. The concept is that the particles nearest the correct pose will have the highest likelihood measure, because their projection of the 3D-edge map will have the best alignment with the camera image. These particles will have the highest probability of being re-sampled for the next iteration. To extract the current pose estimate of the vehicle from the filter, the mean pose of the most highly weighted particles is calculated. Here, the non-weighted mean of the 5% most highly weighted particles is used.

### 5.3.4 Observation Function

The likelihood measure for each particle is generated through a comparison with the edge-image and the 3D-edge map. The map is projected onto the image plane, so a direct comparison can be made. A fast method is presented in [9] that performs this computation on a graphics processing unit (GPU)
counting the number of aligning pixels over the whole image. This section will first describe Klein and Murray’s metric, and then describe modifications that improve performance.

**Klein and Murray**

Klein and Murray’s method was implemented by first placing the undistorted edge-image into the GPU’s texture memory. For each particle, the OpenGL projection matrix was set and the 3D-edge map was called to be rendered for each particle by a custom fragment shader program. The program permits the counting of the visible edge pixels of the 3D-edge map that align with edge-pixels in the undistorted edge-image. The custom fragment shader program only permits pixels to pass through the pipeline that align with edge-pixels in the edge-image. The pixels that pass this custom fragment shader program are counted using the OpenGL occlusion query extension [114].

Klein and Murray present the likelihood measure of the particle, $W_t^{(a)}$, as a ratio between the count of aligning edge-pixels ($a$) and the total number of visible edge-pixels ($v$), calculated as follows:

$$W_t^{(a)} = p(y_t|x_t^{(n)}) \propto \exp \left( \kappa \frac{a}{v} \right) \quad (5.6)$$

where $\kappa$ is a constant that weights the observation function.

Klein and Murray show this metric can successfully track objects, but the simple ratio of pixel counts leads to the situation where large edges, such as the roof-lines of the buildings, dominate other smaller edges, such as the door edges. This is simply because the majority of pixels are in the roof-edges. Smaller edges provide important localisation information and should have more consideration in the observation function.

**Per-edge Function**

The first implementation of the localisation system [109], presented a modification to Klein and Murray’s function that incorporated per-edge measurements instead of a sum over the whole image. The new metric calculates the ratio of aligning-to-visible edge pixels for each edge, $j$. This is calculated using occlusion queries for each edge, giving the two measurements $a_j$ and $v_j$. The first component of the new metric is the original Klein and Murray global ratio, the second component is
the mean of the ratios of each individual edge, and is calculated as follows:

$$W_t^{(n)} = p(y_t|x_t^{(n)}) \propto \exp \left( \frac{\alpha}{\nu} + \lambda \sum_{j=0}^{m} \frac{a_j}{m} v_j \right)$$  \hspace{1cm} (5.7)

where \( m \) is the number of edges. The second component of this equation treats each edge equally, regardless of its size. This penalises particles with edges that are smaller and misaligning, even if the overall count of aligning pixels is high. The filter will prefer particles with the combination of a reasonably high overall count of aligning pixels and smaller aligning edges. This allows the filter to maintain a better track of the smaller door edges in the environment. The new per-edge component has its own constant \( \lambda \) and this has to be tuned in conjunction with \( \kappa \), thus striking a balance between the global and per-edge components, although as shown later in Section 5.5, the function behaves well when these values are equal.

**Nearest-edge Function**

An issue with the above two functions is that the peaks in the functions are narrow – a small change in pose causes a large change in likelihood value due to the binary comparison (aligned or not-aligned) at the core of the functions. Therefore a tight pack of many particles is required to correctly maintain track of the narrow peaks. Such a distribution in the filter will be susceptible to local maxima. This susceptibility will manifest itself in two ways; converging at an incorrect estimate at initialisation and also making the filter unable to recover after slightly losing track of the 3D-edge map.

Klein and Murray proposed two methods to overcome this issue. The first is to dilate the edge image, creating thicker image-edges. However, creating thicker edges will only serve to flatten and plateau the observation function, causing a loss in accuracy. The other solution is to use a two-stage filter with the first stage being performed in low-resolution, essentially creating thicker map-edges. The second high-resolution stage can provide added accuracy; however, the particles re-sampled for the second stage may not be re-sampled near the narrow peaks because of the flat function in the first stage. This will cause the filter to jitter and to lose track, both problems are reported in [9].

We propose a new metric — the distance to the nearest edge pixel, which is an improved measurement regime to the binary alignment. A similar method is proposed in [29], though they used a set of nearest-edge measurements to extract a single pose hypothesis. The algorithm in this chapter incorporates nearest-edge measurements into the multiple hypothesis particle filter. The nearest-edge metric does not only consider pixels aligning with the projected edge, but searches outwards in the image. This provides a wider, sloping, function, that will be easier for the particle filter to remain
converged and recover from divergence and also importantly will be more robust when initialising the filter with a large/sparse particle distribution.

Drummond and Cipolla’s method is to take sample points along each edge at regular pixel increments. At each sample point a search is conducted outwards along the edge normal to find the nearest image edge. Here the same sampling and searching strategy is adopted. A sample is taken every 20 pixels on each 3D-edge, \( s \). A search is conducted in both the positive and negative directions of the 3D-edge’s normal in the image plane. The search distance in the real-world coordinate frame as a constant, \( D_w \).

The search distance in pixels \( D \) is determined according to the focal length \( f \) in pixels and the depth of the sample point \( E_z \):

\[
D = D_w \frac{f}{E_z} \tag{5.8}
\]

![Figure 5.5: Nearest-edge search. The largest lines are the projected edges, the smaller perpendicular lines are the search along the edge normal and the smallest lines indicate the nearest detected edge.](image)

An example of the sample and search for the nearest edge can be seen in Figure 5.5. The distance to the nearest edge, \( d \), which is calculated in pixels and is normalised to a value between 0 (which represents a zero search distance) and 1 (which represents the maximum search distance \( D \)).
Gaussian function converts $d$ to add weight to closer edges:

$$g(d) = \exp \left( -\frac{d^2}{2\sigma^2} \right)$$

(5.9)

where the constant $\sigma$ must be selected to weight the output appropriately. $\sigma$ could be set to $\frac{1}{3}$ to give no importance ($g(1) \approx 0$) for when the nearest edge was found at the end of the search (that is, $d \approx 1$). However, an edge found near the end of the search should hold greater importance than not finding an edge at all. Therefore $\sigma$ is set to $\frac{2}{3}$ to give the output value $g(1) \approx 0.3$ and when no edge is found, $g$ evaluates to 0.

The output of these samples ($g(d)$) are formed into the final observation function by first aggregating the samples for each edge to give a likelihood $l$ for the $s$ samples on the edge:

$$l = \frac{\sum_{i=0}^{s} g(d_i)}{s}$$

(5.10)

The likelihood of all of the $m$ edges are aggregated as:

$$W_{i}^{(n)} = p(y_{t} | x_{t}^{(n)}) \propto \exp \left( \kappa \frac{\sum_{k=0}^{m} l_{k}}{m} \right)$$

(5.11)

This observation function is designed to allow the filter to converge given sparse particle distributions. Particles that are moderate distances away from the correct pose will still provide moderate likelihood scores, whereas the other observation functions will assign very low weights to these particles.

The performance of the three different observation functions presented above are compared in initialisation in Section 5.6. For ease of identification in the remainder of this chapter, the three functions are referred to as follows:

- Equation 5.6 is named *Klein and Murray*,

- Equation 5.7 is named *Per-edge* and

- Equation 5.11 is named *Nearest-edge*. 
5.3 3D-edge Map Localisation

5.3.5 Occlusions

Self-occlusions, where one building occludes another (known occlusions), can be dealt with by the depth buffer. A real-time technique is presented in [9] using a sub-sampled depth buffer. The technique is to render faces of the buildings to the depth buffer, then only edges that are in front of the faces will pass through. The depth buffer is limited in resolution, which leads to the problem of a surface blocking its own edges. To avoid this, the surfaces are recessed back a small distance from the edge. The offset distance between surface and edge needs to be larger than the resolution of the buffer at that depth.

5.3.6 Propagation Model

Motion measurements from the vehicle can be formed into a model that propagates the particle filter. The uncertainty in the motion model \( m_t \) is defined by a Gaussian distribution, \( \varphi \), as follows:

\[
m_t = \varphi(\sigma^2, \mu)
\]

This propagation distribution, \( \varphi \), is defined by the mean, \( \mu \), and variance, \( \sigma^2 \), as follows:

\[
\mu = \delta
\]

\[
\sigma^2 = \beta \delta + \alpha
\]

The vehicle’s wheel encoders and steering encoders form \( \delta \) as a 2D translation, \( t_x, t_y \), and a rotation around the vertical axis, \( r_z \). The model propagates the particle distribution with the odometry estimate, \( \delta \), and perturbs the distribution proportional to the odometry estimate according to the constant \( \beta \). This constant increases the variance in the distribution as the vehicle’s velocity and angular velocity increases, and increases the variance in the direction of the velocity.

The ground in the environment is not perfectly flat and therefore slight vertical translations, \( t_z \), and roll and pitching, \( r_x, r_y \), of the vehicle must be taken into account. No sensors are used to measure these additional degrees of freedom, these unknown degrees of freedom are included in the propagation model by small perturbations across all six degrees of freedom, defined by the constant \( \alpha \).
5.3.7 Initialisation

The initialisation process begins by distributing the particles roughly around the approximated vehicle location. The particle filter is then iterated to converge around a pose estimate.

The variance (uncertainty) in the initial distribution is set according to how accurately the vehicle’s initial pose is known. If the pose is known only approximately, the variance is set high and more particles are needed to cover the larger search space. The larger number of particles slows down the system during this initialisation phase. An adaptive particle filter is required which gradually reduces the number of particles to transition into a faster operating frame rate. Fox [115] presented a particle filter that adapts the number of samples in the distribution according to the spread of the distribution over the state space. Fox discretised the state space into bins and used the number of occupied bins to set the desired number of samples. In this chapter a simpler method is developed that does not require discretising the state space. Here the adaptive filter sets the number of particles according to the variance in the distribution, using the following equation:

\[ n_{t+1} = \max\left(n_0 \frac{v_t}{v_0}, n_d\right) \]  

(5.15)

where \( n_{t+1} \) is the number of particles for the next iteration, \( n_0 \) is the initial number of particles, \( n_d \) is the desired number of particles after full convergence, \( v_t \) is the current translational variance in the particle distribution and \( v_0 \) is the initial variance. The effect of this equation is to reduce the particle count as the filter converges, until the desired number of particles is reached to achieve a processing rate suitable for operation. Admittedly Fox’s method of adapting the particle filter will behave more reasonably in the case of multi-modal distributions that are widely separated but individually are tight distributions. In that case the proposed method will require a large number of particles because a single-mode and will mis-calculate a large variance in the distribution. Whereas Fox’s method captures the true variance of the multiple-modes in the distribution. However, the experiments presented later on demonstrate that the proposed adaptive filter still gives desirable results.

5.4 Intelligent Exposure Control

The lighting conditions of the application environment are harsh with the robot vehicle operating in bright sunlight. The nature of the built environment (tall buildings with gaps in between) results in multiple areas of shadow and of full sunlight. The use of fish-eye cameras also results in the sun itself appearing in the images most of the time. This is a challenge for standard cameras where the built-in
exposure control algorithms use a grey-world assumption. These algorithms aim to control the mean intensity value over the whole image to a predefined intensity value, regardless of the content of the scene. The conventional approach to exposure control causes overcorrection, resulting in an image that contains incorrectly exposed areas. An example of overcorrection is shown in Figure 5.6(a) which shows a lens flare running down the image. But more importantly, Figure 5.6(b) shows that there is too much correction for the sunny sky with a standard exposure control algorithm, leading to underexposure of the building fronts and no door-edges being visible.

Exposure control is a task often undertaken without regard for the specific objects that are in the field of view, and is instead based purely on statistical information, such as in [116]. One example where exposure control is directed towards specific objects of interest is in the work of [117], where a face detection algorithm is investigated to find the areas of the image which are used to control exposure.

In previous work [56] it was demonstrated how a standard camera can be used to create a high dynamic range image. The technique used multiple exposures (low, medium and high) that were combined to form a single image that successfully captured image features in dark shadows and full sunlight. This method is particularly applicable to the application in that the resultant high dynamic range image is an edge image. However, because this technique requires multiple exposures (at least three) to create on image, the effective frame rate of the imaging system is reduced (by at least a third). Here in this work it is shown that a single exposure is sufficient when it is intelligently controlled to correctly expose the specific areas of interest in the image.

Due to the nature of the proposed localisation method, there are specific areas of interest in the scene (that is, the doors and roof lines of the buildings) that must be correctly exposed and because these are being tracked in the system, it is known where in the image the features should lie. Therefore an exposure control algorithm is developed that aims to maximise the strength of image edges corresponding to 3D-map edges, while ignoring all non-essential areas of the image. The algorithm first samples the intensity values of pixels near the tracked edges (the intensity values are taken from the luminance channel of the YCrCb space after conversion from the camera’s native RGB format). The edges are projected into the image according to the current pose estimate and short scans of pixels are taken along the normal of the edges. All edges except the roof-line edges are used for sampling. The roof edges are ignored to avoid sampling pixels from the bright sky that would otherwise heavily weight the sampling against increasing the exposure of the camera. It is assumed here that if the doors are correctly exposed, then so too will the roof line. Figure 5.7(a) shows which pixels are sampled from the scans. The control algorithm is as follows:

\[
\xi_t = \xi_{t-1} + (1.0 - (\epsilon e))
\]
Figure 5.6: Example of the bright lighting conditions. The sun causes a flare in the fish-eye lens and a dark line down the image owing to errors in the sensor’s response. Overcompensation for the sunlight can occur using naive exposure control found on most cameras. 5.6(a) shows an example of overcompensation where the buildings are under-exposed. 5.6(b) is the corresponding edge-image where no edge-features are detected on the doors of the buildings. 5.6(c) shows that with the use of intelligent exposure control algorithm, the buildings are correctly exposed. As a result, the edges are detected on the doors and on the other areas on the buildings in 5.6(d).

where $\xi_t$ is the exposure level at time $t$. The IEEE1394 IIDC [95] compliant digital cameras used in experimentation have two exposure parameters available, an analog-to-digital gain and a digital shutter time. These two parameters are scaled between 0 and 1 and combined linearly into one parameter $\xi$. A constant to determine the rate of adjustment is $\epsilon$ and $e$ is the error, calculated as the ratio between the mean intensity value, $I_m$, from the pixel scans (Figure 5.7(a)) and the goal intensity value $I_d$:

$$e = \frac{I_m}{I_d}$$

(5.17)
Based on the plot seen in Figure 5.7(b) the edge strength is at a maximum when the mean intensity of the sampled pixels is $I_d = 180$, defined on a 8-bit intensity scale. The damping constant $\epsilon$ has been set to 0.02, after empirical tests showed this value to provide a balance between quick response to lighting changes and stable control. This value has not been adjusted since being set and has been used successfully in a wide range of lighting ranging from dark clouds to bright sunlight. Figure 5.6(c) shows a typical result for this intelligent exposure control algorithm, where the buildings are properly exposed. Edges on the doors and other areas of the buildings are now clearly detected (Figure 5.6(d)).

![Figure 5.7: Left: Pixels are sampled along the normal to the tracked door edges. The mean intensity value of these pixels are used as input for the exposure control algorithm. Right: The edge strength sampled from the doors of buildings, plotted against mean 8-bit pixel intensity of the sample. Edge strength is defined as the average intensity difference in a $3 \times 3$ pixel neighbourhood. This graph was recorded over a period of time as the exposure of a stationary camera was incrementally increased.](image)

5.5 Experimental Setup

The vehicle is fitted with two IEEE1394 cameras with fish-eye lenses that are mounted facing sideways on the vehicle (Figure 5.3(b)). The intrinsic and extrinsic camera parameters were calculated and verified by projecting the edge model into the image plane using the ground truth and ensuring the edges were aligned correctly with the recorded video stream. Figure 5.3 shows the undistorted image using the calibrated fish-eye model. An in-house camera driver was implemented with a double-buffer so that current image being processed is at most delayed by one frame and the image transfer time.
The upper and lower hysteresis edge-detection thresholds for the Canny edge detector [112] were adjusted manually to the minimal values that still permitted the edges on the buildings to be reliably detected. The threshold values resulting from these empirical tests were 30 and 100.

