Automated species detection: An experimental approach to kelp detection from sea-floor AUV images

M.S. Bewley, B. Douillard, N. Nourani-Vatani, A. Friedman, O. Pizarro, S.B. Williams
Australian Centre for Field Robotics
The University of Sydney, Australia
Corresponding author: m.bewley@acfr.usyd.edu.au

Abstract

This paper presents an experimental study of automated species detection systems suitable for use with Autonomous Underwater Vehicle (AUV) data. The automated detection systems presented in this paper use supervised learning; a support vector machine and local image features are used to predict the presence or absence of *Ecklonia Radiata* (kelp) in sea floor images. A comparison study was done using a variety of descriptors (such as local binary patterns and principal component analysis) and image scales. The performance was tested on a large data set of images from 14 AUV missions, with more than 60,000 expert labelled points. The best performing model was then analysed in greater detail, to estimate performance on generalising to unseen AUV missions, and characterise errors that may impact the utility of the species detection system for marine scientists.

1 Introduction

Autonomous Underwater Vehicles (AUVs) are being increasingly used to support environmental monitoring programs [10], particularly in the area of monitoring changes to sea-floor (benthic) ecosystems. Benthic ecosystems have been identified as having key importance in understanding how the marine environment responds to various pressures, including fishing activities, invasive species and climate change [2, 33].

The increased use of AUVs is primarily due to their ability to routinely capture high quality geo-refenced imagery of the sea floor in an efficient and repeatable manner. An AUV such as *Sirius* [33] is capable of performing missions autonomously. Over several hours, tens of thousands of images of the sea floor can be captured (along with other sensor data). As part of Australia's Integrated Marine Observation System (IMOS), the AUV IMOS Facility was established to collect repeated measurements of oceanographic data from reference sites around Australia's coastal areas [33]. The data set used in this paper is from a single campaign at a reference site in Eastern Tasmania, conducted in October 2008 [31, 1].

AUVs are enabling researchers to gather visual data with a precision and volume that was previously impossible. The process of manually identifying and labelling species within AUV images is highly labour intensive; in the data set used in this paper, 50 random points were labelled in each image. Understandably, this process is only feasible on a small fraction of the available data set. Scientists use these labels to estimate coverage of species or substrates within an area. There are two natural goals for an automated species detection system: firstly, to reduce or eliminate the manual workload; secondly, to make use of the complete data set in analysis, rather than being restricted to these small subsets.

In this paper, we present results on a series of experiments on automated species detection systems for *Ecklonia Radiata* (kelp). The experiments have two main phases:

A comparison study was performed on the use of local image descriptors to predict the presence of kelp. Several local image descriptors were used at a range of scales to train a support vector machine. Results from these models act as a performance benchmark for kelp detection on a large AUV data set.

The highest performing model was analysed in further detail to estimate the performance when used on unseen AUV dives, and determine the distribution of prediction errors across some environmental variables. . Given that the goal is to build species detection systems that apply to the analysis of the ecology and biology of an area, it is important to estimate the likely performance in a variety of scenarios, rather than simply maximising the classification accuracy on the whole test set.

While the long term aim is to detect multiple species using the AUV data, detection of kelp is useful in its own right. Around much of Australia's sub-tropical and temperate areas, kelp forests act as the primary habitats for many species, and the extent of kelp is used as a proxy for the health and biodiversity of an area [29].

The remainder of this paper is divided into the following sections: Section 2 describes related work done in the area of classification and clustering on benthic images. Section 3 describes the algorithms used to detect kelp in this paper. Section 4 is an overview of the way the AUV images were collected and labelled, as well as the training and testing strategy. Section 5 compares results from a range of descriptors and scales. Section 6 takes the best performing model from the previous section, and subjects it to further examination. Section 7 provides conclusions and directions for future work.

2 Related Work

While this paper frames species detection in the context of the scientific end users of the data, there is also a clear motivation from a direct robotics perspective. Given the limited and unreliable communication bandwidth with AUVs, it is desirable for an AUV to make sense of its surroundings, and respond appropriately. Existing AUV missions typically rely on following predetermined paths, utilising an underwater simultaneous localisation and mapping (SLAM) solution [30]. A long term research goal is for the robot to associate semantic labels with its surroundings as it travels, such that it can adjust its mission, or send the information to human operators.

Whole-Image Approaches

An initial approach that goes some way towards species detection lies in performing clustering and classification at the whole image level. The scale of the images produced by an AUV is fairly predictable (around 1-2 metres across). This is because strong attenuation of light in water requires the robot to travel at an altitude of around 2 metres above the sea floor. Work following the whole-image approach includes online classification of image habitats for adaptive sampling [23], active learning to improve semantic relevance of image clusters [12], and post-processing of large quantities of AUV images for unsupervised clustering [28, 22].

Sub-Image Approaches

The whole-image approaches provide useful semantic information on the image data, but they lack the ability to discriminate between smaller features. As shown in Figure 1, an image can contain examples of multiple occurrences of multiple species. For species detection, it is more accurate to label each occurrence of each species.

In a recent paper, a solution to sub-image detection of coral species from images is presented, with a similar aim of providing coverage estimates for marine scientists [3]. The authors highlight the challenge of recognising amorphous and variable biological organisms, compared to the structured environments often used to test object detection algorithms. In addition, the type of semantic labelling (random image point labels) is the same as presented here.

