AUTOMATED SUBSTRATE CLASSIFICATION IN CARDIGAN BAY THROUGH ANALYSIS OF UNDERWATER VIDEO

Author: Matthew Pugh

Supervisors: Dr. Bernard Tiddeman Dr. Hannah Dee

A thesis submitted in fulfilment of the requirements for the degree of Master of Philosophy

in the

Vision, Graphics & Visualisation Group
Department of Computer Science

March 2015
Mandatory Layout of Declaration/Statements

Word Count of thesis: 29,906

DECLARATION
This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed ................................................................. (candidate)
Date 3 March 2015

STATEMENT 1
This thesis is the result of my own investigations, except where otherwise stated. Where *correction services* have been used, the extent and nature of the correction is clearly marked in a footnote(s).

Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed ................................................................. (candidate)
Date 3 March 2015

[*this refers to the extent to which the text has been corrected by others]

STATEMENT 2
I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ................................................................. (candidate)
Date 3 March 2015

NB: Candidates on whose behalf a bar on access (hard copy) has been approved by the University should use the following version of Statement 2:

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loans after expiry of a bar on access approved by Aberystwyth University.

Signed ................................................................. (candidate)
Date 3 March 2015
“The condition of our maps of the deep sea, however, is about comparable to the maps of North America at the time of the Lewis and Clark Expedition at the beginning of the Nineteenth Century.”

H. B. Stewart
Much of the world’s surface is covered by the oceans. According to the National Oceanic and Atmospheric Administration (NOAA), 99% of the oceans remain unexplored today. This research looks at developing methods to aid in the ecological surveillance of the sea-floor, specifically targeted at the Cardigan Bay area.

We have developed a system for obtaining ground truth data on underwater videos by marine biologists. This data is then used to analyse the videos of the underwater environments using computer vision (including histograms and scene decomposition using wavelets) and machine learning (SVM, kNN) methods, in an attempt to automatically identify visually distinct underwater substrates. Using these trained classifiers, we test whether these generalised methods are able to deal with the known-to-be difficult domain of underwater video.

Initial results are presented for a select subset of videos from two regions of Cardigan Bay, using both full-frame and patch-based analysis, showing the successes and limitations of the approaches. A discussion of the results is given, including the theoretical and practical constraints observed, and the areas upon which this work may be further developed towards a system with which unseen video may be reliably classified automatically.
Acknowledgements

This research is gratefully funded by Knowledge Economy Skills Scholarships (KESS), which is part-funded by the European Social Fund (ESF) through the European Union's Convergence Programme (West Wales and the Valleys) administered by the Welsh Government.

Huge gratitude is given to both Dr. Tiddeman and Dr. Dee for their excellent supervision, advice and constant push to further what could be done with the project, and for inspiring me to carry on in this field at PhD level.

I would like to thank the FoCB for their commitment to this research and continued input, without Phil’s constant input, availability and wisdom, this project would not have been possible. We also wish to thank Dr. Hiddink and his research team at Bangor University for kindly allowing us use of their video collection.

A final thanks is given to both examiners, Dr. Labrosse and Dr. Rosin for their guidance and feedback.
# Contents

Abstract vi

Acknowledgements vii

Contents viii

List of Figures xi

List of Tables xiii

1 Introduction 1
   1.1 Outline ................................. 2
   1.2 Background & Funding ............................. 5
   1.2.1 Friends of Cardigan Bay (FoCB) .................... 5
   1.2.2 Knowledge Economy Skills Scholarships (KESS) ............ 5
   1.3 Scope .................................. 6
   1.4 Goals .................................. 6

2 Literature Review 8
   2.1 Cardigan Bay .................................. 9
   2.1.1 Bathymetry .................................. 10
   2.2 The Physics of Water & Light ....................... 10
   2.2.1 Energy Loss & Attenuation ...................... 12
   2.2.2 Optics ................................... 14
   2.3 Image Acquisition & Formation ..................... 15
   2.4 Key Groups & Research Interests ................. 17
   2.4.1 Fish4Knowledge .............................. 17
   2.4.2 Woods Hole Oceanographic Institute ............... 18
   2.4.3 CIRS / VICOROB ............................ 18
   2.4.4 Monterey Bay Aquarium Research Institute ........... 19
   2.4.5 National Oceanography Centre ............... 19
   2.4.6 IFREMER ................................ 20
   2.5 Capturing Underwater Video ...................... 21
   2.5.1 Sensors .................................. 21
   2.5.2 Manual Capture ............................. 22
2.5.3 Remotely Operated Vehicles .............................................. 23
2.5.4 Autonomous Underwater Vehicles .................................... 27
2.6 Computer-Based Processing ................................................. 31
  2.6.1 Image Processing .......................................................... 31
  2.6.2 Computer Vision ........................................................... 32
  2.6.3 Video Processing & Annotation ........................................ 34
2.7 Applications of Underwater Video .......................................... 42
  2.7.1 Wildlife & Fauna Monitoring ........................................... 42
  2.7.2 Seabed Maps & Bathymetry ............................................. 44
  2.7.3 Humanities, Official & Cultural Heritage .............................. 49
2.8 Discussion, State of the Art & Areas of Expansion ..................... 51

3 Datasets .............................................................................. 52
  3.1 Classification Schema ....................................................... 53
  3.2 Available Datasets ............................................................ 56
    3.2.1 Bangor ................................................................. 56
    3.2.2 Fish4Knowledge ....................................................... 56
    3.2.3 TRIDENT ............................................................ 57
    3.2.4 Friends of Cardigan Bay ............................................. 57
  3.3 Chosen Data ................................................................. 57
    3.3.1 Bangor University ..................................................... 58
    3.3.2 Friends of Cardigan Bay ............................................. 61
  3.4 Original Data Capture ..................................................... 64
    3.4.1 Refinement of FoCB Technique ..................................... 65
    3.4.2 ROVs ................................................................. 66
  3.5 Dataset Index ............................................................... 69

4 Methodology ...................................................................... 71
  4.1 Platform ........................................................................ 72
    4.1.1 Language ................................................................. 72
    4.1.2 Libraries ................................................................. 73
    4.1.3 Hardware ............................................................... 73
  4.2 Performance .................................................................. 75
    4.2.1 Big Data ................................................................. 75
    4.2.2 Libraries ................................................................. 75
    4.2.3 Workload Distribution ................................................ 76
    4.2.4 Parallelisation .......................................................... 76
  4.3 Frame Processing ............................................................ 78
    4.3.1 Full-Frame ............................................................. 78
    4.3.2 Patch-Based ........................................................... 79
    4.3.3 Features ................................................................. 80
  4.4 Ground Truth Acquisition ................................................ 87
  4.5 Frame Selection ............................................................. 88
  4.6 Machine Learning ........................................................... 88
    4.6.1 Support Vector Machines ............................................ 89
    4.6.2 k-Nearest-Neighbour ................................................ 91
    4.6.3 Validation ............................................................... 92
## Contents

4.7 Closing Comments ........................................... 92

5 Results ....................................................... 94
5.1 Full Frame .................................................... 95
5.2 Patch-Based .................................................. 95
5.3 Discussion .................................................... 98
  5.3.1 Comparing Datasets & Schemas, with Chance & Recognition ...... 98
  5.3.2 Perspective Issues ......................................... 100
  5.3.3 Overfitting ................................................. 101
  5.3.4 Incorrect SVM Kernel Tuning ................................ 101
5.4 Plots .......................................................... 104

6 Conclusion ..................................................... 107
6.1 Restrospective ................................................ 109
  6.1.1 Technology .................................................. 109
  6.1.2 Data Collection Timeline .................................... 110
6.2 Future Work .................................................. 111
  6.2.1 Contextual Classification of Seabed Fauna .................... 111
  6.2.2 Frame Selection ............................................ 112
    6.2.2.1 Euclidean Distance .................................. 113
    6.2.2.2 Multi-resolution .................................... 113
    6.2.2.3 Subset ................................................. 113
  6.2.3 Cross-Testing, New Data & Meta-data ........................ 114
  6.2.4 High-dimensional Data Reduction ............................ 115
  6.2.5 ROV automation .......................................... 116
6.3 Critical ....................................................... 116
6.4 Final Words .................................................. 117

A Public Outputs ................................................. 118
A.1 Posters ........................................................ 118

Bibliography ...................................................... 121
List of Figures

1.1 Alexander the Great's first reported underwater voyage ............... 2
2.1 West Wales map - Cardigan Bay ........................................ 10
2.2 The Sarnau of Cardigan Bay ............................................. 11
2.3 Water column & light ..................................................... 13
2.4 Example images illustrating different illumination patterns .......... 16
2.5 ROV control room and live monitoring capabilities ................. 23
2.6 WHOI JASON ROV navigation ........................................... 24
2.7 Different classes of ROVs ................................................. 25
2.8 AUV design example ..................................................... 27
2.9 Examples of AUVs in use in current research ......................... 28
2.10 AUV-captured image ..................................................... 29
2.11 Example of a difficult underwater scene ............................. 32
2.12 Sun-flicker caustic effects on the sea-floor ......................... 33
2.13 SIFT applied to a frame from an underwater video .............. 35
2.14 MBARI VARS software .................................................. 36
2.15 MOSSE tracking of fish in underwater video ...................... 37
2.16 Detections in MBARI’s AVED system ............................... 38
2.17 Example of Optical Flow ............................................... 41
2.18 Colour seabed mosaic .................................................... 44
2.19 Seabed mosaic bathymetric map ..................................... 46
2.20 SLIC Superpixel segmentation of a clown-fish in sea-fauna ...... 48
2.21 Sea-floor mosaics of a Roman shipwreck .......................... 49
3.1 CCW class similarities .................................................. 53
3.2 Examples of class 0 ..................................................... 55
3.3 Sea-floor images for the planning and analysis of the Bangor experiments .... 58
3.4 Bangor data collection method ........................................ 59
3.5 Impact debris from trawler method .................................... 60
3.6 Bangor SITE2 stills ..................................................... 61
3.7 Bangor SITE3 stills ..................................................... 62
3.8 Brittle star obscuring the sea-floor in SITE7 ...................... 63
3.9 FoCB1 stills, showing the harbour environment and reefs near Aberystwyth ... 64
3.10 FoCB2 stills, taken around the Sarn located near Borth and Clarach ... 65
3.11 FoCB data collection method ......................................... 66
3.12 Concept for semi-autonomous method of towed capture .......... 67
3.13 Hydroyview submersible camera ................................... 68
3.14 OpenROV & VideoRay .................................................. 68
List of Figures

3.15 Our OpenROV .................................................. 69
4.1 Distributed work architecture ........................................ 77
4.2 Full-frame analysis artefacts ......................................... 79
4.3 Example viable patches .............................................. 80
4.4 Example of Retinex algorithm use. ................................... 82
4.5 Gabor filter impulse response examples .............................. 84
4.6 Local Binary Pattern (LBP) illustration ............................. 86
4.7 Ground truth interface. ............................................... 87
4.8 Dividing videos into regions of sea-floor substrates over time .......... 88
4.9 SVM margin illustration .............................................. 89
4.10 SVM RBF & linear kernels ........................................... 90
4.11 Training & testing methodology for both datasets .................. 93

5.1 Problems in the FoCB dataset for patch-based experiments ........ 97
5.2 Issues of poor boundaries in CCW schema ......................... 99
5.3 A typical scene from the Bangor dataset, with perspective issues superimposed. 100
5.4 Illustrative example of overfitting on training data .................. 102
5.5 Bangor full-frame results using the simplified schema. ............... 104
5.6 Bangor full-frame results using the CCW schema ..................... 104
5.7 FoCB full-frame results using the simplified schema .................. 105
5.8 FoCB full-frame results using the CCW schema ..................... 105
5.9 Simplified full-frame results ......................................... 106
5.10 Countryside Council for Wales full-frame results .................... 106

6.1 Images captured using OpenROV ................................. 110
6.2 Distribution of training vectors in the feature space ................. 112
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Abridged definitions of the zones present in the water column</td>
<td>12</td>
</tr>
<tr>
<td>3.1</td>
<td>Classes from the CCW schema selected as candidates</td>
<td>54</td>
</tr>
<tr>
<td>3.2</td>
<td>Our texture-based classification schema</td>
<td>55</td>
</tr>
<tr>
<td>3.3</td>
<td>Video dataset technical details</td>
<td>69</td>
</tr>
<tr>
<td>3.4</td>
<td>Video files index</td>
<td>70</td>
</tr>
<tr>
<td>3.5</td>
<td>Video &amp; ground truth relations</td>
<td>70</td>
</tr>
<tr>
<td>4.1</td>
<td>Libraries in use on the project</td>
<td>74</td>
</tr>
<tr>
<td>4.2</td>
<td>Computer hardware used</td>
<td>74</td>
</tr>
<tr>
<td>5.1</td>
<td>Results: full-frame using the Countryside Council for Wales schema</td>
<td>95</td>
</tr>
<tr>
<td>5.2</td>
<td>Results: full-frame using the simplified schema</td>
<td>96</td>
</tr>
<tr>
<td>5.3</td>
<td>Patch-based results using kNN</td>
<td>97</td>
</tr>
<tr>
<td>5.4</td>
<td>Patch-based results using SVCLIN</td>
<td>98</td>
</tr>
</tbody>
</table>
Dedicated to all those who gave me continued words of encouragement.
Chapter 1

Introduction

In this chapter, we discuss the motivation for the project — the problems faced by marine biologists and scientists in gaining (and understanding) data from underwater environments.

We clearly define the scope and objectives of this project, and how it meets the requirements set forth by our industry and financial sponsors.
1.1 Outline

Throughout history, mankind have pushed the boundaries of exploration and discovery. They have taken to the seas on ever-more-impressive vessels, voyaging past the then-accepted edge of the world, finding new lands, forming new colonies and nations. Our quest for exploration and understanding has driven scientific endeavour and technological enhancement, making the previously unreachable areas ripe for new discovery.
We have been able to send humans into space, and keep them functioning away from their home planet for extended periods of time on ageing space stations, sent men to the moon some 40 years ago, and currently have our eyes on our neighbouring red planet, Mars.

Despite all this, on our home planet – whose surface is predominately water based, the oceans of this planet continue to go largely unexplored. The subaquatic environment plays host to a number of challenges for would-be explorers, hampering any discovery due to the requisite technological capabilities and human-expert skills required. One of the earliest underwater explorations is depicted in Figure 1.1.

How can we use the advances in different technologies to begin to solve these issues? Is it now within the realms of possibility that the technological and computation requirements for these tasks have become such a commodity that, with some originality and careful planning, consumer-level available systems could be used to make a contribution in this area. Since taking an interest in underwater habitats, divers have been venturing into the depths and recording their findings. As time went by, their recordings became more empirical – through imaging solutions that meant land-bound marine biologists could also see the underwater environments and form their own conclusions. This process has, historically, been extremely expensive in terms of money, time and opportunity. Finding skilled divers, in the right location, with the right weather and conditions has been a difficult number of variables to balance, leading to the inability to conduct mass surveying of underwater environments.

Today, a number of fully, semi and non-automated techniques are used to gather and process underwater detail. Marine biology, as a subject, has become an area of intense interest for both the computer science and robotics communities, enabling data retrieval missions of ever-increasing duration and fidelity. These applications take the form of Automated Underwater Vehicles (AUVs) - low power semi-to-fully autonomous vehicles equipped with a variety of sensors, through to Remote Operated Vehicles (ROVs) in which an operator above the water directly controls the equipment, motion and sensors of a submersible robot.

Owing to these increasing possibilities, certain large research groups have formed around this area of research. Notable examples thereof include Monterey Bay Aquarium Research Institute (MBARI), Woods Hole Oceanographic Institute (WHOI) and the University of Honolulu in Hawaii. As impressive as these engineering feats are, they require astronomically high budgets, with Southampton’s National Oceanography Centre (NOC) having a custom ROV whose development cost is estimated at £5,000,000.

This is before considering the plethora of bespoke and custom systems in automated analysis of this data. The techniques described until this point have been purely in the obtaining of data for analysis. The process of annotating and analysing the data, to judge benthic densities and populous, has still fallen to the hands of overworked marine biologists (and, typically,
PhD students whose time is free). This workload increases dramatically as we solve the issue of data acquisition; backlogs for manual analysis are created. However, this presents a new opportunity – applying computer processing in an automated manner to either partially, or fully analyse underwater videos for key data. The automated discoveries can then be relayed to the human analyst in whatever form is deemed appropriate.

Computer vision techniques are used everyday to automate, and extend, the collection of knowledge we have of the world around us. An increasing amount of research is being conducted in the application of computer vision techniques to underwater environments. Existing algorithms and works have looked into solving typical real world scenarios, and have some basis in the human visual system in terms of scene understanding. If one were to be able to create a system capable of object recognition as well as the human visual system, it would surely be a breakthrough for any such systems.

As it stands, applying these techniques to video and image data collected from underwater environments present a number of very unique and specific challenges. From illumination, through to organic particles and turbidity, it would appear that it is not as simple as taking what exists and plugging it into a different domain. However, the existing techniques are well documented, as are current endeavours in adapting existing, and developing new approaches to this problem. In conjunction with this research, with the advent of cheaper, open ROVs and the overall lowering cost-of-entry into conducting such research, a number of organisations and groups are leveraging these techniques to increase scientific understanding of the world below using whatever means are at their disposal.

This is what this project plans to do – to evaluate a number of different videos collected in the Cardigan Bay area, use computer vision and machine learning methods and investigate ROVs to start developing an automated system for processing these videos. We are doing this to specifically gain understanding of the Cardigan Bay sea-floor area which, due to its location, has gone relatively unexplored until this point.

We present a computer vision system based on recognising seabed substrates in underwater video, using a number of different metrics and parameters. We use this in conjunction with different models and methods from the machine learning community, comparing the effectiveness of these techniques when applied to this domain.

We also discuss bespoke, original methods of solving some of the issues in data acquisition and processing – where it has been possible, experiments based on these ideas are presented.
1.2 Background & Funding

This project owes its existence to Aberystwyth University, KESS & FoCB, together they have created an opportunity to conduct this research.

1.2.1 Friends of Cardigan Bay (FoCB)

This organisation of volunteers, chaired by Philip Hughes, works to raise awareness of the ecology, fauna and challenges of Cardigan Bay. They perform this through regular monitoring, organised watches, scientific research and engaging community involvement.

Their work is used to influence governmental policy, educate the public and provide an impartial counter argument to any commercial lobbying which may threaten the bay’s animals, habitat or fisheries. Further to this, the work enables them to react quickly to, or outright prevent, new threats.

Land and boat-based surveys carried out by FoCB monitor and investigate the important seabird and marine mammal populations within Cardigan Bay. Ongoing research aims to investigate the habitats and species that exist within Cardigan Bay through collaborations with organisations such as Aberystwyth University and Seasearch. This information contributes to management strategies for Cardigan Bay so that humans can co-exist more harmoniously with this special marine environment.

– Friends of Cardigan Bay’s web-based press release

FoCB have previously successfully partnered with Aberystwyth University in 2008 for an MSc project looking at Bottle-nose dolphins.

1.2.2 Knowledge Economy Skills Scholarships (KESS)

KESS is a programme amongst a number of Welsh Universities, headed by Bangor University that uses ESF to fund postgraduate research in a number of domains which relate directly to Wales and its business interests.

The funding requires an industrial partner and sponsor, with a set project being advertised that has a real-world application on research to Wales. This project is one such KESS project,
that satisfies the requirement by providing a look at how scientific methods in computer science may be applied to the surveying of the Cardigan Bay area.

### 1.3 Scope

It should be noted that this project was undertaken over a 12-month term, starting in October 2013, with the institution focusing on a review of existing literature for the title of MPhil, and not specifically original research. That withstanding, this project has aimed to cover as much material as possible within the small timeframe, and begin conducting original research.

Portions of this work have been either accepted, or already presented at the following events:

- **Marine Imaging Workshop - Southampton NOC - April 2014** Oral presentation on the work-to-date, including the full-frame results on Bangor and FoCB datasets.
- **RIVIC - Swansea - May 2014** Oral presentation on the work to date, including the full-frame results and preliminary patch-based results.
- **PGM16 Public Engagement - Aberystwyth - June 2014** Poster presentation (*awarded best poster prize*).
- **BMVA Summer School - Swansea - June 2014** Poster presentation
- **Aberystwyth Imaging Workshop - Aberystwyth - August 2014** Oral presentation on the work covered in this thesis and OpenROV.

### 1.4 Goals

The project has clear delivery goals:

- **Literature Review** Survey the state of the art in this domain, identifying the key research groups and their output. Discover the specific areas of research these groups contribute to, whether it be primarily in the development of novel or adaptation of existing techniques in robotics, computer vision or machine learning for the domain of underwater video.
• **Apply Existing Techniques** Ascertain the utility of existing techniques against the datasets we have chosen. Specifically, we aim to use computer vision and machine learning techniques to automatically recognise and classify seabed substrates from underwater video. This will take the form of evaluating different methods to derive feature vectors best able to distinguish unique substrates, from both full-frame and patch-based approaches on the captured underwater video frames.

• **Research Data Capture** Evaluate the use of ROVs in data collection. This will include their capabilities and limitations.

• **Develop Software** Create a system that uses the aforementioned techniques, that may be extended with future research.

• **Publishing** Publish and present work to peers in the scientific community, specifically targeting those directly involved with underwater environment monitoring and understanding.
Chapter 2

Literature Review

This chapter covers the state of the art in underwater video processing. This includes investigating a number of existing research groups actively contributing to this domain, and researching a number of different methods and approaches.

Finally, the purported results and effectiveness of selected methods are compared and critically analysed, before highlighting the overall strategy that this research will take.
Marine habitat monitoring and study has been a subject of interest for research as technology has permitted new approaches to this historically manually-performed task (Marcos et al., 2008; Lambert et al., 2013; Riegl, Korrubel, and Martin, 2001). It directly relates to ongoing ecological surveillance methods, the monitoring of climate change and the management of fisheries. Gaining useful information on subaquatic environments has historically been difficult (Riegl, Korrubel, and Martin, 2001). Direct-contact monitoring via marine biologists, either via vehicle or diving, is expensive both in money and time. When considering even specific regions of interest, the sheer scale and area to cover makes this an infeasible task to accomplish.

Collection of subaquatic video provides marine-biologists with both a less expensive and quicker method of retrieval, and a reference recording to monitor the same area over time, whilst simultaneously minimising the human risk. The complexity of ROVs differs substantially between their implementation, with complex, dedicated, commercial systems such as VideoRay through to a simple camera-equipped sled trawling the sea-bed via a boat above.

### 2.1 Cardigan Bay

Cardigan Bay is the largest oceanic bay in Wales. It is located on the western coast, stretching from Bardsey Island in the north, and down to Strumble Head in the west (Figure 2.1).

We target our approach to the marine habitats in this area, working directly with FoCB researchers. As with a number of coastal inlets, most of the subaquatic region at Cardigan Bay has gone relatively unexplored, save for specific mandates. We investigate the use of computer vision and machine learning techniques in order to automatically classify seabed ecology. Building upon work in the areas of texture modelling, understanding and representation Marcos et al., 2008; Soriano et al., 2001, we evaluate their application to underwater video analysis.