The number of particles is a crucial parameter and will ideally be large enough to sample a greater portion of the solution space, but this comes with a larger computation cost. The filter, once converged, is maintained with 500 particles and at initialisation the number of particles is dependant on how well the initial pose is known.

The remaining parameters for the particle filter, the likelihood constants and the motion model parameters, cannot be chosen analytically and require quantitative data for tuning. The approach used to tune these parameters was to record a short sequence of video, odometry and ground truth pose data (see the following section for information on ground truth data) from the vehicle traveling around the environment covering most areas and orientations. The particle filter was then run off-line several times through the same sequence of recorded data to optimise these parameters. Different parameters were tested each cycle through the data and the average pose estimate error was recorded. The three different observation functions behave differently but the parameters need only minor adjustment when switching between the different functions. In fact the parameters are kept the same for the Klein and Murray function and the Per-edge function which optimal values were found to be as follows, the motion model parameters, from Equations 5.12 and 5.14, (one for each pose dimension; $t_x, t_y, t_z, r_x, r_y, r_z$) are calibrated and shown in Table 5.1.

**Table 5.1: Details of Motion Model Parameters**

<table>
<thead>
<tr>
<th>Parameter name</th>
<th>$t_x$</th>
<th>$t_y$</th>
<th>$t_z$</th>
<th>$r_x$</th>
<th>$r_y$</th>
<th>$r_z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$ for Klein and Murray</td>
<td>1.5</td>
<td>1.5</td>
<td>0.1</td>
<td>0.01</td>
<td>0.01</td>
<td>0.3</td>
</tr>
<tr>
<td>$\alpha$ for Per-edge</td>
<td>1.5</td>
<td>1.5</td>
<td>0.1</td>
<td>0.01</td>
<td>0.01</td>
<td>0.3</td>
</tr>
<tr>
<td>$\alpha$ for Nearest-edge</td>
<td>2.0</td>
<td>2.0</td>
<td>0.1</td>
<td>0.01</td>
<td>0.01</td>
<td>0.5</td>
</tr>
<tr>
<td>$\beta$ for Klein and Murray</td>
<td>0.3</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5</td>
</tr>
<tr>
<td>$\beta$ for Per-edge</td>
<td>0.3</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5</td>
</tr>
<tr>
<td>$\beta$ for Nearest-edge</td>
<td>0.3</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5</td>
</tr>
</tbody>
</table>

**Table 5.2: Details of Observation Function Parameters**

<table>
<thead>
<tr>
<th>Parameter name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\kappa$ for Klein and Murray</td>
<td>5</td>
</tr>
<tr>
<td>$\kappa$ for Per-edge</td>
<td>5</td>
</tr>
<tr>
<td>$\kappa$ for Nearest-edge</td>
<td>3</td>
</tr>
<tr>
<td>$\lambda$ for Per-edge</td>
<td>5</td>
</tr>
<tr>
<td>$D_w$ for Nearest-edge</td>
<td>0.5 m</td>
</tr>
</tbody>
</table>
5.6 Initialisation Experiment

The observation functions constants are shown in Table 5.2, where $\lambda$ is specifically for the second component of the Per-edge function, which behaves well when $\kappa$ and $\lambda$ are equal. Noticeably a lot of these parameters are not too different from the other functions’ parameters, and in fact adjusting any parameter up or down by 30% does not significantly change the filter’s behaviour.

5.5.1 Evaluation of Localisation

To evaluate the visual localisation system presented in this paper, we compare its pose estimates to those coming from a laser-scanner localisation system described in [99]. Our laser localisation system was extensively compared with RTK-GPS and shown to give full coverage around the site, whereas GPS was found to experience dropouts and multipath errors in some locations. Given that the laser system is more reliable in terms of coverage and has suitable accuracy in the appropriate areas of the site, it was chosen as the source localisation to evaluate the visual localisation system.

The accuracy of this laser-scanner system is varying dependant on the density of the surrounding laser reflecting beacons. In some areas of the site high accuracy is not required and here the laser-beacons are sparsely placed and, hence, the accuracy of the laser-localisation system is lower in these areas. In some areas the error is as low as 0.97 m in comparison with RTK-GPS, as reported in [99]. Even though the laser localisation system is known to have 0.97 m error in some areas of the site, the area where the experiments are performed in this paper has a dense placement of laser beacons and the laser localisation system is much more accurate here. In this area the laser localisation system has proven to be a reliable basis for closed loop control of precise load transfer maneuvers [118]. These maneuvers have been performed repeatedly for hours and require accuracy on the order of 100 mm to place the pick-up hook inside a small eyelet. Therefore in this area of the site the laser localisation system is known to be an appropriate source of ground truth localisation to evaluate the proposed visual localisation system.

5.6 Initialisation Experiment

The experiments conducted on this system are split into two sections; here the initialisation of the filter is presented and the following section presents results of the filter operating for extended periods outdoors. All experiments are conducted in a 70 m $\times$ 45 m industrial courtyard surrounded by 7 buildings, ranging from 6 m to 9 m in height (see Figure 5.4). The first initialisation experiment demonstrates the initialisation process of the filter using the adaptive particle filter technique described
earlier in Section 5.3.7. In this example, 4000 particles were scattered across a 40 metre diameter area and the full 360 degrees in the orientation. In other words, the pose of the vehicle is known to lie within a 40 metre diameter and the orientation of the vehicle is completely unknown. The centre of the initial distribution is randomly chosen to lie off-centre of the correct pose.

Images of the process are shown in Figures 5.8, 5.9 and 5.10. In this example the Nearest-edge observation function is used, and in the following section the three observation functions are compared by running the initialisation process many times at many locations/orientations. Inspecting the figures, at iteration 10, the particle filter has converged into several distinct distributions, seen in the overlay in Figure 5.9(c). Illustrating that there are several local maxima nearby the correct pose. After further iteration many of these distributions become down-weighted and disappear. Yet later, at iteration 15, there are still two distinct particle distributions visible in Figure 5.9(g). The ability of the particle filter to maintain multiple estimates is a powerful feature lacking in previous single-hypothesis methods. By iteration 40, the particle filter had converged around one hypothesis which was centred within 1 m of the ground truth position (the video attachment named initialisation.mp4 shows this initialisation process).

5.6.1 Comparison of Observation Functions at Initialisation

The three observation functions described earlier in Section 5.3.4 (the functions are labeled; Klein and Murray, Per-edge and Nearest-edge) are compared here to evaluate their abilities on convergence.

The proposed Nearest-edge function is designed to improve limitations of the other two functions. The other two functions consider alignment only at the projected 3D-edge map of each particle, meaning that a small change in pose will cause a large change in the likelihood measured by these functions. By comparison the Nearest-edge function searches outwards from the 3D-edge map projected by each particle to find the nearest image-edge. In theory this will enable the function to correctly converge at initialisation with a large and sparse distribution of particles, whereas the other functions are more likely to converge at incorrect estimates.

To evaluate if the proposed Nearest-edge function does in fact perform better at initialisation an experiment is setup to initialise the particle filter with the three functions 50 times at many locations/orientations over the course of an entire sunny day. The three functions are used to initialise the filter on the same data and the results are compared against the ground truth pose given by the laser-scanner system. The initial distribution is same as the one seen in Figure 5.8, which is a distribution of 4000 particles across a 40 metre diameter area and around the full 360 degrees in the orientation.
5.6 Initialisation Experiment

Figure 5.8: Initialisation of the particle filter showing the image overlaid with 3D-edge-map, at iteration 0 and iteration 5.
Figure 5.9: Initialisation of the particle filter showing the image overlaid with 3D-edge-map, at iteration 10 and iteration 15.
5.6 Initialisation Experiment

![Initialisation Experiment](image)

**Figure 5.10:** Initialisation of the particle filter showing the image overlaid with 3D-edge-map, at iteration 20 and iteration 40.
Table 5.3 present the statistics of the error recorded against the ground truth at the point when the particle filter has converged. The error is calculated against the ground-truth pose (laser-scanner system) from the mean pose of the 5% highest weighted particles. Convergence is defined as the iteration when the desired number of particles (in this case 500) is reached in the filter as controlled by the adaptive particle filtering technique described earlier in Section 5.3.7. The position error is Euclidean distance of the horizontal translation errors.

<table>
<thead>
<tr>
<th></th>
<th>Klein and Murray</th>
<th>Per-edge</th>
<th>Nearest-edge</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Position error (m)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>3.91</td>
<td>1.41</td>
<td>0.71</td>
</tr>
<tr>
<td>25th Percentile</td>
<td>1.52</td>
<td>0.76</td>
<td>0.39</td>
</tr>
<tr>
<td>75th Percentile</td>
<td>8.33</td>
<td>2.65</td>
<td>1.44</td>
</tr>
<tr>
<td>Max</td>
<td>50.167</td>
<td>26.84</td>
<td>19.29</td>
</tr>
<tr>
<td><strong>Rotation error (deg)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>2.8</td>
<td>1.0</td>
<td>0.6</td>
</tr>
<tr>
<td>25th Percentile</td>
<td>0.8</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>75th Percentile</td>
<td>25.2</td>
<td>2.5</td>
<td>1.9</td>
</tr>
<tr>
<td>Max</td>
<td>179.9</td>
<td>90.7</td>
<td>102.9</td>
</tr>
<tr>
<td><strong>Convergence time (sec)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>8.66</td>
<td>6.1</td>
<td>7.9</td>
</tr>
<tr>
<td>25th Percentile</td>
<td>6.46</td>
<td>5.2</td>
<td>6.5</td>
</tr>
<tr>
<td>75th Percentile</td>
<td>10.1</td>
<td>9.1</td>
<td>9.7</td>
</tr>
<tr>
<td>Max</td>
<td>10.6</td>
<td>20</td>
<td>18.8</td>
</tr>
<tr>
<td><strong>Success rate (%)</strong></td>
<td>15</td>
<td>38</td>
<td>71</td>
</tr>
</tbody>
</table>

The table noticeably shows that the Nearest-edge function outperforms the other two in all the statistical indicators of position error. The median and interquartile range of the position error of the Nearest-edge function is 0.7 m and 1 m respectively, whereas for the Klein and Murray 3.9 m and 6.8 m and Per-edge 1.4 m and 1.9 m. The Nearest-edge function also outperforms in orientation error, although the Per-edge function is more similar in orientation and the Klein and Murray function is by far the worst. The median and interquartile range of the orientation errors of the Nearest-edge is 0.6 and 1.6 degrees respectively and for the Per-edge function 1.0 and 2.2 degrees and the Klein and Murray 2.8 and 24.4 degrees. The maximum errors are all quite large and are from the cases where the filter fails to converge and the map is completely misaligned with incorrect edges in the camera image. In these cases the filter can drift unpredictably in the wrong direction and orientation. A success-threshold is defined to separate the situations where the filter fails to converge from the successful convergences. The success-threshold is defined at 1 m and 2 degrees, which are reasonable bounds of the requirements to successfully commence autonomous control of the vehicle. This threshold make the success rate of the three functions; Nearest-edge 71%, the Klein and Murray 15% and the Per-edge 38%.
The time taken for the filter to converge during the experiment is also presented in Table 5.3. The median time and quartile times for convergence of the Nearest-edge function and the Klein and Murray function are longer than the Per-edge function, though, the difference is not significant. If the filter can remain converged for long periods after initialisation then a delay of up to 20 seconds at start-up is not a major concern. The improvements gained in the position and orientation estimates and the success rate, from the Nearest-edge function far outweigh its time penalty.

5.7 Operation Experiments

5.7.1 Bright Sunlight

An experiment was conducted with an extended period of manual operation (30 minutes) of the vehicle at 2 pm on a sunny day. At this time, the lighting conditions were challenging, because the sun was bright and at an angle in the sky. When the vehicle turns and drives into a shadow, the exposure control algorithm had to adjust quickly according to the direction in which the cameras are facing. The vehicle was driven along an arbitrary path for a total distance of 1.5 km during the experiment. The vehicle travelled through a wide range of positions and orientations ensuring the system was well tested.

Figure 5.11(a) presents the heading estimate error of the vision system which was maintained at an average error of 0.62 degrees, as opposed to the accumulated odometry error which drifted to a maximum error of 30 degrees. These errors were recorded from the particle filter using the per-edge observation function. The pose estimate is extracted from the filter as the mean pose of the 5% most highly weighted particles.

Figure 5.11(b) shows the position error of the visual localisation. The vehicle’s position was correctly estimated to an average error of 0.44 m over the 1.5 km run. The maximum error at one stage crept out to 1.4 m and 4.4 degrees. The average errors indicate that the system is sufficiently accurate for autonomous navigation around the site. Example images from the experiment are shown in Figure 5.12 (the video attachment named extended-operation.mpg presents a sped-up sequence from the left camera of this experiment).
Figure 5.11: Results from the extended operation of the visual localisation system. The position error is Euclidean distance of the horizontal translation errors.

Figure 5.12: Undistorted images taken from the extended operation experiment, overlaid with the 3D-edge-map projected from the estimated pose.

5.7.2 All-day Experiment

Here the system is evaluated over the full range of lighting conditions experienced on a bright sunny day. The test consists of a 110 m path which began and ended at the same position and orientation. Two three point turns were completed during the path, simulating the dropping off and picking up
of loads. The path driven by the vehicle is shown in Figure 5.13. Initially, the path was completed twice just after sunrise at 7am. The path was then repeated twice at the beginning of every hour until just before sunset at 5pm. The path was driven manually and only approximately visited the same locations, as seen in the figure.

![Laser Localisation During All-day Test](image1.png)

![Visual Localisation During All-day Test](image2.png)

**Figure 5.13:** Path travelled every hour during the all-day experiment. The path from both the laser-localisation ground-truth system and the visual-localisation system (using the Nearest-edge observation function) are presented.

The two repetitions of the path were completed in approximately 3 minutes at an average velocity of over 1 m/s and maximum velocity of 3 m/s. The path was only an approximate path and was not repeated exactly. Table 5.4 reports the distance travelled, the overall rotation of the vehicle, and the maximum velocity. Table 5.5 presents the error at the end of the path, where the position error is the
length of the hypotenuse of the two horizontal translation errors. These errors are those from the filter whilst it was using the Nearest-edge observation function. The visual localiser can be seen to remove the accumulating odometry error. The odometry error on average was approximately 20m and 25 degrees sampled at the end of the path. The visual localiser’s position error for 7 of the 11 runs was within 0.5m and the maximum error was 1.86m at 10am. The rotation error for 8 out of the 11 runs was within 1 degree and the maximum error was 3.6 degrees recorded again at 10am.