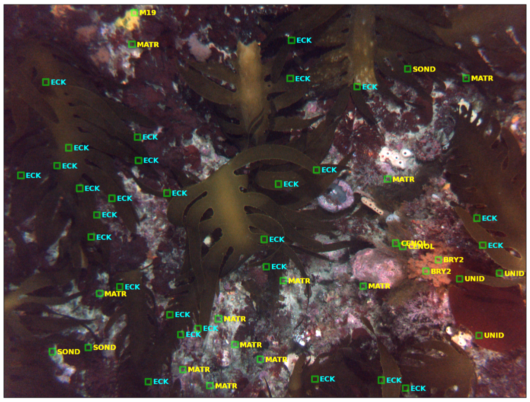

Figure 1: An example AUV image containing multiple dominant species. Expert labels are overlaid (with KELP as *ECK*)

The work performed in [8] is particularly relevant, as it aims to detect exactly the same species of kelp as our study, and is also based on AUV data. In this work, image features were combined with clustering to produce a predictive model, and results were assessed against a set of 45 images. Our paper expands further, using a wider range of image features on a significantly larger data set.

Other sub-image approaches have also been attempted. Several groups have created detection mechanisms for particular species, such as starfish [25, 6, 9] and coral [26]. These studies have typically only tested a small range of features. In addition, the available data was often insufficient to either produce optimal results, or evaluate classifier performance sufficiently to enable confident use of its output in a marine science context.

Some key distinctions exist between the aim of these earlier studies, and the current paper. The data set introduced in this paper presents an opportunity not currently present in the literature. With more than 60,000 labelled points in the set, only the coral data set of [3] is comparable in size (it is actually larger, at 400,000 labelled points). The data set from that study was, however, captured by divers using handheld cameras, rather than an AUV. The potential advantage of using an AUV data set is significant: the ease of capturing a comprehensive data set; spatial registration of the images; availability of additional sensor data (such as 3D stereo) which may boost performance; and the ability to perform much more precise repeated surveys. In addition, the data set used here is from temperate coastal waters off Tasmania, rather than tropical reef images from French Polynesia.

3 Methodology

This methodology applies to both the comparison study (Section 5) and the final model evaluation (Section 6). In summary, experts hand-labelled randomly distributed keypoints on AUV images. Image descriptors were computed on local image patches at various scales centred around the keypoints. Note that throughout this paper, we refer to a feature as a particular type of descriptor computed at a particular scale. Each set of features was used as the input to supervised classification. The goal was to learn to classify the presence or absence of kelp, based on the expert label assigned to the centre pixel of each patch.

3.1 Scale and Keypoints

All the features used for training and testing in this study were derived from local image patches. Square patches of various sizes were used, with a patch centred on each hand labelled keypoint from the data set. This was done to provide some local context around the centre pixel, from which features could be derived. A range of scales were evaluated in the experiments, from 7×7 to 95×95 pixel patches. The image resolution is 1360×1024. At 2m above the sea floor, these patch sizes correspond to approximately 0.7cm to 10cm. Scale trades off between precision (as only the centre pixel is being labelled), and the inclusion of wider context.

3.2 Descriptors

Based on the square patches defined above, various descriptors were used to describe the keypoints. This dimensionality reduction process (where the NxN image patch is reduced to a lower dimensional feature vector) is a commonly used step in machine learning problems; it projects the data into a new, lower dimensional space before the classification algorithm is applied.

Raw Pixels

For the smallest patch sizes (limited by computational tractability), the patch pixels were used directly as a feature vector. As all descriptors are derived from these raw pixel values, it is worthwhile comparing the results.

Principal Component Analysis (PCA)

PCA selects an orthogonal set of coordinates that maximise variance [4]. In the experiments, PCA was applied to a flattened vector of the raw pixel values from each patch, and the lowest variance components were discarded. By experimentation, it was found that using anywhere between 20 and 60 components had little impact on the results. As such, 30 components were selected, and whitening was used in all cases. Greyscale PCA refers to PCA performed on the means of RGB channels. RGB PCA refers to PCA performed directly on all three colour channels.

Grey Level Co-occurrence Matrices (GLCM)

GLCMs were first proposed by Haralick in the 1970s [14, 13]. Where PCA is a naive, commonly used dimensionality reduction technique, GLCMs were created specifically for forming descriptors that represent texture in greyscale images, and have also been extended to colour. The colour GLCM extension was used in [8] for kelp detection. In that study, the GLCM parameters of uniformity, contrast, correlation, local homogeneity and entropy were used. These parameters were computed at angles of 0, 45, 90, 135 degrees and summed to obtain one final rotation invariant matrix. These statistics were computed at four radii, spaced approximately logarithmically between a single pixel offset, and half the patch width (50×50 or 100×100 pixels). To reduce the dimensionality of the matrices, images were reduced to 32 intensity levels on each channel. In order to make use of colour information, the GLCMs were computed between pairs of colour channels (although the red channel was not used, due to the strong absorption of red light in water). By using the same GLCM descriptor in this paper (however at different scales), we evaluated it on a much larger data set, against other common descriptors.