The bay is unique, owing to the presence of distinct habitats such as the Sarnau. The Sarnau (singularly Sarn) is a Welsh name for glacial stony reefs which run along the West Wales coastline. These areas provide a rich habitat for a number of different fauna species.

From an economic perspective, Wales has a thriving fisheries industry which depends on sustainability and manageability of the local area. Research and monitoring goes on to support long-term-viable plans for these interests, although much of the data provided is based on limited sampling and the validity thereof remains contested.
Chapter 2. Literature Review & Discussion

2.1 Bathymetry

Bathymetry refers to the depth, formations and structures of underwater environments. It can be considered an analogue to Topography. Cardigan Bay is a shallow inlet, with the altitude from water-surface to sea-floor varying between 3 m to approximately 10 m deep. This is true for up to 15 – 20 m from the coast line, at which point a sharper gradient is experienced, quickly dropping to > 40 m. The last bathymetric survey of Cardigan Bay was performed by the National Oceanic and Atmospheric Administration (NOAA) in 1992. The data for which remains available to the public using their online tools.

The Sarnau offer particularly shallow areas in the bay, this is illustrated in Figure 2.2.

2.2 The Physics of Water & Light

The water column refers to the distance from the sea surface, down to the seabed. Figure 2.3 illustrates this concept, and how light can behave within the column. Depending on the geographical location, the altitude of this can reach approximately 11 km at its deepest point. The structure of the column is separated into 5 zones, based on depth as defined in Table 2.1.

The chemical, and physical make-up of seawater is a complicated state to define. Based on numerous environmental variables, the level of salinity and turbidity can fluctuate wildly.
based on geographical location, seasonal fauna dispersal, weather and time of day. These effects include the distribution of gases within the water, the abundance of organic particles, and their ultimate effect on the way in which light may travel through the medium. Turbidity,
TABLE 2.1: Abridged definitions of the zones present in the water column. Adapted from public material made available by NOC.

<table>
<thead>
<tr>
<th>Name</th>
<th>Depth (m)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epipelagic</td>
<td>0 - 200</td>
<td>Sunlight present, supports 90% of oceanic life</td>
</tr>
<tr>
<td>Mesopelagic</td>
<td>200 - 1,000</td>
<td>Limited light for photosynthesis, Octopus, Squid and bio-luminescent species present</td>
</tr>
<tr>
<td>Bathypelagic</td>
<td>1,000 - 4,000</td>
<td>No natural light, cold temperatures. Supports 1% of oceanic life</td>
</tr>
<tr>
<td>Abyssopelagic</td>
<td>4,000 - 6,000</td>
<td>Covers approximately 75% of the sea-floor, water temperature near freezing</td>
</tr>
<tr>
<td>Hadalpelagic</td>
<td>6,000 - 10,000</td>
<td>Deepest points, such as the Mariana Trench (10,911 m). Life does exist here also.</td>
</tr>
</tbody>
</table>

specifically, refers to the amount of organic material and back-scatter that occurs in the water column. This is particularly an issue when attempting to illuminate those zones that are unable to receive natural light.

2.2.1 Energy Loss & Attenuation

White light is the product of the visible light spectrum combined, ranging from violet at approximately 400 nm to red at 700 nm. As the components that produce this white light, by definition, vary in their wavelength — any attenuation or distortion that occurs to it will therefore be a function of the wavelength in question, and further parameters of the environment. We can think of this as saying the energy of ray $r$ at time $t$ can be defined as the following, where $\lambda$ is the wavelength of $r$, $\delta$ is the altitude underwater and $s$ some metric of salinity and/or organic matter present:

$$ r = f(\lambda, t, \delta, s) $$

(2.1)

By way of explanation, we can say that as $t$, $\delta$ and $s$ increase, the energy $r$ will decrease. The wavelength $\lambda$ will determine the amount of attenuation these parameters will provide. Newton's famous prism dispersion experiment demonstrates this effect in a more concrete manner; the spread of the spectrum is larger in spatial terms, the further away from the point of dispersal it is observed. The effect of increasing depth, salinity and the presence of organic particles both obfuscate the path of any given ray, and absorb the energy that is required to continue travelling. Image sensors, as commonly configured, capture red, green and blue (RGB) channels independently as individual receptors. A matrix of these
values is created, taking the individual RGB components into a representative form of the eventual pixel formation, typically RRGGBB in a joined hexadecimal format. The data is then normalised against some pre-defined camera matrix, producing the resultant image.

Illustrating this point, consider the weak assumption that the RGB components have evenly-distributed intensities across the full intensity-range. The values of $r_R$, $r_G$, and $r_B$ obtained from Equation 2.1 will be different when considering the same values for time $t$, altitude $\delta$ and salinity $s$. Visually, this will result in images that demonstrate a colour bias in one specific wavelength range subset, where the attenuation function is minimised.

This model is highly abstract, and assumes perfect uniformity in the source lighting and diffusion thereof in the target environment. The type of illumination, how it is formed, projected and radiated are further parameters of this difficult model that must also be measured, calibrated and compensated for.
Pope and Fry (1997) have conducted research in this domain to measure the levels of attenuation that does occur, and to provide corrective models. This is in order to aid marine biologist examinations, by providing accurate colour reproductions of the projections in question. A method of comparatively approaching this issue based on real-world assumptions on the measurements of absorption from these experiments exists, this method (Rock, 2011) shows that higher-wavelengths corresponding to red light are those which have the highest level of transmittance, where blue light suffers between approximately 50-75% absorption. These experiments were calibrated and tested based on different projection materials for the xenon-based flash. The greatest variance in the different glasses used occur when $\lambda < 300\text{nm}$. Given that the research centres around the visible light spectrum, this appears to have no effect on the projected light for scene capture.

Singh et al. (2007) describe this process as being modelled by an exponential function with an attenuation coefficient that is a function of absorption and scattering. By expanding upon the existing work (Mobley and Mobley, 1994; Mertens, 1970), the model is defined based on the depth in question, with the water column being segmented into stages of different models. Factoring these further considerations into the simplistic model, they propose a method of decomposing an image $F$ into its reflectance image $R$, and illumination image $I$, where $(x, y)$ are the spatial pixel coordinates, and $\lambda$ the light wavelength (colour channel).

$$F(x, y, \lambda) = I(x, y, \lambda)R(x, y, \lambda)$$

(2.2)

### 2.2.2 Optics

In underwater recording, typically a commercial camera unit is enclosed in a waterproof housing of appropriate size. Many of these camera units have been designed with land-based applications as their primary use. As such, the introduction of transparent housing around this equipment is, in effect, the introduction of a second lens through which the light must pass. It has been shown that the artefacts recorded in the reflectance properties of the glass can create a caustic, whose intensity is directly proportional to the distance between the camera lens and the shell plane (Shortis et al., 2013; Pope and Fry, 1997). Domed lenses are an example of a way to counteract this effect, by mitigating the caustic development through a larger direction of ray reflection.

The use of a domed lens, however, mandates high precision in the manufacturing thereof. By introducing this curved lens, the possibility of distortion before light reaching the image sensor is higher. Due to this, and the inevitable more complex manufacturing process, suitable domed lenses are traditionally far more expensive to procure. The effect and influence of
caustics in flat lens apparatus can be significantly reduced, however, as research has shown that it is directly related to the distance between it and the imaging sensor lens.

Rock (2011) shows that by modelling the interaction of the different lens stages using photon-mapping and ray-tracing, it is possible to create a simulation of the effect that altering the parameters of the above model exhibits. The phenomena of absorption and associated characteristics have been discussed, with respect to attenuation and the effect of water salinity.

### 2.3 Image Acquisition & Formation

Numerous factors affect the utility of the resultant video and images, both natural and mechanical. The physical properties of sea-water mean that obtaining uniform illumination is difficult, specifically when the depth of the sea increases.

Organic particles can obscure frame clarity, and add motion to a scene which can be difficult to cleanly disregard. Colour information varies both within and between videos due to differences in lighting at different depths. Visibility can be substantially affected given the turbidity (the cloudiness of a liquid given particles within, shown in Figure 2.11) present in the field of view. Finally, limitations of the recording hardware itself, such as low resolution or low frame-rate capture can also adversely affect any attempts to manually and automatically analyse the video content.

As discussed in Section 2.2.1, the physical properties of light in water generate unique issues in trying to capture underwater images. Examples of these illumination issues are given in Figure 2.4
Figure 2.4: Contrast between varying illumination patterns of seabed substrates in source video. The four frames all correlate to class III, as defined in Table 3.2 in Section 3.1. All video frames come from videos in the Bangor dataset (Lambert et al., 2013) (see Section 3.2.1 for further information).
2.4 Key Groups & Research Interests

Existing work in this area has investigated the improvement of capturing, cataloguing and understanding video data. Research into the effects of climate change on species richness of marine fishes and the optimisation of fishing strategies using these methods continues to be performed.

There exist many groups and institutions, large and small, that contribute research and insight to underwater understanding. Although too numerous to catalogue and discuss all, this section looks at those who are publishing and presenting work in current (this year) conferences, and those with key contributions in recent history. Those omitted are more concerned with the marine biology aspects of oceanography, and that do not have a direct bearing on the computational perspective of this project, such as Oceans Network Canada.

2.4.1 Fish4Knowledge

Fish4Knowledge was a European Union funded project under the Seventh Framework Programme\(^1\). It was an international collaboration between the University of Edinburgh (Scotland), Universitat di Catania (Italy), Centrum Wiskunde & Informatica (Netherlands) and National Applied Research Laboratories (Taiwan). The project ran from October 1, 2010 - September 30, 2013.

Led by a team of researchers including Prof. Daniela Giordano (Italy), and Prof. Robert Fisher & Dr. Yun-Heh Chen-Burger (Edinburgh), with an advisory board of Marine Biology experts such as, Prof. Konstantinos Stergiou (Univ. of Thessaloniki) & Prof. Kwang-Tsao Shao (Academica Sinica). The objective of the project was to investigate the full scope of problems and techniques in monitoring underwater fish habitats. These included methods of data capture, automated analysis, storage and semantic querying thereof. This research involved interdisciplinary work from theoretical computer science, including machine learning & computer vision, through to intelligent systems engineering and biology.

This project describes itself as part of the new-wave of computing metaphors, where science will be data driven. The research question being tackled is how to provide the tools for this analysis and data extraction, without the marine biologist in question being a programmer (Nadarajan, Yang, and Chen-Burger, 2013; Spampinato et al., 2012).

\(^1\)http://groups.inf.ed.ac.uk/f4k/
2.4.2 Woods Hole Oceanographic Institute

The Woods Hole Oceanographic Institute (WHOI) in Massachusetts, America was established in 1930 as a private research and higher-education establishment. Dr. Hanumant Singh is part of the Applied Ocean Physics & Engineering\(^2\) (AOPE) department, which concerns itself with the research and development of numerous deep-sea level AUVs, ROVs and image processing techniques. Most notably, the group partners with MIT to provide PhD training, with post-doctoral and resident researchers also key contributors to the area.

The scope of AOPE’s research ranges from designing ROVs and AUVs (discussed further in Section 2.5), underwater communications and telemetry, optics and illumination corrective models. WHOI maintains two primary research vessels, Atlantis and Knorr, and collaborates with institutions worldwide.

2.4.3 CIRS / VICOROB

After many years doing quality research, the Underwater Vision and Robotics Research Centre has become a benchmark in Europe for the design and construction of autonomous underwater vehicles, and the development of cutting edge software for the processing of visual and acoustic data.

– CIRS

The quotation above is taken from the CIRS website\(^3\). The group was founded by Dr. Rafael Garcia at the Universitat de Girona (UDG), as part of the Vision and Robotics department, this group work with developing technologies for underwater exploration. Through the development of AUVs, ROVs and enhanced vision algorithms, including seabed mosaicking and caustic reduction, they are a big contributor to the area.

This group frequently contributes to international projects, that focus on specific areas of research in underwater scenarios. Two examples of which are of particular relevance to the current state-of-the-art:

TRIDENT

In collaboration with various European Universities, including Genova & Portugal, TRIDENT focuses on addressing needs of specific underwater tasks such as archaeology, and industrial applications. The project builds upon existing AUV and computer vision methodology to

\(^2\)http://www.whoi.edu/main/aope
\(^3\)http://cirs.udg.edu/
collect and amalgamate optical and sonar data. Based on the visual understanding of the scene, the AUV will alter its navigational method to search for specific objects of interest.

Once it has recognised the object of interest, a robotic arm is used to automatically collect, retrieve and return it to the mission-designated area.

**EUROFLEETS**

The EUROFLEETS project is a collaboration of 22 out of 27 European Union member states, and has a total budget of €9,000,000. CIRS is one of many European institutions (including IFREMER as the project lead) working on capitalising the marine environments to increase its economic output for all member states.

The method for this appears to be the refinement of current methods in the deployment of underwater vehicles, and the use of high-performance computation to find areas of interest, highlighting those areas which may bring economic benefit.

### 2.4.4 Monterey Bay Aquarium Research Institute

David Packard (of Hewlett Packard) founded the MBARI institute in 1987, and it continues to be a private institution funded by The David and Lucile Packard Foundation. With education and research of the science and technology of the world’s oceans as its primary focus, the institute develops instruments, platforms, robotics and scientific method to contribute to this domain. Approximately 220 staff are located on-site, the majority of whom are classed as engineers, researchers or scientists.

MBARI aims to identify areas within the community where progress is lacking via technological or standardisation means, and put its knowledge and skill into rectifying these areas. They classify their research into eight themes, the most relevant for this project being remotely operated vehicle enhancements, infrastructure support and information dissemination.

### 2.4.5 National Oceanography Centre

Based in both Southampton and Liverpool, the National Oceanography Centre (NOC) focuses on the biological, physical, engineering and public engagement aspects of marine exploration. It is the largest institution in the UK that focuses on research in this domain. It was formed in early 2010 by the Natural Environment Research Council (NERC) to be the focal point of industry, academia and government based research and monitoring for the UK.
Its responsibilities include Royal Research Ships RRS James Cook & RRS Discovery, which are used on cruises to fulfil the institution's objectives in deploying or conducting specific missions around the world. The NOC is partnered with the University of Southampton, with this partnership alone covering approximately 1,200 students and staff, many of whom function in a research capacity. Importantly, NOC’s status as responsibility over the UK's input to sea-level science means that they offer centralised services, equipment and expert knowledge to any researchers covered under their remit.

Major projects of the NOC include ISIS, the UK's first deep-sea diving ROV and a series of AutoSub AUVs for long-term analysis (Figure 2.7(a)).

2.4.6 IFREMER

The Institut Français de Recherche pour l'Exploitation de la Mer (French Institute of Research and Exploitation of the Sea) is headed in Issy-les-Moulineaux and has 26 sites world-wide, with 5 main centres of research and education in Brest, Boulogne, Nantes, Toulon and Tahiti. The institute has approximately 20 separate research departments, monitoring fisheries resources, economic use of coastal seas, exploration and the advancement of oceanographic engineering.

IFREMER can be considered the equivalent of NOC in France, albeit at a larger scale owing to its longer history. Like the NOC, it has a number of its own research vessels, though mostly in the form of submersibles. As previously noted, the EUROFLEETS project, of which CIRS is contributor, is headed by IFREMER.
2.5 Capturing Underwater Video

This section details the methods, both historic and current, used to capture underwater data for analysis. There is an abundance of literature in this field, with the computer vision and robotics elements being intertwined throughout. This review aims to focus on the methods of capturing data so as to illustrate the technical difficulties which result in less-than-ideal material for processing.

Discussions of the physical robotics issues is limited, as it is not the primary goal of this research. However, where these limitations directly impact the collection of data, they are noted to a lesser degree.

2.5.1 Sensors

Images are captured via an image sensing panel of \( m \times n \) pixels in a matrix assembly. This raw data accumulates the number of photons during the time of exposure. The data of this matrix is then applied to a pre-calculated camera matrix within the imaging system, to produce the visual representation by some assumed model of real-world conditions. These conditions, through which the camera matrix is calibrated, tends to be one of the above-water variables. As such, the colour and contrast composition of the resultant images when this equipment is used underwater may not be desirable.

When dealing with the difficulties of underwater video, the primary consideration in selecting an appropriate sensor has been stated as being a wide dynamic contrast range. HDR cameras deal with varying illumination areas of a frame by taking multiple shots at different exposures, and then forming them into one image. Low dynamic, high-resolution sensors are the norm in consumer devices. This is colloquially referred to as the ‘mega-pixel war.’ In underwater imaging, typically off-the-shelf devices are purchased and subsequently modified to be fit for the task at hand. This includes mounting a Pixelfly 1024 × 1280 12 bit CCD in a flat glass plate housing to minimise caustics in the capture (Singh et al., 2004), through to the use of Apogee Instruments AP47p CCD sensors with a Homeyer interference filter wheel for multispectral image capture (Gleason, Reid, and Voss, 2007).

A large number of high-resolution images of the sea-floor environments are constructed from a number of smaller images covering a given region. Within these smaller images, the size of the image taken is generally less important than the quality of the information contained within. This means that having a smaller image with a higher dynamic range and accurate representation of small nuances in the visual appearance of the underwater scene is considered more important than a larger image containing less detail in the nuanced differences.
Therefore, a compromise of dynamic range and resolution must be made, where the primary consideration is the contrast.

2.5.2 Manual Capture

Historically, data has been captured manually via trained divers, or manned submersibles. Examination of the sea-floor by divers requires substantial training, certification, insurance, equipment and opportunity. The cost alone is extremely high, having to pay for the aforementioned requirements, but another factor that must be considered is the weather. Surface and subsurface temperature, volatility of the vessel and illumination levels are safety considerations which drastically reduce the available time in any given period for a successful dive. The problem of adequately water-proofing capturing equipment still exists, with the added complexity of enabling free underwater use of the equipment, and ensuring any power requirements are met.

Typically, divers can reach 30 m below surface level, or up to 50 m in exceptional circumstances. Their motion and area of interest must be precisely pre-calculated based on prior knowledge, or best-guess based on available data to isolate areas of interest. Divers are unable to spend extended periods of time underwater freely exploring, due to the pressure implications on the human body, fatigue and oxygen supply. Another key consideration for the safety of the divers is the wildlife present in the target locations which could, at worst, be lethal.

Submersibles, also submarines, are fully enclosed vehicles that travel under the water surface and can house one or more operator within. The technological advancement in hull integrity, energy storage and thrust / propulsion mechanisms have led to compact, high-efficiency submersibles which may carry a high number of technological apparatus to aid in exploration.

Due to the inclusion of human operators, a submersible is still limited by the tolerances of the operator, oxygen levels and calculated risk of the hull’s integrity to withstand the high pressures of deep-sea diving. Any potential errors in a mission can result in the death of the human operators within.

Typically, data was captured using conventional film-based cameras and an artificial light source. Once successfully completed, the images would be manually stitched together to create a photo-mosaic of the area in question. A well known example of this technique is Ballard’s work in using submersibles to capture a sunken submarine in 1960 (Ballard, 1975) and later, his work in constructing a mosaic of RMS Titanic (Ludvigsen et al., 2007; Singh et al., 2007). For fauna monitoring, manually-captured data has been used in coral reef
assessment, through methods such as Line Intercept Transect (LITR) and In-Situ Mapping (ISMP) which typically requires a diver (Marcos, Soriano, and Saloma, 2007), though limiting them to the reef at the diver's maximum range. Modern examples of submersibles being used to survey and capture data of the sea-floor include IFREMER's Neptune submersible and WHOI's ALVIN.

2.5.3 Remotely Operated Vehicles

ROVs are remotely operated, robotic platforms for scientific exploration. By design, they mandate the skilled knowledge of a human operator who remains in a control room above-surface. The complexity of this control room could be a single laptop computer with one operator, or an entire control room with a team of operators assigned to individual components of the ROV working together.

![Figure 2.5: A typical ROV control room, in this picture - that for the WHOI JASON ROV (Whitcomb et al., 1998).](image)

Typically, an ROV will have a guidance image sensor capable of delivering real-time visual feedback to the operator(s) (Figure 2.5), an all-axis propulsion system, its own internal power supply and extensions / platforms for scientific apparatus as required by the current mission. A key trait that all ROVs have, at present at least, is that it is physically attached to the control room at all times via a tether, sometimes known as an umbilical, which provides bidirectional data communication for feedback and control. Figure 2.7 shows a number of ROVs.

Due to the varying nature of experiments, and the depths of different locations, ROV classes exist. WHOI have used ROVs of different classes and configurations to tackle many underwater problems. In their 2000 paper, Singh et al. (2000) presented the results of equipping
JASON (Figure 2.6), an ROV rated for up to 6,000 m, with an image sensing module and a 675 kHz pencil-beamed sonar sensor to build topographic maps of a scene at 800 m depth in the Mediterranean, the applications of such a combination is discussed further in Section 2.7.2. In contrast, smaller ROVs such as the Sperre AS Sub-Fighter 7500 can be deployed up to 650 m, at 400 kg weight and still maintain sufficient manoeuvrability, whilst the underwater currents and scientific demands for deeper work require ROVs weighing upwards of 1,500 kg and improved payload capacity, such as the WROV (Whitcomb et al., 1998; Pizarro and Singh, 2003; Ludvigsen et al., 2007; Whitcomb et al., 1999)

ROVs are used in interdisciplinary research often, where immediate feedback resulting in altering mission parameters is vital. The control room engineers will immediately alter the function of the robot, to gain samples or visual data of areas that prove interesting to the marine biologist / geologist in question. Their deployment in 2007 under the Arctic Gakkel Vents Expedition (AGAVE) project, in conjunction with WHOI, was an effort to find underwater thermal vents in the area. Upon viewing the live data, geologists and glaciologists noted unexpected glass formations on the seabed, using the ROV’s sampling capabilities, glass samples were taken which provided hitherto unknown properties of underwater volcanoes. Further discoveries aided by the use of ROVs in the AGAVE project were the discovery
of 12 distinct new forms of microbial mat lifeforms, sampled by the robot and later DNA sequenced by biologists.

In the UK, ISIS is an ROV developed at NOC in conjunction with WHOI, based on their JASON II platform, with an estimated cost of £5,000,000. It is a deep-water work-class ROV, weighing 3,000 kg and capable of diving to 6,500 km. Its tether extends 10 km and is modular in its design to be configurable to different research goals. ISIS has been used by researchers to study sea-floor sediments in Marguerite Bay, and to study the fauna present in the bay. As equipment of the NOC, it is made available to UK-based researchers with an accepted proposal. As well as WHOI, in the United States of America, MBARI develop and maintain a fleet of ROVs made available for use to US-based researchers.

![Figure 2.7: Examples of shallow and deep-water work-class ROVs in use in current research.](image)

Although ROVs have many benefits over manual data capture and submersibles, the common issues and areas for improvement tend to be shared within the context of performance and image acquisition. Due to the trade-off of stability, manoeuvrability and expandability, there
are inherent issues with the quality and robustness of the data captured by ROVs. These issues stem from their stability in the water column, along all axes, which can mean that the physical position of the ROV is not where it is assumed to be. Subsurface currents, which work against the weight of the ROV based on its depth, as well as the effect on currents exerting force upon up to 10 km of tethered cable, will cause great fluctuations in actual position (Whitcomb, Yoerger, and Singh, 1999). This margin of error can vary wildly, due to the difficulty of providing exact instrumentation at the depth, pressures and environments in which they operate. Due to the ever-diminishing levels of light and energy transmission from the Epipelagic zone onwards, accurate position deduction through methods such as GPS are impossible. Instead, IMU sensors, as well as acoustic based methods are used to approximate location (Singh et al., 2007; Ludvigsen et al., 2007).