Table 5.4: Details of Paths Travelled

<table>
<thead>
<tr>
<th>Time</th>
<th>Distance Travelled(m)</th>
<th>Total Rotation(deg)</th>
<th>Max Velocity(m/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:00</td>
<td>220</td>
<td>669</td>
<td>2.8</td>
</tr>
<tr>
<td>8:00</td>
<td>211</td>
<td>771</td>
<td>2.7</td>
</tr>
<tr>
<td>9:00</td>
<td>214</td>
<td>719</td>
<td>2.4</td>
</tr>
<tr>
<td>10:00</td>
<td>211</td>
<td>697</td>
<td>3.1</td>
</tr>
<tr>
<td>11:00</td>
<td>214</td>
<td>735</td>
<td>3.3</td>
</tr>
<tr>
<td>12:00</td>
<td>223</td>
<td>720</td>
<td>3.2</td>
</tr>
<tr>
<td>13:00</td>
<td>229</td>
<td>709</td>
<td>2.5</td>
</tr>
<tr>
<td>14:00</td>
<td>225</td>
<td>664</td>
<td>2.7</td>
</tr>
<tr>
<td>15:00</td>
<td>228</td>
<td>721</td>
<td>3.4</td>
</tr>
<tr>
<td>16:00</td>
<td>211</td>
<td>790</td>
<td>2.3</td>
</tr>
<tr>
<td>17:00</td>
<td>218</td>
<td>708</td>
<td>2.4</td>
</tr>
<tr>
<td>Overall</td>
<td>2176</td>
<td>7903</td>
<td>3.4</td>
</tr>
</tbody>
</table>

Table 5.5: Error at the End of Path

<table>
<thead>
<tr>
<th>Time</th>
<th>Final Position Error (m)</th>
<th>Final Rotation Error (deg)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Visual Localiser</td>
<td>Odometry</td>
</tr>
<tr>
<td>7:00</td>
<td>0.41</td>
<td>22.33</td>
</tr>
<tr>
<td>8:00</td>
<td>1.40</td>
<td>21.25</td>
</tr>
<tr>
<td>9:00</td>
<td>0.25</td>
<td>19.46</td>
</tr>
<tr>
<td>10:00</td>
<td>1.86</td>
<td>21.82</td>
</tr>
<tr>
<td>11:00</td>
<td>0.71</td>
<td>22.47</td>
</tr>
<tr>
<td>12:00</td>
<td>0.28</td>
<td>21.83</td>
</tr>
<tr>
<td>13:00</td>
<td>0.64</td>
<td>21.69</td>
</tr>
<tr>
<td>14:00</td>
<td>0.47</td>
<td>22.19</td>
</tr>
<tr>
<td>15:00</td>
<td>0.07</td>
<td>26.06</td>
</tr>
<tr>
<td>16:00</td>
<td>0.31</td>
<td>18.22</td>
</tr>
<tr>
<td>17:00</td>
<td>0.32</td>
<td>22.93</td>
</tr>
<tr>
<td>Mean</td>
<td>0.61</td>
<td>21.81</td>
</tr>
</tbody>
</table>
Figure 5.14 presents example images from the experiment (the video attachment labeled `allday.mpg` is a 10× sped-up movie of the entire all-day experiment including both the left and right video streams).

The example images show the successful tracking of the buildings across the entire day. The system can maintain track of the buildings even in the early morning and late afternoon when the low angle of the sun in the sky causes large lens flares blocking out most of the image. The exposure control algorithm samples pixels within the lens flare and, as a result, door edges not in the flare are left underexposed. The system does not fail in this situation because the roof edges are still visible and, more importantly, the camera facing in the other direction is facing away from the sun.

At 8am, the right camera entered a peculiar state where the captured images are grey and washed-out, this can be seen in the second row and second column of Figure 5.14. The camera was power-cycled when the vehicle completed the path and subsequently the images appeared normal. This peculiar camera state did increase the error in the system as seen in figure comparing the three observation functions, Figure 5.15. However the increase in error was not drastic and there were other times during the day where the error is similar.

### 5.7.3 Comparison of Observation Functions

A comparison is made between the three observation functions on the all-day data; the Per-edge function (Equation 5.7), the Nearest-edge function (Equation 5.11) and the Klein and Murray function from [9] (Equation 5.6).

The rates at which the three functions can be processed with the 500 particles are listed in Table 5.6. All timings are recorded on a laptop with a Intel dual core 2.33GHz CPU and a NVIDIA Quadro FX 350M GPU which is carried onboard the vehicle in the cabin and powered through the vehicle’s power supply. The video was captured at 15Hz and therefore only Klein and Murray’s metric is efficient enough to process the video at the full 15Hz. The other two functions are slower, and process a smaller number of frames. The software implementation of the Nearest-edge function is by no means optimal in terms of computation. Aspects of the software implementation that can be optimised are:

- square root calculations, moving from one sample point on the edge to another, could be avoided
- matrix multiplications, that project the 3D-edge points onto the image plane, could be optimised
- calls to access the pixels in edge-image could be more efficient
- conversions between normalised image coordinates and image pixel coordinates could be avoided.
Figure 5.14: Tracking results from all-day experiment. The two left columns of figures are from the morning hours and the two right columns are from the afternoon hours. The 3D-edge map is projected from each particle and overlaid on the camera image.

Table 5.6: Comparison of Likelihood Functions on Processing Rate

<table>
<thead>
<tr>
<th></th>
<th>Processing Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nearest-edge</td>
<td>4.41 Hz</td>
</tr>
<tr>
<td>Klein and Murray</td>
<td>15.27 Hz</td>
</tr>
<tr>
<td>Per-edge</td>
<td>8.33 Hz</td>
</tr>
</tbody>
</table>

The three different metrics were each evaluated against the laser-based localiser on the same logged sensor data from the all-day test. The statistics of the recorded errors over the whole day are shown in the box plots shown in Figure 5.15 and the statistics of the individual times are shown in Figure 5.16. The error is recorded at each iteration and during the entire day there were a total of 29378, 15633 and 7369 iterations for the Klein-Murray, Per-edge and Nearest-edge versions of the system respectively. The position error is the length of the hypotenuse of two horizontal translation errors.
In a previous paper [109] the Per-edge function was shown to be better at orientation estimation than Klein and Murray’s original function from [9]. This is because the Per-edge function enables better tracking of the smaller door edges and, therefore, better orientation estimation. However, in the experiments presented here, the orientation estimation of the two functions was more comparable. In earlier reported experiments [109], where the Per-edge function performed better, both functions were operated in an offline manner, one frame per-iteration, essentially being processed at the same frame rate. Here the experiments were run in an online manner, where if the function operated slower than frame rate then frames were dropped. The Klein and Murray function is more efficient, processing twice the number of frames than are processed by the Per-edge function and it is thought that this is why the orientation estimation is more comparable, as seen in Figure 5.16. Here the median orientation error over the whole day for the Klein-Murray function is 0.8 degrees and the Per-edge is 0.65 degrees.

The observation function proposed in this chapter, the Nearest-edge function, was by far the better of the three at initialising from particle distributions with large uncertainty, as seen earlier in Section 5.6.1. The improvements of the Nearest-edge function in pose estimation during operation is not quite as profound. After inspecting Figures 5.15 and 5.16 there are noticeable improvements in using the Nearest-edge function, especially in position error where the median error is 0.45 m as contrasted to 0.7 m and 0.8 m of the Per-edge and Klein-Murray functions, respectively. The orientation error of the three is similar where the median errors of the three are between 0.6 and 0.8 degrees. Importantly, the median pose estimation errors of the three functions are all sufficient to be used as a basis of navigating the vehicle around the site.

The system is not accurate enough to repeatedly perform precise load pick-ups. However, another vision system (using a camera pointing towards the load carrying point of the vehicle) has been developed specifically for this task [119], which can provide the level of accuracy required to pick-up a load.

The outliers seen in Figures 5.15 and 5.16 drift to several metres and several degrees of error, which is a concern for a vehicle navigating near buildings. These errors occur for short periods on the order of 5 to 10 seconds where one of two things occur with the filter’s particle distribution; either the distribution drifts into and out of a local maxima located away from the correct pose, or the filter forms multi-modal distributions – such as the one seen in Figures 5.9(c) and 5.9(g) – and the filter’s pose estimate (mean of the 5% highest weighted particles) switches between the distinct distributions. Most often these situations happen during sharp turns and the distribution returns to the correct pose estimate once the turn is complete – as reflected by the relatively low median errors and also reflected in the errors at the end of the path displayed in Table 5.5. An obvious solution to this issue is a
Figure 5.15: Boxplots of the three different observation functions during the all day test. The position error is the length of the hypotenuse of the horizontal translation errors. The boxes in the plots represent the interquartile range, the whiskers the minimum and maximum values, the line inside the boxes is the median and the crosses outside the whiskers are the outliers which are values more than 1.5 times outside the interquartile range.
5.7 Operation Experiments

Figure 5.16: Box plots of the pose estimation errors of the different times during the all day test of the three different observation functions. The position error is the length of the hypotenuse of the horizontal translation errors. The boxes in the plots represent the interquartile range, the whiskers are the minimum and maximum values, the line inside the boxes is the median and the crosses outside the whiskers are the outliers which are values more than 1.5 times outside the interquartile range. The error is calculated against the ground-truth pose (laser-scanner system) from the mean pose of the 5% highest weighted particles at each iteration of the filter. Some of the outliers are cropped off this figure in order to zoom in on the interquartile ranges, though all the outliers are visible in Figure 5.15.

more accurate odometry source that can better propagate the filter during quick turns, which will be investigated in future work.

5.7.4 Robustness of Image-point Feature Matching

The all day vehicle based experiment described in this section allows us to test the performance of image-point features compared to the performance of the 3D edge-based localisation algorithm.

Figures 5.17(a), 5.17(c) and 5.17(e) present an evaluation of the performance of SIFT on images recorded during the all-day test and Figures 5.17(b), 5.17(d) and 5.17(f) present an evaluation of the performance of SIFT on images recorded during the all-day test. Images captured at locations A, B and the pick-up point from 7 am to 5 pm were used to evaluate the SIFT technique. These locations can be seen in Figure 5.13. Examples of the images at location A can be seen in Figure 5.14, although for the SIFT matching the raw images were used without the overlaid edges. SIFT matches between images recorded at the same location were calculated one hour apart (Figure 5.17(a)), then two hours apart (Figure 5.17(c)), and finally, three hours apart (Figure 5.17(e)) and likewise for SURF features in Figures 5.17(b), 5.17(d) and 5.17(f).
The images were taken when the vehicle was close to the listed positions where the field-of-view was almost identical. However, because the vehicle was not in exactly the same location, the pixel location of the matches could not be used to differentiate inliers and outliers. Even if all matches are inliers, which manual inspections indicate they are not, there are not enough matches to evaluate pose from a map of features recorded three hours earlier. Many more matches would be needed to use a robust algorithm that removes outliers. Similar results are recorded for the SIFT and SURF features with the exception of slightly better performance of SURF for the 3 hour separation. These results indicate that SIFT and SURF are far from being lighting invariant and not ideal as the basis for a permanent map of the environment. The poor performance of the SIFT and SURF technique highlights the strength of the results the proposed 3D-edge system achieved on the same data, summarised in Table 5.5.

5.7.5 Rain Experiment

In addition to the difficulties of direct sunlight, vision systems operating outdoors also face the problems of rainy weather. Here a brief experiment showing that it is possible to operate the proposed system even with rain drops sitting on the lenses of the cameras disrupting the view. The vehicle is driven a path around the industrial courtyard whilst it is raining.

The estimated path can be seen in Figure 5.18(a) and the errors recorded against the ground truth can be seen in Table 5.7. The median position error is 0.5 m and the maximum error is 1.7 m. The median orientation error is 0.7 degrees and the maximum error 4.7 degrees recorded during a turn. These errors are similar in comparison to the experiments in clear conditions presented in earlier sections and are promising for the possibility of operating the visual-localisation system in a wide range of outdoor conditions. Example camera images are seen in Figure 5.18 where the rain drops on the lenses can be seen to dramatically obscure the view (a video attachment of this test is named rain.mp4).
Figure 5.17: Graphs of the number of SIFT and SURF matches between images taken at the same place one hour, two hours and three hours apart. The bottom axes shows when initial image was captured and groups results by the three different places.
Figure 5.18: Estimated path travelled during the rain experiment and example camera images. Rain drops on the lenses are noticeably disrupting the camera’s view.

Table 5.7: Errors recorded during the rain test

<table>
<thead>
<tr>
<th></th>
<th>Visual Localiser</th>
<th>Odometry</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Position error (m)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>0.45</td>
<td>1.54</td>
</tr>
<tr>
<td>25th Percentile</td>
<td>0.29</td>
<td>0.33</td>
</tr>
<tr>
<td>75th Percentile</td>
<td>0.63</td>
<td>4.80</td>
</tr>
<tr>
<td>Max</td>
<td>1.71</td>
<td>6.73</td>
</tr>
<tr>
<td><strong>Rotation error (deg)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>0.71</td>
<td>10.8</td>
</tr>
<tr>
<td>25th Percentile</td>
<td>0.32</td>
<td>2.2</td>
</tr>
<tr>
<td>75th Percentile</td>
<td>1.23</td>
<td>16.2</td>
</tr>
<tr>
<td>Max</td>
<td>4.79</td>
<td>22.3</td>
</tr>
</tbody>
</table>
This chapter has investigated the use of vision-based localisation for a ground vehicle operating in an outdoor industrial setting. The system developed here fits into the project’s over-arching plan of utilising several independent localisation systems based on different physical properties enabling a level of redundancy, confidence and robustness in localisation required for truly long term autonomous operation.

The visual localisation system uses a manually surveyed 3D-edge map of the permanent buildings in the environment. The sparse 3D-edge map is not specific to any lighting condition and includes only permanent information. Two sideways facing fish-eye cameras are calibrated and used to project the 3D-edge map onto the image plane for comparison with the detected edges in the camera-image. A particle filter is used to localise the vehicle and is proven reliable for initialisation given a very coarse initial pose estimate and also for extended operation in changing outdoor lighting conditions.

The first experiment of the localisation system evaluated the ability of the system to initialise given only a coarse estimate of the vehicle’s location and no indication of the vehicle’s initial orientation. The experiment provided the system 50 different opportunities to initialise the vehicle from a range of poses and across many different times during a sunny day. A novel observation function was compared to two existing functions on the 50 different examples and performed far better at initialising the filter. Not only was less error in the filter’s converged estimate but also was successful at converging far more often.

Three experiments were conducted to evaluate the performance of the localisation system in challenging outdoor conditions. One demonstrated the system on a vehicle while it was driven around a test site for an extended period where the vehicle covered a total distance of 1.5 km. The pose estimates from the vision-based localiser were compared to a laser-based localiser which acted as a ground truth. The pose of the vehicle was estimated by the visual localiser to an average position error of 0.44 m and average rotation error of 0.62 degrees over the 1.5 km path. The other experiment was carried out over the period of an entire day. At every hour during the day, the localisation system was evaluated as the vehicle was driven along a 220 metre path. The localisation system was evaluated and shown to successfully localise the vehicle in all the lighting conditions over the whole day. The final experiment evaluated the system operating in rainy weather with rain drops sitting on the lenses obscuring the view.

In addition to the particle filter an intelligent exposure control algorithm was presented that enabled operation in the highly non-uniform outdoor lighting conditions. The algorithm developed uses
knowledge of the scene to adjust the camera exposure and hence improve the quality of the important information in the image.

In conclusion, the combination of a sparse but high quality invariant map, a robust localisation algorithm and an intelligent exposure control algorithm all combined to produce dependable visual localisation in difficult outdoor lighting conditions.
The previous chapter presented a conventional visual localisation system; in that the process is to extract visual information from an image and localise the vehicle through comparison with a map. In this chapter a novel route for localisation is proposed which adds a light source model to the process to simulate the appearance of the scene in highly-dynamic and non-uniform lighting conditions. This is necessary when the appearance of the scene changes so drastically with the lighting that the visual system can not operate without knowledge of how the lighting is interacting with the scene.

Robotic vision systems operating in uncontrolled lighting conditions cannot escape the fact that an image of the object/environment changes when the lighting changes. Modern visual localisation approaches deal with this problem by attempting to extract visual information from images that is invariant to lighting, essentially, attempting to factor out the specific lighting condition from the visual process.
To date, there is no type of visual information extracted directly from images that is guaranteed to be decoupled from the effects of lighting for all situations. An image is highly dependent on the lighting in the scene.

This chapter describes a novel method which explicitly models the light source and incorporates the model within the system, instead of trying to factor out the effects of lighting. Even though an understanding of the lighting is not of primary interest for robotic applications, it can play an important role when processing the content of images. This is a new and exciting idea for visual localisation that has the potential to solve some challenging visual scenarios that were previously not possible with the conventional approaches that factor out lighting. Effects, such as the subtle shading on curved surfaces or specular reflections, are heavily dependent on the light direction and are difficult to deal with when using existing methodologies.