Local Binary Patterns (LBP)

Local Binary Patterns (LBP) [21] are a texture descriptor that has been widely used for underwater image interpretation. [19, 6, 26, 12]. They can be computed at multiple scales and made rotation invariant. Combining multiple LBP scales has proven useful for improving classification performance [21].

A number of studies have been performed comparing LBPs to other descriptors (including GLCMs), and LBPs were usually found to provide superior results [11, 24, 6, 27].

In brief, a ring of pixel intensities are compared to the intensity at the centre of the ring. A threshold is applied to the differences, to create the local binary pattern. A histogram of the occurrence of these patterns is computed for each image patch, and used as the descriptor. In this study, a ring radius of 1 pixel, with 8 points, was used to create the patterns.

Histogram of Oriented Gradients (HOG)

The HOG algorithm was designed for detection of humans in images, but has also been used in other applications [7]. It breaks an image patch into cells, and computes gradients within those cells. It also uses blocks (adjacent cells) to perform some relative normalisation for illumination invariance. For this application, HOG with 7×7 cells was used, with a block consisting of 3×3 cells. This means that the smallest patch used (7×7) consisted of a single cell, whereas the largest patch (95×95) consisted of a grid of 13×13 cells. As the descriptor was produced by concatenating the cell histograms, the length of the descriptor increased with the square of the patch edge size.

3.3 Classification

The key focus of this paper is on a comparison of features, and the system level solution of the kelp detection problem. A single classifier was used - a Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel. While further optimisation may be possible using alternative classification algorithms, RBF-SVMs are known to provide good results on a wide range of problems [15]. In brief, an SVM finds the globally unique linear separating hyperplane that separates the two classes with the maximal margin. As an extension, the RBF kernel first transforms the input vector to infinite dimensional Hilbert space, so that non-linear separations can be obtained. Only two parameters need to be tuned. C is the cost parameter (large C is less tolerant of non-separable data sets, and forces a model that is more accurate on the training set). γ adjusts the radius of the kernel (large γ decreases the kernel radius, which reduces the distance over which support vectors have influence, also creating a more complex decision boundary).

In this study, parameter selection was performed by observing cross-validation results of a grid search over C and γ. Following the guidance of [15], the parameter ranges used were $C = [2^{-5}, 2^{15}]$, and $\gamma = [2^{-15}, 2^{3}]$, with a logarithmic density of powers of 2.

4 Data Set

An additional objective of this paper is to establish a benchmark for species detection in a large, high quality AUV data set. Future research can then assess the benefit of using AUV sensors and positioning, compared to image processing alone.

The primary use of the labelled data set used in this study is to estimate percentage cover and distribution of various species and features in an area. For kelp, this is done by taking all the labelled images from the area of interest, and calculating the percentage of labelled points that are labelled as *ECK*.

4.1 Classes and Labelling

The entire data set is comprised of 14 dive missions conducted by the AUV *Sirius* off the South-East coast of Tasmania in October 2008 [32]. To capture images, Sirius uses a calibrated stereo pair of 1.4 megapixel cameras, and uses strobes to illuminate each capture. From the data set containing over 100,000 stereo pairs of images, marine scientists selected every 100th colour image, and used the CPCe software package [18] to label 50 random points on each (as shown in Figure 1). Images were taken at

A wide range of class labels were used, indicating biological species (including types of sponge, coral, algae and others), abiotic elements (types of sand, gravel, rock, shells etc.), and types of unknown data (ambiguous species, poor image quality, etc.). Precise details of the labelling methodology can be found in [1]. The KELP (ecklonia radiata) class was used as-is, and all other points were relabelled as OTHER. These labels were then used as training and testing instances, with features derived from localised image patches. Many of the classes can be arranged in a hierarchy, as shown in Figure 2. Of particular relevance are the groups “Brown Macroalgae” (containing kelp and other brown aquatic plants), and “Macroalgae” (containing brown, red and other macroalgae). Figure 3 shows some examples, where the yellow crosses indicate the labelled points, with the surrounding patch on which the descriptors were computed.

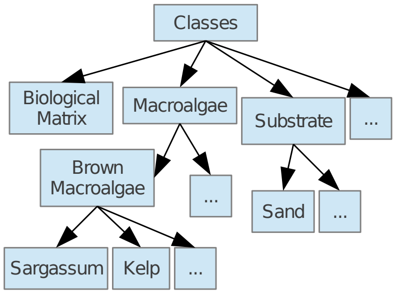

Figure 2: Class hierarchy summary

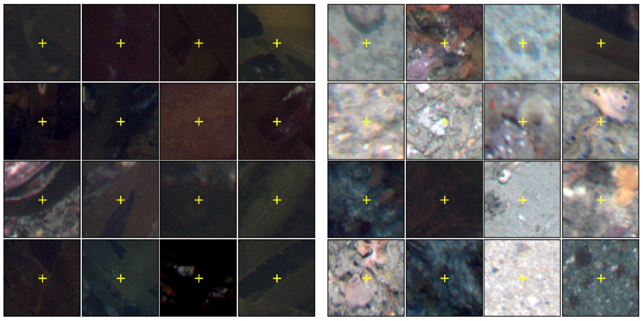

Figure 3: Example 95×95 patches of kelp (left) and all classes (right)

4.2 Training, Validation and Test Sets

As shown in Figure 4, data was divided into training, validation and test sets. This is described in more detail below.