The effects of this instability manifest themselves in the uniform continuity of the data acquired by the ROV. For example, a series of patch based images, designed to be aligned in one direction, may indeed fall within some $\epsilon$ variance of that target area. This makes image alignment and stitching far more difficult (Singh, Howland, and Pizarro, 2004; Ludvigsen et al., 2007; Elibol et al., 2014; Eustice, Singh, and Howland, 2000; Borgetto, Rigaud, and Lots, 2003; Pizarro and Singh, 2003). Approaches to rectify this include the semi-autonomous nature of self-correction, and by making the process of sampling more efficient.

In Section 2.3, the illumination problems of underwater video were discussed. The efficiency of illuminating and capturing a large area directly compensates for positional instability, as capturing a wider area of the sea-floor will result in fewer points to sample, and larger overlap between them. This has been highlighted as a potential area for improvement on existing methodology (Ludvigsen et al., 2007), stating that more powerful light sources will allow for greater distance between camera and subject, whilst conceding that the inherent issues of backscatter would likely be more pronounced. Singh et al. (2004) have researched the applicability of different light sources to compensate for these phenomena, including RED-biased lighting or 150 Hz strobe lighting.

Another key limitation is to the typical ROV set of a top-down image capture of the sea-floor is the geometric distortion incurred due to the non-planar surface observed. Research has been conducted into this limitation, stating that the ideal altitude for an ROV to be surveying the sea-floor is 3 – 4 meters. However, due to the variance of altitude of the sea-floor, the 2D projection of areas of interest will contain different relations between key points (Singh, Howland, and Pizarro, 2004). This issue affects registration of images for mosaicking.
2.5.4 Autonomous Underwater Vehicles

AUVs, unlike ROVs, do not require a control room nor an operator. They are designed to be low-power, high-duration robotic surveillance units which have a mission time that can range up to almost 6 months at the pinnacle of current technology. As they are autonomous, they have no tether attaching them to a separate system. They are capable of carrying an array of sensors, including image capture (typically in a top-down orientation, sample image captures are presented in Figure 2.10), spectrometers and sonar modules. A typical AUV construction is shown in Figure 2.8, with various in-use models shown in Figure 2.9. Due to the varying scale and mission requirements, on-board power typically require around 2 kWh battery packs for up-to 24 hour missions. Typically, the embedded control system runs a low power computer platform, using various distributions of Linux as the OS (Singh et al., 2004).

For long term monitoring experiments, AUVs propose a unique solution. Being able to lay dormant on the seabed until some stimuli awake them, whether time-based or environmental, before returning to a pre-determined area for collection at mission-end for expert analysis of the data. Due to the inability to transmit information via light or radio due to energy absorption in water, AUVs communicate with mission control through acoustics and satellite (Iridium) communication (when in shallow water) to relay observations, status, and receive new orders. The physics and logistics of these methods are not within the scope of this literature review, as they do not influence the quality of the data collected.

REMUS, Remote Environmental Monitoring UnitS, are a new-class of AUV being developed at WHOI for the long term monitoring of the sea-floor whilst only requiring a single laptop computer to interface with remotely, and low purchase cost. AUVs such as REMUS can be
used for sea-bed mapping by being provided a path to follow, manoeuvring themselves to cover an area methodically, recording image and acoustic data. These AUVs are capable of autonomous operation up to 600 m, REMUS AUVs are highly flexible in their application, including mine detection in Iraq under the US Military. Strobe lighting is used to illuminate the areas or objects of interest in order to mitigate the effect known as marine snow, which describes the organic particles suspended in the water column that reflect light back towards the imaging sensor, generating high-intensity reflections that visually resemble snow in the frame.

Due to the lack of a tether, and typically torpedo shape, the positional issues which affect ROVs are not so pronounced on AUVs. They are about to ascertain their position using a series of sensors, including INUs (Internal Navigation Unit) which can determine the orientation across all 3 axes. Speed and altitude are determined using Acoustic Doppler Current Profilers (ADCP) to permit autonomous corrections, and obstacle avoidance through sonar units.

AUVs have been purchased by many worldwide institutions based on this research, including the Australian Center for Field Robotics, using a high-resolution stereo camera system and
obstacle avoidance sonar - surveying protected marine areas. UPRM in Puerto Rico use AUVs for long-term coral reef health monitoring. The National Sun Yat-Sen University in Taiwan use a WHOI-provided AUV as the basis of their oceanographic research.

There are some key issues with using AUVs, as they have no hard connection to the land-side mission operations. If there is no functioning beacon mechanism, and any number of autonomous, mission critical features fail, finding a stranded AUV will be extremely difficult. In the example of AGAVE, Dr. Singh of WHOI describes some of the challenges experienced in a TEDx talk\(^4\). The situation is described as placing the AUV into a small hole in the glacial ice, it performs its mission, returns to near-surface 5 hours later and begins signalling its position to be collected. In this time, the glacial ice has moved at a constant velocity, moving the mission vessel approximately 10 km from its original destination. Despite moving to within 50 m of the AUV, the team were unable to locate it for 24 hours.

\[\text{FIGURE 2.10: Unprocessed and colour-corrected image of sea-bed fauna captured (Armstrong et al., 2006)}\]

\(^4\)TEDxWoodsHole (WHOI) = Exploring the Arctic Ocean with Robotic Vehicles (4/12/2010).
MBARI develop and produce two AUV classes: Dorado & Tethys. They posit that AUVs will see increasing importance in the future, due to the relatively low cost of deployment and reuse, compared to manned retrieval, whether via submersible or ROV (Walther, Edgington, and Koch, 2004). Exploiting this area, the two classes are designed for deep-sea exploration within maximum expandability for instrumentation. NOC currently produce three models: AutoSub 3, AutoSub 6000 & AutoSub LR, with ongoing research on the latter, aiming to surpass 6 months mission time underwater.

CIRS / VICOROB are engaged in COMAROB, a multi-institution project whose aim is to, within three years, carry out substantial improvements to AUV technology in terms of functional and operative levels. They are also involved with TRIDENT, which aims to create new forms of cooperation between autonomous surface craft and an I-AUV (Intervention Autonomous Vehicle). By creating a tandem system for exploration, the use of computer vision, pattern recognition and machine learning are used to identify targets to being back to the surface.
2.6 Computer-Based Processing

Thus far, this literature review has covered the physics, logistics and expense of data capture in underwater environments. This section aims to look at the computational techniques used in processing that data. Whether it be simply improving the visual quality of the imagery for human observers, from colour correction through to artefact suppression, or through to automated information inference and retrieval.

The target domains of interfacing with underwater data vary hugely, and this section aims to focus on those most concerned with the use of marine biologists and labelling underwater video and imagery.

2.6.1 Image Processing

This area deals with the modification of a digitised image to alter its representation, but not to make inferences based on the data within. Practical examples include the mapping of colour distributions present to more preferable ones, based on models of the human visual system (Provenzi et al., 2005; Land, 1986) or otherwise. Typically, image processing will be used to rectify issues in the image capture caused by lens and image sensor characteristics, and any environmental variables which will alter the clarity of the subject captured. The difficulties of capturing image data underwater are present in Figure 2.11

In underwater video, all of these characteristics mandate specific models for the rectification thereof. Underwater video capture and analyses is researched, primarily, to aid marine biologists and ecologists in the understanding of Earth’s habitats. This distinction is paramount, as Computer Vision (Section 2.6.2) deals with the fully or semi-automated extraction of data contained within images or video.

A major application of image processing in underwater video is colour correction. Section 2.2.1 discussed the effects on light in water. Work in this area has modelled these phenomena (Shihavuddin, Gracias, and García, 2012) and by building a model of the patterns observed, creating compensatory functions has been a big topic of research.

A similar approach to the modelling methods of those at WHOI (Singh et al., 2007, also noted in Equation 2.2) has been published by UDG. García, Nicosevici, and Cufi (2002) propose a model that has individual component functions for the contribution of gain \( g(x, y) \) and camera offset \( o(x, y) \) at the current spatial location \( (x, y) \). They also propose a revised equation that can be considered the reflectance function of the scene \( f(x, y) = c_m(x, y)r(x, y) + c_a(x, y) \), where the component \( c_m(x, y) \) is the multiplicative factor – a Gaussian-smoothed version of the image acquired to approximate lighting – and \( c_a(x, y) \) is a shading additive. This model
is then used to derive the contrast of the image, and the final illumination field. In practice, this may then be subtracted from the sensed image to produce evenly-illuminated frames.

Gracias et al. (2008) have conducted research into the removal of the caustic projection (sun-flicker) onto the sea-floor. Figure 2.12 shows the effects of the sun’s rays in shallow water sea-bed environments. Working on the basis that the same scene is viewed many times in different video frames, the method involves the creation of a median image over a temporal subset with close neighbourhood approximation. Two key components are described; the illumination field and the registration errors experienced. Sun-flicker removal finds its basis in shadow determination and deduction from land-based images, the techniques described build upon these by modelling the radiance of sunlight over the scene, per pixel, and over time. By combining registration errors within the scene, and the radiance variance of a given point \((x, y)\), the method is able to isolate the points in the images which deviate from the illumination model described. Tested over a number of environments and hardware, the algorithm is demonstrated to perform well.

**Figure 2.11:** Typical scene demonstrating difficulties of automated analysis, specifically when underwater conditions include high turbidity.

### 2.6.2 Computer Vision

Computer vision studies the methods of pattern and object recognition in digitised images. By analysing relationships in the spatial, colour or statistical domain, models of parameters and features that define, and are shared by, classes of objects can be drawn. These models are mathematical in nature and, as such, using advances in generalised machine learning techniques – automated classification of certain classes is possible.
A digitised image is a series of points that show the spatial relationships of individual picture elements (pixels) in the order in which they were captured. An image, in this context, is a matrix of coloured pixels. The distribution of colours, or the significance thereof, is one metric that computer vision can employ in extracting data from a digitised scene. Colour information, however, is not always indicative of what a scene contains. Disregarding colour information, and working in the greyscale domain, brings the patterns of the digitised objects. This is the texture(s) of the object or scene.

The spatial relations of texture, or the colour and intensity distributions of a scene are the foundation for a number of techniques aimed at identifying the unique combinations or arrangements of characteristics which make items distinct or similar. Texture is the scale and rotation invariant pattern that defines the visual characteristics of a given part of an object of interest, and is used to characterise its visual appearance.

**Geometry & Spatial**

A key area where computer vision is used in underwater video is to develop established algorithms to deal with the geometry of the sea-floor whilst compensating for the complexities the domain brings. A key example is the application of Leonard’s SLAM algorithm, used to detect structure and topography from video (Leonard and Durrant-Whyte, 1991) to ROVs
and AUVs. This problem has been tackled by using a combination of topological information from SLAM, and image features using SURF from the areas of interest (Aulinas et al., 2011a).

**Features**

Structure from motion (SFM) is the process of identifying key points in any given frame of a video or series of still images, and tracking them over the subsequent frames. By mapping the spatial correlations of the images, and, specifically, their variation over time (over the frames), the 3D topography of the scene can be inferred. This has found commercial and research uses in many fields by building 3D models of real-world objects from digital photographs and video.

The identification of key-points that are useful from a given frame or scene can be achieved through a number of approaches. A key limitation of these approaches is the assumption of uniformity in scale and rotation, this is particularly an issue for image registration, that requires uniformity of this nature to perform pair-wise registration in a hierarchical format, in order to converge on a single map.

Scale Invariant Feature Transform (SIFT) is an algorithm proposed by Lowe (Lowe, 2004; Lowe, 1999) and aims to identify the central points of dominant features in a given frame or scene. In recent times, this has proven to be a highly robust method of key-point feature detection, and is used throughout current research (Shihavuddin et al., 2013; Rodrigues et al., 2010; Eustice, Singh, and Leonard, 2005; Lowe, 1999; Shihavuddin, Gracias, and García, 2012; Singh et al., 2007; Eustice, Pizarro, and Singh, 2008). Figure 2.13 demonstrates SIFT key-points on a frame taken showing the sea-bed in an underwater scene. The point with the large radius towards the middle of the frame remains constant throughout the subsequent frames, and is due to the constant illumination pattern observed.

### 2.6.3 Video Processing & Annotation

The key benefit of using video sources versus image datasets is the capture of temporal information relating to the scene within the data itself. By obtaining a sequence of images at a short interval, major and minor differences in appearance, state and position may be inferred. To a human observer, the pattern recognition is natural – being able to map the scene and its contents through changes over time.
Figure 2.13: SIFT applied to a frame from an underwater video taken by Bangor University. Key-points are superimposed, including their size (visualised as circle radius).

Annotation Systems

The annotation of videos can be considered a manual, semi-automated or fully automated task. Typically, unless weakly-supervised or deep-learning is taking place, it is preferable to have a series of labels provided by a human-expert that map real world concepts to individual frames, or sequences over a subset of the video. These labels can take many forms, in both their application, format, level of detail, interface and intended audience.

VARS (Video Annotation & Reference System) is a video annotation system developed by MBARI (Figure 2.14. By standardising the method of storing labelled data, it is hoped to act as a common platform for annotations across distinct research groups. Built in Java, the software is cross platform and made open-source as of November 20055. The software exposes a tree-like structure of organism classification, with the ability of different organisms to be classed within shared frames. Due to its ambitious goals, the complexity of input is high, with dedicated groups looking at simplifying work-flows based on specifically targeted organism research.

Annotating huge datasets doesn’t necessarily have to be boring. Di Salvo, Giordano, and Kavasidis (2013) have developed a method of crowd-sourcing the annotation process by creating an on-line Flash-based game, ‘Flash the Fish’. By playing videos from the dataset, the objective is to photograph fish as they swim. This version of crowd-sourcing aims to

5http://vars.sourceforge.net/
Figure 2.14: MBARI’s VARS annotation system, showing the tree structure of annotations, the main annotation window and querying interface.

bring amusement into the acceptedly-tedious annotation process. The player clicks on a fish, accumulating points for doing so, and is then able to progress through levels.

In fact, the point of selection is then subject to region-growing algorithms to ascertain the shape and boundaries of the fish. The locations of the clicks themselves are stored and clustered using K-Mean Clustering, in order to realise the premise that the most-clicked areas are most likely to contain the fish themselves. The GrabCut algorithm is then used to segment the fish from the background, giving a best-case precision of 80% accuracy in the annotation method.

Event Detection & Object Recognition

The detection of objects, from proto-object through to recognised instance of some trained class, leads to the desire to track this identified region of video over the temporal domain. In generalised computer vision, the current state-of-the-art in object tracking comes in the form of Kernel-based tracking (Comaniciu, Ramesh, and Meer, 2003), IVT (Ross et al., 2008) & MOSSE (Bolme et al., 2010) (Figure 2.15)

MBARI’s Automated Visual Event Detection (AVED) system deals with recognition and classification of different fish species, by identifying regions of interest in a frame through the use of visual saliency, before submitting this area of interest for further analysis (Walther, Edgington, and Koch, 2004; Edgington et al., 2006). The algorithms deal with the low-contrast and poor-illumination of the source video by applying a bottom-up saliency based approach. By using a subset window of the video frames, such that the length is approximately 0.33 s / 10 frames, they are able to accurately determine motion patterns that are likely to be objects of
interest. Building upon the work of Itti, Koch, and Niebur (1998) in saliency determination, firstly background subtraction is applied through modelling areas in motion as foreground, and removing the remainder of the frame. The foreground data is then decomposed into seven channels covering colour and four spatial orientations. By focusing on oriented edges in the proto-objects, they are able to deduce those which are marine animals with reasonable certainty. The outline of the object(s) of interest is captured, and the centroid of their region tracked in the temporal domain, using Kalman filters to estimate their position. The success of this system is approximately 50% over the two datasets, with a false-positive rate of 10-20%. At the time, performance on powerful 16-core Xeon workstations, working at 720×480 pixel resolution was, at best, 3 fps.

The development of AVED continued, and improved results were published in 2006 (an example of object detection and classification in AVED is demonstrated in Figure 2.16), with the inclusion of rudimentary unassisted object classification. Following the same work-flow, salient objects deemed to be possible marine life are tracked, with each frame being decomposed into several greyscale sub-images that contain examples of the target object classes. The training data was modelled using a mixture of Gaussians, with three distinct classes

**Figure 2.15:** MOSSE tracking of fish in underwater video implemented in our system. Due to rotation and turbidity, the fish is lost in the final frame.
chosen, and classification of new samples was achieved based on a majority vote from these models. This method returns a 90% correct detection rate, where 94% of those had a correct species classification (Edgington et al., 2006).

![Figure 2.16: MBARI’s AVED system automatically detecting and classification of sea-floor marine life. Taken from publicly available footage from MBARI (Edgington et al., 2006).](image)

Nadarajan et al. (2011) have developed a more generalised framework for the automated composition of video analysis as part of the Fish4Knowledge project. In contrast with existing knowledge-based systems such as LLVE, CONNY, and OCAPI, the proposed system does not require 
\textit{a priori} knowledge and is, therefore, claimed to be more generalised. Their method involves the distinction of background and foreground objects. Using a three-tier system of adaptive Gaussian modelling and moving average models, the basis for this Fish4Knowledge project is from fixed-position camera platforms. As such, the resultant motion observed must be an area of the foreground, as changes in the background models may simply be caustics from sea-surface, or fauna movement due to current.

Once the regions have been identified, the contours of the shapes are extracted. These models are applied over the temporal domain to create binary representations that continue over time to segment foreground and background. Regions of the computed foreground image are grouped together, with CAMSHIFT being applied to them to predict their next state. This is different to other work from the Fish4Knowledge group, that used a bounding-box approach to track objects in video. The performance of the system is not given in terms of fish accurately recognised and counted, but rather on the performance and modularity of the tools themselves, with an emphasis on those of non-programming backgrounds being able to use the systems (Nadarajan et al., 2011; Nadarajan, Chen-Burger, and Fisher, 2009).
Given this approach, it is clear that the subject of tracking and analysis is not important to the system, and that it is similar to prior work that has used a bounding-box method for fish tracking (Shiau, Chen, and Lin, 2013). Within the Fish4Knowledge project, another research group discuss similar methods of tracking proto-objects on the consecutive frame(s), and applying both foreground segmentation and tracking in one step. The precision of this technique is reported as up-to 89.93% where time $\tau = 2$, meaning the interval of time considered in the step, with $\tau = 1$ relating to 1 second, or 20 frames.

The study of fish movement and behaviour is another application that has typically been performed by human visual inspection. Given the vast number of videos, locations and species present in any given research project, the process can be long, arduous and fatiguing. In Fish4Knowledge’s remit, is the automated detection, extraction and extrapolation of fish trajectories (Beyan and Fisher, 2013; Palazzo, Spampinato, and Beyan, 2012; Spampinato and Palazzo, 2012b). The extraction, cataloguing, and subsequent analysis of fish behaviour is used as a key indicator of environmental changes.

Spampinato et al. (2010) propose a method for the extraction of this data. Their approach is based on fixed-array cameras is justified by stating that the use of manual capturing methods are invasive, and will therefore affect the observed behaviour being studied. This research is based on Fish4Knowledge’s vast dataset, and builds upon work on curvature scale space (CSS) analysis (Mokhtarian, Abbasi, and Kittler, 1996).

Like other work from the group, they use a combination of Gaussian mixture and moving average models (Nadarajan et al., 2011; Nadarajan, Chen-Burger, and Fisher, 2009; Spampinato et al., 2008; Beyan and Fisher, 2013; Palazzo, Spampinato, and Beyan, 2012), although tracking over the temporal domain is performed using the Adaptive Mean Shift Algorithm. Affine invariant features are used to model the 3D fish and its varying rotation and orientation, such that the contours of the fish boundary are extracted through a series of affine transformations. A number of metrics are applied, which are derived from Gabor filters and Grey-Level Co-occurrence Matrices (GLCM) (Haralick, Shanmugam, and Dinstein, 1973). These include calculating mean, standard deviation, variance, difference entropy, using statistical moments for representative values of the resultant histograms and others. The resultant feature-vector $x \in \mathbb{R}^{70}$ for a given texture incorporates these statistics, and encodes the unique visual information associated with the detected fish.

Morphological operations are applied to remove noise from the resultant fish contours, which are expressed as a series of points. The feature space is bounded (reduced), as is computational complexity, by using Fourier descriptors of the contours to describe the frequency variability of the shapes. From these Fourier descriptors, a histogram is generated of their modulus (with 30 values) due to their high number of points, this histogram is then invariant to affine transformations. This contour is then smoothed using CSS, and the final number
of extracted shape features is 50. Combined with the 70 previous points, the feature vector \( x \in \mathbb{R}^{120} \) – PCA (Principal Component Analysis) is applied, resulting in a final feature vector of \( x_{PCA} \in \mathbb{R}^{24} \). With this feature vector encoding the detected fish, the accuracy of fish detection and tracking is as good as 85%. It also performs well for fish classification, reaching up to 100% for many species, with Chromis viridis performing worst at 80%. The reason given for the low classification rate of this species is the visual appearance of the fish, being near-white, and difficult to detect against an underwater background (Spampinato et al., 2010).

In their later work, Spampinato et al. (2012) acknowledge the theoretical limitations of the CAMSHIFT algorithm acting only upon colour information which encounters difficulty with fish-fish and fish-background occlusion. The approach is modified to use co-variance matrices (CVM) for fish representation, rather than metrics derived from local histograms. This is in order to maintain the spatial relations of pixels that are lost in histogram analysis. CVMs on a pixel level, with a surrounding \( 5 \times 5 \) window, and act as an encoding that describes the detected object. This can then be compared with CVMs on subsequent frames to track objects over the temporal domain.

CVMs do not lie in the Euclidean space, therefore Förstner's distance metric is used, as it acts upon generalised eigenvalues to determine similarity (Spampinato et al., 2012). The distance metric works on positive-definite CVMs, and is affine-transform and inversion invariant (Förstner and Moonen, 2003). Occlusion is handled by giving the CVM metrics a time to live\(^6\), after which if no sufficiently similar CVM is encountered, the tracked object is considered to have left the scene, and not simply be occluded.

Tested on 30,000 frames, with measures of correct counting, recognition and trajectory matching, the CVM method is compared with the earlier work, where the new method performs better than all metrics included in the comparison (Nadarajan et al., 2011; Nadarajan, Chen-Burger, and Fisher, 2009; Spampinato et al., 2012; Spampinato et al., 2008).

Methods exist for the categorising of fish motion patterns / trajectories in underwater video, through the use of fine-grained event classification (Amer et al., 2011). Training and testing data was collected at three sea depths, at half a day frequencies, to collect data on six behavioural stages. Classification of fish is performed using histograms and Random Forests (Breiman, 2001). Fish detection is performed through the use of Optical Flow (Figure 2.17), from which over \( n \) frames, the dominant areas of motion are identified.