The framework proposed in this chapter uses a 3D-surface map in conjunction with a light model. A 3D-reflectance map has the advantage of being explicitly separate from the lighting and viewing pose. This allows the generation of synthetic images of the 3D-reflectance map according to pose hypotheses and a lighting model. Localisation is performed through comparisons between synthetic images and a real camera image.

The rest of the chapter progresses as follows; an overview of the target application, the motivations for this work from the human visual system, a discussion of the related work, the proposed framework, experiments proving the suitability of the method for the target robot application and then the conclusions.

### 6.1 System Design Motivations

#### 6.1.1 Motivations from the Application Environment

The specific task targeted in this chapter is localising an autonomous submarine from underwater structures that will perform inspection tasks, see Figure 6.1. It is desirable to have robots performing such tasks, when it is either too expensive or too dangerous for human divers. This target environment is visually difficult for a number of reasons. Firstly, poor water quality reduces visibility. Secondly, there is minimal or no natural lighting deep underwater, thus requiring an artificial light source to be mounted on the submarine. Thirdly, the visual appearance of the structure in this scenario is highly dependent on the incident angle of the light source. The light source is constantly moving and consequently the visual appearance of the structure varies dramatically. This is quite different from...
regular environments, where the light source is far less dynamic and also where there is a significant level of ambient lighting.

![Autonomous underwater vehicle](image)

(a) Autonomous underwater vehicle

(b) Camera image of an underwater structure

(c) Synthetic image rendered from a 3D-surface map of an underwater structure using a model of the light source

**Figure 6.1:** Underwater structure inspection with an autonomous underwater vehicle.

An example camera image taken of an underwater structure can be seen in Figure 6.1(b). The visibility is poor and the light mounted on the submarine illuminates only a part of the structure in view. The part of the structure that is illuminated is also illuminated non-uniformly. The dominant visible features are specular highlights, the locations of which vary as the submarine moves; this means it is
Visual Localisation Incorporating A Light Source Model

Figure 6.2: SIFT matches between two images from the submarine. The submarine translated 0.2 m between the two image captures. Only a single match was detected which is also an incorrect match, indicating that SIFT features are not usable for this application.

not feasible to employ a feature tracking technique to localise the vehicle, demonstrated in Figure 6.2. In this figure, two images are captured from the submarine where it has moved only 0.2 m between captures. The two images are processed with the popular SIFT matching technique. The result is one false match and no correct matches, indicating that the SIFT technique is not suitable for this application.

The underwater structure being navigated in this application is permanent and, therefore, it is possible to obtain an a priori map. The previously presented localisation technique in Chapter 5 tracks a 3D-edge map which is surveyed a priori. However, the 3D-edge approach would not be applicable here, since the edges of the structure cannot be reliably detected in the camera images because of the lighting (See the edge image comparison in Figure 6.3).

A localisation system based on stereo vision would also not to be a suitable option for this task. The specular reflections on the metal structure would appear different in each of the stereo pair and cause problems in estimating accurate disparities.

Literature surveys did not reveal any existing visual localisation methods which would clearly provide a solution to the difficult visual task. It may be possible to tailor an algorithm which specifically takes advantage of the simple shape and symmetry of the structure. However, a new algorithm would need to be designed for a new structure, or even another part of the same structure that has a different shape. Furthermore, it is not possible to tailor an algorithm for arbitrarily complex structures. There is nothing inherent, in the system developed in this chapter, which is specific to the structure (apart
6.1 System Design Motivations

Figure 6.3: Example comparing the edges extracted from a camera image and the actual expected edges. Top-Left: Camera image. Top-Right: Camera edges. Bottom-Left: Expected edges. Bottom-Right: Camera edges overlaid on expected edges. There are only minimal edges extracted from the camera. Furthermore the edges that are extracted are around the light reflections, not where the actual structural edges are, indicating that a 3D-edge map of the structure could not be tracked in such conditions.

from the input \textit{a priori} model of the structure). Therefore the system can generalise to other structures (if an \textit{a priori} model of the new structure is available).

The method presented in this chapter brings the novel use of a lighting model incorporated into a visual localisation system, and provides a solution to the difficult visual task. The framework presented in this chapter uses a 3D-surface map of the underwater structure, which is generated \textit{a priori}. The spotlight on board the submarine is modeled and used in conjunction with the 3D-surface map to generate synthetic images of the environment, see Figure 6.1(c). The synthetic images can be compared with the real camera image to localise the submarine. The specifics of this system are discussed in Section 6.3.
6.1.2 Motivations from the Human Visual System

A further motivation for the study in this chapter is the human visual system’s obvious understanding of the lighting in a scene. Simple experiments, such as Figure 6.4, which is from a study by Adelson [120], confirm that humans use a light model when evaluating the structure and surface properties of the scene. In the Figure, square A is painted darker than square B, however, the light reflected off the two is the same. This is due to the square B being in the shadow of the cylinder. To a human viewer it is easy to distinguish that the two squares actually have different surface properties. There is no doubt that this is because a human viewer has the ability to understand and calculate the 3D geometry, the light source and the corresponding shadows and shading from the scene.

Cornsweet [52] presents another study which illustrates the human visual system’s ability to estimate and use a model of the light source in the scene. The experimental setup Cornsweet used was of multiple surfaces of varying shades. He asked the test subject to identify which surfaces had similar shades. When the subject was aware of an obvious light shining on certain surfaces they could correctly adjust their perception of the reflectance properties of the surfaces. However, when a secret light was shone on surfaces without the subject being able to detect the new light source, their perception of surface reflectance was incorrect. Cornsweet’s experiment demonstrates that the human has an internal model of the light source, and when this model is estimated correctly, a human can use it to process and derive other information from the scene.

**Figure 6.4:** Optical illustration taken from [120]. A viewer can distinguish the two squares, ’A’ being darker than ’B’, even though the same amount of light is being reflected. This is a demonstration of the human visual understanding and estimation of scene properties; the 3D geometry, the light source and the surface reflectance.
Robotic vision has not yet reached this level of sophistication. Although a computer did generate the image in Figure 6.4 from given environment maps, robotic vision is not able to competently perform the reverse: estimating 3D-surface maps and a light model from the image. A human can do the reverse and the replication of this ability is a goal for ongoing research. This chapter discusses an intermediate step towards this goal, by bringing a light model inside the robotic visual process. This shows how to solve difficult visual tasks once light and surface models are obtained.

6.2 Related Work

There are some existing visual localisation works in underwater environments. Recent works by Petillot et al. [121] and Williams et al. [122] propose different methods for an underwater vehicle with downward facing cameras to localise the vehicle from the seabed using SIFT features and also to use a stereo unit to compute a dense 3D map. Fundamentally their work is not related to this chapter, because this chapter is looking at localisation from a structure, not from the seabed. Furthermore, as demonstrated earlier in Figure 6.2, an image-point feature approach like SIFT is not useful for the application presented in this chapter.

Kondo et al. present two methods of navigating underwater structures in [123] and [124]. In [123], two laser beams are directed at the structure which are detected in the camera images to triangulate the relative distance and orientation of the vehicle. In [124], Kondo et al. use a light stripe to illuminate a 2D profile of the structure which is detected in the camera images. A common feature of the two systems developed by Kondo et al. is the use of active lighting. In this chapter an artificial light source is also used, but, unlike the focused beams or light stripes of Kondo et al., the light source in this chapter is unfocused.

Stolkin et al. [125] present work for a submarine localising from a similar structure to the structure in the experiments in this chapter. Stolkin et al. also use an explicit 3D model of the structure and project the model onto the image plane to predict the shape of the structure. However, they model the structure as a single Gaussian distribution on the image intensity value, and do not take into consideration the non-uniformities of the structure’s appearance in the image.

The works of Gerard and Gagalowicz [21], Noyer et al. [46] and Ho and Jarvis [22] also present pose estimation systems based on 3D-surface maps. They perform correspondences between real and synthetic images that are more closely related to the work in this chapter. At the beginning of the 1990’s, Gagalowicz [48] predicted that, in the future, computer graphics technology will be used within robotic vision to solve hard visual problems; now, authors are beginning to develop techniques
Face Identification
1) Database of 3D face models
2) Fixed (known) pose of face w.r.t. camera
3) Render an image of each face model
4) Pick the face model which produces the best match to the real image

Visual Localisation
1) A single 3D environment model
2) Unknown pose of robot
3) Render images from pose hypotheses
4) Resample pose hypotheses according to a probability that is calculated by the match to the real image

Figure 6.5: Framework comparison between the two tasks of face identification and visual localisation

which make this concept possible. Both Noyer et al. [46] and Ho and Jarvis [22] estimate pose with a probabilistic particle filter, which is an efficient means of sampling the solution-space, whereas Gerard and Gagalowicz [21] present a more brute force evaluation of the solution-space. None of these 3D-surface based methods consider reflectance and lighting properties in their work – they only use textured 3D models – which do not generalize to any lighting condition. A textured model would not suffice for the application presented in this chapter, because the structure is made of one material and, therefore, is essentially texture-less. The images of the structure are also highly dependent on the light source, indicating that both reflectance properties of the structure and a light model should be known.
The work of Kee et al. [51] and Blicher et al. [50], in the domain of face identification, introduces the idea of using a 3D-surface model together with a light model. They show how to perform face identification in unknown lighting conditions by first estimating the current light source, then generating synthetic images of each face model using the estimated light source model. This idea of estimating and incorporating a light model has not yet been applied to visual localisation. This chapter aims to bring this idea from the domain of face identification to a visual localisation framework.

This related face recognition work uses a database of many different 3D-surface face models. A single fixed pose of the faces with respect to the camera is assumed, then multiple synthetic face hypothesis images are matched to the real image. Whereas, for the localisation work in this chapter, there is a single 3D-surface map of the environment and then many pose hypotheses are matched to the real image. The pose hypotheses with the best image match to the camera image from the robot will provide the pose estimate. The relationship between the face identification framework and this chapter’s proposed visual localisation framework is illustrated in Figure 6.5.

6.3 Localisation Framework

This section presents a localisation framework that incorporates a light source model which is used to estimate the pose of a camera relative to a known structure. The system is tested on a submarine with a forward looking camera and spotlight that are both mounted rigidly to the vehicle. Synthetic images of the structure are generated using a 3D-surface model of the structure and a lighting model for a given pose. The localisation is facilitated in a probabilistic multiple pose hypothesis framework – a particle filter – where observations are calculated by comparison between the real camera image and synthetic images generated from each pose hypothesis.

This section describes the aspects of the system in the following order:

1. submarine and camera-setup
2. 3D-surface model of the structure being navigated
3. synthetic image generation and light source model
4. particle filter used to estimate the vehicle’s pose
5. likelihood function used by the particle filter that compares real and synthetic images
6. implementation of the likelihood function using a graphics processing unit (GPU).
6.3.1 Submarine-setup

Figure 6.6(a) shows the submarine, the coordinate system, the camera and spotlight-setup. The submarine is a twin hull design, with detailed information on the vessel found in [126]. The coordinate system of the vehicle is defined by the positive $x$ axis pointing forwards, the positive $y$ axis to the port side and the positive $z$ axis upwards. The world coordinate system is also defined with an upwards positive $z$ axis. Two large thrusters at the rear of the vehicle provide the fast motions of the vehicle. These fast motions are predominantly in the horizontal translations in the world frame. Secondary thrusters produce smaller vertical movements, and also adjust the orientation of the vehicle. There are certain situations where rolling and/or pitching motions are produced, these occur when the vehicle is descending and moving forward and/or turning. In calm water a model of the thrusters can derive an estimate of the vehicle’s motion. However because the localisation system is designed for offshore operation, the ocean currents can cause significant disturbances to the model. The ocean current is assumed to be unknown and so this visual localisation system does not use a model of the thrusters.

A single forward-facing camera is used, the field-of-view of the camera, is shown as a blue viewing volume in Figure 6.6(a). The extrinsic pose of the camera is calculated with respect to the vehicle, and is facing along the vehicle’s positive $x$ axis. The camera’s intrinsic parameters are calibrated 	extit{a priori} and the camera images are undistorted at run-time by the calibrated distortion model. The 3D model of the structure is then projected directly onto the calibrated image plane.

Sunlight does not penetrate water as well as it penetrates air and, as a result, only a small percentage of sunlight reaches underwater environments that are tens of metres deep. In such dark underwater environments, artificial lighting is required. Here a torch is mounted on the vehicle facing forwards to illuminate part of the field-of-view of the camera. Figure 6.6(a) shows the spotlight as a yellow cone.

6.3.2 Oil-rig Scaffolding

The application of this work is an autonomous submarine inspecting underwater structures. However, the system presented here is designed in a generic manner, so that it would function with a range of different structures. The specific structure that the system is tested with is a steel scaffolding representative of structures found supporting offshore oil-rig platforms. A similar structure was investigated by [125] who also use an 	extit{a priori} model of the structure to predict its location and shape in the image. They attempt to model the structure’s reflectance as a single Gaussian on the image intensity value, not taking into consideration the non-uniformities of the structure’s appearance in the image. In this chapter’s work a model of the reflectance of the structure is used. The surface
6.3 Localisation Framework

Figure 6.6: (a) The submarine, the coordinate system, the camera-setup, the spotlight-setup and the type of structure being inspected by the submarine. The submarine has a forward facing camera with a field-of-view depicted in the figure by a pyramidal volume. The spotlight, depicted by a cone, is also facing forward and partially illuminates the field-of-view of the camera. (b) Illustrates the distance and angle attenuation parameters in the lighting model.

normals, the diffuse and specular reflectance properties all combined with an explicit light model (the light model is discussed later in Section 6.3.3) to generate synthetic images of the structure. The combination of the light source model and surface-reflectance model can accurately predict the lighting-dependent non-uniform appearance of the structure in the image, as seen in the synthetic image presented in Figure 6.1(c).

The structure is comprised of three steel tubes, linked together by smaller rungs at approximately 45 degree angles. This structure is modelled as a polygon mesh of each of the tubes by defining the number of slices and stacks in the mesh of each tube. Section 6.4.1 discusses the selection process.
that was performed to decide the ideal number of slices and stacks as a compromise between render
time and accuracy.

The mesh includes the reflectance properties required to calculate an interaction with the light source
model. The properties include the surface normal, and the diffuse and specular reflectance. Meshes
using such detailed surface properties have rarely been applied to visual localisation.

### 6.3.3 Synthetic Image Generation and Light Source Model

The photometric properties of the structure and a light source model are incorporated into the Blinn-
Phong reflectance model [127], which is chosen for its simplicity, speed of computation and prevalent
implementation on most graphics processors. The rendering is performed with per-fragment lighting
calculations, and shading interpolation to the pixels within the fragment.

The light source model includes the position and orientation of the light source with respect to the
vehicle’s frame and other parameters, including; the ambient, diffuse and specular components of the
light source and the angle attenuation and distance attenuation parameters.

The light is mounted slightly above and to the port side of the camera, and therefore the light is
oriented slightly to the starboard side of the forward direction vector, ensuring that a large portion of
the field-of-view receives light.

The distance and angle attenuation models are discussed in [128], and are a function of the following
factors: the distance the object being lit is from the spotlight, $\beta$, and the angle, $\alpha$, of the object from
the spotlight direction vector, $\lambda_v$. These two factors are shown in Figure 6.6(b).

All of the light source parameters are calibrated before run-time, but there is scope for future work to
investigate the calibration of these parameters online.

The angle attenuation parameter, $k_\alpha$, is employed to progressively reduce the brightness of the light
the farther away from the centre vector of the light cone. The angle from the centre of the light cone,
$\alpha$, as seen in Figure 6.6(b), is used to calculate the angle attenuation factor, $\phi$, as follows;

$$
\alpha = \cos^{-1}(\lambda_v \cdot \omega_L) \tag{6.1}
$$

$$
\phi = (\cos(\alpha))^{k_\alpha} \tag{6.2}
$$
Where $\lambda_v$ is the light centre vector, and $\omega_v$ is the light incident vector on the surface. The angle attenuation factor reduces the light reaching the object;

$$L' = \phi L$$

(6.3)

Where $L$ is the light intensity along the direction vector of the light source and $L'$ is the light intensity at angle $\alpha$ away from the centre vector.