Test Set

A random 20 percent of the images were reserved for use in evaluating the final model. Data from these images was not used in any stage of the model selection or parameter optimisation.

Training and Validation Sets

The images not reserved in the test set were used for training and validation. 67% of the randomly labelled image points were used as the Training Set, and the remaining 33% as the Validation Set. A fixed validation set was used for model comparison, instead of cross validation. This greatly reduced the computational requirement (by approximately an order of magnitude compared to 10-fold cross validation). Training set sensitivity results on the final model (Figure 7) demonstrated that peak performance could be obtained with significantly smaller training sets, rendering cross-validation unnecessary.

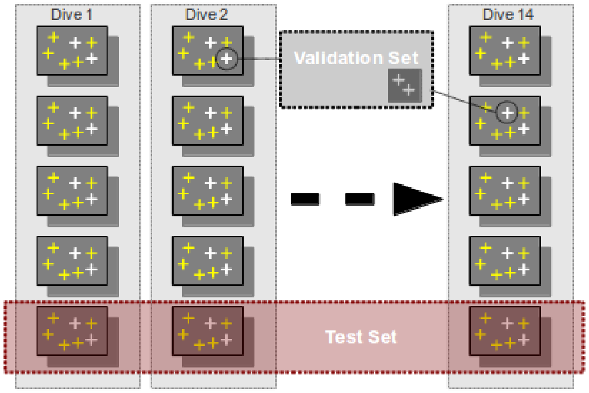

Figure 4: Breakdown of data set into Training Set (yellow +'s), Validation Set (white +'s) and Test Set (red box)

Training and Testing Strategy

For the Model Comparison stage, 3-fold cross validation was performed within the Training Set to optimise the SVM parameters C and γ. Models were then retrained on the full training set, and performance results reported on the Validation Set. Once the best model had been selected, detailed analysis was performed on the Test Set results.

4.3 Dive Characteristics

Figure 5 shows the most prominently featured classes in each dive. It is worthwhile noting: Overall, only around 5% of the points are KELP; the largest component of OTHER is sand; 6 of the dives included no KELP at all.

While dives in the campaign were conducted in a similar geographic area (the South-East coast of Tasmania), and at a similar time (over several weeks) they vary greatly in depth and content. It is important to train and test across multiple dives, as the detection system needs to be robust enough to manage anything from a deep dive over sand, to a shallow dive with rock and kelp.

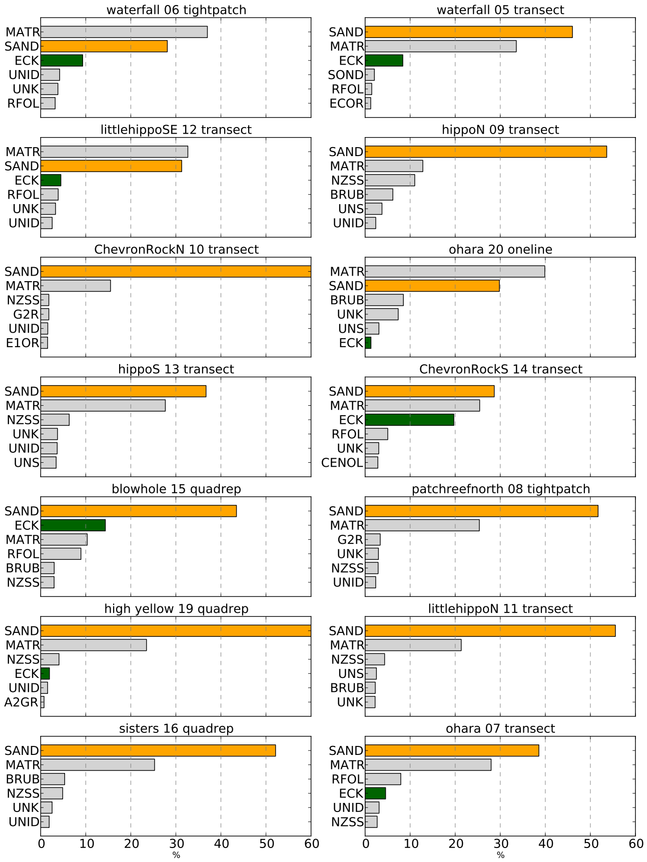

Figure 5: Percentage distribution of most commonly occurring class labels in each dive mission. The mission names consist of a number (indicating the chronological order), and keywords describing the local area. Key classes are SAND (sand), ECK (kelp) and MATR (biological matrix, typically a range of biological organisms in one area).

5 Model Comparison Experiments

A series of experiments was run to test the performance of classifiers trained using the various local image descriptors and scales described in Section 3.

	Raw	PCA	GLCM	HOG	LBP
7 × 7	0.63	0.63	0.41	0.17	0.38
15 × 15	0.60	0.61	0.48	0.30	0.54
31 × 31	-	0.57	0.51	0.44	0.61
63 × 63	-	0.54	0.57	-	0.62
95 × 95	-	0.57	0.63	-	**0.65**

Table 1: f1-scores for greyscale descriptors

5.1 Patch, Feature and Classifier Combinations

Greyscale Comparison

A classifier was trained to predict KELP vs OTHER, using the descriptors extracted from greyscale patches at each scale. This allows both a comparison of the descriptors, and an assessment of which scales are most useful.