The resultant vectors are clustered to identify motion groups, which then permit the deduction of the background for subtraction. This method is not perfect, and indeed false positives

\(^6\)Time to live (TTL) is a software-engineering term, it is a value of time for which, after creation, the object is valid and usable. Once the TTL has expired, the object is discarded.
are reported. Background / foreground analysis is then performed to provide the required subtraction. Those determined to be fish are placed in a boundary box, which is then inspected for internal gradient changes to be greater than some threshold $\epsilon$, in an attempt to avoid false positives. This approach isn’t resilient to sun-flicker, plant life which may also move between frames, and other foreground changes such as compression artefacts, as the foreground segmentation will still detect pixel regions that are not considered background. Some fine tuning of $\epsilon$ must be made to balance the true positive and false positive ratio, or some secondary processing on the detected fish must be applied to attempt identification of false positive examples in order to exclude them. The threshold value must also take into consideration environmental changes, including turbidity and illumination. These possible future directions have yet to be researched by the Fish4Knowledge group.

Tracking of the boundary boxes around the identified objects is performed by the Maximum Weight Clique problem, that is an NP-hard graph traversal algorithm. Histograms that are collected are classified in Random Forests in frame-pairs from the video. Precision rating is high at 66.2%, stating better results than prior methods using linear-kernel based SVMs. Detection of fish (but not their classification) achieves 79.6% precision. SVM with linear-kernels as provided by libsvm (Chang and Lin, 2011) were also used in place of Random Forests, but resulted in precision of 58.7%. The authors attribute this to the problem not being a linear-issue, but acknowledge that the default parameters of libsvm were used. The performance of an SVM based system is directly related to the parameter choices made, specifically $C$, this is discussed more in Section 4.6.1 and may be more of an issue than the linearity (or linear separability) of data in the feature-space.
2.7 Applications of Underwater Video

This section aims to discuss where the automated analysis of underwater video is used and actively developed to solve real-world issues. Whilst not exhaustive, the groups and papers discussed cover the major pillars of research that drive research at the key groups discussed in Section 2.4.

2.7.1 Wildlife & Fauna Monitoring

Fish behaviour is studied as part of Fish4Knowledge, building upon their work in the accurate modelling, storage and representation of the fish classifications and trajectories serves as the basis for work in classifying trajectories themselves (Spampinato et al., 2010; Nadarajan et al., 2011; Nadarajan, Chen-Burger, and Fisher, 2009; Spampinato et al., 2008; Beyan and Fisher, 2013; Palazzo, Spampinato, and Beyan, 2012). Beyan and Fisher (2013) extend this work by building a hierarchical classifier to analyse what constitutes abnormal behaviour. Using the CVM approach for obtaining these trajectories, a multi-level classifier is built by labelling said trajectories and introducing new features in the selection step of each iteration. Using 9-fold cross validation and Affinity Propagation (AP) clustering, the method is compared with kNN and SVM classifiers. AP tries to identify the cluster centres from the observed data points, it performs this through a message passing system that checks the agreement between points $i$ and $j$, where at some point $j$ will be found to be an exemplar; an observed data point that acts as a cluster centre. This differs from kNN as the number of clusters does not need to be defined before running the algorithm, the number of required clusters is determined within the AP algorithm. Results show that it has the highest successful detection rate of abnormal trajectories, although kNN was the best performing for normal trajectory detection.

Stewman et al. (2006) describe a method for automated dorsal fin recognition in bottlenose dolphins. Each dolphin has a unique shape to its dorsal fin, a number of which exist as labelled training examples in the DARWIN (Digital Analysis and Recognition of Whale Images on a Network) database. The process pipeline begins by using unsupervised learning to obtain a threshold value, in order to isolate the fin outline. A wavelet transform is applied to obtain the characteristics of the fin, which is then categorised as a series of points to define an individual. The purpose of this research is to save the time of manual annotation. The author notes promise in the paper, but does not offer any results.

Another application of these methods is the monitoring of coral reefs. The condition of these is assessed by estimating the population of biotics and abiotics in the area, typically based on visual appearance into 6 visually-distinct classes (Marcos, Soriano, and Saloma, 2007).
Software has been developed to aid the annotation of coral reef video\textsuperscript{7} but still requires frame-by-frame annotation by a human-expert. It is stated that, typically, alive corals in images and video appear blue, yellow, brown or green, whereas dead coral appear white (Soriano et al., 2001).

Marcos, Soriano, and Saloma (2007) use a combination of colour and texture metrics with machine learning models to attempt the automation of coral reef assessment in underwater video, and build upon earlier work where the colour-texture approach was used whilst taking into consideration the 3D structure of the coral (Soriano et al., 2001), they reflect and refine upon poor initial results. The original work used a very small dataset of 50 images of Australia’s Great Barrier Reef and whole-image classification into 5 distinct classes of abiotics and biotics was performed. These frames were then segmented equally, with training classes being chosen at regular intervals, before manually segmenting the remainder into 115 samples.

To work with the colour data, the digitised samples were transposed from the Red-Green-Blue (RGB) colour space, typically associated with CCD sensors and digital monitors, into the Normalised Chromaticity Coordinates (NCC). This transposition is based on the relations of the RGB channels, and is achieved by obtaining $I, r, g$ through $I = R + G + B, r = R/I, g = G/I$. Soriano et al. (2001) state computational complexity reduction, separation of brightness and chromaticity and more linear combinatorics of colours are possible by doing this. The colours of the samples are binned in approximate ranges (thus, there are fewer bins than observed colours) by the authors’ judgement. For example, grey and white are both determined to be white as ‘grey is merely an instance of white’. As a textural descriptor, LBPs are used with $p = 8, r = 3$ (LBPs are discussed in-depth in Section 4.3.3). The experiments were evaluated using the leave-one-out-method, due to the small sample size and using a kNN classifier. These resulted in excellent recognition of dead-coral at up to 91.30%, whereas living coral as low as 17.39%. Amalgamated together and based on recognition across all 5 types, the best performing combination achieved 47.83% with $k = 5$ for the kNN classifier (Soriano et al., 2001). Using the NCC colour space for extracting texture using LBP is a precarious idea. Given the very different illumination patterns that can occur underwater, due to scattering, occlusion, etc., it was likely the case that the local patterns were indeed too localised for be applied generally to the scene owing to the non-uniformity of the contrasts within.

Marcos, David, and Soriano (2008) describe a method of using the textural information as part of the feature vector, with the colour. The same sample set of the Great Barrier reef as Soriano et al. (2001) was used, with further definition such that the frame sizes were 640×480 in resolution. The textural information is obtained by discarding colour information and using the grey level image, $I$ using the same methodology as the original work, then the

\textsuperscript{7}PointCount 99 - http://www.cofc.edu/ \textbackslash coral/pc99
region of interest is given a value based on the variance of the grey values noted within. This value is interpreted as a small variance indicating a fine texture, and a large value indicating coarse texture. No results are provided, and the proposition is left to the scrutiny of its peers. The problem is being approached in a simplistic manner, by simply mapping the intensity variance of a given frame as being indicative of its texture, any illumination inconsistencies or obfuscation is being omitted. Given that the frames are of distinct make-up, it may be a better approach to use multi-resolution sampling and a hierarchical classifier to match generalised shapes of coral, and then texture. Belongie, Malik, and Puzicha (2002) discuss this method from a generalised vision perspective.

Given the restrictions on divers up to around 30 m of depth, the monitoring of complete coral reefs is somewhat restricted. Reefs have been shown to extend up to 200 m in depth. WHOI have used ROVs to provide this possibility, the results of which, from a biological standpoint, challenged the status quo and accepted knowledge – the observations were in contradiction with popular beliefs as the reefs were healthier than expected.

2.7.2 Seabed Maps & Bathymetry

**Figure 2.18:** Labelled seabed mosaic by Ludvigsen et al. (2007), by combining the techniques of alignment, topology correction, intensity scaling and tone mapping, a uniformly-illuminated representation of the sea-floor is obtained.
Topographic data for Bathymetric maps of the sea-floor have historically been taken via manual capture, or through the use of sonar to approximate the local depths of specific areas. A major application of underwater video analysis is the generation of high-resolution, 2D, 2.5D or 3D Bathymetric maps (a 2D example is shown in Figure 2.18, and 3D shown in Figure 2.19). Typically, the algorithm used follows four stages (Borgetto, Rigaud, and Lots, 2003):

1. Preprocessing & feature extraction
2. Matching & registration
3. Motion estimation
4. Full-map rendering

Singh, Howland, and Pizarro (2004) covered the many issues of light propagation in seawater, and potential methods for dealing with back-scatter of light. They dismiss the need to attenuate forward scatter, as it contributes to the dominant terms. This is imagined as light rays scattering in all directions in the water column, where those in a forward direction will only serve to further illuminate the area of interest (AOI). However, those scattering in a backwards direction cause reflectance against debris, and obscure the AOI. The total energy $E_{\text{total}}$ is defined, where $E_d$ is a direct component expressed in terms of the incident radiance pattern, $E_{fs}$ forward-scattered, and $E_{bs}$ backscattered:

$$E_{\text{total}} = E_d + E_{fs} + E_{bs}$$  \hspace{1cm} (2.3)

Later, Singh et al. (2007) describe a method of sea-bed map construction using an ROV in their 2007 paper. By performing illumination correction, a series of evenly illuminated patches of the sea-floor are obtained. Their method involves the use of SFM from Pose Instrumented Cameras. This local-to-global approach builds upon the work of others used in conjunction with the navigational and altitude information available to them. Locally, features are derived independently as, due to the 3D features appearing in only a few images, a global approach would result in weakly correlated local solutions. Using a simple pin-hole camera model, they project the points of interest as a function of the pin-hole camera and a normalised homogeneous representation. Their methodology permits them to recover the 5-degree-of-freedom pose from correspondences between these points of interest.

These points of interest are gathered using SIFT (Lowe, 1999) and Harris (Harris and Stephens, 1988) feature extraction, the latter being surrounded with Zerkine moments. Registration is performed as a minimisation problem; median of least squares. The resultant registration has consistent inlier correspondence. Spatial relationships are inferred by their algorithm.
through the correspondence of overlap from parallel lines or loop closures. The correspond-
ing 3D points have outliers removed through application of a RANSAC algorithm (Fischler
and Bolles, 1981). This process is continued until all possible correspondences have been
exhausted, or some upper limit $\epsilon$ has been reached.

![Figure 2.19: 3D Bathymetric data extracted from photo mosaicking from underwater video
by Singh et al. (2007) showing the navigational route taken by the ROV, the extracted to-
pography, and the resultant topography mapped with the post-registration sea-bed.]

Borgetto, Rigaud, and Lots (2003) of IFREMER describe the area of mosaicking as a key
area of current research, and propose two new pre-processing algorithms to deal with the
associated illumination problems. They describe their image retrieval using ROVs at an
altitude of 5 m, which creates the typical issue of non-uniform lighting. The first method they
describe is taken from remote sensing and astronomical imaging; CCD camera radiometric
correction. This model requires the estimation of a darkness reference (DR), and a uniformly
lit reference (ULR) image. Due to practical reasons of obtaining a calibrated DR at the
temperature and depth of underwater video, its contribution is omitted in their method. The
ULR is also difficult to acquire due to organic sediment on the seabed. The proposed solution
to this is the use of convolution with a Gaussian kernel to obtained a visually-smoothed copy
of the original image. Although these methods are shown to, in principal experimentation,
show some benefit in creating seamless mosaics, the results are not quantitative. Due to
these results being qualitative nature, they prove difficult to discuss. The authors note the
inclusion of these methodologies into IFREMER’s commercial mosaic software MATISSE for
further experimentation.

It has been stated that there exist problems with relying on RGB data for the collection,
and automatic analysis of, coral reefs. Gleason, Reid, and Voss (2007) investigated the use of alternative imaging sensors, capable of acquiring digital image data with different colour band configurations. One such project looked at the use of a multi-spectral camera (MSCAM) which was built to cover six bands, with footage gathered to be compared to an RGB baseline. Their MSCAM is based on previous work at investigating the use of multi-spectral imaging on biotics underwater, with simplified optics giving a smaller FOV. Experiments were conducted using two datasets collected from coral reefs, and two synthetic sets obtained from an underwater tank in an academic context. Comparison is performed using Grey-Level Co-occurrence Matrices (GLCM) of the captured images, with metrics extracted being contrast, correlation, energy and homogeneity. They state that high-spectral resolution does improve classification of coral reefs, but the narrow bands separately were not a complete solution over RGB. This leads to interesting ideas about further characteristics of all wavelengths (visible & invisible) and the potential meta-data (data from non-imaging sensors in conjunction with the image data) that can be collected to automate classification.

Extending their work on feature extraction and adapting SLAM to underwater environments (Aulinas et al., 2011a), Aulinas et al. (2011b) present findings of applying this method to the SPARUS AUV in capturing Bathymetric data. By using Leonard’s concept of point features (Leonard and Durrant-Whyte, 1991), they are able to construct on-the-fly geometric representations of the sea-floor to aid in navigational drift experienced by the AUV in the currents. They demonstrate that by implementing this revised SLAM method, the end point of the mission is not as susceptible to offsets that would occur using traditional sonar & doppler-based methods.

It is not only the sea-floor topography that is of interest, techniques of mapping and registration have been applied to biotic habitat image sensing also. Benthos is a term used to describe the biotics present in the benthic zone of the water column. Gonzalez-Mirelis et al. (2009) describe traditional techniques of mapping these organisms in a Swedish Fjord. It describes the extraction of statistics based on human-expert analysis from underwater video data.

A recent publication from UDG is based around the automated classification and mapping of bacterial mats. Bacterial mats are grouped colonies of microbial life that occupy areas of the seabed. Their algorithm includes a preprocessing step of subdividing the images into several subregions of 192 × 192 pixels. Each of these regions is then subjected to Contrast Limited Adaptive Histogram Specification (CLAHS) to improve the visibility of the area. Superpixels (Achanta et al., 2012) (Figure 2.20) are used to isolate areas of the image in the form of Turbopixels (Levinshtein et al., 2009), to deal with areas on a higher-abstraction than per-pixel data (see Figure 2.20). These regions are then subjected to feature extraction through the use of Gabor filters of 4 sizes and 6 rotations, LBPs and GLCM. By extracting features
such as standard deviation and mean from these metrics, a feature vector is obtained that describes the superpixel within the overall mosaic structure. This complex approach results in reported overall accuracy of 99.7% with a runtime of 43.5 minutes (Shihavuddin et al., 2013). Although higher-precision than existing techniques, it is one of the slowest methods computationally.

Following on from this research, UDG are currently involved in 2 major European projects, MUMAR and PICMAP. The former, MUMAR, is specifically targeted to monitor marine habitats with a view to manage ocean resources, both from a scientific and conversational standpoint. By building upon the work discussed (Gracias et al., 2008; Campos et al., 2014; Shihavuddin et al., 2013; Armstrong et al., 2006; Elibol et al., 2014; Aulinas et al., 2011a; Campos et al., 2013; García, Nicosevici, and Cufi, 2002; Shihavuddin, Gracias, and García, 2012; Campos, García, and Nicosevici, 2011; Aulinas et al., 2011b), MUMAR aims to develop new high-resolution automated mapping software for large-scale mosaicking of the sea-floor. PICMAR takes a more industrious perspective, aiming to improve upon the state of art in multi-modal data registration, 2.5D\(^8\) mapping and congealment to aid in marine infrastructure planning with its focus on increasing the economic strength of Europe in this regard through the use of intelligent software systems.

\(\text{\footnotesize\textsuperscript{8}}\text{Where estimation of the third dimension is obtained as a projection of 2D sources.}\)
2.7.3 Humanities, Official & Cultural Heritage

We do not know how many beautiful museums exist on the bottom of the Mediterranean.

– Hanumant Singh, TEDxWoodsHole Talk

Earlier work conducted at WHOI discusses the opportunities of digital imaging for examining sunken shipwrecks. Focusing on a Roman shipwreck in the Mediterranean, they manually define key-points across mosaics before using co-ordinate mapping to align the images based on these points (Eustice, Singh, and Howland, 2000; Whitcomb et al., 1999; Singh et al., 2000). The resultant mosaic is seen in Figure 2.21.

Later, Singh, Howland, and Pizarro (2004) built upon this work, comparing the originally captured data versus that obtained from the JASON ROV, noting the clear improvement in fidelity. The manual key-point definition and spatial registration methods remain the same, as the challenges are not overcome by the increased image capture resolution.
Recent developments in this area come from UDG looking at merging Bathymetric and optical cues for underwater shipwrecks in 2013 (Campos et al., 2013). This paper was unavailable for review.
2.8 Discussion, State of the Art & Areas of Expansion

This chapter has reviewed the theory, advances and applications of computer science in the underwater environments. We have discussed the fundamental problems that have been, and continue to be researched. The fullest extent of the applications of underwater video analysis has not been covered, only the subset of that which is most applicable to this area.

The time-line of research into the capturing of data seems to be favouring a move towards AUVs over ROVs. This seems sensible, as on a monetary and time basis, the cost of deployment is vastly reduced with lower requirements of skilled-operators to continually manage the hardware. On missions approaching months in length, the benefits of this approach are clear. There does seem to be an opportunity to further develop ROV technology to reach semi-autonomy in certain circumstances. A potential hybridised approach, whereby an ROV could be shifted into autonomous mode to follow a transect could be realised. However, given the complexities of ROV navigation and the subsequent required remote processing power, achieving this goal as a sole researcher is not realistic.

Identifying a novel area to research in this domain is difficult, not only due to the wide spread of existing work and the key institutions participating therein. Typically, the researchers are numerous, with projects that cover many years and encompass researchers of all levels. Indeed, many of them are permanent departments within their academic structure. KESS projects have a budget allocation for spending on equipment, up to £1,000, as well as limited time to perform the research of twelve months. It is important to also consider the time-line of this research at an institution with no prior history in the domain.

The approach of this research is targeted at the application of general computer vision techniques, and machine learning classifiers to address the problem of substrate recognition in Cardigan Bay. The focus of the project is the visual identification of unique substrates, using new data and evaluating cost-effective methods of data retrieval. Due to the time and monetary constraints on this project, relative to those discussed already, intelligent planning and allocation of resources will be key in ensuring the viability of the project. This area is relatively unexplored, with the focus of existing research being monitoring and assessment of fauna (Section 2.7.1) and creating 2.5D or 3D bathymetric maps (Section 2.7.2).

Further research is conducted, and discussed, into different methods in computer vision, machine learning and original data capture in Chapters 3 & 4. Extensions to this work, that were simply too large to undertake given the constraints, are discussed in Section 6.2.
Chapter 3

Datasets

A number of datasets have been collected through various projects worldwide. The reason and project purpose for these datasets are varied, from ecological surveying, disaster management and seabed mapping. Many of these were not planned with applying computer vision & machine learning techniques in any meaningful way, due to their biological basis.

This chapter describes the datasets chosen for these experiments, their methodology and purpose in collection, their applicability to our goals and any technical or ecological constraints they present. In conjunction with stating the chosen datasets, the classification schemas used are defined and justified.

Collecting data via OpenROV specifically for testing the chosen methods is discussed, in construction and strategies.
3.1 Classification Schema

There exists a document published by the then Countryside Council for Wales (CCW) which defines 31 distinct classifications for marine and coastal habitats around the country’s waters Ramsay, 2010. These 31 classifications consider many factors in their definitions, most notably:

- **Composition** - the biological make-up of the sea-floor deposits, both surface-visible and sub-surface.

- **Fauna** - the wildlife present in any given area, including densities thereof.

- **Ecology** - habitats of the area, including fish species.

- **Location** - distance from shoreline, depth of sea-floor and temperature.

Due to the biological and ecological bases of this classification schema, the differences between classifications may not be immediately apparent nor even distinguishable from a computer vision perspective.

![Figure 3.1: Images taken from CCW classification schema, illustrating the visual similarities between distinct classes (Ramsay, 2010).](image)

As illustrated in Figure. 3.1, the visual appearance of the three supplied images is not hugely different from a human observer’s perspective. The key difference in definition relates to the depth of the substrate and its subsurface composition – two parameters that computer vision is not yet able to infer.

In conjunction with this ambiguity between classes for this project’s goals, certain classifications are not applicable for this problem. This project is concerned only with underwater video, and, specifically, those substrates which are present in the area of Cardigan Bay. Marine biologist Philip Hughes, of FoCB, identified 13 applicable substrates for consideration in this project’s work. These 13 are noted, with their original CCW description, in Table. 3.1, with those omitted being related to beach composition, cliffs and land-substrates.
TABLE 3.1: Classes from the CCW schema selected as candidates.

<table>
<thead>
<tr>
<th>CCW ID</th>
<th>Visual Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No relevant data present</td>
</tr>
<tr>
<td>14</td>
<td>Vertical sub-tidal rock with associated community</td>
</tr>
<tr>
<td>16</td>
<td>Coarse sands and gravels with communities that include large and/or long lived bivalves</td>
</tr>
<tr>
<td>17</td>
<td>Maerl beds</td>
</tr>
<tr>
<td>18</td>
<td>Stable predominantly sub-tidal fine sands</td>
</tr>
<tr>
<td>19</td>
<td>Sub-tidal stable muddy sands, sandy muds and muds</td>
</tr>
<tr>
<td>20</td>
<td>Predominantly sub-tidal rock with low-lying and fast growing faunal turf</td>
</tr>
<tr>
<td>22</td>
<td>Shallow sub-tidal rock with kelp</td>
</tr>
<tr>
<td>23</td>
<td>Kelp and seaweed communities on sand scoured rock</td>
</tr>
<tr>
<td>24</td>
<td>Dynamic, shallow water fine sands</td>
</tr>
<tr>
<td>27</td>
<td>Biogenic reef on sediment and mixed substrates</td>
</tr>
<tr>
<td>28</td>
<td>Stable, species rich mixed sediments</td>
</tr>
<tr>
<td>29</td>
<td>Unstable cobbles, pebbles, gravels and/or coarse sands supporting relatively robust communities</td>
</tr>
<tr>
<td>31</td>
<td>Seagrass beds</td>
</tr>
</tbody>
</table>

By anticipating the problem illustrated in Figure 3.1, a second classification schema was established, based on the CCW and grouped by similar visual appearance. Similarity of visual appearance was agreed upon by the author and Philip Hughes (FoCB), based on studying the examples in the CCW document (Ramsay, 2010) and reviewing the video footage contained within the datasets, where the overlaps occurred.

Table 3.2 maps six concrete classes to their approximations in the CCW schema. Five of these classes are concrete visual categories with shared appearance and biological basis, with 0 being placed for no relevant data.

In both classification schemas, a required catch-all category classification of 0 is required. Due to the nature of recording video, especially across transects underwater with launching & surfacing, there will be periods of the video that do not contain relevant data.

Specifically, when the apparatus is in motion and there is limited visual information, it is not possible to accurately classify the seabed substrate at the moment in time. Using temporal information this may indeed be possible, but not on a per-frame basis.
### Table 3.2: Our texture-based classification schema.