### Reduced Visibility

Underwater environments often have poor visibility caused by the sedimentation of sand, silt, or microorganisms suspended in the water. These tiny particles reduce the visibility by absorbing and scattering light. The reduction of visibility is dependent on the observation distance and the density of the sediment in the water.

Figure 6.7 compares the images of a structure recorded in poor and good visibility conditions and also from different viewing distances. In Figure 6.7, (a) and (b) are taken in clear conditions, and (c) and (d) are taken in poor visibility conditions. The reduction in visibility between (a) and (c) is less than the reduction between (b) and (d), because of the higher density of the obstructing medium.

Our system provides two mechanisms to handle reduced visibility. Firstly, the light model is defined with distance attenuation (Equation 6.6) that has the effect of modelling sediment by reducing the power of light with viewing distance according to visibility parameters. Figure 6.8 presents a comparison between synthetic images rendered according to clear conditions and poor visibility conditions.

The distance attenuation parameters, $k_c$, $k_l$ and $k_q$, are used to calculate the distance attenuation factor, $\eta$, as follows;

$$\beta = |\lambda_p - \omega_p|$$

(6.4)

$$\eta = \frac{1}{k_c + k_l \beta + k_q \beta^2}$$

(6.5)

Where $\lambda_p$ is the position of the light source and $\omega_p$ is the position of the surface. The distance attenuation factor linearly reduces the light reaching the object;

$$L'' = \eta L'$$

(6.6)
Figure 6.7: Comparisons of real camera images between clear and poor visibility conditions. Also shown is a comparison of near and far observations. The farther the viewing distance the more the visibility reduces in poor conditions.

The second mechanism to provide robustness in poor visibility conditions is in the design of the synthetic-to-real image matching, described in Section 6.3.5. The matching is performed in the intensity-gradient domain which is more robust to synthetic images that do not precisely match the intensity values in the real camera images.

6.3.4 Particle Filter Localisation

The use of a particle filter is described in detail by [113]. The particle filter is a set of $N$ pose hypotheses (particles) $X_t = x_t^{(1)}, x_t^{(2)}, x_t^{(3)}, \ldots, x_t^{(N)}$. The set is sampled from the previous set $X_{t-1}$ using a propagation model $m_t$ (discussed in Section 6.3.4) and a corresponding set of weights
6.3 Localisation Framework

Figure 6.8: Comparison of synthetic images rendered of the structure in clear and poor visibility conditions. The poor visibility is modelled using a distance attenuation in the light source model.

(probabilities), \( W \). The weights are calculated from an observation of the environment, \( y \), as follows:

\[
W_k^{(n)} = p(y_k|x_k^{(n)})
\]  \hspace{1cm} (6.7)

the observation of the environment is a camera image, \( y \), which is compared with each pose particle \( x \) by rendering a synthetic image. The measurement of probability is provided from an image matching technique (discussed in Section 6.3.5). The concept is that a synthetic image generated from the particles nearest the correct pose will give the best matches with the real image. These particles are then the most likely to be resampled for the next iteration. The current pose estimate of the vehicle is extracted from the filter as the mean pose of the particles with the highest weights. Here, the mean of the 5% most highly weighted particles is used.

The vehicle can potentially move with six degrees of freedom, three translational \( t_x, t_y, t_z \) and three rotational \( r_x, r_y, r_z \). Without any information of where the vehicle is moving, the filter would require a large amount of particles to correctly track these motions and therefore impose a significant strain on processing power.

Propagation Model

A propagation model is defined to reduce the uncertainty in localisation.
The submarine has three different internal sensors that could help propagate the filter and therefore reduce the number of particles required in the filter. The sensors are an inertial measurement unit (IMU) which estimates pitch and roll angle ($r_x$ and $r_y$), a magnetic compass which estimates heading ($r_z$) and a pressure sensor which estimates depth ($t_z$). The IMU is evaluated using a highly precise industrial articulated arm to have a mean error within 0.5 degrees and it is thought that this would be representative of the accuracy of the IMU in the application environment. The magnetic compass and the pressure sensor are not used in this chapter’s experiments because their performance in the test laboratory is most likely not to be representative of its performance in the eventual application environment. On this basis only the IMU is used in the experiments presented in this chapter, ensuring that the system does not receive overly accurate sensor measurements in the test environment.

The IMU measurements have a sufficient level of accuracy in absolute measurements, and are taken explicitly as each particle’s roll and pitch estimate. The remaining four degrees of freedom are propagated using a constant velocity model calculated from a set of previous pose estimates extracted from the particle filter (as mentioned before, the pose estimates are extracted as a mean of the 5% highest weighted particles). The purpose of the velocity model is to propagate the particle filter in the correct direction of movement, to reduce the uncertainty of the particle distribution.

### 6.3.5 Gradient-domain Image Matching

The comparison process between the camera image and a synthetic image provides a likelihood measure for each particle in the filter. The works of [50, 21, 51, 46], also use image matching techniques which compare real and synthetic images. All of these techniques are variants of the intensity-domain mean absolute difference (MAD):

$$\text{MAD}(I_r, I_s) = \frac{1}{N} \sum_{k=0}^{P} |I_r(k) - I_s(k)|$$

where $N$ is the number of pixels in the image, $I_r$ is the real image and $I_s$ is the synthetic image.

This simple image matching technique assumes that it is possible to generate pixel values for the synthetic image that are equal or close to those in the real image. Matching intensity values directly requires the environment model and light model to be accurate estimates of the actual physical properties. It is difficult to model these properties and also difficult to simulate images at a rate fast enough for this framework.
For this reason, intensity-based matching is abandoned and, instead, a gradient-domain image matching technique is developed. The gradient-domain removes the absolute intensity levels whilst capturing the subtle shading in the environment – which is different to edge detectors that identify drastic boundaries of intensity.

The first step is to pass the real-image through a Gaussian filter, which removes the effects of noise. The synthetic and real images are then both passed through a Sobel operator, to generate gradient images in both the horizontal, $x$, and vertical, $y$, directions; $G_x$ and $G_y$ are the real Sobel images and $g_x$ and $g_y$ are the synthetic Sobel images. Example synthetic and real images are shown in Figure 6.9.

![Figure 6.9: Top Left: Real camera image. Top Right: Real Sobel image. Bottom Left: Synthetic image. Bottom Right: Synthetic Sobel Image. Horizontal gradient is shown in red and vertical in green; therefore pixels with diagonal gradients are yellow.](image)

To compare the real and synthetic Sobel images, it would be possible to turn these two images into a gradient magnitude image and a gradient orientation image, which would enable a more logical means for comparison. But to avoid the expensive square root and arc tan computations, the images are compared directly in $x$ and $y$ gradients.

147
Firstly, a sum is taken of the gradient magnitude in the real, $S_r$, and synthetic, $S_s$, images:

$$S_r = \sum_{p=0}^{N}(|G_x(p)| + |G_y(p)|)$$

$$S_s = \sum_{p=0}^{N}(|g_x(p)| + |g_y(p)|)$$

(6.9)

secondly, a sum of the difference in gradients between real and synthetic is calculated in each direction, $D_x, D_y$;

$$D_x = \sum_{p=0}^{N}(|G_x(p) - g_x(p)|)$$

$$D_y = \sum_{p=0}^{N}(|G_y(p) - g_y(p)|)$$

(6.10)

the final image matching score is derived as a ratio between the sum of gradient magnitude and the sum of gradient difference:

$$D_\mu = \frac{(S_r + S_s) - (D_x + D_y)}{S_r + S_s}$$

(6.11)

This result equates to the observation $y$ and the particle $x^{(n)}$ from Equation 6.7. The better the match between the images the larger the value of $D_\mu$. This score can be incorporated into the observation $y$ and the particle $x^{(n)}$ from Equation 6.7 as follows:

$$W_k^{(n)} = p(y_k | x_k^{(n)}) \propto e^{\rho D_\mu}$$

(6.12)

Where $\rho$ is a positive constant that adjusts the convergence of the particle filter.

### 6.3.6 GPU Implementation

Here the issue of implementation of the above gradient-domain image matching is considered. Some parts of the processing pipeline are fixed; the synthetic images are to be generated on a GPU and the camera images arrive on the CPU. Though, it is undecided which processor will perform the remainder of the processing. Timing tests of the various possible CPU and GPU implementations are conducted to decide the most efficient implementation of the above functions. Figure 6.10 presents mean timings from 1000 iterations of each of the possible operations. All timings are recorded on a
laptop with a Intel dual core 2.33GHz CPU and a NVIDIA Quadro FX 350M GPU, using 320 × 240 images. These timings are dependent on the model resolution, and the selection of an appropriate resolution is discussed in Section 6.4.1.

![GPU vs CPU Timings](image)

**Figure 6.10:** Processing timings for various CPU and GPU functions

Transferring each synthetic image from the GPU to the CPU and carrying out the matching on the CPU for each particle would involve the following functions, a GPU to CPU image transfer, CPU Sobel, CPU Sobel difference and a CPU mean image. The two big consumers of time are the CPU-based Sobel function and the differencing, taking 7.68 ms and 9.13 ms per image respectively.

GPU implementations of these functions are much quicker, 1.44 ms and 0.69 ms. Computing these functions on the GPU in turn causes another issue of getting the match result off the GPU. It is not possible to sum (or calculate the mean) an image directly on the GPU, as Equations 6.9 and 6.10 require. One possible method to extract the mean value of the difference image from the GPU is to recursively reduce the size of the image on the GPU, shifting the image back and forth in the texture memory, eventually arriving at a single value. The other option is to copy the resulting Sobel difference image to the CPU and compute the mean on the CPU. The two options take a surprisingly similar amount of time, the GPU averaging taking 1.38 ms and the image transfer with CPU averaging taking 1.34 ms. Choosing the more efficient option, the GPU Sobel differencing with CPU averaging option, the total processing time per particle is now 3.47 ms. However, the 1.34 ms to perform the GPU averaging is still 38% of the entire 3.47 ms.
To remove this final 1.34 ms from the processing time, an idea can be borrowed from [9]. The idea is to use an OpenGL occlusion query extension [114] to extract from the GPU a binary count of pixels in an image that pass a certain criteria. This would remove the need to copy the Sobel difference image to the CPU and calculate the mean. The criteria of Klein and Murray’s technique was edge-alignment between 3D-edge map and camera image. The criterion of this chapter’s system is pixels between the synthetic and real images that have matching gradients. However, with this occlusion query extension, only a binary result can be recorded per-pixel. This slightly alters the behaviour of Equations 6.9, 6.10 and 6.11. These need to be reformulated to incorporate binary per-pixel information, rather than scalar. The image summing Equation 6.9 for the GPU implementation becomes a binary count using a gradient magnitude threshold $\kappa$:

$$s_t = \sum_{p=0}^{N} \begin{cases} 
1, & \text{if } (G_x(p) > \kappa \lor G_y(p) > \kappa) \\
0, & \text{otherwise}
\end{cases}$$  \tag{6.13}

$\kappa$ is empirically set to 10% of the intensity range. The intensity range is 0 to 255 with 8-bit images output by the camera, but in the GPU programs, 0 to 1 float values are used. Equation 6.10 for the GPU implementation becomes a binary determination of whether two pixels have matching gradients:

$$d_x(p) = \begin{cases} 
\text{true}, & \text{if } (|g_x(p) - G_x(p)| < \gamma M_x(p)) \\
\text{false}, & \text{otherwise}
\end{cases}$$

$$d_y(p) = \begin{cases} 
\text{true}, & \text{if } (|g_y(p) - G_y(p)| < \gamma M_y(p)) \\
\text{false}, & \text{otherwise}
\end{cases}$$

$$d(p) = \begin{cases} 
1, & \text{if } ((m_x(p) > \kappa \land d_x(p)) \lor (m_y(p) > \kappa \land d_y(p))) \\
0, & \text{otherwise}
\end{cases}$$  \tag{6.14}

where $\gamma$ is a constant that adjusts how similar the synthetic and real gradients have to be classified as a match. $M_x(p)$ and $M_y(p)$ are the maximum values between the synthetic and real image and $m_x(p)$ and $m_y(p)$ are the minimum values:

$$M_x(p) = \max(|g_x(p)|, |G_x(p)|)$$  \tag{6.15}

$$M_y(p) = \max(|g_y(p)|, |G_y(p)|)$$  \tag{6.16}

$$m_x(p) = \min(|g_x(p)|, |G_x(p)|)$$  \tag{6.17}

$$m_y(p) = \min(|g_y(p)|, |G_y(p)|)$$  \tag{6.18}
These functions are implemented on the GPU using OpenGL’s shading language. Pixels where \( d(p) = 1 \) are allowed to pass through the GPU pipeline and pixels where \( d(p) = 0 \) are discarded. The OpenGL occlusion query provides the CPU with a sum of all pixels which pass through the pipeline:

\[
s = \sum_{p=0}^{N} d(p) \tag{6.19}
\]

The final synthetic image to real image match score is then calculated as:

\[
d_{\mu} = \frac{s}{s_t} \tag{6.20}
\]

Therefore, the eventual likelihood equation, Equation 6.7, becomes:

\[
W_{k}^{(n)} = p(y_k|x_k^{(n)}) \propto e^{\rho d_{\mu}} \tag{6.21}
\]

Where \( \rho \) is a positive constant that adjusts the convergence of the filter.

**Figure 6.11:** Probability density calculated with the scalar function from Equation 6.12 and compared to those calculated with the binary function from Equation 6.21. These calculations are recorded between a single camera image captured from an arbitrary pose and synthetic images rendered from different locations along the three translation axes.
This match score is based on binary per-pixel information and is a more efficient variant of Equation 6.11 which is based on scalar per-pixel information. This formulation saves 38% of the processing time that would otherwise have been spent deriving a mean value of the difference image. However, it is also important to ensure that it does not come with a reduction in accuracy and a test is devised to evaluate the comparative accuracy of the two functions. The test takes images from several arbitrary locations which are known with ground truth. Then probability density plots are sampled from each image using sets of images rendered along each of the pose axes (centred around the ground truth pose) using both Equation 6.12 and Equation 6.21. An example of one of the plots recorded at one of the camera locations, which is representative of the form of the plots at other locations, is presented in Figure 6.11. The figure illustrates almost identical form between the two formulations, confirming that the efficiency of the binary formulation does not compromise the accuracy in likelihood estimation.

6.4 Results

The eventual application environments are remote and inaccessible and, hence, conducting on-site experiments are expensive and it is not trivial to gather ground truth data there to evaluate the system. Therefore, the proof-of-concept experiments presented here are conducted in robotics laboratories which are easier to control and ground truth. The experiments are in environments that have little, to no, ambient light and are conducted with a real camera, a real spotlight and a real steel structure, all of these aspects of the experimental setup would be the same, or at least similar to those used in the target environment. It is a goal of future work to move the experimentation to offshore environments.

The first experiment (Experiment A) is conducted in a dark laboratory filled with fog, generated by a fog machine, which creates an environment with poor visibility, similar to what is expected in offshore underwater environments. The camera and light are mounted to the end effector of an industrial articulated arm, which serves to provide ground truth results for the tracking system. The results from Experiment A are presented in Section 6.4.3.

The second experiment (Experiment B) is conducted in a pool with no ambient light, the only light being the spotlight mounted on the submarine. The pool is clean and clear, so in terms of visibility this is an easier test for the system. However, the goal of Experiment B, is to move the submarine freely in all six degrees of freedom to evaluate if the IMU can suitably estimate the pitch and roll angles while the visual localisation system estimates the other four degrees of freedom. The experiment will also ascertain if there are disruptions to the passage of the light as it travels through several different mediums; the light initially travels through water, then through the glass window at the front of the
submarine, then through a small section of air before reaching the camera lens. The results from Experiment B are presented in Section 6.4.4.

Besides the camera, the only sensory information used is from an IMU, measuring roll and pitch angles. The system is responsible for the estimation of the remaining degrees of freedom.

First, the implementation and initialisation process of the particle filter is discussed in the following two sections.

