UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Automated Annotation of Coral Reef Survey Images

Permalink

<https://escholarship.org/uc/item/0rd0r3wd>

Author

Beijbom, Oscar

Publication Date

2015

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Automated Annotation of Coral Reef Survey Images

A dissertation submitted in partial satisfaction of the
requirements for the degree of Doctor of Philosophy

in

Computer Science

by

Oscar Beijbom

Committee in charge:

David Kriegman, Chair
Serge Belongie
Sanjoy Dasgupta
Brian Gregory Mitchell
Nuno Vasconcelos

2015

Copyright
Oscar Beijbom, 2015
All rights reserved.

The Dissertation of Oscar Beijbom is approved and is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2015

DEDICATION

I dedicate this to my family, and my love.

TABLE OF CONTENTS

Signature Page	iii
Dedication	iv
Table of Contents	v
List of Figures	viii
List of Tables	x
Acknowledgements	xi
Vita	xiv
Abstract of the Dissertation	xvi
Introduction	1
Chapter 1 Background	4
1.1 Data Collection	4
1.2 Point Annotations	5
1.3 Computer Vision Methods For Automated Annotation	7
1.4 Computer Vision Coral Ecology Data-sets	8
Chapter 2 Texture and color descriptors	10
2.1 Introduction	12
2.1.1 Related work	13
2.1.2 Contributions	15
2.2 Dataset	15
2.2.1 Moorea Labeled Corals	16
2.2.2 Labeling Ambiguity	16
2.3 Method	17
2.3.1 Preprocessing	17
2.3.2 Texture and Color	18
2.3.3 Descriptors	19
2.3.4 Machine Learning	21
2.4 Experiments	21
2.4.1 Patchwise Results	22
2.4.2 Coverage Results	22
2.5 Discussion	24
2.6 Conclusion	24
Chapter 3 Variability of Human Experts and Operational Modes of Automation	26
3.1 Introduction	27
3.2 Materials & Methods	30

3.2.1	Study Locations	31
3.2.2	Label-set	33
3.2.3	Manual Annotations	34
3.2.4	Automated Annotations	35
3.2.5	Post-processing of Annotations	36
3.2.6	Estimating Annotator Variability	37
3.2.7	Modes of Automated Annotation	38
3.2.8	CoralNet	40
3.3	Results	41
3.3.1	Annotator Accuracy	41
3.3.2	Modes of Automated Annotation	43
3.4	Discussion	48
3.4.1	Annotator Accuracy	48
3.4.2	Modes of Automated Annotation	50
3.5	Conclusions	51
3.6	Supporting Figures & Tables	53
	Appendices	57
3.A	Details of Coral Reef Survey Locations	57
3.A.1	Moorea	57
3.A.2	Northern Line Islands	57
3.A.3	Nanwan Bay, Taiwan	58
3.A.4	Heron Reef, Great Barrier Reef	58
3.B	Classification Using Linear Support Vector Machines	58
3.C	Importance of Training Size for Alleviate	60
3.D	Refine: a Supplementary Operation Mode	61
Chapter 4	Cost-Effective Sampling Designs for Automated Annotation	64
4.1	Introduction	65
4.1.1	Related Work	67
4.2	Problem Setup	68
4.3	Conventional Sampling	69
4.4	Hybrid Sampling	70
4.5	Auxiliary Sampling	73
4.6	Experimental Verification	76
4.6.1	Applicability of Hybrid Sampling	77
4.6.2	Validity of Assumptions	79
4.7	Discussion	82
4.7.1	Sampling under Domain Shift	82
4.7.2	Estimation of Multiple Quantities	83
4.8	Conclusion	84
	Appendices	84
4.A	Proof of Theorem 2	84
4.B	Proof of Theorem 3	85
4.C	Two-Stage Auxiliary Sampling	86

Chapter 5	Fluorescence imaging	90
5.1	Introduction	92
5.2	Materials & Methods	94
5.2.1	Image Collection	94
5.2.2	Post-processing	95
5.2.3	Automated Annotation	97
5.2.4	Statistical Analysis	101
5.3	Results	102
5.3.1	Analysis of Image Intensities	102
5.3.2	Automated Annotation Using Convolutional Neural Networks	103
5.3.3	Training a Five-channel Network	104
5.3.4	Comparison to Traditional Automated Annotation Methods	105
5.4	Discussion	106
Chapter 6	CoralNet	110
6.1	CoralNet Design Overview	110
6.2	Snapshot of CoralNet Vision Backend Performance	112
Chapter 7	Conclusion	115
7.1	Final Thoughts	115
7.2	Future Work	117
Bibliography		119