<table>
<thead>
<tr>
<th>ID</th>
<th>CCW</th>
<th>Visual Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>No relevant data present</td>
</tr>
<tr>
<td>I</td>
<td>18, 19, 24</td>
<td>Fine sands</td>
</tr>
<tr>
<td>II</td>
<td>16, 28, 29, 31</td>
<td>Coarse sands with occasional rocks and fauna</td>
</tr>
<tr>
<td>III</td>
<td>20</td>
<td>Pebbled seabed with occasional rocks</td>
</tr>
<tr>
<td>IV</td>
<td>14</td>
<td>Predominately large-boulders</td>
</tr>
<tr>
<td>V</td>
<td>17, 22, 23, 27</td>
<td>Coral &amp; rich in organic life</td>
</tr>
</tbody>
</table>

Figure 3.2 contains a number of scenes from within the Bangor dataset that have been classified as 0. Within the classification system as it is presented in this project, these frames are ignored.

![Figure 3.2: Class 0 frames from videos in the Bangor dataset. Frames deemed to have no information relevant to the application of computer vision in ascertaining seabed substrate.](image)

Due to the huge variety of possible scenarios present that would be classed 0, the technical complexity of the hierarchical classifiers required to correctly label it is beyond what was achievable in the time-frame.
3.2 Available Datasets

There exist several other datasets available for free use in research, such as Fish4Knowledge and TRIDENT (Section 2.4). The selection of the data used in this paper was taken based on those most appropriate for our target system, and those which were available at the time. Variety of substrate data is of key importance, whereas the focus of other datasets may specifically be fish monitoring from a fixed-position array or the development and testing of underwater robotics. As a final note of justification, this research is specifically aimed at Cardigan Bay, Wales, which served as a key selector in dataset choice.

3.2.1 Bangor

Originally collected by Lambert et al. (2013) at Bangor University, these videos have been made available to this project. In fact, the work conducted at Bangor University studying the benthic habitats of Cardigan Bay serves as a precursor to this project itself. These videos have not been released to the public domain, and are the basis of a biologist’s survey into the area. The data was collected via an agreement with a local fishing vessel, and is made up of 10 separate videos at different transects. It follows on previous work with a similar set-up used with laser lines to ascertain the topology of the seabed.

3.2.2 Fish4Knowledge

Over a period spanning years, the Fish4Knowledge group amassed a collection of approaching (in total) nearly 200 TB of video data (320 × 240 resolution at 5 fps). The project has made all this data available to other researchers\(^1\).

The primary issue with considering this dataset is that it simply was not available at the time the project began. In fact, it became available quite close to the end of the project and, as such, it was not possible to include this in the consideration for the datasets chosen. Another key issue with this dataset is simply that the primary goal of the Fish4Knowledge research was to investigate the fish species present at the video sites, with video being taken at the same time everyday at specific intervals via fix mounted cameras.

The research question we are proposing is looking at the seabed substrates and habitats that are present, and (in the future) how they change over time. As such, the utility of this data must be questioned as a given video from this dataset does not contain any substrate variation within it. Theoretically, it could be used to test existing classifiers in order to see

\(^1\)http://groups.inf.ed.ac.uk/f4k/F4KDATASAMPLES/INTERFACE/DATASAMPLES/search.php
whether the backgrounds of the videos could be correctly classified. Although, even with this goal in mind our research is targeted at Cardigan Bay, and not the underwater habitats present in the sites recorded in Taiwan.

### 3.2.3 TRIDENT

This dataset was collected by the TRIDENT group at UDG, and the videos focus on the research that was conducted into underwater robotics at the group. As with the Fish4Knowledge project, these were not made with substrate variation in mind and do not contain the requisite data. They are also not recorded in Cardigan Bay.

### 3.2.4 Friends of Cardigan Bay

Experimenting with consumer technology, FoCB have recorded some underwater footage from the Cardigan Bay area as part of their monitoring work.

This dataset is not in the public domain, but has been made available for the purposes of our project. The locations where the footage has been taken are in the geographical area in which we looking to conduct research, and therefore contains the substrate variations that we are looking to find. Specifically, we will look at two locations in Cardigan Bay; Aberystwyth and Clarach. Both of these locations have underwater footage present in this dataset.

Upon reviewing the footage, it becomes clear that there is sufficient substrate variation present within the recordings, however the quality of the videos themselves is very low and, as such, presents the technical difficulties discussed in the previous analysis of underwater video.

### 3.3 Chosen Data

Two different video sets are used in this research, collected via distinct methods. We focus only on these two sets in this project, and select specific videos from within each dataset. It is not possible to use every video, as many have the same substrate throughout, so here only those videos containing multiple substrate changes are selected to gain quantitative data for analysis.
3.3.1 Bangor University

These videos were generously offered by Bangor University collected whilst researching the effects of trawling on the seabed (Lambert et al., 2013), with Figure 3.3 showing the differences between estimated and observed substrates. Typical video scenes are shown in Figures 3.6 & 3.7, showing the class diversity in the substrates observed.

![Figure 3.3: Sea-floor images for the planning and post-experiment analysis of the Bangor experiments (Lambert et al., 2013).](image)

Methodology

The data was captured using a GoPro camera attached to a sled. The camera faces forward on the sled, and each video begins with the operator (a fisherman) recording the current GPS location by showing a hand-held unit to the camera. The sled is then dragged along the sea-floor by the vessel above, being attached by a tether (in this case, rope). As there is no monitoring on the vessel’s deck, the operator(s) does not know what data is being captured. The vessel continues along a set transect, before the sled is removed from the water and its final GPS location recorded.

The motion of the boat whilst pulling the sled in the water causes the raising and lowering of the sled down, moving it further along the transect, capturing data (as illustrated in Figure 3.4). As such, the continuity and predictability of the data is difficult – this is due to the fact that there are certain erratic movements in the cameras motion field, and in certain circumstances we can see dragging across the sea-floor depending on the altitude of sled, relative to the boat. With no means by which to judge the altitude from the vessel’s deck, the data capture itself contains a lot of random elements. Due to the lack of predictability, accurate modelling of the data held within it is difficult.
As with the FoCB dataset, this is of our target area Cardigan Bay and it does show the present substrates. The capturing from the GoPro itself is also very high-quality, although this is mitigated somewhat by the quality of planning across the transect the motion difference noted.

**Constraints**

Unfortunately, due to the lack of substrate variety in the videos, we were unable to use all of the dataset in our experiments. It was the case that many videos within the dataset would contain only one substrate throughout, which would not provide any basis by which to build a classifier to differentiate substrate classes.

Using SITE7 as an example, the entirety of the video shows the sea-floor covered by a large collection of brittle star. It is not possible for a human observer to see enough of the sea-floor to make a confident judgement of the substrate being observed, ignoring the fact even if it were possible for a ground truth to be established for this video, that the system would then learn the brittle star pattern as that class. If it were the case that some form of template matching, or brittle star detector / excluder were available, it may have been possible to include the video. Due to constraints such as these, we selected sites 2, 3 and 10, as they had been confirmed by a marine biologist to contain variation in the substrates.
Due to the nature of underwater video captured via trawler, seabed disturbance is to be expected upon the trawler’s impact. Figure 3.5 demonstrates this in effect, with a trawler maintaining its position on the seabed long enough for the debris to settle. During the marking up phase, the start and end sections of a classification range are set to omit these frames.

(a) Debris from impact

(b) Viable frame

Figure 3.5: Selecting viable frames for use as training data involves manually discarding those with trawler-impact debris present.
Other videos in the dataset do not contain sufficient variation in the substrates within, or the substrates themselves are obfuscated for other reasons. Considering SITE7, Figure 3.8 shows that, by chance, the location chosen was covered in Brittlestar. Ascertaining the substrate type in this video is, therefore, impractical at best.

Although a great candidate for organism recognition, such work is outside the scope of this project.

3.3.2 Friends of Cardigan Bay

As part of ongoing surveying of Cardigan Bay, FoCB use a variety of methods to gain information. Typically these involve diving and capturing imagery using hand-held cameras, or
simply reporting findings textually.

The organisation has experimented with different methods of data retrieval, and performed exploratory work using a towed camera system in the bay both for itself and the CCW.

**Methodology**

FoCB used a commercial DVD recorder to gain underwater footage. The method was to drop a waterproof camera over the boat’s edge, with real-time viewing via LCD on the deck. The operator would use their knowledge of the seabed to manually raise and lower the camera (weighted to achieve stability in the water column). Figure 3.11 illustrates this technique, and discusses the issues encountered due to instability.
Chapter 3. Underwater Video Datasets

Figure 3.8: SITE7 frames demonstrating the lack of variety present in the video, and the obfuscated nature of the seabed within. Thousands of brittle star can be seen throughout the video, making classification of the seabed beneath impractical.

Originally encoded in PAL resolution of effective $768 \times 576$, as an interlaced signal at 25 fps. Digitisation was performed using consumer TV transfer software, resulting in a series de-interlaced $320 \times 240$ resolution video files. Due to lower resolution of these files, the unknown de-interlacing algorithm and the low-bitrate of the encoding, the image fidelity is very low.

Constraints

Though this dataset directly covers the targeted area of sea-floor and, from a biologist’s perspective, contains much sea fauna and substrate variation, there exist a number of issues in analysing it algorithmically. Firstly, the nature of the data acquisition was purely down to curiosity with no set parameters for areas covered nor duration. The result of this is that a number of the videos contained in this set are simply too short to gain meaningful data or perform analysis.

Further issues come from the methodology chosen, in conjunction with the sensor’s low resolution and low contrast ratio — there is significant motion seen in the video itself. This is due to the apparatus’ instability in the water column, specifically when influenced by currents. This unwanted motion causes problems in the temporal modelling of the underwater scene, as no associated telemetric data could have been captured using the set-up noted. The angle and field of view also result in the upper-half of the majority of frames containing only turbid sea-water, and no substrate information (Figures 3.9 & 3.10).

Due to these factors, creating any set of seabed maps from the FoCB dataset is difficult, but it does serve as a challenging testing target for the classification systems.
3.4 Original Data Capture

Part of this project aims to investigate the use of bespoke methods to gather data from Cardigan Bay. By reviewing state-of-the-art with respect to underwater robotics and imaging hardware, we propose two methods which are both reasonable in terms of time and cost to gain original data from the Cardigan Bay area. With the help of Philip Hughes (FoCB), we have access to a boat which will take us to two areas of interest, specifically the southernmost two of the Sarnau, Sarn-y-Bwch and Sarn Cynfelin (Figure 2.2), where we hope to use our chosen method to gather new data to test our classifiers built from the chosen training datasets.
3.4.1 Refinement of FoCB Technique

The varied nature of the content captured in the FoCB dataset led to investigations in how to better the setup. Principally, the points that needed to be addressed were the stability in the water column, image quality, and reliance on an expert-operator.

Figure 3.12 is a mock-up of a design which would incorporate a mechanical winch that would be mounted to the vessel, with two data-capable cables travelling down to a weighted platform that goes underwater. Upon this platform, 3× GoPros would be mounted on the same plane and aligned height-wise in the water-column. The reasoning for this was to produce video sequences covering a far wider field of view by aligning frames in post processing.
Figure 3.11: Another tethered method, FoCB used a submersible camera that was attached to recording apparatus on the boat’s deck. The camera module itself was suspended in the water column by its tether. The altitude was manually managed by the operator. The direction of the image sensor cannot be guaranteed as it is freely rotating according to current and influence of boat motion, whilst the vessel travels along the surface.

A sonar module underneath would relay the current altitude relative to the sea-floor, and an embedded Arduino system would then automatically alter the altitude by engaging the winch in either direction.

Conceptually, and theoretically, this would provide an excellent, stable and consistent capturing of a given transect across the sea-floor. In practice, however, its motion would be tied directly to the boat’s, which would lead to issues of altitude depending on surface stability, and the currents induced from operating the vessel.

3.4.2 ROVs

ROVs have been, historically, expensive and out of reach of smaller research groups. Many research institutions, as outlined in the literature review (Sections 2.4 & 2.5.3) offer their designs and fabricated ROVs for sale to both academia and industry.

Commercially, ROVs are also available at different budgets. HyrdoView is a series of underwater cameras marketed by AquaBotix, whose purpose is described as for curiosity whilst boating, or equipment inspection underwater. Figure 3.13 shows the Sport model, which comes with either 8GB or 32GB internal memory, a 1MP image sensor and LED illumination banks capable of delivering 80 lumens. The system is controlled via an iPad app, where still
images can be saved to the HydroView’s internal memory for later extraction. Its run time is stated as 2 – 3 hours. This model is sold for $5,500. As such, using this device for the acquisition of underwater video is not part of its intended use, and the specifications thereof do not permit this mode of operation. It is, therefore, a highly restricted ROV from a research perspective.

VideoRay offer ROVs at varying specifications to meet many mission objectives. The VideoRay Explorer offers a depth-rating of up to 75m, with various extension possibilities. The VideoRay Pro 4 is what is considered to be the de-facto entry-level commercial ROV, finding use in law-enforcement, disaster recovery and marine surveillance alike. It is a micro ROV that weighs 6 kg, produces 21 lb of forward thrust with a custom brushless-thruster design and costs approximately $50,000. It is equipped with a wide-dynamic-range colour camera module offering standard definition (SD) 570p resolution, two high-powered LED units capable of emitting 3,600 lumens, and is controlled via software or a dedicated hardware remote system from an operator on the surface.

Identifying this high cost-of-entry, American entrepreneurs created a crowd-sourced project for an open source and affordable ROV. Priced at approximately $1,000, the ROV is delivered as a kit that, once assembled, is capable of depths of up to 100 m, with LED units for illumination and a tiltable 1080p camera module, and live video-feed. The system uses software running via a web-browser on the operator’s laptop to control the device in the water.
The OpenROV achieves its low cost re-purposing various components from remote-controlled vehicles and distributing the research and development, testing and design process amongst the open source community. By running Ubuntu Linux on a Beaglebone Black, modification and extension of the internal system software is trivial. Figure 3.14 shows the difference in size, and optical lensing of both the OpenROV and the VideoRay Pro 4.

Due to the equipment budget made available to this project, the OpenROV was the only option available that would permit the use and evaluation of ROVs in this domain. The final, constructed OpenROV is shown in Figure 3.15.
Figure 3.15: Our final, v2.6 OpenROV constructed and pictured at testing up Nant-y-moch, Wales. Note the roll of solder underneath the main capsule, responsible for variable ballast in waters of different salinity.

Table 3.3: Technical details of the digitisation of the datasets. All encodings make use of variable bitrates, with ranges noted. FPS = frames per second.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Resolution</th>
<th>FPS</th>
<th>Bitrate (Mbps)</th>
<th>Encoding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bangor</td>
<td>1280 × 720</td>
<td>60</td>
<td>0.5 - 35</td>
<td>H264</td>
</tr>
<tr>
<td>Friends of Cardigan Bay</td>
<td>320 × 240</td>
<td>25</td>
<td>0.4 - 0.65</td>
<td>WMV</td>
</tr>
<tr>
<td>OpenROV</td>
<td>1280 × 720</td>
<td>21.5</td>
<td>0.7 - 30</td>
<td>H264</td>
</tr>
</tbody>
</table>

3.5 Dataset Index

The tables provided in this section (Tables 3.3, 3.4 & 4.3) detail the final videos that were available at the time of the project, the datasets technical specifications, and the ground truth information obtained for selected videos in the experiments.

Although the OpenROV is mentioned, its videos are not in the Table 3.4 due to the inability of capturing sea-floor data with it. However, in order to aid discussion, its specifications are noted.
Table 3.4: An index of the files available in both datasets used. LEN = length, NFR = approximate number of frames, GT = ground truth present, TRN = used in training.

<table>
<thead>
<tr>
<th>REF</th>
<th>Filename</th>
<th>LEN (s)</th>
<th>NFR</th>
<th>GT</th>
<th>TRN</th>
</tr>
</thead>
<tbody>
<tr>
<td>SITE1</td>
<td>Site 1.MP4</td>
<td>16:04</td>
<td>57,840</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SITE2</td>
<td>Site 2.MP4</td>
<td>15:38</td>
<td>56,280</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SITE3</td>
<td>Site 3.MP4</td>
<td>20:54</td>
<td>75,240</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SITE4</td>
<td>Site 4.MP4</td>
<td>14:52</td>
<td>53,520</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SITE5</td>
<td>Site 5 (2).MP4</td>
<td>17:46</td>
<td>63,960</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SITE6</td>
<td>Site 6.MP4</td>
<td>18:49</td>
<td>67,740</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SITE7</td>
<td>Site 7.MP4</td>
<td>15:03</td>
<td>54,180</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SITE10</td>
<td>Site 10.MP4</td>
<td>16:20</td>
<td>58,800</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Friends of Cardigan Bay Dataset

<table>
<thead>
<tr>
<th>REF</th>
<th>Filename</th>
<th>LEN (s)</th>
<th>NFR</th>
<th>GT</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABER1</td>
<td>aber harbour 26th sep 08 tape 2.wmv</td>
<td>36:17</td>
<td>54,425</td>
<td>✓</td>
</tr>
<tr>
<td>ABER2</td>
<td>clarach consti and harbour pollock.wmv</td>
<td>36:50</td>
<td>55,250</td>
<td>✓</td>
</tr>
<tr>
<td>ABER3</td>
<td>pinnacles july18.wmv</td>
<td>9:14</td>
<td>13,850</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 3.5: Expands upon Table 3.4 of those selected, showing the present CCW and simplified classes in each video. Numbers in parenthesis under the Simplified column, represent the number of the encompassed CCW classes present in the video.

<table>
<thead>
<tr>
<th>REF</th>
<th>Annotator</th>
<th>CCW</th>
<th>Simplified</th>
</tr>
</thead>
<tbody>
<tr>
<td>SITE2</td>
<td>Philip Hughes</td>
<td>16, 22, 29</td>
<td>II (2), V (1)</td>
</tr>
<tr>
<td>SITE3</td>
<td>Philip Hughes</td>
<td>16, 22, 28, 29, 31</td>
<td>II (4), V (1)</td>
</tr>
<tr>
<td>SITE10</td>
<td>Philip Hughes</td>
<td>16, 18, 24, 27</td>
<td>I (2), II (1), V (1)</td>
</tr>
</tbody>
</table>

Friends of Cardigan Bay Dataset

<table>
<thead>
<tr>
<th>REF</th>
<th>Annotator</th>
<th>CCW</th>
<th>Simplified</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABER1</td>
<td>Philip Hughes</td>
<td>18, 22, 23, 24</td>
<td>I (2), V (2)</td>
</tr>
<tr>
<td>ABER2</td>
<td>Philip Hughes</td>
<td>22, 23, 24, 29</td>
<td>I (1), II (2), V (2)</td>
</tr>
</tbody>
</table>
Chapter 4

Methodology

This chapter details the design decisions taken in creating and conducting the experiments. The rationale for technology choices and approaches is discussed, and the experiment hardware and software platforms are clearly defined.

The methodology of the experiments themselves is then discussed, defining the approaches, theory and anticipated results. Where applicable, shortcomings and compromises in decisions are also stated, including their influence on the final results.

This project looks at developing a practical solution with an industry partner, as such there is a discussion of software engineering considerations taken, including choices made in order to achieve the results present in Section 5. This chapter looks at both these practical challenges, and the scientific methods employed in the automated analysis of underwater substrates.
4.1 Platform

Key considerations in platform choice were cross-platform compatibility, including any and all third party libraries being used. Software performance was another large area of consideration, as the datasets themselves would be so large in RAM (post decompression) that the overhead of any language or system must be taken into consideration at the time of design.

In conjunction with these parameters, it had to be assured that the requisite libraries existed on the target platform. And, time to develop (language and platform complexity) were other aspects that were considered, due to the short nature of this project.

Given that the ground truth aspect of the system was an integral part built-in, it was key that the technologies chosen would be portable enough to run on the marine biologists’ systems. Therefore, some capacity to distribute OS-independent binaries became paramount.

4.1.1 Language

Several languages and their respective platforms were considered for this project, this subsection aims to discuss the main ones, their applicability and constraints.

Powerful, and cross-platform (based on compiler and library choices) C++ and its parent, C, are typically go-to languages for performance-critical systems. Developing programs quickly using these technologies depends on the author’s prior experience and debugging skills, either gdb or IDE based. Due to the varied nature of dependencies across different platforms, and catering to the idiosyncrasies that platform-specific implementations require, this approach was deemed to be too time-consuming development-wise for this project.

Java is another cross-platform technology with many key libraries available. It does mandate the requisite Java Virtual Machine (JVM) and associated libraries be installed on the host machine, therefore the portability that it champions becomes environment and configuration specific. Distribution of binaries that enclose their own JVM is non-trivial. Another typical strategy of research in this domain is to use dedicated, proprietary systems such as Simulink’s MATLAB, the benefit of this being that a host of methods and approaches are available as so-called 'tool-boxes', which enable the rapid prototyping of methods. Although convenient, the system is extremely expensive for non-academic use, and given its proprietary nature, not easily scaled horizontally\(^1\) and is not easy to generate and distribute binaries to third parties.

\(^1\)Horizontal scaling refers to the distribution of work across several machines. Conversely, vertical scaling refers to the distribution of work across more hardware (RAM, CPU / GPU cores) within the same system.
In recent years, there has been a move towards using Python more in both industry and academia. Researchers in institutions such as Google, Amazon and CERN use the language. An increasing library of scientific-orientated modules are becoming available. A key trait which permits this crossover of usability and time-to-develop is Python’s built in ability to wrap around pure C extensions; providing the trade-off of performance and time requirement. Due to the nature of the Python interpreter and specification, building self-contained binaries is a simple process — encapsulating internal and external libraries in a transparent manner. For these reasons, Python was chosen as the basis of this project.

### 4.1.2 Libraries

A number of computer vision and machine vision libraries exist, with many options available depending on the platform chosen. As this project is based on Python, we considered the use of those languages which have up-to-date and maintained Python bindings.

These options included EVE\(^2\), SimpleCV, mahotas (another open source library looking at pattern recognition) and OpenCV. We decided to use OpenCV due its large collection of implemented algorithms, its permissive license and its ubiquity within the image processing domain (Nadarajan et al., 2011).

As a final note, graphical interfaces were created using Qt4 (with Python bindings) to ensure ease of use across multiple platforms. For reference, all third party libraries used, including their versions and notes, are defined in Table 4.1.

### 4.1.3 Hardware

Development and testing was performed across a number of different computing devices, all running different OSes. The configurations of these systems are detailed in Table 4.2. The subtle differences in the configuration of these machines is important, when considering the performance characteristics that were being evaluated.

\(^2\)EVE is a small library for image processing provided by Dr. Adrian Clarke of the University of Essex — http://vase.essex.ac.uk/software/eve.html
Table 4.1: Libraries in use on the project.