Colour Comparison

Some descriptors (Raw, PCA, GLCM) can be extended to work in the RGB colour space. This comparison is less complete (as it includes fewer descriptors), but it allows an assessment of how useful the colour information in the AUV images is used for species detection. Note that for colour GLCM, only the green and blue channels were included, to maintain consistency with [8].

5.2 Performance Evaluation

One key attribute of the data set is that it is significantly skewed against KELP. The OTHER class contains approximately 20 times the number of instances. During classification, the instances were weighted [16] in the RBF-SVM, so as not to bias the solution against the minority class.

Class weighting reduces the inherent classifier bias, but it is still important to choose an appropriate performance metric for the classifiers, that is insensitive to class imbalance. The f1-score is commonly used for biased classes, and is used as the primary performance metric in this paper. It is defined as the geometric mean of precision and recall on the minority class. Precision (the percentage of instances classified as kelp that are actually kelp) and recall (the percentage of genuine kelp that is classified as kelp) must both be high in order to obtain a good f1-score.

5.3 Results

Table 1 and Table 2 are the final results on the Validation Set, after training the RBF-SVM classifier on each descriptor and scale. Note that where a '-' is listed, the model was unable to be run, usually due to dimensionality of the feature vector, or memory limitations. Figure 6 shows both greyscale and colour descriptors across all scales.

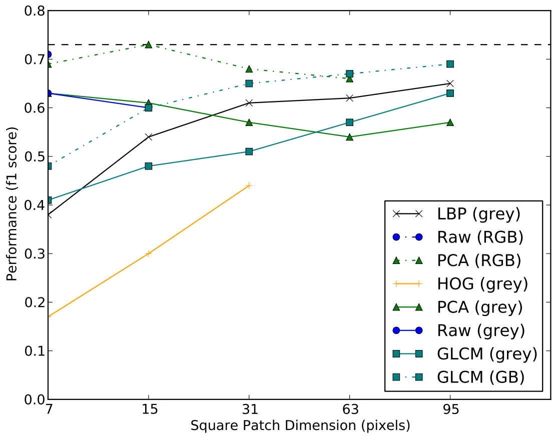

Figure 6: f1-scores for colour and greyscale descriptors. The dashed line highlights the performance of the peak performing model)

	Raw	PCA	GLCM
7 × 7	0.71	0.69	0.48
15 × 15	-	**0.73**	0.60
31 × 31	-	0.68	0.65
63 × 63	-	0.66	0.67
95 × 95	-	-	0.69

Table 2: f1-scores for colour descriptors

5.4 Discussion

Greycale Comparison

Both Raw pixels and PCA performed best at the smallest scale (7 × 7). The similarity indicates that at the smaller scales, very little useful information is discarded by PCA. As the scale was increased, training the SVM on raw pixels became computationally intractable, and PCA showed a slight decrease in performance. This decrease in performance at larger patch sizes for PCA suggests that the descriptor takes its information from the immediate local area, and larger patch sizes are simply adding noise, weakening performance.

HOG, GLCM and LBP all performed relatively poorly at smaller scales, with steadily increasing performance towards the larger patch size. As these descriptors were explicitly constructed to capture textural information, this strongly suggests that the textures that are best suited to kelp detection are on the scale of the larger patch sizes.

It is worth noting that the PCA model operating on very localised patterns performs similarly to LBP and GLCM at the largest patches (the LBP having a slightly higher performance). Given the difference in scale used, it is possible that a combination of the 7 × 7 PCA and 95×95 (or larger) LBP would have superior performance.

This suggests that the combination of image descriptors at multiple scales could be an interesting area for further work. In addition, although LBP gave the best greyscale performance, no colour extension was tested. A colour extension of LBP could provide superior results to the PCA and GLCM descriptors that were tested, and would be a worthwhile area of future work.

Colour Comparison

Colour correction in underwater imagery is a notoriously difficult problem. Water absorbs the red much more strongly than other parts of the spectrum, so there is strong dependence on depth of both illumination and colour. The strobes on the AUV are limited in range, and objects below the AUV can vary from well illuminated, to strongly over or under exposed.

Despite this, all of the colour descriptors exhibited a clear performance increase over their greyscale counterparts, at all scales. This indicates that even with the basic colour correction used in this data set [17], including colour information in the descriptors is beneficial for kelp detection. A method for more advanced colour correction has been proposed [5], which may result in further improvements to performance.

6 Final Model Evaluation

The primary aim of separating this study into "Model Comparison Experiments" and "Final Model Evaluation" was to enable an in-depth analysis of the quality and error on the final classifier. This section includes retraining the model on various subsets of the Training Set, and then an analysis of its properties on the Test Set.

The model chosen for evaluation was the PCA (RGB) at the scale of 15×15 . This model was selected for its combination of high performance, and low computational load (only requiring the linear combination of a small image patch, where other descriptors have greater computational complexity, and require computation over much larger patches).