6.4.1 Implementation Notes

Two cameras were used the experiments presented here. For the industrial articulated arm experiment, a Unibrain Fire-i 1.2 camera is used, and in the underwater experiment, the submarine is fitted with a Dragonfly Point Grey Research camera. Both cameras have a IEEE1394 connection and capture $640 \times 480$ images.

The spotlight is mounted slightly differently in the two experiments. Each time the location and orientation must be calculated for the purpose of rendering the synthetic images. In Experiment A the light is mounted on the end effector of the articulated arm 50 mm above the camera sensor and is mounted flush with a mounting plate, therefore, it is pointing directly along the principal axis of the camera. In Experiment B, the light is mounted approximately 100 mm to the port side of the camera and 50 mm above the camera, and in this situation the light is oriented approximately 5 degrees off the principal axis of the camera in the starboard direction.

The other light model parameters are calibrated manually. The angle attenuation parameter, $k_a$ from Equation 6.3, is 50. The linear distance attenuation parameter, $k_l$ from Equation 6.6, specific to millimetre units, for Experiment A in the foggy conditions is 0.0028 and in the clear pool of Experiment B is 0.001. The likelihood constant that adjusts the convergence of the particle filter, $\rho$ from Equation 6.21, is set to 40.

The camera and IMU data is processed, and integrated into the system using the DDX [129] and DDXVideo [130] frameworks. These frameworks enable the data to be logged and replayed at a later time at the original rate as if it was live sensor data.
Tests are conducted to calculate the desired number of polygons in the 3D-surface model. The decision on the final polygon count becomes a compromise between render time and accuracy of the resulting image. Worst case render times are calculated for a selection of different model resolutions (different polygon counts) using the light model defined in Section 6.3.3. The accuracy of the resulting synthetic images is also evaluated to select an appropriate resolution.

A reduced polygon count will reduce the accuracy of the synthetic image and hence, will presumably reduce the accuracy of the localisation. Therefore, a test is devised to evaluate varying accuracy with varying polygon counts. Probability density plots are generated according to image matching equations defined in Section 6.3.5 and are qualitatively analysed manually to judge the performance. The plots are calculated from a set of camera poses near the known ground truth pose.

Ideally, these plots will have peaks near the ground truth and slope off gradually. Counter to intuition, some of the lower resolution models tested produced a higher peak than the models with the very highest resolutions. This unexpected result is thought to be due to the positive effects of inter fragment shading-interpolation calculated by the graphics card. However, in the plot calculated from the very lowest resolution there is a local maxima located away from the ground truth, which may adversely affect the particle filter. This local maxima is not present in the plot of the second lowest resolution tested and, on this basis, the second lowest resolution is selected as the ideal resolution. It has a total of 3120 polygons and equates to a time of 0.325 ms to render a synthetic image for each particle.

This study of different resolutions sampled and generated probability density plots for only a small set of camera poses. However, the test is sufficient to have confidence that this particular resolution of model will not reduce performance in comparison with other resolutions and also provide efficiency in computation.

### 6.4.2 Initialisation

The first experiment is to evaluate the initialisation of the particle filter. An arbitrary camera pose is selected and the ground truth of the pose is measured and an image is recorded, seen in Figure 6.12. Realistic bounds of the uncertainty in the vehicle’s initial pose in the filter are set as follows: the roll and pitch angles to within two degrees, the heading to within 10 degrees and the depth, $t_z$, to within 200 mm and the other translations, $t_x$ and $t_y$, to within 1000 mm.
A large number of particles (2000) is used in the initial distribution of the particle filter to correctly account for the bounds of the 6 degrees-of-freedom. This initial distribution is too large to process the system near real-time frame rates. It is assumed the vehicle is stationary or near stationary during this phase, because it will take a number of seconds to converge the filter.

As in Chapter 5, Section 5.3.7, an adaptive particle filter is required which gradually reduces the number of particles to transition into a faster operating frame rate. Fox [115] presented a particle filter that adapts the number of samples in the distribution according to the spread of the distribution over the state space. Fox discretised the state space into bins and used the number of occupied bins to set the desired number of samples. In this chapter a simpler method is developed that does not require discretising the state space. Here the adaptive filter sets the number of particles according to the variance in the distribution, using the following equation:

\[ n_{t+1} = \max(n_0 \frac{v_t}{v_0}, n_d) \]  

(6.22)

where \( n_{t+1} \) is the number of particles for the next iteration, \( n_0 \) is the initial number of particles, \( n_d \) is the desired number of particles after full convergence, \( v_t \) is the current translational variance in the particle distribution and \( v_0 \) is the initial variance. The effect of this equation is to reduce the particle count as the filter converges, until the desired number of particles is reached to achieve a processing rate suitable for operation. Admittedly Fox’s method of adapting the particle filter will behave more reasonably in the case of multi-modal distributions that are widely separated but individually are tight distributions. In that case the proposed method will require a large number of particles because a single-mode and will mis-calculate a large variance in the distribution. Whereas Fox’s method captures the true variance of the multiple-modes in the distribution. However, the experiments in this chapter demonstrate that the proposed adaptive filter performs adequately for the application at hand.

The image used as input to begin the particle filter initialisation process is seen in Figure 6.12. Examples from the commencement of the initialisation can be seen in Figure 6.13 – where the centre lines of each particle are projected and overlaid over the raw images. The initial distribution is not centred around the correct pose, which is clearly visible, whereas by the second iteration, the projected centre lines are more densely overlaid on the structure. By the fifth iteration shown in Figure 6.14, the general form of the structure is seen in the projected lines and the convergence is completed by the 20th iteration. The final distribution can be noticed to be slightly off the ground truth. This is thought to be caused by one of two reasons; either there was inaccuracies in the camera/lighting model calibration or the filter converged at a local maxima near the ground truth pose.

The initialisation can be accelerated to be completed faster than 20 iterations by increasing the value of the constant in Equation 6.21 to increase the probability that fewer particles are resampled at each
iteration. However, this increases the probability of the filter converging to a local maxima. The other method to converge the filter faster is to utilise an IMU to estimate the roll and pitch axes and, therefore, the size of the initial distribution can be reduced. The convergence in Experiment B uses an IMU for estimating the roll and pitch angles and the convergence was completed in 5 iterations.

The initial distribution in this example is 2000 particles and is reduced with adaptive particle filtering to 800 particles – the required number needed in the following experiments. The goal of the system was to localise the vehicle with 200 particles which would have been processed at 2 Hz. However with 800 the frame rate to 0.5 Hz. At this this rate the submarine needs to be moving slowly for real-time operation. Future implementations with additional sensors, more powerful processing hardware or more optimized algorithms will quite possibly increase the frame rate of the system to allow operation on a fast moving submarine.

The following tracking experiments also use the same initialisation procedure, however detailed images and graphs of the procedure such as those in Figures 6.13 and 6.14 were not produced for initialisation of those experiments.
Figure 6.13: Initialisation of the particle filter, iterations 0, 2 and 4 are shown here. Left: images with centre lines of the rungs and columns of the structure projected from each particle. Right: graphs of the particles for the $t_x$ and $t_y$ axes.
Figure 6.14: Initialisation of the particle filter continued, iterations 5, 10 and 20 are shown here. Left: images with centre lines of the structure projected from each particle. Right: graphs of the particles for the $t_x$ and $t_y$ axes.
6.4 Results

6.4.3 Experiment A

After the convergence in the initialisation phase, the particle filter is ready to begin operation on a moving vehicle. An experiment is devised in a dark room that is filled with fog generated by a commercially available fog machine. The fog reduces the visibility in the room, similar to what is expected in underwater environments. A camera and light are mounted on the end effector of an industrial articulated arm. Photographs of the articulated arm and the steel structure are displayed in Figure 6.15.

The arm is an ABB IRB-2400/16 which has the ability to move its end effector into precise positions and orientations. The robot specifications are quoted with position repeatability of 0.06 mm, which is more than adequate to provide ground truth measurements.

The tracking experiment begins with the camera about 1500 mm away from the structure. The vehicle approaches the structure maintaining a constant heading. Once close to the structure it descends whilst rotating around the structure. A variety of viewing angles, from both near and far locations are provided by this path, giving a good set of images for the inspection task.

A set of 467 images are recorded during the path. The process to acquire the images is to instruct the robot to move the arm consecutively through poses on the path, when the robot reaches the demanded pose, the robot sends a software signal indicating its location. Then a camera image is recorded and tagged with the pose of the robot. The camera initially approaches the structure by moving upwards 500 mm (z-axis) and 750 mm towards the structure (y-axis). Then the camera descends whilst moving back and forth inspecting the structure. The total movement of the camera from the highest point to the lowest point is 1300 mm, it moves back and forth 500 mm in the y-axis (parallel to the structure) and rotates its heading 30 degrees repeatedly as it descends. Before finally rising 500 mm and moving 750 mm backwards to rest near the initial position. The entire length of the path is 5.2 m.

The heading of the camera moves port or starboard to keep the structure in view and the accumulated rotation is 211 degrees. The pitch and roll assumed known during this experiment, and in the following experiment (Experiment B) these angles are estimated by an IMU on the real submarine in Experiment B.

The particle filter is passed over the recorded data, and the estimated pose of the vehicle is extracted from the filter as the mean of the highest weighted particles. Here the mean pose of the top 5% most highly weighted particles is used. To maintain track of the structure in the experiment 800 particles are required. This is more particles than initially planned in the design phase and, consequently, the frame rate of the system is limited. The system operates at 0.5 Hz in the worst case during the experiment when there are more pixels and polygons of the structure in view.
Figure 6.15: Photos from the robotics laboratory used for Experiment A. The camera and light are mounted on an industrial articulated arm that serves to provide ground truth. The room is filled with fog to provide poor visibility conditions.
6.4 Results

This implementation of the system could only track a slow moving submarine in real time. The velocity of the vehicle would need to be at an average rate of 50 mm per second. This velocity is calculated considering the 0.5 Hz frame rate, the entire distance travelled of 5.2 m and the 467 images in the sequence. As mentioned before, with more optimal algorithms or additional sensors may well a future implementation of this system may be able to operate on a faster moving vehicle.

A figure of the operation of the particle filter during the tracking is shown in Figure 6.18. The figure presents the raw image, the synthetic image, an overlay of the projected centre lines of the structure from each particle and has a plot of the x and z translation axes of the particle filter, with the ground truth path (an attached video of this sequence depicting the raw video overlaid with the centre lines of the structure projected weighted mean pose can be found named \textit{auv.fogtest.mpg}).

Graphs of the ground truth and estimated pose of the camera are presented in Figure 6.17. The error of the pose estimated by visual localisation system is calculated from the ground truth and presented in the graph in Figure 6.16(a). An indication of the pose estimation uncertainty can be seen in standard deviation plot in Figure 6.16(b). A discussion of these results is presented later in Section 6.5.

6.4.4 Experiment B

The second experiment is conducted underwater in a pool, seen in Figure 6.19. The primary goal of this particular experiment is to evaluate if the visual localisation system in combination with the IMU can localise the submarine as it moves freely in all six degrees of freedom. The IMU will be responsible for measuring the pitch and roll angles whilst the visual localisation system is estimating the other four degrees of freedom. There is no ground truth data collected during this experiment, so therefore the accuracy is not quantitatively analysed. The performance of the system is checked manually by inspecting the projected centre lines of the structure from the estimated pose, and confirming they align correctly in the raw camera image.

The experiment is conducted at night in the dark with the only light coming from the spotlight mounted on the submarine. The pool is clean and clear, so in terms of visibility this is an easier test for the system. The previous experiment provided a more rigorous test of performance in poor visibility. The challenges unique to this experiment are the jerky movements in all six degrees of freedom and the potential disturbances in the passage of the light as it travels through several different mediums. The light which begins its journey from a waterproof spotlight mounted atop the submarine, travels through the water, is reflected off the structure and returns through the glass window at the front of the submarine, then through a small section of air before reaching the camera lens. This experiment
Figure 6.16: Error and uncertainty of the visual localisation system in Experiment A. The error is calculated against the ground truth provided by the articulated arm. The uncertainty is calculated as one standard deviation of the different dimensions of the particle distribution.
Figure 6.17: Estimated and ground truth path of the camera during Experiment A. The camera approaches the structure whilst rising. Then descends moving sideways back and forth whilst rotating to inspect the structure. Finally rising and returning to the starting position.

will give an indication of whether the system’s performance degrades as a result of the light’s passage being refracted as it passes through several mediums.

The experiment begins with the submarine approximately 1.5 m away from the structure. The submarine then approaches the structure, strafes side to side, descends and rotates around the structure. A second run of this experiment is conducted where similar motions are generated, and this time the vehicle is rotated around the opposite side of the structure. At the time of experimentation the submarine’s tail was broken and the thrusters were not available to control the movement of the submarine. Instead, the movement was supplied by a human in the water guiding the submarine along its path. The movements are not exactly the same as those produced by the submarine’s motors. However, there are rapid changes of acceleration and the movements are rougher than an untethered submarine. Therefore, the movements are more difficult to estimate than the movements normally produced by a coasting submarine and, hence, is a rigorous test for the system. The number of
Figure 6.18: Images showing the tracking of the oil rig structure in Experiment A. Top Left: Camera image with overlaid centre lines of the structure from each particle. Top Right: Camera image. Bottom Left: $t_x$ and $t_z$ axes of the particle filter, green line is ground truth, red line is estimated pose. Bottom right: Synthetic image of the structure.
6.5 Discussion

Figure 6.19: Photograph of the pool where Experiment B is conducted.

particles used to track the structure is 800, the same number as in the previous tracking experiment. Tracking images from the first run are presented in Figure 6.20.

The visual localisation system maintains accurate track of the structure for 440 frames in the first sequence. There are significant changes in scale, orientation and translation. The system then makes a mistake 440 frames into the sequence when one of the columns of the structure disappears behind another, then reappears on the other side. The system estimates the column reappearing on the same side, and does not recover from this error. The frame just before and just after the disappearing/reappearing column can be seen in the bottom row of Figure 6.20. This error is discussed in more detail in Section 6.5.1.

In the second sequence, the system maintains correct track of the structure until frame 340. Afterwards, the filter drifts in and out of convergence from frames 340 until the final frame, 600 (the video attachments labeled `auv_pooltest1.mpg` and `auv_pooltest2.mpg` show the tracking results of both sequences.).

6.5 Discussion

The experiments presented in this chapter are promising for the incorporation of a light source model within a visual localisation framework. The highlight of the results is that the system maintained track of the structure through a wide range of translations and rotations, which is a significant achievement considering the difficulty of the task. The poor visibility conditions in Experiment A reduced the
Figure 6.20: Images showing the tracking of the oil rig structure in Experiment B. Real camera image is overlaid with the centre lines of the structure projected from the estimated pose. Bottom left corner of each image is a synthetic rendering of the structure. The frame number is shown at the top left of each image.
quality of information available from the images but, even so, the system could track the structure. The other difficulty was the lack of other sensory information, which would normally assist the propagation of the particle filter; an IMU and the camera were the only forms of sensory information utilised.

Despite the successes of the system, there are still areas which must be improved before a practical and functioning localisation system can be implemented at an offshore underwater site as a basis for closed-loop navigation of such structures.

### 6.5.1 Divergence

In Experiment B the particle filter diverges in an ambiguous visual scenario where the rear column disappears behind the front column and reappears on the other side. The system estimates that the column appears on the same side that it disappeared and leads to a complete divergence of the filter and it does not recover.

The bottom right of Figure 6.20 shows this divergence, evident in the difference between the synthetic and real images, where the rear column has appeared in the real image on the opposite side to the synthetic image. The system is run several times over this same data, and occasionally the system correctly estimates that the rear column appears on the other side. However, it fails more times than it succeeds and a solution to this problem has not yet been implemented.

A potential solution to this problem is to scatter particles either side of the structure when confusing situations arise. This would create a multi-modal distribution which will hopefully converge to the correct solution when the visual information is less ambiguous. The particle framework can intrinsically form multi-modal distributions, however, in this failure situation the multi-modal distribution did not naturally occur. A higher-level layer operating above the filter that detects the ambiguous situation and scatters particles in multiple locations would enforce the desired multi-modal distribution.