<table>
<thead>
<tr>
<th>Name</th>
<th>Version</th>
<th>Notes</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpenCV (Nadarajan et al., 2011)</td>
<td>2.7</td>
<td>cv2 module for FFT convolution and general image processing</td>
<td>GPL</td>
</tr>
<tr>
<td>scikit-learn (Pedregosa et al., 2011)</td>
<td>0.13</td>
<td>Basis of all classifiers</td>
<td>BSD</td>
</tr>
<tr>
<td>matplotlib (Jones et al., 2001)</td>
<td>1.3.1</td>
<td>Used in research and evaluation of algorithms, including final plot generation</td>
<td>BSD</td>
</tr>
<tr>
<td>SciPy (Jones et al., 2001)</td>
<td>0.13.0</td>
<td>Linear algebra and statistics package, used in conjunction with NumPy. Permissive Enthought license</td>
<td>CUSTOM</td>
</tr>
<tr>
<td>NumPy (Jones et al., 2001)</td>
<td>1.8.0</td>
<td>Numerical algebra library used for fast matrix manipulation. Permissive Enthought license</td>
<td>CUSTOM</td>
</tr>
<tr>
<td>Mahotas</td>
<td>1.0.4</td>
<td>Used for LBP implementation</td>
<td>CUSTOM</td>
</tr>
<tr>
<td>Billiard</td>
<td>3.3.0.16</td>
<td>Drop-in replacement for native multiprocessing module, with community fixes</td>
<td>BSD</td>
</tr>
<tr>
<td>PyQt4</td>
<td>4.2</td>
<td>Used for ground-truth parts</td>
<td>GPL</td>
</tr>
</tbody>
</table>

Table 4.2: Computer systems used in the development, training and testing of these experiments. HT = Hyper Threading.

<table>
<thead>
<tr>
<th>Model</th>
<th>OS</th>
<th>CPU</th>
<th>RAM</th>
<th>SSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMBP</td>
<td>OS X 10.9.2</td>
<td>Core i7 (8-core HT)</td>
<td>16GB</td>
<td>✓</td>
</tr>
<tr>
<td>Z600</td>
<td>Windows 8.1 Pro</td>
<td>Dual Xeon (16-core HT)</td>
<td>12GB</td>
<td>✓</td>
</tr>
<tr>
<td>X220</td>
<td>Debian Jessie</td>
<td>Core i5 (4-core HT)</td>
<td>8GB</td>
<td>✓</td>
</tr>
<tr>
<td>VIGLEN</td>
<td>Xubuntu 14.04</td>
<td>Core i5 (4-core HT)</td>
<td>4GB</td>
<td></td>
</tr>
</tbody>
</table>
4.2 Performance

A key target in the computer vision world is approaching real-time analysis and information retrieval/extraction from (digital) images. The mammalian visual system accomplishes real-time processing, and in remote operations of underwater surveillance, it would be preferable to have immediate and correct feedback.

Realising this within the time, financial and computational constraints present in this project is not a simple task. This section aims to discuss the steps taken to break down large problems, address areas of poor performance and demonstrate the tangible, technical issues encountered.

4.2.1 Big Data

This is a term in research vogue at the moment — the idea of Big Data processing. What actually quantifies Big Data is very much open to interpretation.

Consider a given video, SITE2 from the Bangor dataset. At 1280 × 720 resolution, 60fps and 15:38 duration, we can estimate that there are 56,280 frames (ignoring lost frames in the decompression process) contained in this video. Each of these frames has three channels in our analysis, using RGB colour space. We now have 168,840 separate 1280 × 720 matrices available, effectively 155,602,944,000 pixel components.

Conservatively, if we assume that 30% of the video has been marked as class 0 (see Section 3.1), 118,188 matrices remain for processing. Extrapolating this over every video analysed, for every metric, every set of parameters, every classifier and every window type – it becomes clear that, computationally, this is a very expensive and demanding process.

4.2.2 Libraries

Library selection is a key consideration for performance. In conjunction with considering the big-data being worked with, we must consider the call-stack of each operation of deriving the feature vector, and using the classifiers. If we consider SITE2 once more as $n$ frames, where $n = 56,280$ assuming no class 0 present, our metric for calculating the feature vector will be called up to $n$ times.

Clearly, at very least this is $O(n)$ complexity. If we are to subdivide each frame in $n$ into a set $P$ of patches, we have a worst-case scenario of $O(n^2)$ computational complexity. Assuming $|P| = 25$, we will have $|P| \times n = 1,125,600$ unique calls to the metric function, traversing the call stack. For every selected video, with every metric and classifier, this becomes a big task.
It is therefore paramount that the code used to derive these feature vectors be as platform-optimised, numerically-stable\(^3\) and robust as possible. By using open source software where possible, these risks and weak points are mitigated somewhat by community involvement and inspection.

### 4.2.3 Workload Distribution

In order to ensure that the programs complete and algorithms converge in time, it was necessary to delegate different combinations of experiments to different pieces of hardware. No specific preference was given to a dataset or metric per machine. Job allocation was simply based on a FIFO queue that was manually managed based on experiment inclusion.

The datasets and libraries were constant amongst the platforms, as defined in Tables 3.4 & 4.1. Once the experiments had been completed, the resultant CSV files were collated and sorted into a spreadsheet manually.

### 4.2.4 Parallelisation

Distributing the work across separate machines is one method of segmenting workloads via horizontal scaling. Within each system, there is the possibility of vertical scaling using the number of simultaneous threads or CPU & GPU cores available to the host OS. For these experiments, the workload was distributed across the systems defined in Table 4.2.

Parallelisation of this nature features heavy in these experiments. Wherever a problem is linear, in the sense that we are principally concerned with a tuple \(N\) that consists of \(N = (F_n, F_c, F_k)\) where \(F_n\) is the current frame, \(F_c\) is the actual substrate classification and \(F_k\) is the predicted class. The basic methods applied in these experiments do not consider temporal information from the visual pattern recognition perspective. It is therefore possible to state that the problem of learning or classifying \(n\) frames, may be performed over \(n\) simultaneous processes, resulting in a computational runtime of \(O(1)\).

In reality, the technical and financial requirements of parallelisation at this scale are far beyond the scope of this project. Therefore, a dynamic system was implemented to use synchronised lists of jobs, which would automatically determine the maximum possible simultaneous tasks on a given system. This would be derived through processor cores, and available memory. The allocation of its workload per machine is then further distributed on a per-experiment basis. Figure 4.1 shows this architecture.

---

\(^3\)The numeric stability of an algorithm notes whether it is susceptible to hardware-based bugs, such as integer roll-over or floating-point precision errors. The presence of which may cause anomalous results.
Chapter 4. Methodology

Figure 4.1: Illustration of experiment definitions, ground truth and source videos being distributed by machine. Further subdivision of the work happens through enhanced multiprocessing using the Billiard package. Eventually, all experiments converge by generating separate CSVs ready for analysis.

Training

The training stage is distinct from testing, and does not automatically lead to the latter. Using the parallelised processes mentioned, all combinations of video, classifier and metric
are trained in parallel. These are then saved to disk using a serialisation format native to the SciKit-Learn module (Pedregosa et al., 2011).

This file is saved with a particular filename that details the parameters and configuration used in training that module. Serialisation occurs using NumPy C-based data-structures, and therefore requires little caching and writes straight to disk (Jones et al., 2001). The operation is very quick.

**Testing**

Once a host of classifiers have been trained and the distributed processes finished, the application may begin testing the data. Initial testing is against the video to which it was trained. This is again performed by creating a synchronised queue, and determining the machine's available resources.

Experiments are performed, with data written to a CSV file per experiment, with each row relating to the testing of a frame or a window. This row also uses the tuple $N = (F_n, F_c, F_k)$.

Once these have completed, the CSV data is parsed using scripts within the software and statistics based on the simplified and CCW schemas are compiled. These are then entered manually into a spreadsheet for storage and analysis.

### 4.3 Frame Processing

In these experiments, we look at the characteristics and information that can be derived from individual video frames, whether in testing or training. Video analysis provides temporal information that is a series of time-based information that can aid analysis. Extensions to this work using the temporal data for classification are discussed in Section 6.2.

This section outlines the approaches of resizing the feature space, feature extraction and processes used to obtain our feature vectors.

#### 4.3.1 Full-Frame

The initial experiments focused on the use of the entire video frame as the source of the feature vector creation. No segmentation or input masking were performed. The motivation for this was to establish a baseline upon which further enhancement of the methodologies could be considered.
Using this approach has some clear drawbacks which are evident before consulting experiment results. Such issues as the formation of continuous patterns cast by artificial illumination may provide a sufficiently continuous pattern that could prove more important to the feature vector than the textural data within. Conversely, it could be treated as a constant and the differences in the texture within be enough to separate the classes.

Another key issue, particularly with the Bangor and OpenROV datasets, is that components of the apparatus may be visible throughout the video. Once again, the question remains whether their continued presence will affect the overall performance, or prove to be non-contributory. Figure 4.2 illustrates these issues.

![Figure 4.2](image)

**Figure 4.2**: This image highlights the constant presence of the sled either side of the frame, and the laser pointers reflecting particles in the water-column, and their reflected & scattered points of reflection on the seabed.

### 4.3.2 Patch-Based

Designed after-the-fact of the full-frame analysis, a more interactive method of finding textural information is through the decomposition of a video frame into patches of $m \times n$ pixels. To reduce auxiliary and irrelevant textural information from being considered in either the training or testing phase, we use a viability metric on each patch $x$ to ascertain whether it will be included in the experiments or not. This is simply performed by calculating the the mean intensity of the patch $\bar{x}$, such that if $i_{\text{MIN}} \leq \bar{x} \leq i_{\text{MAX}}$ then it is considered a viable patch, and is included in the set of patches for the current video frame.

The values $i_{\text{MIN}}$ & $i_{\text{MAX}}$ are calculated as the median of the possible frame intensity range $m$ (dependent on colour resolution) $\pm i$ where $1 \leq i \leq (\bar{m} - 1)$, we select $i = 70$ to approximate
the interquartile range\(^4\). For these experiments, any frames that do not satisfy this condition are discarded as not containing useful information, although extensions to this approach are possible (see Section 6.2).

In determining an appropriate window size, smaller experiments were run on shorter and less visually distinct videos from the FoCB dataset. These experiments evaluated the values of \(m, n \in [30, 35, \ldots, 60]\). In these tests, it was observed that values less than 45 suffered notable deterioration in classifier generality, whereas for those values greater, the improvement was marginal, as such, values of \(m = 50, n = 50\) were chosen. Figure 4.3 has a sample of patches taken of two different window sizes, showing the differences in texture capture between the two. Further to this, based on the conclusions drawn from the full-frame analysis where the inclusion of colour information did not significantly increase accuracy (as discussed in Section 5), colour information is discarded in the patch-based approach.

![Figure 4.3: Patches taken at 70 & 50 pixel wide windows. The difference in texture capture was not found to be relevant in small-scale testing, and the latter then chosen for having more representative training samples.](image)

Initially, a normalisation of the patch histograms was applied, however this was found to reduce the distance between distinct features in texture. The contrast between specific features caused weakened boundaries between classifications.

### 4.3.3 Features

A number of metrics are used to obtain the feature vectors used in the training and testing stages of the experiments. These metrics range from statistical representations, through to wavelet-based textural decomposition.

#### Histograms

The histogram is a statistical representation of data, which determines the occurrences of data points which are collected into bins of an arbitrary size. Pearson describes it as an estimate of the probability distribution of a continuous variable. The intervals at which

\(^4\)8-bit colour information, 256 levels of intensity
frequencies of data points can be recorded do not necessarily need to be uniform, but the boundaries of the bins must be touching.

We use $c$ bins, where $c$ is the number of intensity levels present in the image. Therefore, $c = 256$ in all experiments where histograms are taken as the feature vector.

**Greyscale Histogram Equalisation**

The process of histogram equalisation redistributes the intensity values present over the full-range of intensities available, the effect of this is contrast enhancement in the digital image (Pizer et al., 1987). The technique is used over many domains in digital image processing, and has found use in underwater imagery (Borgetto, Rigaud, and Lots, 2003; García, Nicosevici, and Cufí, 2002; Pizarro and Singh, 2003; Shihavuddin et al., 2013).

The purpose of using histogram equalisation in greyscale is to alleviate problems related to non-uniform illumination, and the dominant colours present in the source videos. By discarding the colour information in a pre-processing step, only the textural information remains as a basis for feature vector encoding.

**RGB Histogram Equalisation**

In the case of colour histograms, RGB sub-channel histograms were taken after equalisation was performed on each channel separately. The three resultant histograms were then horizontally-concatenated into a feature vector of the form $\mathbb{R}^{768}$. All colour information for each channel is preserved in this form.

This representation loses all colour channel correlation between channels. Dimensionality reduction is not applied. This decision was taken to create said baseline data, and to avoid the potential computational overhead as described in Section 4.2 that feature selection (Guyon and Elisseeff, 2003) or PCA (Jolliffe, 2005) would require. It would also provide a baseline comparison with any future experiments with different colour spaces.

**Retinex**

Developed by Land (1986), the Retinex (*retina plus cortex*) algorithm attempts to map the dominant colours present in a given image into those which are more likely to be interpreted by the human visual system. The theory of this method relates to the way in which the visual system attempts tone-mapping based on local colour dominance in small areas, given the visible light available, as well as approximating the way in which the human brain tries to infer
colour based on expectation and inherent bias and subconscious processes (the Logvinenko illusion is an excellent example of how the same pixel intensities are perceived differently by humans, based on neighbouring luminance). Retinex considers the idea that illumination to commonly vary gradually over space, but reflectance varies abruptly based on the object being reflected. Spatial derivatives are taken to identify areas of large gradients, where over some threshold they are considered changes in reflectance, or if below, changes in illumination. Random paths are then taken in the image to identify a series of points, where the average is taken of their values to calculate the value of the current pixel in question (Morel, Petro, and Sbert, 2009).

In effect, Figure 4.4 shows how the visual appearance of the frame is altered from taking a heavily green hue into a representation that appears more like what we would expect the colour values to be.

The purpose of including this method in the analysis is to determine, in comparison with the simplistic histogram equalisation, whether the highlighting of certain features or transposition of intensity energies into different areas has any tangible effect on the classification accuracy.

Once the Retinex algorithm has been applied to either the full frame or patch, a histogram is then taken of that area, which becomes the feature vector.

**Gabor Wavelets**

A different approach to the problem is the use of textural image descriptors. This involves the use of a number of Gabor filters defined at a number of different orientations. It has been shown that the use of Gabor filters in this way approximates the way in which simple cell
receptors in the mammalian visual system process textures (Vitaladevuni and Domke, 2005; Lee, 2008; Jin, Angelini, and Laine, 2005; Meyer, 1993).

The Gabor filter $g(x, y, \lambda, \theta, \psi, \sigma, \gamma)$ used is defined in Equation 4.1, where $x' = x \cos(\theta) + y \sin(\theta)$ and $y' = x \sin(\theta) + y \cos(\theta)$. The parameters of this filter are $\lambda$, the wavelength of the sinusoid, $\theta$, the angle of filter rotation, $\psi$, the phase offset from centre, $\sigma$, the standard deviation of the Gaussian enveloping the filter, and $\gamma$, which is the spatial aspect ratio (Movellan, 2002).

$$g(x, y, \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + y'^2}{2\sigma^2}\right)\exp\left(i\left(2\pi\frac{\lambda}{\lambda}x + \psi\right)\right)$$ (4.1)

For our experiments, we use the parameters $\theta \in [0, 45, 90, 135], \sigma = 5, \psi = 90, \lambda = 50$ and a kernel size of 21. These values were decided upon through testing the impulse response given by Equation 4.1 against viable frames with different parameters until a whole-texture pattern was obtained. This was performed through consultation with similar approaches (Itti, Koch, and Niebur, 1998; Spampinato et al., 2010; Palazzo, Spampinato, and Beyan, 2012; Huang, Boom, and Fisher, 2012; Huang, Boom, and Fisher, 2013), analysing resultant impulse responses for distribution of energy and through visualisation routines implemented to view distributions of responses and their relation to the input data. Once we have processed the frame or window, sub-band histograms taken of the impulse responses at each rotation are used as a feature vector of the target training frame.

Finding a library with an existing implementation of the kernel proved an interesting challenge. OpenCV does support Gabor filters, although only as part of a predefined image processing method. Below is the code developed, modified from existing libraries, to produce the Gabor kernel we use in these experiments.

```python
def gabor_2d_kernel(xy=21, sigma=5, theta=0, lmbda=50, psi=90):
    if not xy % 2:
        # Kernel size must be odd to ensure central point
        exit(1)

    theta = theta * np.pi / 180
    psi = psi * np.pi / 180

    xs = np.linspace(-1., 1., xy)
    ys = np.linspace(-1., 1., xy)

    lmbda = np.float(lmbda)
    x, y = np.meshgrid(xs, ys)
    sigma = np.float(sigma) / xy

    x_theta = x * np.cos(theta) + y * np.sin(theta)
    y_theta = x * np.sin(theta) + y * np.cos(theta)

    term1 = np.exp(-0.5 * (x_theta ** 2 + y_theta ** 2) / sigma ** 2)
    term2 = np.cos(2. * np.pi * x_theta / lmbda + psi)
```
Figure 4.5: Impulse responses based on input frame, the four rotation values of $\theta$ used, and the resultant output.
In signal processing, a fundamental principle is that the signal in question, even if discrete, is cyclical in nature. This is an important point when considering the edges of an image, where the central point of the kernel with which it will be convoluted, occurs where one or more sides of the kernel window are outside of the pixel-area. There are many approaches in dealing with this issue, such as reflecting pixels or wrapping around the axes. The solution used for this implementation is one that also provides a significant performance increase, computationally. The kernel is not treated as a sliding window across the frame or patch. Instead, the area in question is transformed into the Fourier domain via Fast-Fourier-Transform function $\text{FFT}(x)$, where $i, j$ are the row and column of the original image $x$, with $k, l$ being their frequency domain analogue in the final output matrix.

$$\text{FFT}(x)_{kl} = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} x_{ij} e^{i2\pi(\frac{ki}{N} + \frac{lj}{N})} \quad (4.2)$$

Transforming an image from the spatial domain to the Fourier frequency domain deconstructs the image into the approximation of fundamental signals; sine and cosine variants (Bradski, 2000; Lee, 2008; Meyer, 1993).

Within the frequency domain, these frequencies are multiplied by the kernel produced, where $\otimes$ represents the equivalent convolution, to obtain the convolved image $I'_{\text{FFT}}$ in the Fourier domain. The inverse Fourier transform $\text{IFFT}(x)$ is applied, producing the required impulse response $I''$ in the spatial domain. The output of this process is visualised in Figure 4.5.

$$I_{\text{FFT}} = \text{FFT}(I) \quad I'_{\text{FFT}} = I_{\text{FFT}} \otimes K \quad I'' = \text{IFFT}(I'_{\text{FFT}}) \quad (4.3)$$

We apply the inverse Fourier transform to take $I'_{\text{FFT}}$ into the spatial domain, and the resultant matrix $I''$ is equivalent to the aforementioned sliding-window approach.

**Gabor Wavelets with LBP**

Visualised in Figure 4.6, a neighbourhood $N$ of $p$ pixels around radius $r$ from the pixel of interest $N_0$ is compiled. This sequence is followed, either clockwise or anticlockwise, and a bar-code is generated as defined, where $I_i$ is determined based on a comparison against
the impulse response value of the central pixel around which the LBP is calculated (Marcos et al., 2008).

\[
I_i = \begin{cases} 
1 & \text{if } N_i \geq N_0 \\
0 & \text{if } N_i < N_0
\end{cases}
\]  

(4.4)

A histogram \(H\) is then taken of \(I\) to determine the distribution of the local intensities. This is the final LBP value.

Using the four Gabor impulse responses together, LBPs (Marcos et al., 2008; Soriano et al., 2001) are used to model local texture information on the final texture representation. Initially, three sets of values for radius \(r\) and points \(p\) were evaluated: \((r = 1, p = 8)\), \((r = 2, p = 12)\) & \((r = 3, p = 16)\). In the case of full-frame analysis, initial experimentation showed that no significant difference was present in the varying LBP parameters, as such only \((r = 1, p = 8)\) was evaluated.

A key limitation of the implementation used results in poor time-performance when using this metric. Given the chosen language, Python, and the library used for LBP functionality (mahotas), there is no capacity to pre-compute pixel mappings for LBP bar-code generation. Due to this, interpolation of pixel selection based on chosen values of \(n\) and \(p\) is performed for every pixel, of every frame or window, of every video.

This was accepted as a short-coming in this implementation, as a performance trade-off in the preliminary nature of this research. Were it the case that a core objective was to achieve near or real-time performance, it would be preferable to pre-compute and store these maps to enable fast lookups based on pixel values in their RAM locations.

**Figure 4.6:** This example contains the same \(9 \times 9\) pixel window, with the current pixel \(P = (4,4)\) marked by the blue circle. Showing the different parameters \((r = 1, p = 8)\), \((r = 2, p = 12)\) & \((r = 3, p = 16)\), where the points in the circle at radius \(r\) fall, the selected pixel is shaded pink. This series of pink shaded pixels represents the neighbourhood for analysis. Image based on a diagram on Wikipedia released under a GNU 1.2 license.
4.4 Ground Truth Acquisition

To obtain ground truth data in a usable format, a specific module with a GUI was developed, to be used by marine biologists. The system permits the part and full annotation of any given video, by any schema that it was a record of. In this case, classification is performed only with the CCW schema. The expert then can annotate substrates over time, noting the transitions between them as they occur. This GUI, and its inputs, are shown in Figure 4.7.

![Ground truth interface.](image)

**Figure 4.7**: Ground truth interface.

There are many competing ontologies and specifications for the storing and distribution of ground truth data (Di Salvo, Giordano, and Kavasidis, 2013; Spampinato, Boom, and He, 2012; Kavasidis, Spampinato, and Giordano, 2013; He, Ossenbruggen, and Vries, 2013a), the chosen method for this system is the JSON specification. By mapping frame ranges to specific classes, we create a simple text file that can be parsed and used in the majority of modern languages and platforms.

Once saved and marked as complete, the JSON can then be loaded into the experiments for training and testing, it is parsed such that the internal representation of a given video can be visualised as Figure 4.8.

This method of direct expert opinion removes ambiguity from the translation of notes by those not qualified, and offers per-frame accuracy to the annotator’s specific standards.
4.5 Frame Selection

The frame selection process is performed at random. Once the ground truth is loaded into the system, all ranges of the same class are combined into a single list. From which, 10% of the frames of each class are selected, at random, from this list and passed through the chosen metric and classifier.

Each experiment keeps a record of its training frames, this is so that the future experiments in the n-fold cross validation will not use the same frames in their training steps. In the future experiments, the list of available frames simply removes those already selected from the pool of available frames before the randomised 10% are chosen again.

This methodology is the same for both full-frame and patch-based analyses. There is an acknowledged bias in this frame selection process, Section 6.2.2 discusses this and potential solutions.

4.6 Machine Learning

Our models for capturing the features of the feature space are two distinct methods of machine learning algorithms used within this, and other fields. This discipline concerns itself with using mathematical definitions for attempting to differentiate between discrete sets of data classes, to approximate the uniqueness in order to provide automatic classification of previously-unseen data.

There exist three forms of machine learning classification systems: Supervised, unsupervised and weakly supervised. As we have labelled data, we approach the problem using supervised classification systems. Using the expert-provided ground truth, we have an accurate set of labels to apply to each frame in each target video.
4.6.1 Support Vector Machines

Support Vector Machines (SVM) are binary classifiers, but can be used to classify multiple classes; in order to achieve this, a cluster of SVMs are trained to form a SVC (Support Vector Classifier), using the one-versus-many approach (Chang and Lin, 2011). This results in a SVM-based multi-class classification pipeline.