6.1 Training Set Sensitivity Tests

In this group of tests, we retrained the classifier on subsets of the Training Set, to determine the impact on performance. Determining an appropriate training set size is of great interest for automating species detection. If the training data is found to be insufficient, more data could be manually labelled in order to improve performance. However, if a classifier of similar performance can be built with less training data, then this puts less burden on the scientists to manually label large amounts of data.

The size of the training set was varied in several ways:

Discarding a percentage of random patches. This tests whether the training set was of sufficient size to permit optimum performance on the model

A "single dive" approach. This trains on one dive at a time, and demonstrates how optimistic a single-dive model is when generalised to many dives

A "leave one out" approach. This trains on all but one dive, simulati the performance on an additional, unseen dive being performed in the same geographic area.

6.2 Error Analysis Tests

The experiments in Section 6.1 were designed to test whether the data sets were of sufficient size to provide good generalisation performance. The question remains, however, as to *where* the classifier makes mistakes. As stated previously, the primary intended use of the kelp detection system is to estimate percentage cover of kelp in an area. Unevenly distributed error between dives, or relative to other variables could result in coverage estimate errors that are significantly above what the overall model performance would suggest. Any variation in performance for some subset of the data (such as a bias for another type of macro-algae to be recognised as kelp) needs to be understood by scientists making use of the automated detection system, or their conclusions may be biased by the error of the detection system. To provide insight into this problem, the error on the Test Set was broken down by dive, and by groups of some of the original expert labelled classes.

6.3 Results and Discussion

Overall Performance

Training	Testing	Precision	Recall	f1-score
Training Set	Training Set	0.79	1.0	0.88
Training Set	Validation Set	0.69	0.77	0.73
Training Set	Testing Set	0.64	0.74	0.69

Table 3: Precision, Recall and f1-score on each testing set, using PCA (RGB) on 15×15 patches

Results are shown in Table 3. The f1-score of KELP decreases when assessed on the Training, Validation and Testing Sets. The higher performance on the Training Set could indicate slight over-fitting. The slightly lower performance on the Testing Set compared to the Validation Set could be attributed to one of two effects.

Firstly, the Testing Set contained keypoints from completely unseen images, whereas the Validation Set is made up of keypoints that occupy different areas of the same images as the training set. This makes the Validation Set more optimistic; unusual illumination, colour balance or content of an image is likely to be represented in both Training and Testing Sets.

Secondly, there is a known effect when a large number of models are compared on a validation set. If the model with the best performance is selected, the performance quoted on that same set is biased to be slightly optimistic [20].

Training Set Sensitivity

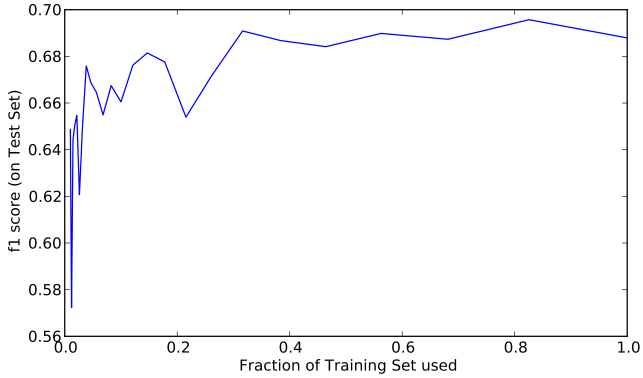

Figure 7: Performance on random subsets of training data

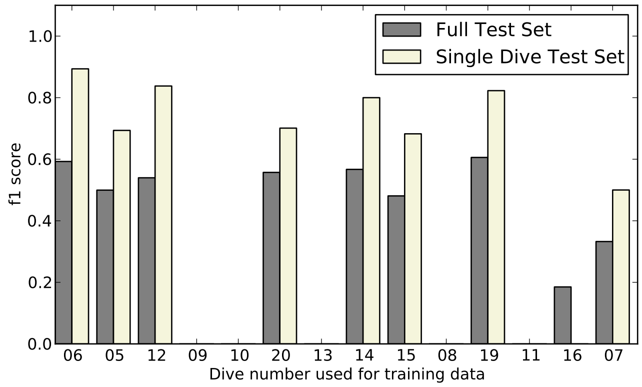

Figure 8: Performance on single dive training sets

The Training Set percentage results in Figure 7 demonstrate that the model has no significant performance improvement when trained on more than around 30% of the original Training Set. This supports the earlier assumption that the training set was of sufficient size that a separate Validation Set and Test Set could be reserved (rather than using cross-validation). It also gives an indication of how much hand-labelling is required to make an optimum model (at least one based on local image features). As discussed earlier, the onerous nature of hand labelling makes this a valuable contribution. 30% of the original Training Set corresponds to 10,086 hand labelled points (700 of which are kelp). This is far smaller than the full data set of more than 60,000 hand labelled points.