### 6.5.2 Efficiency

Efficiency improvements are an important aspect of future work to produce a practical system for the target application. The current system could not operate real-time for a fast moving submarine. Effort was placed during design of this system to enable it to be processed quickly. The resolution of the 3D-surface was minimised, and the image processing and matching algorithms were implemented on a
GPU, which all brought the processing times much closer to real-time rates. The system was tested on a laptop computer, which uses small mobile processors and graphics hardware. The achieved frame rates are a good indication of the processing hardware that could be carried on-board a submarine. However, even with the processing abilities of a GPU there is more work required to reach real-time processing of a fast moving submarine.

The present implementation of the system would require the submarine to be moving slowly to be able to correctly localise in real time. During the Experiment A the camera moved 5.2 m and 467 images were captured. This sequence could be tracked at worst case at 0.5 Hz. The processing rate increased if there were less pixels and polygons of the structure in view. Considering the worst case rate the vehicle would need to be travelling at an average velocity of 50 mm per second. This slow velocity is not ideal for a practical implementation, because the slower the vehicle is moving, the less area of the structure can be inspected on the limited battery life of the submarine. Therefore, future work would need to optimise the processing of the system.

The two obvious methods to achieve higher efficiency, would be to further reduce the render times of the synthetic images, and also to reduce the number of particles in the filter. Reducing the render time could involve further reductions in the polygon counts, only passing sections of the model that are in view to the graphics pipeline, optimising the lighting calculations or with the use of improved/multiple GPUs. Future improvements to reduce the number of particles could include using a two-stage coarse-to-fine particle filter, such as used in the work of [9], or to develop a better propagation model.

The depth sensor and the magnetic compass are two sensors which could be included in the propagation model. However, it would need to be confirmed that these sensors are locally consistent in the desired environment (that is, if their inter-frame motion estimates are accurate). Another possible method of improving the propagation model is to use a visual odometry system, [131] present such an approach. Their system tracks the motion of the object in 2D space to assist the absolute localisation in 3D space. However, Marchand et al. track the silhouette of the object which may not be possible in this environment because the silhouette is not clearly visible. Also, other popular visual odometry approaches that use image-point features are not applicable here as image-keypoints cannot be reliably found and matched, as seen earlier in Figure 6.2. Developing a suitable visual odometry method for this difficult application environment would be an interesting future research topic.

### 6.5.3 Accuracy

In Experiment A the accuracy of the system was measured against the robot arm ground truth. The error graphs from the experiment can be seen in Figure 6.16(a). The mean position error in the x,
y and z axes respectively were 44 mm, 31 mm and 8 mm and the heading error was 4.21 degrees. Figure 6.16(b) shows the standard deviation of the particle distribution during the experiment. This plot gives an idea of the uncertainty in pose estimation. The standard deviation is on the order of the actual error of the system which might indicate that the particle filter is incorrectly tuned to be overly confident in its tight distribution and it is trapped in local minima away from the correct pose estimate. However another possibility is that the particle filter is actually correctly tuned but the imprecise lens calibration or light source calibration are the causes of the recorded inaccuracy.

The system was most accurate in the z axis, due to the smaller diagonal rungs that are useful for estimating the height of the camera. The worst case position accuracy was 200 mm and worst case heading was 10 degrees which is not favourably comparable to other visual localisation frameworks. However, other visual localisation systems are generally tested in easier scenes where features can be accurately located in the image. In the tests performed using this structure the visual features change as the vehicle moves and, therefore, it is difficult to precisely locate the structure in the image. However, considering the mean and worst case accuracy, the system is suitable to be used as a basis for closed-loop navigation in the inspection application.

One step for future work to improve the accuracy would be to improve the lighting model to generate a more precise synthetic images of the structure. Improvements to the lighting model could include an automated calibration procedure to accurately estimate the parameters or, other lighting models could also be investigated which may better represent the reflectance of the structure. One extension would be to model the light as an area light source instead of a point light source. Another extension would be to employ the model of [132] which uses two specular components, a specular spike and a specular lobe. However, such lighting models come with a computational burden and may not be able to be processed efficiently.

6.6 Chapter Summary

This chapter has presented a new approach to visual localisation with the use of a light source model and a 3D-surface map. A system is designed and implemented for the application of localising an autonomous submarine from underwater structures. The target environment has no natural light penetrating from the surface and an artificial light mounted on the vehicle is required. This spotlight causes the structure to be lit non-uniformly and causes the appearance to change dramatically as the vehicle moves. Through some brief tests, it seemed that the task was too difficult for the conventional edge-tracking or image-point feature tracking methods. These methods attempt to decouple the effects
of lighting from image-level information, which is not possible in this scenario because the image-level information is highly variant with the constantly moving light source.

The proposed method does not attempt to decouple the effects of lighting, instead uses a model of the light to predict how the appearance of the scene is changing with the moving light source. The developed system successfully localises the camera during a set of experiments using a real steel structure, camera and spotlight. The results display that this new idea of incorporating a light source model within a visual localisation framework can help to solve visually-challenging tasks that are too difficult for traditional methods that ignore the lighting.
This document has made an in-depth investigation of, and provided practical solutions to, the extremely challenging problem of visually localising in dynamic non-uniform lighting. After the literature survey presented in Chapter 2 it was clear many authors had identified lighting as a major problem, and most solutions relied on factoring-out the effects of lighting when extracting information from the image. Chapter 3 studied the extraction of image-level information in dynamic non-uniform lighting and presented experiments evaluating various types of information. The chapter raised doubts about whether a visual localisation system could autonomously build a visual map that could be used through major changes in the lighting.

The following chapter, investigated the dynamic range limitations of conventional cameras. A system was developed that collected and merged multiple differently-exposed images in real-time. This system increases the information available to a visual localisation system operating in highly non-uniform lighting.
Chapter 5 presented a system to localise an industrial vehicle operating in an outdoor industrial setting. The system used a surveyed 3D-edge map of the buildings, which included only invariant information and proved to be useful for localisation in the full range of lighting conditions experienced on a sunny day.

Chapter 6 developed a localisation system for a submarine inspecting structures in dark underwater environments. The system incorporated a lighting model of a spotlight mounted at the front of a submarine which accurately accounted for the dynamic non-uniform visual appearance of the structure.

The conclusion now lists the research contributions made throughout the document and finishes with a discussion linking the various pieces of work together into a cohesive argument with implications and future directions for visual localisation research.

### 7.1 Contributions

The research contributions are presented here with the most important contributions listed first.

- Chapter 6 developed a localisation technique that explicitly incorporates a light model to render realistic synthetic images of an environment to compare with the real camera images. Visual localisation systems rarely – if ever – have used a light model to predict the appearance of the scene.

- Chapter 5 designed a 3D-edge map localisation system using a particle filter, based on the work of Klein and Murray [9] and applied in an outdoor industrial environment. Two new likelihood functions were proposed and compared. One new function displayed favourable recovery properties in the situation of particle filter divergence. Successful localisation results were presented from experiments in a large range of challenging outdoor lighting conditions. In comparison, the alternative SIFT technique, often used in visual localisation systems, performed poorly on the same data, thus confirming the strength of the proposed system’s results.

- Another novel aspect of the 3D-edge system in Chapter 5 is an intelligent exposure control algorithm, which uses knowledge of where the pixels of interest are in the image and controls the intensity of these pixels of interest.

- Chapter 4 developed a real-time system which captures and merges multiple differently-exposed images for extending the dynamic range of any standard IEEE1394 digital camera.
• Section 3.4 evaluated the performance of the SIFT technique in outdoor environments. The experiments provided rigorous quantitative results revealing the sensitivities of the SIFT features to lighting.

• Section 3.2 studied the problem of correlating image intensities and object reflectance in dynamic lighting conditions and was framed in the context of image segmentation algorithms. An algorithm was proposed that used a novel concept of edge-based motion history images, which could ignore appearing and disappearing shadows cast by static objects.

• Section 3.3 studied various chrominance transforms in dynamic lighting. It was discovered that the most suitable transform is hue from HSV. Certain capture and transmission scenarios that should be avoided were identified because they result in unexpected chrominance variations.

7.2 Research Questions and Answers

What are the properties of the various types of low level visual information in dynamic non-uniform lighting? Chapter 3 studied low level visual with a number of different experiments in dynamic non-uniform lighting. The experiments demonstrated that trying to have constant perception of the properties of objects in the world is very difficult in uncontrolled lighting conditions. Specifically, descriptions based directly on intensity values are extremely sensitive to varying lighting. Relative-intensity descriptions eliminate the overall light level, but will still be disrupted by non-uniform lighting changes. Chrominance information can eliminate the lighting effects, besides spectral changes in the light source. However, the correct colour-space must be used and the camera capture and transmission must be controlled carefully. Furthermore, chrominance is only useful for brightly coloured objects/environments. Image-point features are found to be sensitive to outdoor lighting changes and there are not any obvious developments that would make the image-point descriptions more robust.

How can this thesis improve the performance of robotic vision systems operating in highly non-uniform outdoor lighting? Three sections of this thesis dealt with this question in three different ways. Chapter 4 developed a multiple-exposure technique to gather information across non-uniformly lit scenes that would otherwise be lost because it lay outside the sensitivity range of the camera. Chapter 5 presented an intelligent exposure technique which controlled the camera’s exposure according to the pertinent areas of the image; this confronted the non-uniform lighting by leaving unimportant areas of the image saturated or underexposed. Chapter 6 showed that, by
modelling the light source, the non-uniformities in the scene’s appearance caused by the lighting could be predicted and dealt with by a robotic visual system.

**How can visual localisation systems operate reliably in dynamic lighting conditions?** Chapter 3 firstly evaluated and found the categories of low-level visual information that are most stable in dynamic lighting. However, the Chapter also showed that the extraction of relatively stable low-level information is not the complete solution to the lighting problem. What is also important is to use an invariant map within a visual localisation system that is not specific to any lighting condition. Two types of invariant maps were utilised in this thesis; a 3D-edge map in Chapter 5 and a 3D-reflectance map in Chapter 6. The final – and most novel – method that this thesis proposed to enable visual localisation system to operate reliably in dynamic lighting conditions is to use a light source model within the process to predict the changing appearance of the scene, this was demonstrated in Chapter 6.

### 7.3 Discussion

The two main chapters of this thesis, Chapters 5 and 6 presented two different visual localisation systems that were developed for two completely different application environments. Neither system would be effective for the other application environment, showing that there is no one visual technique that can perform all localisation tasks. The 3D-edge map technique from Chapter 5 is not appropriate for the underwater structure localisation in Section 6.1.1, because the edges of the underwater structures are not reliably detected in the image, owing to the limited lighting provided by the spotlight. Likewise, the 3D-surface map technique incorporating a light model is not suitable for the building environment, because it is suited to localising from objects with subtle shading, not from large, flat, homogeneously-coloured building walls.

Despite the differences, the systems had common elements. Both used maps based on the 3D properties of their respective environments that included permanent information that was not specific to any lighting condition. Both systems localise from the map in a top-down approach, in that they both project the invariant map onto the image and discover the best alignment. This is opposed to the bottom-up approach, which extracts features and simultaneously localises and builds a map. The bottom-up approach relies on the learned features to be robust to lighting changes, otherwise the map will be specific to the lighting condition under which it was created and is no longer useful when the lighting changes significantly. Section 3.4 presented experiments with SIFT features – commonly used for visual localisation systems – that revealed that these image-point features cannot be reliably matched after significant light changes. The results indicate that geometric localisation based on SIFT
features would not be dependable over long periods, because of the very few feature-matching inliers found with images taken only hours apart. With the top-down approach, it has been shown that, even with lighting specific information extracted from the image, robust localisation can be achieved with a good invariant map in place. However, it must be noted that these types of maps are not applicable in environments that are unknown. They do not attempt to solve the mapping problem, they are pure localisation systems. The targeted robotic applications presented in this thesis have been in environments with permanent structures for which it is acceptable to manually create maps \textit{a priori}.

Another common attribute of the two localisation systems presented is the use of relative-intensity information. Relative-intensity information is intrinsically robust to uniform lighting changes, as first discussed in Chapter 3, though the problem of non-uniform lighting must be solved. Each system approaches the problems of using relative-intensity in non-uniform lighting in a different way. Chapter 5 uses a surveyed map of the structural edges of the buildings that has only permanent edges in the map. This enables the spurious edges detected at shadow boundaries to be ignored with a robust tracking filter. Furthermore, the intelligent exposure control technique developed adjusts the camera exposure to detect building edges that are in shadows. Chapter 6 deals with non-uniform lighting, by explicitly incorporating a light model to replicate the non-uniform lighting with accurate synthetic images of the environment used to compare with the real camera images. Section 3.2 presents another method of using relative-intensity information, for the application of detecting moving pedestrians and vehicles during fluctuating lighting. Through the use of motion history, non-uniform lighting changes that cast shadows from static objects will be ignored. However, quickly-moving shadow edges will be detected, but these are assumed to be associated with moving pedestrians or vehicles. All of these relative-intensity systems described above present a variety of different solutions to deal appropriately with the non-uniform lighting issue.

Exposure control, an often overlooked aspect in robotic vision systems, is covered by two sections of this thesis to improve the quality of visual information provided by cameras. Chapter 4 developed a multiple-exposure technique to increase the dynamic range of standard IEEE1394 cameras. Then, in Section 5.4, it was shown that with knowledge of the areas of interest in the image, it is possible to intelligently control a single exposure to improve the pertinent visual information.

Another important implication that this thesis has for visual localisation research is from the experiments revealing the sensitivities that the commonly-used image-point features have to outdoor lighting changes. Localisation techniques based on image-point features can perform powerful and impressive functionalities; however, it must be acknowledged that the features are separate from lighting conditions. Shadows and shading are tangled into their descriptions. Therefore, maps of image-point features collected under one lighting condition may not be useful when the lighting changes. It is not obvious how to extend the techniques to completely factor out lighting and it would
appear that autonomously extracting lighting invariant information from images remains, for the most part, an unsolved problem. This thesis has shown that the actual 3D properties of an environment (for example the 3D structural edges or a 3D surface map) are an alternative basis for robust visual localisation systems, with a key aspect being that these representations are completely separate from lighting conditions. Furthermore, image-point features and other image-level descriptions are limited in what they reveal about the environment. It is no doubt advantageous for computer vision systems to have 3D representations, not only for localisation, but also for robots to effectively interact with the world.

Another direction that this thesis recommends to future visual localisation research is the incorporation of a light source within the visual framework. This is a logical idea, because images are fundamentally measurements of light; therefore, it is unreasonable to expect that lighting can be factored out when processing images. Modelling the light source seems to be an essential factor in enabling future robotic vision systems to not only be robust to lighting, but also to actually understand the effects of lighting. Humans have an obvious knowledge of the light source and its interaction with the 3D geometry in a scene and can use this understanding to their advantage. A light source model within robotic vision systems will enable a more intimate connection between the visual system and the light measurements of the camera image.

The two localisation systems developed during this thesis have shown practical methods to perform successful visual localisation in extremely challenging dynamic non-uniform lighting. The promising results of both systems have brought interest from industry partners and will continue to be developed into the future, with the aim to turn these into functioning systems in the field.
References


Redundant Localisation

This chapter will describe experiences with two independent localisation systems, one using a laser scanner and the other using cameras, both developed for a large forklift type vehicle. A navigation system is developed to take input from the multiple localisation systems, and compare and arbitrate the input information to decide the appropriate navigation actions. The system is designed to combine these independent and unrelated localisation systems to give redundancy, which provides improved levels of dependability in operation. Results from experiments demonstrate continued performance of the vehicle when one localisation system degrades. This system was published in [118].