SVMs work by separating classes using a hyperplane, whose distance (margin) between class boundaries is maximised by the selected support vectors. Figure 4.9 illustrates this. The black and white points represent two distinct classes. The hyperplane, which has equidistant margins which cross through one or more points belonging to each class, at their boundaries. In Figure 4.9, the data points are linearly separable, and the noted equations and definitions denote the basis of this.

The use of SVMs and SVC is prevalent in fish species identification (Shortis et al., 2013; Spampinato et al., 2008; Beyan and Fisher, 2013; Amer et al., 2011; Huang, Boom, and Fisher, 2012; Boom et al., 2012a; Rova, Mori, and Dill, 2007; Shiau et al., 2012b; Shiau, Chen, and Lin, 2013), seabed point recognition (Singh et al., 2004; Gonzalez-Mirelis et al., 2009) and in various other computer vision domains including the categorisation of proteins, and handwriting recognition. The use of a SVMs with non-linear kernels tend to result in good recognition of non-linear classification (Hsu, Chang, and Lin, 2003; Chang and Lin, 2011).
This demonstrated capability comes at the cost of performance, where the training stage with its error handing (incorrect classifications & penalty term) can occupy large amounts of memory, and where the trained classifiers can have a (relatively) poor query time.

Kernels

![Figure 4.10: These figures demonstrate how two distinct groups of data points are separable by both linear and non-linear SVM kernels. In the case of non-linear separation, a kernel-trick (Chang and Lin, 2011) is applied to project high-dimensions mappings into the 2-dimensional domain, in which the SVM operates. This is the reason that the margin width separating classes fluctuates, as the support vectors exist in higher-order polynomial space. For the linear separation, the support vectors lay within the constraints of the plane illustrated, and are not a projection. Image adapted from a public domain image obtained from a Wikipedia entry on SVMs.]

Two SVM kernels were chosen; linear (SVCLIN) and radial basis function (SVCRBF). These are visualised in Figure 4.10. SVCLIN is typically the de-facto kernel used, to linearly separate the data classes in each SVM. SVCRBF, however, functions in higher-dimensional space, projecting the contour of separation into the plane in which the two classes exist (Hsu, Chang, and Lin, 2003; Beyan and Fisher, 2013; Chang and Lin, 2011; Cortes and Vapnik, 1995). These approaches are both evaluated to ascertain the differences that dimensions considered makes.

This approach correlates with the literature in the classification of Fish species (Spampinato et al., 2010; Huang, Boom, and Fisher, 2012; Rova, Mori, and Dill, 2007; Huang, Boom, and Fisher, 2013), motion & trajectories (Beyan and Fisher, 2013; Amer et al., 2011; Ross et al., 2008) and the automated analysis of benthic environments (Shihavuddin et al., 2013).
Parameter Tuning

Obtaining satisfactory performance from the chosen classifiers requires testing and configuration of the input parameters that will dictate their tolerances, and the eventual hyperplanes separating the classes. SVM parameter $C$ is the penalty of the error term, representing the compromise of training error versus generality. Higher values of $C$ can result in overfitting, forcing much tighter fitting of the plane to the datasets. Parameter $\gamma$ is used in SVCRBF only, and is the RBF kernel coefficient (Hsu, Chang, and Lin, 2003; Chang and Lin, 2011; Cortes and Vapnik, 1995).

Values $C = 1.0, \gamma = 0.0$ were selected. These were obtained through using short subsets of data from chosen videos to test their performance in a smaller test. This was necessary as the processing time for whole videos was simply too long to be practical in the time period.

Implementation

The implementation of the SVC systems used in SciKit-Learn are based around libsvm, which underpins many SVM / SVC interfaces in different libraries (Chang and Lin, 2011). It provides native, optimised machine-code and, thus, mitigates much of the computational overhead incurred in using a dynamic language such as Python.

This also ensures consistency of the data, and numerical stability of the required algorithms across related peer-work, systems and publications.

4.6.2 k-Nearest-Neighbour

We also use k-Nearest-Neighbour (kNN) classifiers using the ball tree algorithm with $k = 5$, where $k$ refers to the number of neighbours to the data-point that are considered in the search algorithm. Higher values of $k$ have been shown to reduce noise in classification performance, but also make the boundaries between distinct classes less clear (Bhatia and Ashev, 2010).

The algorithm works through the generation of boundaries which encapsulate the labeled data relating to a given class. When classifying new data, it will consider those neighbours in the region for classification. Where there exist numerous classes present that satisfy this distance condition, a vote is cast whereby the majority class in the target area is deemed the correct classification.

A value of 5 was determined for $k$ to be a compromise through the numerous CCW classes present for the data, whilst attempting to maintain generality due to the potentially unknown numbers of classes present for testing in unseen video.
Chapter 4. **Methodology**

### Implementation

SciKit-Learn uses a number of different algorithms for traversing the internal representation of kNN objects. In these experiments, the ball tree algorithm (Bhatia and Ashev, 2010) is used for search. We found that this performed queries the quickest in smaller-scale testing.

Internally, the kNN neighbourhood is represented as a binary tree in native machine code (not Python), and therefore has acceptable performance.

#### 4.6.3 Validation

A testing & training strategy of 10:90 n-fold cross validation was used, where 10% of a given class’ frames from within a video were selected at random and used as its training data. Section 4.5 discusses and defines this.

### 4.7 Closing Comments

This chapter has detailed the methodology used, in both the practical aspects of the software-engineering goals of the project, through to the scientific methods employed to answer the research question. Due to the high number of possible combinations of metrics, values and videos, it has been established that distribution of tasks that scale both horizontally and vertically is required.

Figure 4.11 illustrates the work-flow as a complete system, presuming that each component is part of the synchronised lists that make these experiments possible. It shows the datasets, experiment configuration, and the video subdivision into the following categories: *training*, *testing* and *other*. *Other* is for any frames of class 0 which have been recorded as not containing relevant information.

As noted previously, it is accepted that many of these approaches are simplistic in their execution. This is by design, due to the exploratory nature of the research in the specific problem domain we are addressing.
Figure 4.11: Training & testing methodology for both datasets
Chapter 5

Results

This chapter presents the gathered data for full frame, and patch-based analysis of the Bangor and FoCB datasets. Discussions are made based on the contrasts between them, and between the classification schemas used within them.

Further discussion revolves around explanations for the trends noticed in the data, including data constraints and experiment parameter choices.
TABLE 5.1: Experiment Results: all values are percent (%) of testing frames correctly classified as per the Countryside Council for Wales schema (where B. refers to the Bangor dataset.)

<table>
<thead>
<tr>
<th>FEATURE &amp; CLASSIFIER</th>
<th>VIDEO</th>
<th>B. SITE 2</th>
<th>B. SITE 3</th>
<th>B. SITE 10</th>
<th>FoCB 1</th>
<th>FoCB 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Histogram (Greyscale)</td>
<td></td>
<td>52 59</td>
<td>65 80</td>
<td>30 78</td>
<td>54 71</td>
<td>30 71</td>
</tr>
<tr>
<td>Histogram (Colour)</td>
<td></td>
<td>66 66</td>
<td>66 88</td>
<td>30 88</td>
<td>54 79</td>
<td>30 72</td>
</tr>
<tr>
<td>Retinex Histo (Colour)</td>
<td></td>
<td>65 78</td>
<td>66 78</td>
<td>30 88</td>
<td>54 79</td>
<td>30 78</td>
</tr>
<tr>
<td>Retinex Histo (Greyscale)</td>
<td></td>
<td>52 61</td>
<td>52 66</td>
<td>30 86</td>
<td>54 79</td>
<td>30 66</td>
</tr>
<tr>
<td>Gabor Sub-band Histogram</td>
<td></td>
<td>52 71</td>
<td>52 66</td>
<td>30 55</td>
<td>54 60</td>
<td>30 60</td>
</tr>
<tr>
<td>Gabor LBP ($r = 1, p = 8$)</td>
<td></td>
<td>52 71</td>
<td>52 66</td>
<td>30 55</td>
<td>54 60</td>
<td>30 60</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>MIN</th>
<th>MEDIAN</th>
<th>MAX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Histogram (Greyscale)</td>
<td>52 59</td>
<td>65 80</td>
<td>66 88</td>
</tr>
<tr>
<td>Histogram (Colour)</td>
<td>66 66</td>
<td>66 88</td>
<td>66 88</td>
</tr>
<tr>
<td>Retinex Histo (Colour)</td>
<td>65 78</td>
<td>66 78</td>
<td>66 88</td>
</tr>
<tr>
<td>Retinex Histo (Greyscale)</td>
<td>52 61</td>
<td>52 66</td>
<td>52 66</td>
</tr>
<tr>
<td>Gabor Sub-band Histogram</td>
<td>52 71</td>
<td>52 66</td>
<td>52 66</td>
</tr>
<tr>
<td>Gabor LBP ($r = 1, p = 8$)</td>
<td>52 71</td>
<td>52 66</td>
<td>52 66</td>
</tr>
</tbody>
</table>

5.1 Full Frame

All preliminary experiment results are noted in Tables 5.1 & 5.2, grouped by video. Reviewing these, kNN classifiers perform with greater accuracy than SVCs. The difference between a kNN approach and SVCLIN is as high as 35% increase in successful classification. Using the same features and the same video, SVCLIN outperforms SVCRBF by up to $\approx 81\%$ success rate (FoCB1 using RGB Histograms, SVCLIN = 90%, SVCRBF = 53%) using the simplified schema (Table 5.2). The same metric and video in the CCW schema (Table 5.1) yields an increase in accuracy for SVCLIN over SCVRBF of $\approx 293\%$ (SVCLIN = 88%, SVCRBF = 30%). These findings demonstrate that the problem both responds well to a clearly linear and isolated classification system, which remains constant irrespective of the feature being trained, and that the visually similar (yet distinct) classes present in the CCW schema cause far more confusion in SVCRBF than in SVCLIN.

This difference in results between the CCW and simplified schemas present interesting observations. It stands to reason that the simplified schema would perform better than, if not equal to, CCW in every case. In one example, SITE10, the variance between the two schemas is not significant. This is not due to the granularity of the CCW schema being sufficiently varied, but due to the lack of variety in substratum observed within the video.

5.2 Patch-Based

The patch-based approach was developed based on the observations gathered from the full-frame experiments. A number of assumptions were made as to what the differences between
Table 5.2: Experiment Results: all values are percent (%) of testing frames correctly classified as per our classification schema (where B. refers to the Bangor dataset). Many results using this schema are high, due to the low variance of the data in this schema.

<table>
<thead>
<tr>
<th>Feature &amp; Classifier</th>
<th>B. Site 2</th>
<th>B. Site 3</th>
<th>B. Site 10</th>
<th>FoCB 1</th>
<th>FoCB 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SVCRBF</td>
<td>SVCLIN</td>
<td>SVCRBF</td>
<td>SVCLIN</td>
<td>KNN</td>
</tr>
<tr>
<td>Histogram (Greyscale)</td>
<td>96 93 98</td>
<td>100 100 97</td>
<td>65 81 86</td>
<td>53 93 97</td>
<td>54 71 83</td>
</tr>
<tr>
<td>Histogram (Colour)</td>
<td>96 93 98</td>
<td>100 99 100</td>
<td>66 86 88</td>
<td>53 96 98</td>
<td>54 79 86</td>
</tr>
<tr>
<td>Retinex Histo (Colour)</td>
<td>96 96 97</td>
<td>100 99 100</td>
<td>66 87 83</td>
<td>53 96 99</td>
<td>54 81 89</td>
</tr>
<tr>
<td>Retinex Histo (Greyscale)</td>
<td>96 96 97</td>
<td>100 100 100</td>
<td>65 78 85</td>
<td>57 92 96</td>
<td>54 46 79</td>
</tr>
<tr>
<td>Gabor Sub-band Histogram</td>
<td>96 96 96</td>
<td>100 100 100</td>
<td>65 66 79</td>
<td>53 85 99</td>
<td>54 68 91</td>
</tr>
<tr>
<td>Gabor LBP ($r = 1, p = 8$)</td>
<td>96 96 95</td>
<td>100 100 99</td>
<td>65 65 74</td>
<td>53 72 99</td>
<td>54 61 69</td>
</tr>
</tbody>
</table>

- **Min**: 96 93 95 100 99 97 65 65 74 53 72 96 54 46 69
- **Median**: 96 96 97 100 100 100 65 80 84 53 93 99 54 70 85
- **Max**: 96 96 98 100 100 100 66 87 88 57 96 99 54 81 91

the two approaches would be. It was expected that the CCW Bangor results should improve, and that the FoCB results should degrade. It was also noted that comparing the simplified schema results would not be beneficial, as on the full-frame approach these were already approaching 100%, even if overfit on the training data, the frame selection process between the two approaches were the same and, as such, exhibit the same aforementioned bias.

Another point to note is that in these results, multiple parameters are tried for the LBP approaches, whereas this is not the case for the full-frame results. These parameters were included in the initial batch of experiments for the full-frame approach, but were found to have no significant different over the values retained in the presented full-frame results. This was concluded to be due to the size of the frame meaning no further information was being captured & encoded using more points and a higher radius, but on a smaller image (as with the patches), any benefit would be more likely to manifest. Further to this, a number of methods (specifically those encoding colour information) were also discarded, due to the observation that the inclusion of colour information did not significantly improve the accuracies recorded.

The final judgements on which methods and parameters to include in the patch-based experiments were ones based on expectations derived from the full-frame results set versus computation power available.

The reason for these assumptions were that the Bangor dataset was uniform in that, where not class 0, the entire frame was of the sea-floor albeit at an angle. Figure 5.1 shows that in the FoCB dataset, this was not the case.
Chapter 5. Experiment Results & Discussion

Figure 5.1: These three scenes, taken from the FoCB dataset, show the issues with extremely high turbidity and sea-floor to sea-water proportions of frames.

Table 5.3: Patch-based results using kNN (– denotes no data as experiments did not complete before project end)

<table>
<thead>
<tr>
<th>SIMPLIFIED</th>
<th>SITE 2</th>
<th>SITE 3</th>
<th>SITE 10</th>
<th>FoCB 1</th>
<th>FoCB 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Histogram (Greyscale)</td>
<td>–</td>
<td>–</td>
<td>81.00</td>
<td>25.00</td>
<td>68.00</td>
</tr>
<tr>
<td>Gabor Histogram</td>
<td>95.65</td>
<td>99.52</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Gabor LBP ($r = 1, p = 8$)</td>
<td>95.62</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Gabor LBP ($r = 2, p = 12$)</td>
<td>100</td>
<td>100</td>
<td>–</td>
<td>47.19</td>
<td>14.69</td>
</tr>
<tr>
<td>Gabor LBP ($r = 3, p = 16$)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>52.81</td>
<td>54.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CCW</th>
<th>SITE 2</th>
<th>SITE 3</th>
<th>SITE 10</th>
<th>FoCB 1</th>
<th>FoCB 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Histogram (Greyscale)</td>
<td>–</td>
<td>–</td>
<td>81.00</td>
<td>5.00</td>
<td>68</td>
</tr>
<tr>
<td>Gabor Histogram</td>
<td>45.65</td>
<td>23.26</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Gabor LBP ($r = 1, p = 8$)</td>
<td>53.22</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Gabor LBP ($r = 2, p = 12$)</td>
<td>94.00</td>
<td>79.00</td>
<td>–</td>
<td>25.11</td>
<td>14.69</td>
</tr>
<tr>
<td>Gabor LBP ($r = 3, p = 16$)</td>
<td>51.67</td>
<td>–</td>
<td>–</td>
<td>23.28</td>
<td>54.02</td>
</tr>
</tbody>
</table>

Given the viability metric for selecting patches of frames, it was assumed that the upper section of each frame that consists of turbid water would be classified as patches of the same class as the visible sea-floor.

Some attempts were made to mitigate this by approaching the issue with rejection if the hue was ultimately blue, but a number of problems with this approach arose. Firstly, it is a highly specialised solution and not the generalised approach required for this project. Secondly, the computational overhead simply could not be afforded given the time and technological constraints present.

The preliminary patch-based results do indeed confirm these suspicions (as the results in Table 5.3 show), with greyscale histograms and kNN performing badly in both schemas. FoCB2, however, demonstrates much better performance due to the frames consisting almost entirely of dense reef areas at a lower altitude than FoCB1. Table 5.4 shows comparable performance between the SVM-based classifiers and those using kNN in Table 5.3, where equivalent metrics have been tested.
Table 5.4: Patch-based results using SVCLIN (– denotes no data as experiments did not complete before project end)

<table>
<thead>
<tr>
<th></th>
<th>Simplified</th>
<th>Site 2</th>
<th>Site 3</th>
<th>Site 10</th>
<th>FoCB 1</th>
<th>FoCB 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Histogram (Greyscale)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Gabor Histogram</td>
<td>–</td>
<td>99.52</td>
<td>66.23</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Gabor LBP (r = 1, p = 8)</td>
<td>95.65</td>
<td>99.52</td>
<td>65.40</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Gabor LBP (r = 2, p = 12)</td>
<td>–</td>
<td>99.52</td>
<td>–</td>
<td>53.00</td>
<td>18.42</td>
<td>–</td>
</tr>
<tr>
<td>Gabor LBP (r = 3, p = 16)</td>
<td>–</td>
<td>99.52</td>
<td>–</td>
<td>56.98</td>
<td>31.46</td>
<td>–</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>CCW</th>
<th>Site 2</th>
<th>Site 3</th>
<th>Site 10</th>
<th>FoCB 1</th>
<th>FoCB 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Histogram (Greyscale)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Gabor Histogram</td>
<td>–</td>
<td>–</td>
<td>29.89</td>
<td>66.21</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Gabor LBP (r = 1, p = 8)</td>
<td>43.97</td>
<td>33.26</td>
<td>65.40</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Gabor LBP (r = 2, p = 12)</td>
<td>51.12</td>
<td>33.32</td>
<td>–</td>
<td>29.30</td>
<td>17.41</td>
<td>–</td>
</tr>
<tr>
<td>Gabor LBP (r = 3, p = 16)</td>
<td>43.98</td>
<td>33.32</td>
<td>–</td>
<td>31.95</td>
<td>8.05</td>
<td>–</td>
</tr>
</tbody>
</table>

5.3 Discussion

The use of Retinex as a preprocessing step before using greyscale and colour frame histograms as a feature vector does not show a clear advantage over omitting this preprocessing step from the data present. The observable difference may be attributable to increased contrast in certain features, as a form of preliminary histogram equalisation. This finding corroborates other work stating that colour information correction in such scenes is not a meaningful step in this feature space (Singh, Howland, and Pizarro, 2004; Eustice, Singh, and Howland, 2000; Pizarro and Singh, 2003).

Features which are dependent on texture result in successful classification on par, or below, the statistical frame histogram metrics. LBPs do not follow a clear correlation of effectiveness based on the radius and number of points used, and sub-band histograms perform better than LBP where comparable data is present.

A consequence of the resultant data on full-frame experimentation led to the reduction of using varying parameters of the LBP algorithm. The variance noted was high which contradicted expectation, but may be accountable to the illumination texture present in all substrates, rather than the seabed texture itself.

5.3.1 Comparing Datasets & Schemas, with Chance & Recognition

By comparing the CCW and Simplified schemas, with the classes observed in each video as previously noted, it is clear that, in many cases, the metrics are indeed unable to differentiate CCW classes. Amongst the Bangor dataset specifically, there is a very clear consistency to the accuracy using both schemas. This is likely due to the fact that the cross-classification crosstalk of similar CCW classes (in the example of SITE3, Simplified II has 4 separate CCW
classes present) results in a poor-performance metric for CCW ($\approx 50\%$) versus an excellent result for Simplified schema ($\approx 100\%$). This problem is best visualised in Figure 5.2, which shows the difficulty in ascertaining unique defining features of the classes, but why the Simplified schema is so highly successful.

![Image]

**Figure 5.2:** These patches are taken from CCW classes 16, 18 & 24. With four columns per class, in their respective order. The first row is the raw patch, second is the histogram equalised variant, then the visualisation of the results of using the 4 Gabor filters.

In these results, the Bangor dataset responds to the techniques used more favourably than the FoCB dataset. This fact is more apparent in the simplified schema than the CCW. As noted in Chapter 3, a number of constraints are present in both datasets, although the videos belonging to a particular dataset which may have common issues, may not share them with the other dataset.

Reviewing the Bangor footage (and visible in Figure 2.4(a)) it becomes clear that the illumination pattern cast on the seabed is providing a constant filter response which is influencing the feature vector uniqueness per frame. In conjunction with this, care was given to select videos from each dataset that contained sufficient variety in the captured seabed substrates to obtain quantitative, relevant data. Within the chosen videos of the Bangor dataset, the visual difference between those classes recorded were not sufficiently different to generate the desired results.

The simplified schema was created to specifically address the difficulties inherent to underwater video processing, however the results have shown that where the captured video lacks sufficient class variety, this results in a gross over-simplification and irrelevant results, where the distinctions recorded amongst a given video may simply be attributable to chance based on the lack of visual distinction of the classes present. This is true for the FoCB set to a lesser degree than that of Bangor’s, although with results frequently greater than even 90%, the validity of these must also be questioned.

Using the original CCW schema, the results are more congruent to those originally expected. The variance of the results, per video and per classifier, appears to be less due to the metrics used but more closely related to the classifier configuration used. This observation is
FIGURE 5.3: A typical scene from the Bangor dataset, with perspective issues superimposed.

discussed more thoroughly in Section 5.3.3.

The trends noted suggest that, even using a full-frame as the basis of a feature vector, there is enough visible difference in the textures of distinct CCW classes to achieve satisfactory results, and to serve as the basis for further refinement. This fact calls into question the necessity of formulating simpler schema, at least for the data analysed.

5.3.2 Perspective Issues

Figure 5.3 illustrates the effect of perspective due to the camera’s angle of view. The green grid represents an approximation of the surface plane, with two example patches designated by red squares. The issue arises that objects of similar size and texture will appear differently in the two patches, due to this transformation.

In the patch-based work, this will present a key issue that the scale of the features being modelled will differ, thus the extracted features will be susceptible to blurring the boundaries within the distinct classes.

Due to the problems of the results stability, as discussed in Section 5.3.3, it is unknown to what degree this effect affects the results.
5.3.3 Overfitting

Overfitting is something that can happen with machine learning algorithms when the parameters have been incorrectly set. In this context, incorrectly means that the model that has been generated in the training stage of the classifier(s) has been tuned to match the training examples too strongly. Generality is a key point in machine learning, this means that the solution should deal with unseen data in the testing stage, by segmenting the feature space such that the class boundaries will encompass the majority (this does not necessarily mean all) of the observations. By having an overfit separation boundary, misclassifications are generally more common, correct classifications tend to only occur if the new observed data lie to very similar points as those in the training data.

Figure 5.4 illustrates the problem. These four synthetic plots consider the training points illustrated by red and blue markers for a two class dataset in the classification examples. There are two green testing points, $x_1$ & $x_2$. $f_1(x)$ is an overfit model to the training data. The line it produces does segment the feature space, going through all training points. Visually, it is clear that the contours of the line may not be what is desirable. $f_2(x)$ also satisfies the condition of going through every training point. In classification terms, which side of the line values $x_1$ & $x_2$ fall, depending on the function, show different classifications.

In the case of overfitting, only samples which are near identical to those which were used in the training phase will be correctly classified, as they will fall in close proximity to those which were used to fit the original model. As such, overfitting represents a real problem in building a generalised classifier.