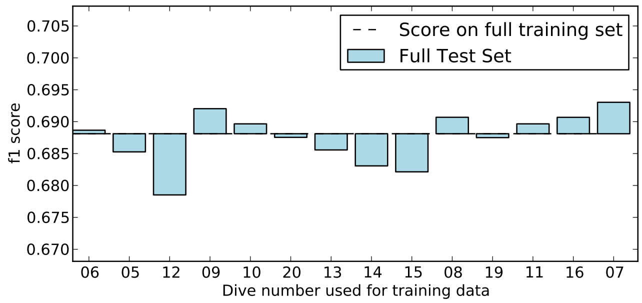

Figure 9: Performance on discarding individual dives from training set

In Figure 8, the right hand columns show the performance one gets looking solely at data from one dive (both for training and testing). The left hand columns show the performance of the exact same model on the full Test Set (all dives). It is interesting to note a significant drop for most dives when tested on the full Test Set. The marked drop (of around 20% in kelp f1-score) show that generalisation from a single dive is poor even on dives done at a similar time, in a nearby geographic location.

The results after training on all but one dives (Figure 9) are very similar to the performance when the full training set is used.

The most interesting deduction comes from combining these results. Together, they suggest that although training on a single dive has poor generalisation, it is reasonable to expect good generalisation results in a geographic area with sufficient training data. This means that once a species detection system has been trained on a number of dive missions, it may not be necessary for experts to perform any further hand labelling on subsequent dives in the area.

Error Analysis

Table 4 shows an analysis of performance on each dive, on the model trained on the full Training Set. It is clear that there is significant performance variation across the dives (note, however, that the f1-score for KELP is misleading on subsets where the true amount of KELP is very small, and undefined with 0% kelp).

Table 5 shows a breakdown of where errors occurred in the test set, on various subsets of the data. This table reveals a highly skewed error distribution. Despite comprising nearly half the set, only one sand point was incorrectly labelled as kelp. In contrast, the error rates on the macro algae (particularly the brown group, of similar colour to kelp) were significantly higher. Col-

Dive	Total Accuracy	f1-score	% Kelp
14	92.92%	80.15%	18.96%
12	91.63%	73.79%	17.52%
15	86.18%	60.66%	12.48%
20	94.74%	71.67%	9.12%
6	97.70%	87.10%	9.06%
5	93.53%	57.92%	5.55%
19	97.45%	69.84%	5.38%
7	98.95%	40.00%	0.53%
16	99.68%	-	0.26%
13	99.28%	-	0.00%
11	99.54%	-	0.00%
10	99.33%	-	0.00%
9	100.00%	-	0.00%
8	99.84%	-	0.00%

Table 4: Performance on each dive, using PCA (RGB) on 15 × 15 patches

Test Set subset	N	Accuracy	Errors
Kelp	663	74.21%	171
Brown M.A. (except kelp)	34	44.12%	19
All other M.A.	556	73.38%	148
Sand	5279	99.98%	1
All other classes	5780	98.15%	107
Total	12312	96.38%	446

Table 5: Performance on subsets of the data ('M.A.' refers to macro-algae), using PCA (RGB) on 15 × 15 patches

lectively, the combined macro algae groups contribute 61% of the false positives, despite the fact they constitute only 5% of the OTHER instances in the test set. While it is unsurprising that the model has most difficult with other types of macro algae, the magnitude of the problem should provide a focus for further research.

7 Conclusion

In this paper, we have introduced a large data set of expert labelled AUV images, computed benchmark results on a number of different techniques, and examined the behaviour of an automated species detection system in detail. Results on detection of kelp were promising, and the work could be used to develop a practical system to assist marine scientists. For the kelp detection problem, future work that may improve performance includes: combining multiple scales; finding superior descriptors; and making use of the additional non-image information embedded in the AUV data.

For the evaluation of the model, we provided evidence to suggest that a fixed species detection model with good generalisation to at least a local geographic region was feasible. The skewed error distribution in Table 5 highlights the importance of creating detailed performance metrics for end-users, to avoid biasing their scientific outcomes. Finally, the problem should be expanded beyond *Ecklonia Radiata* to allow discrimination between multiple species and features.

References

[1] N.S. Barrett, L. Meyer, N. Hill, and P.H. Walsh. Methods for the processing and scoring of AUV digital imagery from South Eastern Tasmania. Technical Report 11560, University of Tasmania, August 2011.

[2] N.S. Barrett, J. Seiler, T. Anderson, S. Williams, S. Nichol, and S.N. Hill. Autonomous underwater vehicle (AUV) for mapping marine biodiversity in coastal and shelf waters: Implications for marine management. In *IEEE OCEANS*, pages 1 –6, May 2010.

[3] O. Beijbom, P.J. Edmunds, D.I. Kline, B.G. Mitchell, and D. Kriegman. Automated annotation of coral reef survey images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1170, 2012.

[4] C. M Bishop. *Pattern recognition and machine learning*. Springer Science+Business Media, LLC, New York, 2006.

[5] M. Bryson, M. Johnson-Roberson, O.R. Pizarro, and S. B. Williams. Repeatable robotic surveying of marine benthic habitats for monitoring long-term change. In *Robotics: Science and Systems Conference (RSS)*, July 2012.

[6] R. Clement, M. Dunbabin, and G. Wyeth. Toward robust image detection of crown-of-thorns starfish for autonomous population monitoring. In *The Australasian Conference on Robotics and Automation (ACRA)*, 2005.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 886 – 893, June 2005.

[8] A. Denuelle and M. Dunbabin. Kelp detection in highly dynamic environments using texture recognition. In *The Australasian Conference on Robotics & Automation (ACRA)*, December 2010.