A.1 Related work

The area of dependability in outdoor (terrestrial) field robots did not gain significant attention until the early-to-mid 2000’s when the DARPA Grand Challenge events were held. See the Journal of Field
Redundant Localisation

Robotics Special Issue on the DARPA Grand Challenge ([133] and [134]) for a comprehensive set of papers by some of the successful and unsuccessful teams. The more recent DARPA Urban Challenge again focused teams into the area of dependability. However, the research results from this event have yet to be published and it is unclear as to whether any teams had redundant localisation systems. The first Grand Challenges have relied heavily on the use of GPS – which is something that cannot be used much of the time in the application of interest. In building environments, urban canyons cause reductions satellite coverage and GPS signal is not received reliably – if at all – on-board the vehicle.

The use of multiple sensors for localisation has been well researched and has been widely applied in the area of field robotics. For the most part, multiple sensor information is fused to form a single localisation system. This approach can improve the situation where the sensors individually cannot provide enough information for continuous and/or reliable localisation. In multi-sensor data fusion (Figure A.1(a)), the aim is to provide a single localisation system with a more complete set of input sensor data by fusing all available sensor information. However, when sensors fail, provide erroneous readings or have a limited view of the world, the accuracy and confidence of the localisation estimates degrade. Hence, the data fusion process is not focused on providing redundancy. Examples of sensor fusion in the literature are Majumder et al. [100] who fuse sonar and camera information for an
underwater vehicle, Miura et al. [101] fuse laser and stereo camera data into an obstacle map and Arras and Tomatis [102] fuse tracked features extracted from laser and camera data into a single EKF.

In this work lasers and cameras are used in outdoor environments; however, most previous laser and camera systems were developed for indoor situations, Newman et al. [25] is one of the only examples of outdoor localisation using both a laser and a camera. They use these sensors in a single localisation system, whereas this work presents two individual and unrelated localisation systems.

Examples of redundant sensing are high-integrity inertial sensing with pairs of inertial sensors to achieve high levels of reliability[135]. In this case, the sensors are duplicated and the sensor readings themselves compared (that is, they are not completely unrelated sensors). Scheding et al. [136] use multiple redundant sensors, a laser and a gyro to identify system faults. They assert that the probability of identical sensor fault modes is much lower using sensors with different physical principals, as opposed to using multiples of the same sensor. In their work the only sensor that can perform localisation is the laser, the gyro is just measuring motion and detecting faults. A similar technique is used in standard GPS processing engines that use more than the required minimum number of satellites to obtain a reliable position estimate.

The use of multiple, and often independent sensing and control systems has been widely used by spacecraft engineers since the beginning of human spaceflight. [137] describes the Saturn V guidance and control system that used complete subsystem duplication in many of its operations to achieve the required reliability. Similarly, the Space Shuttle exploits four primary computers at the heart of its fly-by-wire control system[138, 139]. It is thought that it will require similar practices to achieve the required reliability for certain field robotics applications – especially those of heavy machinery operating in human populated environments.

### A.2 Proposed System

The system presented here uses multiple sensors in an alternative and more dependable manner. The unrelated sensors are used by independent localisation systems, which provide redundancy to the navigation system. To the authors’ knowledge, the use of multiple sensors for multiple-independent-localisation systems has rarely been investigated in the area of field robotics research. Figure A.1 shows the fundamental difference in this approach. A system using independent localisation systems (Figure A.1(b)), uses an additional process – an arbitrator or comparator – to monitor the pose estimates from the multiple localisation systems and cross-checks them for consistency. It is only in very recent times that field roboticists have had the ability to compare pose estimates from
independent localisation systems, because until now it has been difficult to deploy more than one working localisation system on a field robot. Two high-reliability localisation systems have been developed to work in large outdoor industrial environments. One is a system based on the use of multiple 2D laser scanners and reflective beacons. The other uses a vision system to estimate the vehicle’s pose based on an *a priori* edge map of the buildings in the environment. Both these systems have been operating on the autonomous HMC and both can be used to guide the HMC around the test site. The remainder of this chapter describes the two localisation systems and shows the results from experiments where one of the localisation systems (the laser-based system) was disabled.

### A.2.1 Laser Localisation

The laser localisation system for the vehicle, previously published in [99], comprises four laser rangefinders placed on the four corners of the vehicle (Figure A.2(a)). The lasers detect reflective beacons that are placed around the environment on the posts and walls at surveyed locations (Figure A.2(b)). The beacons’ locations are used to triangulate the vehicle’s position to a site-referenced (global) coordinate system when detected.

![Laser setup and coverage](image1.png)

**Figure A.2:** Laser localisation system. Four lasers are placed at each corner of the vehicle and (a) demonstrates the coverage of the laser scans, (b) is an image of the industrial setting, reflective strips on the posts and walls of the site are detected by the laser scanners.
A.2 Proposed System

A.2.2 Camera Localisation

The vision-based localisation system described in detail in Chapter 5, uses two fish-eye cameras mounted sideways on the vehicle. A sparse 3D-edge-map of the building environment can be tracked in the camera images giving the pose of the vehicle. The 3D-edge map tracking is facilitated in a particle filter.

The incoming fish-eye images are first corrected for distortion and then passed through an edge filter. The 3D-edge-map can then be projected onto the undistorted edge images for direct comparison. The comparison score is calculated as the alignment between the 3D-edge map and the camera edge-image and is computed for every particle. This comparison score gives an indication of the likelihood a particle is at or near the correct pose estimate and is used by the filter to re-sample the particles each iteration.

A confidence measure of whether the particle filter is still correctly tracking the buildings is calculated as the mean alignment score of the best 5% (with the highest likelihood) of the particles. This confidence value is used by the vehicle’s navigation system for decisions regarding when and how to use the vision-based localisation.

A.2.3 Navigation System

The vision and laser localisers are two independent systems that are each able to provide the inputs for navigation of the vehicle. However, when combined together, there is redundancy in localisation. An independent process is used – an arbitrator or comparator – which accepts these two inputs, evaluates a confidence in each system and determines the appropriate pose estimate for the navigation system (Figure A.1(b)). Four modes are proposed for vehicle operation post-failure of the localisation system:

1. Termination of operation to an immediate safe state (fail-safe behaviour)
2. Termination of operation where the vehicle defaults to “limp home” type navigation after which it can be investigated and repaired
3. Continued operation with a degradation in operational performance (for example, slower speed operation)

Ultimately, Mode 4 is the target of the research outlined in this chapter, where vehicles can continue to operate, even after one localisation system fails. The system failure would then be repaired by
Redundant Localisation

a maintenance crew at the next available opportunity. However, even the development of Mode 1 is a challenge, because this requires that the AGV system correctly detects the localisation system failure. Apart from a partial or complete sensor failure, it can be difficult detecting when a single localisation system becomes inaccurate. In particular, cases where a localisation system’s pose estimate slowly drifts from the correct solution, a second, and independent localisation system is required for comparison. This sort of functionality is much better performed with multiple localisation systems.

A.2.4 Localisation Arbitration

The autonomous HMC’s primary global localiser is the laser-based system. It provides accuracies within 100 mm for navigation and crucible operations – docking and drop-off. Both systems provide internal estimates of their confidence of operations which has been determined to be reasonable metrics that reflect the system’s accuracy. To provide redundancy and reliability in localisation, an arbitration module is used to provide the most accurate pose estimate by comparing the systems and making decisions about which is providing the highest confidence and most accurate estimates to pass through to the navigation system. Furthermore, the arbitrator also passes through the confidence value which the navigation system can use as a dynamic guide for setting the upper limits for velocity control – if the confidence in localisation is low, then the vehicle’s maximum forward and reverse speeds should be reduced (Mode 3 in Section A.2.3).

The input parameters for the arbitrator are the vision and laser pose estimates ($v_{\text{pose}}$ and $l_{\text{pose}}$), and their confidence measures ($v_{\text{conf}}$ and $l_{\text{conf}}$). Currently, the arbitrator will always choose $l_{\text{pose}}$ and $l_{\text{conf}}$ as output values unless either of the following cases occur:

1. $l_{\text{conf}}$ and $v_{\text{conf}}$ are low
2. $l_{\text{conf}}$ is low and $v_{\text{conf}}$ is high

The choice of low and high thresholds for these evaluations are currently empirically determined based on previous testing of individual systems. In case 1, the navigation system will slow the vehicle to a stop, because it has assumed inaccurate localisation from all available sources (Mode 1 in Section A.2.3). In case 2, the arbitrator will switch to the visual localiser and use its confidence and pose estimate as an output (Mode 4 in Section A.2.3).
To test the idea, an experimental trial in devised which the HMC was tasked to perform a normal crucible pickup, transit and drop off. The trials were run outside in a compound area – surrounded by large industrial sheds as shown in Figure A.4(a). The mission of the HMC was to:

1. From a parked position, drive to the crucible position (known from a previous autonomous mission)
2. Pick up the crucible
3. While carrying the crucible, complete a circuit of the compound area
4. Return to the crucible pick-up position and drop off the crucible
5. Return to the parking position.
Redundant Localisation

The failure mode tested was a loss of the laser-based localiser triggered by a simulated power failure to the lasers. The failure was timed to occur during the transit phase of the HMC (just after the crucible pick-up). A simple arbitrator was created (Figure A.3) that took input from the two localisation systems and output the pose of the system that it trusted most. The output value was then used by the navigation system. It did this by continuously monitoring the confidence values of the localisation systems ($l_{conf}$ and $v_{conf}$). For this experiment, the arbitrator was programmed to trust the laser localiser more than the vision localiser as long as the laser localiser’s confidence was greater than 0.4 (on a scale of 0.0 to 1.0). If the laser localiser’s confidence dropped below this threshold then the arbitrator used the vision localiser’s output and continued the mission. It should be noted that because each system is independent, then so are the confidence values. Both systems report confidence in the 0.0 to 1.0 range but confidence estimates are not calibrated. The speed of the vehicle changes depending on the confidence value of the arbitrator as follows:

\[
1.0 \leq aconf > 0.75, \quad speed = 100\%
\]
\[
0.75 < aconf > 0.5, \quad speed = 75\%
\]
\[
0.5 < aconf > 0.3, \quad speed = 50\%
\]
\[
0.3 < aconf > 0.0, \quad speed = 0\%
\]

Figure A.4(b) shows the confidence values plotted against time. The initial high confidence values in the Figure are derived from the laser localiser. The simulated laser failure occurred at approximately the 80 second point in the Figure. At this point, the arbitrator switched to the vision localiser. The confidence values from that point onwards are from the vision localiser. It is clear from Figure A.4(b) that when the vision localiser takes over, it reports relatively low confidence values between the 80 and 160 second marks. During this phase, the HMC was performing a full 360 degree loop around the compound. This was the most difficult section for the vision-system to track the buildings owing to the rapid change in vehicle orientation. While the confidence reported by the vision system was low, the navigation system moderated the speed of the vehicle as per the rules indicated above. The HMC had completed the turn at the 160 second point, and from then until the end of the mission, the vision localiser reported a high confidence. As a result, the HMC was able to complete its mission at full speed, dropping off the crucible and returning to the parking position. Figure A.4(c) shows the HMC’s path and indicates the location of the parking position, crucible pickup and drop-off point and the location of the simulated laser power failure (the video attachment named redundant-localisation.mpg shows the entire sequence of this experiment).
A.4 Conclusion

(a) The compound area navigated during the experiment

(b) Localiser confidence values during the experiment

(c) The path of the HMC during the experiment. The vehicle uses the laser-based system until it fails then the remainder of the path is the vision-based system.

Figure A.4: Figures showing the experiment area and experimental results
A.4 Conclusion

This chapter detailed the development of an autonomous navigation system for a heavy duty industrial transport application – that of the movement of molten aluminium around a smelter. The system incorporates a number of independent localisation systems. The idea of sensor fusion in field robotics has been widely exploited over the past decade. However, the motivation for this sensor fusion has often been to achieve the reliable operation of a single localisation system. Algorithms, sensors and computing hardware has now reached a point where it is possible to deploy multiple localisation systems that can work throughout a mobile robot’s environment. This finally allows the opportunity to compare the estimates from these systems and start to investigate the best methods of choosing the most reliable, the most trusted or developing methods to optimally combine them to achieve a more dependable outcome.
Many of the experimental results of this thesis are best presented in video format. Following are lists of the videos that accompany this thesis on a CD. The videos are organised into sub-directories according to the corresponding chapter in the document. The videos are encoded in MPEG-1 and MPEG-4 standards.
Table B.1: Table of video files attached to this thesis on a CD

<table>
<thead>
<tr>
<th>Chapter-3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>mhiWalk.mpg</td>
<td>Examples of the edge-motion-history-images generated using</td>
</tr>
<tr>
<td>mhiVan.mpg</td>
<td>the algorithm described in Section 3.2.3</td>
</tr>
<tr>
<td>outputWalk.mpg</td>
<td>Foreground segmentation videos generated using the region</td>
</tr>
<tr>
<td>outputVan.mpg</td>
<td>growing segmentation process described in Section 3.2.3. The</td>
</tr>
<tr>
<td></td>
<td>final segmented foreground regions are displayed in blue.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter-4</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>regular.mpg</td>
<td>Video showing limitations of a single exposure</td>
</tr>
<tr>
<td>merge-not-registered.mpg</td>
<td>Merged chrominance and edge information from multiple-different-</td>
</tr>
<tr>
<td></td>
<td>exposure system without image registration</td>
</tr>
<tr>
<td>merge-registered.mpg</td>
<td>Merged chrominance and edge information from multiple-different-</td>
</tr>
<tr>
<td></td>
<td>exposure system using the image registration process described</td>
</tr>
<tr>
<td></td>
<td>in Section 4.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter-5</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>initialisation.mp4</td>
<td>Example showing the initialisation process described in</td>
</tr>
<tr>
<td></td>
<td>Section 5.6.</td>
</tr>
<tr>
<td>extended-operation.mp4</td>
<td>Sped-up video showing localisation from the 30 minute extended</td>
</tr>
<tr>
<td></td>
<td>operation experiment presented in Section 5.7.1. Localised 3D-</td>
</tr>
<tr>
<td></td>
<td>edge map is projected in red.</td>
</tr>
<tr>
<td>allday.mp4</td>
<td>Sped-up video showing the localisation of the vehicle every</td>
</tr>
<tr>
<td></td>
<td>hour from 7 am to 5 pm, described in Section 5.7.2. Images</td>
</tr>
<tr>
<td></td>
<td>from both left and right facing cameras are shown. The 3D-</td>
</tr>
<tr>
<td></td>
<td>edge map is projected from each particle in green and mean</td>
</tr>
<tr>
<td></td>
<td>pose of top 5% most highly weighted particles is projected in</td>
</tr>
<tr>
<td></td>
<td>white.</td>
</tr>
<tr>
<td>rain.mp4</td>
<td>Video showing the visual localisation system operating in rainy</td>
</tr>
<tr>
<td></td>
<td>weather from the experiment in Section 5.7.5</td>
</tr>
</tbody>
</table>

200
Table B.2: Table of video files attached to this thesis on a CD

<table>
<thead>
<tr>
<th>Chapter-6</th>
<th>Appendix-A</th>
</tr>
</thead>
<tbody>
<tr>
<td>fogtest.mpg</td>
<td>Experiment of closed-loop autonomous control of an industrial vehicle picking up, carrying and then putting down a load. During the experiment the laser localisation system failed and the navigation system continued seamless operation by using the 3D-edge map visual localisation system; demonstrating the success of the arbitration system. The flashing green light on the front of the vehicle indicates when the navigation system is closed-loop using the visual localisation system. This experiment was described in Section A.3.</td>
</tr>
<tr>
<td>pooltest1.mpg</td>
<td></td>
</tr>
<tr>
<td>pooltest2.mpg</td>
<td></td>
</tr>
<tr>
<td>In all three videos the centre lines of the structure are projected from the estimated pose in yellow and in the bottom left corner is a synthetic image rendered from the estimated pose</td>
<td></td>
</tr>
</tbody>
</table>

| pooltest1.mpg | Underwater localisation experiment presented in Section 6.4.4 |
| pooltest2.mpg | Underwater localisation experiment presented in Section 6.4.4 |

Localisation experiment in poor visibility conditions from Section 6.4.3