Owing to the lack of variety in the datasets chosen for these experiments, it becomes clear that the limited variation in underwater substrates and even the limited variation inside each video of different visual qualities of the substrate within, has resulted in overfitting in many of the example cases.

Due to this limited variety in the Bangor dataset, there is evidence of classifier overfitting due to the lack of generality in the source material. In the FoCB dataset, whose content is more varied, this issue is not as prominent, though still easily observed across all experimental combinations. It is clear that, for the purposes of these experiments, the FoCB results in simplified classification classes represent a more real-world result expectation than Bangor's.

5.3.4 Incorrect SVM Kernel Tuning

Support vector machines benefit greatly by having been trained using sparse examples in the feature space, and having been subjected to prior tuning of any and all parameters which
affect the specific combination of kernel and implementation used. To elaborate, the SVM looks to maximise the margin between the hyperplane and the support vectors of the two classes, with sparse data one can imagine that the likelihood of maximising these margins is increased, assuming there is not a large density of data points for both classes near the hyperplane location.

In these results, it is clear that SVCRBF has consistently performed poorly, with the highest
recorded accuracy being \( \approx 54\% \). The graphs in Figures 5.5, 5.6, 5.7 & 5.8 visualise this difference, where, with the exception of GABORLBP in Figure 5.5, it performs far more poorly consistently. In contrast to Tables 5.1 & 5.2, this view of the data emphasises just how large the difference in recorded accuracy against the many combinations of values tested is. The reason that this huge accuracy difference is occurring, irrespective of the metric used to obtain the feature vector encoding, is that incorrect parameter tuning has been performed. It is reasonable to assume that the potential accuracy of this classifier is higher than that observed in these experiments, but that during the training stage incorrect parameters were discovered which define the RBF kernel function that acts as the hyperplane in the SVMs. The combined results, in Figures 5.9, & 5.10 illustrate that this phenomena is consistent amongst all configurations of the experiments performed.

This poor selection of parameters manifests itself in this way, such is the effect that when the original feature space where the training points exist was not optimally-separated, due to the parameter \( C \) and \( \gamma \) selection. This results in a hyperplane that is not suitable to sufficiently separate the classes, and may cross over unique (but potentially overlapping) class boundaries (Hsu, Chang, and Lin, 2003; Chang and Lin, 2011; Cortes and Vapnik, 1995; Pedregosa et al., 2011).

The fact that this has been observed suggests that re-tuning of parameters, for example using a grid search such as the one in SciKit-Learn, and subsequent re-running of the experiments would provide a better view of the true classifier performance in this domain. Unfortunately, due to the wide number of configurations and experiments that needs to be performed it has not been possible to redo this entire set of experiments during the duration of this project. Across the experiments combinations, SVC-RBF performed the worst consistently. This contradicts the results of others (Campos et al., 2014; Spampinato et al., 2010; Beyan and Fisher, 2013), where the use of RBF kernels in SVCs have performed well. An explanation for this is the phenomena of incorrect kernel configuration (Hsu, Chang, and Lin, 2003), whereby the \( C \) and \( \gamma \) parameter tuning of the SVC was not optimised for the feature space correctly. As such, the resultant classifier will either be too generalised, or suffer from overfitting. In this case, this incorrect configuration contributed to the overfitting problem.
5.4 Plots

**Figure 5.5**: Bangor full-frame results using the simplified schema.

**Figure 5.6**: Bangor full-frame results using the CCW schema.
Figure 5.7: FoCB full-frame results using the simplified schema.

Figure 5.8: FoCB full-frame results using the CCW schema.
Chapter 5. Experiment Results & Discussion

**Figure 5.9:** Average results for full-frame analysis across both the Bangor and FoCB datasets using the simplified classification schema.

**Figure 5.10:** Average results for full-frame analysis across both the Bangor and FoCB datasets using the CCW classification schema.
Chapter 6

Conclusion

This chapter discussed the observations, issues and outputs of the project. By comparing the outcomes with the intended outputs, retrospective analysis is applied to discover areas where improvement could be applied, both in method and management.

A closing discussion on the future work that this project may inspire is given.
We have evaluated several machine learning and computer vision techniques as preliminary steps in understanding underwater environments. The complexity of our simplified classification schema is sufficient for the identification of seabed substrate types, however appears to be susceptible to significant overfitting.

It was believed that the CCW schema requires further additional non-visual information. The results, however, noted that there were enough differences to have a meaningful effect on the statistics recovered, but also that there was, as anticipated, significant cross-talk between similar classes as per Simplified schema. Approximating this extra information could potentially be performed by identifying regions for further analysis, and observing specific fauna that is present within the frame / temporal region that a substrate occurs in. More empirical methods could be with the use of extra sensory equipment, such a spectroscope for evaluating composition based on known parameters of different sea-floor substratum.

Fundamentally, we have shown that it is possible to use existing computer vision and machine learning methods to achieve suitable classifiers on underwater video, albeit with a number of caveats. The results gathered indicate a higher than anticipated rate of success using full-frame classification methods and our simplified schema, in many instances, 100% testing success is noted. A key reason for this observation is that the videos selected from each dataset were those deemed to be most complex by a marine biologist, with complex being defined as most changes in the observed sea-bed substrates. However, even in these chosen videos the range of visually-distinct substrates as per our schema is low. This is due to the short length of the videos themselves and their coverage not being in areas pre-selected for likelihood of substrate complexity. The results validate the use of these methods to build a more generic underwater substrate classification system but do not yet give any indication as to what its accuracy could be.

A prime example of this is SITE10, which as the results show demonstrates little difference in using the CCW or Simplified schema. The reason of this is simply that not enough variation exists between the classes present in SITE10 to generate the confusion as noted in SITE2 & SITE3.

Performance wise, the system underwent significant optimisation throughout its use yet continued to be very slow, due to the high number of parameter, metric and source combinations, plus the cross-validation method.


6.1 Restrospective

At the project's conclusion, certain decisions taken in the design and execution of its components have resulted in a less-than-desirable outcome. This section highlights these shortcomings, and offers potential improvements in future work.

6.1.1 Technology

This discussion once again must refer to the software-engineering aspects of the project, as it was a key consideration in achieving the results presented, and so that the suggested platform may be extended with future research.

The largest bottleneck in running these experiments has been the time of execution. Section 4.2.1 discusses the problems on dealing with such large data, in so many configurations of experiments, parameters and metrics.

Whilst the project was ongoing, numerous optimisation efforts were attempted. These include the profiling of operations comparatively with optimised routines of abstraction removal, algorithm changing and re-structuring of key data-structures to permit more parallelisation. At the point at which the code-base was settled, it had reached the point where no further optimisation could realistically be performed using the chosen technology stack. Further optimisation would have come by moving key algorithm components to higher-performance platforms or languages.

There were efforts to do this, whilst maintaining the work-flow established. Experiments using NVidia's CUDA SDK were performed. These included the vectorisation of simple histogram equalisation methods. Although a significant performance increase was noted, the amount was not measured empirically. This experiment was abandoned due to the complex nature of GPU programming, in establishing the development & execution environments. In addition, only systems RMBP and Z600 had the hardware capable of performing the experiments. The decision was taken that the time-to-develop and re-implement the existing logic established in Python, was not feasible the remaining time.

What was clear was the performance improvement of using a low-level language such as C++. Although Python and C-based libraries have been shown to perform well, the number of execution calls per experiment proved the Achilles heel in reality. CUDA was chosen as an experimental framework due to simple research in the domain, looking at where specific parallel-processing cards, such as the Tesla card, were used. Had more research been performed in this area, settling on a technological stack such as C++11 and OpenCL would have
been a better option. The resultant code would have been highly portable, and completely parallelised.

Taking the distributed metaphor further, using the technologies of C++ & OpenCL, high performance clusters (HPC) could have been used as the target platform, distributing instances of the software across multiple nodes. Small experiments of this nature were conducted on HPC Wales’ platform, but were ultimately rejected due to time available, and cluster resources permitted. Using technologies such as MPI (Message Passing Interface), and substantial engineering & testing, the final system would have been as optimised as possible given current technology. Ultimately, this issue comes down to the balance of time-to-develop versus time-to-execute.

6.1.2 Data Collection Timeline

Acquisition of the OpenROV took a lot longer than had been anticipated, due to the nature of procurement within the institution and direct-from-manufacturer business model of the kit distribution. Authorisation and payment was not complete until almost 6 months into the project. Initially, progress on developing the OpenROV was good. However, due to the inexperience of those assisting with the build in electronics, initial testing was not successful.

Subsequent rebuilds of the OpenROV required delicate deconstruction, rectification of incorrect assembly and thorough testing. This process was highly time consuming. After extensive repair work, waiting for a weather window to test became the next delay. The originally planned window for exploration was missed due the cumulative delays noted. Further issues arose during the testing of the ROV, due directly and indirectly to the original errors in its construction. Such problems included, but were not limited to, the splitting of a battery.
enclosure whilst in-situ in Cardigan Bay. Some video was acquired, as shown in Figure 6.1, but it was either not of sufficient depth (altitude in the water column) nor length to perform analysis upon, or was not taken in Cardigan Bay and, instead, shows the bottom of the lake used during testing.

It would have been preferable to have approached the procurement of the kit in a different manner, to save needed time and gather more data with the ROV. If this were the case, a more direct approach to building it could have been established.

### 6.2 Future Work

The work that has been conducted is acknowledged as being preliminary, and the basis for future techniques to be developed upon. This section aims to discuss areas which present a clear opportunity for improvements in efficiency and performance.

#### 6.2.1 Contextual Classification of Seabed Fauna

Work was conducted with a summer student under my supervision into the automated detection of fish in the FoCB dataset. By analysing the scene for motion, segmenting based on outlines and deriving a metric for the smoothness of the detected proto-objects, we were able to successfully detect a sea-bass in one video. The methodology was extremely basic and served as a proof-of-concept for obtaining unsupervised fish detection using simple computer vision methods.

Once isolated, the targeted fish was placed in a bounding box and passed through to the integrated MOSSE tracker (Bolme et al., 2010), which managed to track it as well as if the box were defined manually using the mouse pointer.

This project, in mapping the sea-floor over the temporal domain, was planned to minimise the feature space of identifying fauna in underwater video. By combining automated substrate classification, we can build assumptions on the probabilities of certain species being present in a given substrate, and thus weight the probably classification of said automatically discovered fish to (theoretically) produce more accurate results.

Building upon this will be a substantial area of growth, and provides an analogue, whilst introducing new perspectives, to the work on systems such as MBARI's AVED (Section 2.6.3).
6.2.2 Frame Selection

The problem of overfitting can be visualised in Figure 6.2, whereby the training samples are closely bunched within the boundaries of any given class. More involved frame selection in each iteration should alleviate this problem, by ensuring near-identical frames are not used in the same training pass.

**Figure 6.2:** This example shows the distribution of training points in the feature space. The classes used are arbitrary, and do not relate to actual classification schemas. Classes 0 & 3 are expanded upon, in which we can see the boundaries of the class ranges denoted by the blue line. Within this range, the training points are represented by grey-filled rectangles. The green lines denote the actual learned-range, based on the sparseness of the training data in the field. The red lines denote the undesirable margin between where the training points lay, and the true class boundary.
6.2.2.1 Euclidean Distance

By comparing the Euclidean distance between two frames, \( f_1 \) & \( f_2 \), a resultant matrix of distance \( f_d \) should be acquired. A number of approaches are available with \( f_d \) to determine if it is of sufficiently high difference to permit the next frame as a training frame, including the mean of the difference, or spatial mapping. By using templates, the amount of difference over areas of the frame or window can be described as a shape. By having permitted shapes & areas of difference, this could be achieved.

6.2.2.2 Multi-resolution

Another approach to reducing the problem-space size, and computational overhead is the use of multi-resolution, pyramid representations of textures and features. This method of representation is abundant in generalised computer vision, and underwater video literature (Marin-Jimenez and De La Blanca, 2006; Spampinato and Palazzo, 2012a; Vitaladevuni and Domke, 2005; Lowe, 1999; Freeman and Adelson, 1990; Lee, 2008).

By generating multi-resolution pyramidal representations, it should be possible to begin approaching real-time performance. This depends on the separation of classes in the problem-space, but given sufficient confidence through testing, it is a key opportunity for development. This is a bit different to the standard pyramid approach in computer vision, such that the proposed structure could have a confidence value of correct classification based on smaller scaled texture samples, should some threshold be passed at one scale then a successful classification is recorded. If ambiguity remains, that is no classifier has surpassed this threshold, the next level of the pyramidal structure is used at a higher-scale (and thus more computationally intensive). This process would repeat until either the scales of the structure have been exhausted (in which case, a majority vote would win), or until at some scale, the threshold is passed by a classifier.

6.2.2.3 Subset

A relatively simple improvement on the method is to avoid classifying on a per-frame basis, but instead segment the video into sections in the temporal domain. By doing so, a statistically-relevant sampling of the frames contained within could be analysed to determine an overall ‘vote’ for the classification of the region.

This task could be further subdivided by having the subsequent frames decomposed into patches, with another statistically-viable section being used to vote per frame. This approach would mitigate much of the performance issues, and given an appropriate viability metric
on the patches, could drastically increase the accuracy versus the baseline presented in this work for patch-based analysis.

### 6.2.3 Cross-Testing, New Data & Meta-data

Testing of the experiments was limited to within individual videos, though the intent was to build a hierarchical classifier. The move towards a patch-based methodology was the first step to this process, abstracting the textural information from the confines of the whole-frame original analysis.

The goal of this project was to build a generalised classifier, capable of training on one or more videos, and be applied to unseen videos and produce acceptable results. Testing this given the current technological stack of Python, SciPy (Jones et al., 2001) and SciKit-Learn (Pedregosa et al., 2011) make this difficult. By building a number of patch-based classifiers, per video, a technique such as AdaBoost (Zhu et al., 2009) could be applied to amalgamate the machine learning models into a strong classifier for unseen data. It was still possible and principal experimentation proved this. Issues came in determining how to group classifiers in RAM, given the relatively large sizes the serialised classifier objects required.

How this would perform based on the perspective issues that manifest in scale and rotation variation would be a key point to overcome, but would result in the realisation of this project’s goals. These issues are discussed in Section 5.3.2 and illustrated in Figure 5.3. The numerous factors in orientation (pitch, yaw, roll) of the camera on the sea surface, these distortions vary greatly in their visual appearance in the captured video. Due to these reasons being present in the data, the use boosting algorithms was not possible. Some faux-boosted methods would have needed to have been implemented in a decision-tree-like structure, capable of selecting the based trained classifier for the current problem. This proposed faux-boosted model could perhaps take the form of taking the saved classifiers we have generated in this project, and consulting each for some new observed sample \( x \) (whether patch-based or full-frame), and taken a probabilistic value from each of the classifiers denoting their predicted class label \( y_i \) along with a confidence of classification \( P(Y = y|X = x, M_j) \) where \( M_j \) is one of the classifiers already trained and stored. Within this structure, a highest-probability-wins approach could begin to emulated boosted algorithms. This solution is not elegant, nor is it simple to create in a robust manner, but it is presented as a possible direction for future refinement.

Using the OpenROV correctly, with the Inertial Motion Unit (IMU), the collection of meta-data including depth, water temperature and GPS co-ordinates during ROV survey could enable research into more detailed mapping of the sub-aquatic environment. The use of GPS
technology underwater is difficult, but given the shallow nature of Cardigan Bay and an altitude of 3 – 4 m of the ROV above the sea-floor, the GPS hardware would typically be within 1 – 2 m of the sea-surface. It is therefore conceivable to imagine that some success with this technology could be managed, even if it were not fully robust (maintaining a constant and accurate location via GPS).

Sea-bed depth and temperatures are also key factors in identifying habitats for species of interest to marine biologists which, in conjunction with the work presented in this project, could lead to accurate, automated systems for surveying in the future. The presence of a compass and pressure sensor could also lead to the mapping of sea-bed substrate(s) in an automated manner. This would require significant work on the stability of the OpenROV in the water column, or sufficiently robust models to compensate for the inevitable drift in post-processing.

6.2.4 High-dimensional Data Reduction

Although preliminary in nature, many of the feature vectors used in these experiments were of extremely high dimensionality. Justified as being baseline work, a number of techniques were researched but not implemented due to the performance overhead in doing so due to the computational complexity already experienced, as described in Section 4.2.1.

PCA is used throughout the literature (Campos et al., 2014; Ross, Lim, and Yang, 2004; Spampinato et al., 2010; Kuncheva, 2004; Beyan and Fisher, 2013; Amer et al., 2011; Lim et al., 2004; Riegl, Korrubel, and Martin, 2001; Shihavuddin, Gracias, and García, 2012; Campos, García, and Nicosevici, 2011) for reducing the complexity of feature vectors by deriving its principal components, and the amount of variance thereof based on the data’s eigenvectors and eigenvalues, respectively (Jolliffe, 2005).

As a post-processing step, the input is the high-dimensional feature vector $\mathbf{x}$, with a lower-dimensional representation (typically the first $n$ principal components of $\mathbf{x}$, where $n$ is chosen for the specific problem, based on the amount of unique information encoded as more principal components are considered) being the output. Algorithmically, this would have added another computationally-expensive step on every iteration, given the large number of videos being sampled. Feature selection may have provided the required element in the work-flow, by automatically deducing relevant features (Guyon and Elisseeff, 2003) but was, again, expensive in execution. The problem was one of performance, not implementation, as both are part of the SciKit-Learn library (Pedregosa et al., 2011).
6.2.5 ROV automation

The OpenROV’s controls do not mandate specific hardware, and are performed using Web-Sockets within a modern web browser. As such, it is a realistic expectation that desktop-native software could be developed that, with a computer vision & machine learning backend, could control the OpenROV with semi-autonomy.

Use-cases for this could be where a fish has been located, by using the MOSSE tracker as per our basic implementation, the region could be tracked based on its relative position to the frame with corrective orders automatically issued to the ROV to follow it in all three dimensions of motion. This paves the way towards a hybridised AUV / ROV, albeit tethered at all times, and represents an exciting direction that the open source software and hardware could be advanced in.

6.3 Critical

This project achieved its goals, by proving that simple vision techniques are applicable to this domain. What was not proved, however, was the robustness of the techniques across varying datasets. The use of different image descriptors and evaluation of different texture extraction methods (and parameters thereof) could provide more successful classification with further work.

In particular, shape and approximation of motion against a static background could theoretically eliminate issues in existing video relating to trawler-impact debris via automated dismissal, instead of simply ignoring as class 0. The way in which class 0 was approached is justifiable given the constraints of the project, but is impractical for unseen video. Though it is unclear, at the end of the project, how classifying everything that is not a substrate could be realistically achieved without a significantly large trained classifier on approximate foreign objects.

As discussed in this chapter, the choice of technology was justified for the range of platforms and time-to-develop, but flawed in the permitting of extra work to be conducted. Time division was made as best as possible, but parts of the execution of the project could simply not be completed whilst resources were otherwise occupied.


6.4 Final Words

During the undertaking of this project, two new workshops for underwater video analysis began. NOC organised a 3-day workshop based at their premises in Southampton, and Ocean Networks Canada organised the first Underwater Video workshop at ICPR 2014 in Stockholm, Sweden. This momentum is due to continue with NOAA organising a similar workshop for WACV 2015 in Hawaii.

The publication and significant number of public presentations of the work confirm its validity in the scientific community, and also represent the growing interest in the area in current research. There is, as with any project, a number of threads that appear unfinished, and will be continue to be pursued.

In conclusion, we have focused on the use of generalised methods in underwater video, and have compared with the existing literature, whilst having identified areas of improvement. The knowledge gained from this MPhil project will now serve as the basis of PhD study in underwater video analysis, both in direct and indirect continuation.
Appendix A

Public Outputs

A.1 Posters
AUTOMATED ANALYSIS OF UNDERWATER VIDEO
DETECTING SUBSTRATES AND SEABED FAUNA

CAN COMPUTERS AND ROBOTICS HELP AUTOMATE THE UNDERSTANDING AND RETRIEVAL OF UNDERWATER VIDEO?

WHY? HOW?

The underwater environment of Cardigan Bay is largely unexplored. Previous surveys are site based with specific goals in mind. There are known to be areas that are important to marine biologists, ecologists and industry alike. These are areas of intense interest to marine biologists, ecologists and industry alike.

MOTIVATION

WHY? HOW?

Why? How?

THE PROPOSED SYSTEM

We are aiming to automate the process of information discovery in underwater video. The video itself is extremely difficult to annotate, even by human observers, due to the fluctuating clarity caused by natural variations in turbidity, fluctuations in depth, light attenuation and fouling.

Natural patterns. Numerous studies have been, and continue to be, published in the reduction of these issues.

We are looking at a large scale, of initially trying to apply a number of computer vision techniques with machine learning mechanisms, so that we can train a computer system to recognize substrates present in videos automatically.

This is performed through comparative experiments; using quantitative data provided by our industry expert, the CCW, the above frame is the resulting comparison results against the prediction.

Once this system has reached a point of viability, it will move to a patch-based system where tuning will be revisited to provide a better match to the real world. The final goal is to be able to detect changes in the environment using visual and sonar-based information to help automatically identify new substrates.

OPENROV

REMOTE OPERATED VEHICLE (ROV)

The bay extends from Borth to Strumble Head. Our research is looking at the southern part of the bay. Our interest is in Cardigan Bay, the FoCB have been exploring the environment in the bay, collecting regular surveys and data directly with local government.

LEARNING

VCIRL badges, compared with SVCLINE (a simple linear function) and SVCLINE (a simple linear function) and SVCLINE (a simple linear function) and SVCLINE (a simple linear function) and SVCLINE (a simple linear function) shown.

PRELIMINARY RESULTS OBTAINED, A PROMISING START

1. Observations

2. Openrov and FoCB simplified schema

3. Bangor & FoCB CCW schema

4. Combined simplified & CCW

5. Computer vision and machine learning

6. Gabro waves

7. Retinex

8. Scales

9. Motivation

10. Why? How?
Can computers and robotics help automate the understanding and retrieval of underwater video?

CARDIGAN BAY
The bay extends from Bardsey Island down to Strumble Head. Our research is looking at the southern part of the bay.

Local organisation, Friends of Cardigan Bay (FoCB) monitor the environment in the bay, conducting regular surveys and deal directly with local government.

This project is in direct collaboration with the FoCB, to learn more about them, go to: http://friendsofcardiganbay.org

LOCAL WILDLIFE
We know there are many interesting species in the bay, but finding and monitoring them is difficult. We’re looking at using computers to help this task.

Currently, most of the surveying is done by human divers equipped with cameras and other equipment. This is really time and money consuming, and is weather dependent.

By combining all this together, and trying the system on a few different videos from different origins, we’ve found that certain approaches seem to work well.

The finer answer is a bit more complex, as there are more factors involved in making this system better. Research continues...

GABOR WAVELETS
We define what makes a texture unique to the computer, and use different ways of comparing this data to find the most effective combination.

By combining all this together, and trying the system on a few different videos from different origins, we’ve found that certain approaches seem to work well.

The finer answer is a bit more complex, as there are more factors involved in making this system better. Research continues...
Bibliography


Bibliography