[9] V. Di Gesu, F. Isgr, D. Tegolo, and E. Trucco. Finding essential features for tracking starfish in a video sequence. In *International Conference on Image Analysis and Processing*, pages 504 – 509, 2003.

[10] M. Dunbabin and L. Marques. Robots for environmental monitoring: Significant advancements and applications. *Robotics Automation Magazine, IEEE*, 19(1):24 – 39, March 2012.

[11] I Fogel and D Sagi. Gabor filters as texture discriminator. *Biological Cybernetics*, 61:103 – 113, 1989.

[12] A. Friedman, D. Steinberg, O. Pizarro, and S. B. Williams. Active learning using a variational dirichlet process model for pre-clustering and classification of underwater stereo imagery. In *Intelligent Robots and Systems (IROS)*, pages 1533 – 1539, 2011.

[13] M. Hall-Beyer. The GLCM tutorial home page. http://www.fp.ucalgary.ca/mhallbey/tutorial.htm, February 2007.

[14] R.M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, 3(6):610 – 621, November 1973.

[15] C.W. Hsu, C.C. Chang, and C.J. Lin. A practical guide to support vector classification, April 2010.

[16] Y.M. Huang and S.X. Du. Weighted support vector machine for classification with uneven training class sizes. In *International Conference on Machine Learning and Cybernetics*, volume 7, pages 4365 – 4369, August 2005.

[17] M. Johnson-Roberson, O.R. Pizarro, S.B. Williams, and I. Mahon. Generation and visualization of large-scale three-dimensional reconstructions from underwater robotic surveys. *Journal of Field Robotics*, 27(1):21 – 51, 2010.

[18] K. E Kohler and S. M Gill. Coral point count with excel extensions (CPCe): a visual basic program for the determination of coral and substrate coverage using random point count methodology. *Computers & Geosciences*, 32(9):1259 – 1269, 2006.

[19] M.S.A. Marcos, M Soriano, and C. Saloma. Classification of coral reef images from underwater video using neural networks. *Optics Express*, 13(22):8766 – 8771, 2005.

[20] A.Y. Ng. Preventing "overfitting" of cross-validation data. In *International Conference on Machine Learning (ICML)*, pages 245 – 253, 1997.

[21] T. Ojala, M. Pietikaeinen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971 – 987, 2002.

[22] O.R. Pizarro, S.B. Williams, and J. Colquhoun. Topic-based habitat classification using visual data. pages 1 –8, May 2009.

[23] P. Rigby, O.R. Pizarro, and S.B. Williams. Toward adaptive benthic habitat mapping using gaussian process classification. *Journal of Field Robotics*, 27(6):741 – 758, 2010.

[24] C. Shan, S. Gong, and P.W. McOwan. Robust facial expression recognition using local binary patterns. In *IEEE International Conference on Image Processing*, pages 4 – 7. IEEE, 2005.

[25] D. Smith and M. Dunbabin. Automated counting of the northern pacific sea star in the derwent using shape recognition. In *The Australasian Conference on Robotics & Automation (ACRA)*, pages 500 – 507, December 2007.

[26] M. Soriano, S. Marcos, C. Saloma, M. Quibilan, and P. Alino. Image classification of coral reef components from underwater color video. volume 2, pages 1008 –1013, 2001.

[27] M. Soriano, T. Ojala, and M. Pietikainen. Robustness of local binary pattern operators to tilt compensated textures. *Workshop on Texture Analysis in Machine Vision Oulu Finland*, 1999.

[28] D. Steinberg, A. Friedman, O.R. Pizarro, and S.B. Williams. A bayesian nonparametric approach to clustering data from underwater robotic surveys. In *International Symposium on Robotics Research (ISRR)*, August 2011.

[29] R.B. Taylor. Seasonal variation in assemblages of mobile epifauna inhabiting three subtidal brown seaweeds in northeastern new zealand. *Hydrobiologia*, 361(1):25–35, 1997.

[30] S.B. Williams, P. Newman, G. Dissanayake, and H. Durrant-Whyte. Autonomous underwater simultaneous localisation and map building. In *International Conference on Robotics and Automation (ICRA)*, volume 2, pages 1793 –1798, 2000.

[31] S.B. Williams, O. Pizarro, M.V. Jakuba, D. Steinberg, A. Friedman, and N. Ahsan. Autonomous underwater vehicle monitoring of benthic reference sites. In *Intelligent Robots and Systems (IROS)*, August 2010.

[32] S.B. Williams, O.R. Pizarro, M. Jakuba, and N.S. Barrett. AUV benthic habitat mapping in south eastern Tasmania. In Andrew Howard, Karl Iagnemma, and Alonzo Kelly, editors, *Field and Service Robotics*, volume 62, pages 275 – 284. 2010.

[33] S.B. Williams, O.R. Pizarro, M.V. Jakuba, C.R. Johnson, N.S. Barrett, R.C. Babcock, G.A. Kendrick, P.D. Steinberg, A.J. Heyward, P.J. Doherty, I. Mahon, M. Johnson-Roberson, D. Steinberg, and A. Friedman. Monitoring of benthic reference sites: Using an autonomous underwater vehicle. *Robotics and Automation Magazine, IEEE*, 19(1):73 – 84, March 2012.