LIST OF FIGURES

Figure 1.1.	Evolution of methods for benthic cover estimation	6
Figure 1.2.	Sample-point-annotations	7
Figure 2.1.	Example images from the Moorea Labeled Corals dataset.	11
Figure 2.2.	Sample patches from the Moorea Labeled Corals dataset.	14
Figure 2.3.	Illustration of the patch size selection problem.	19
Figure 2.4.	Importance of using appropriate patch-size for recognition.	20
Figure 2.5.	Confusion matrices between automated and human annotations.	23
Figure 2.6.	Scatter plots of benthic coverage estimation.	25
Figure 3.1.	Sample photoquadrats acquired and annotated as part of four long-term monitoring projects.	32
Figure 3.2.	Classification accuracy.	42
Figure 3.3.	Relationship between mean percent cover and annotation accuracies.	43
Figure 3.4.	Accuracy as measured by Cohen’s kappa for the task of discriminating corals from other labels, κ_{coral} at various levels of Alleviation for the four studied locations.	45
Figure 3.5.	Percentage cover estimates of functional groups (coral, macroalgae, CCA, and turf algae) and the three most abundant coral genera in each location as determined by different annotation methods.	46
Figure 3.6.	Differences in percent cover estimates between the Host and Visiting experts as well as the Abundance and Alleviate operational modes and the Archived annotations.	47
Figure 3.7.	Screenshots from CoralNet.	53
Figure 3.8.	Confusion matrices for Moorea, Line Islands, Nanwan Bay and Heron Reef	54
Figure 3.A.1.	Sample photoquadrats drawn from the long-term coral reef projects that served as test data for the present analysis	59
Figure 3.C.1.	Training set size analysis.	62
Figure 3.D.1.	Performance evalutaion of the Refine operational mode.	63

Figure 4.1.	Amount of auxiliary (n_b), and primary (n_a) annotation required to achieve error $d \leq 0.02$ at 95% confidence for $\sigma_p = 0.2$, $\sigma_a = 0.02$, and $\sigma_b = \{0.25\sigma_p, 0.5\sigma_p, \sigma_p\}$ under the hybrid sampling design.	72
Figure 4.2.	Difference in Total Sampling Cost ($\text{TSC}_{\text{conventional}} - \text{TSC}_{\text{hybrid}}$) for $\sigma_p = 0.2$, $\sigma_a = 0.02$, $\sigma_b = \{0.25\sigma_p, 0.5\sigma_p, 0.75\sigma_p\}$, $c_b = 0$ and $c_c = 1$, as a function of the relative cost of primary annotation: $k = c_a/(c_c + c_b)$	74
Figure 4.1.	Sample coral survey images	79
Figure 4.2.	Sample plankton images.	80
Figure 4.3.	Simulations on the coral reef image dataset.	81
Figure 5.1.	Co-located images captured using a custom framer.	95
Figure 5.2.	Three registered image pairs.	96
Figure 5.3.	Histogram of point annotations.....	97
Figure 5.4.	Cropped patch-pairs from the data set.	98
Figure 5.5.	Beta distribution estimation of image channel intensities.	99
Figure 5.6.	Training-set preprocessing procedure for the proposed automated annotation method.	101
Figure 5.1.	Accuracy of the naïve Bayes classifier for different combinations of image channels.....	102
Figure 5.2.	Confusion matrices for the naïve Bayes classifier.....	103
Figure 5.3.	Confusion matrices for the proposed convolutional neural network method.	104
Figure 5.4.	Precision vs. recall curves for f^{REF} , f^{FLR} , and f^{JOINT}	105
Figure 6.1.	Accuracy and alleviation levels of CoralNet	114
Figure 6.2.	Scatterplots of CoralNet accuracy and training time.	114

LIST OF TABLES

Table 1.1.	Thesis data-sets overview.	9
Table 2.1.	Comparison of color spaces and preprocessing methods.....	18
Table 2.2.	Result summary table.	23
Table 3.1.	Summary of image resources.	33
Table 3.2.	Label-set used in this work, and definition provided to human annotators. .	34
Table 3.3.	Tabulated accuracies.	55
Table 3.4.	Tabulated p-values.	56
Table 4.1.	Empirical evaluation of hybrid sampling for two marine ecology data-mining tasks.....	77

ACKNOWLEDGEMENTS

As I conclude this PhD and my formal education I'd like to take a moment to express gratitude towards those who have inspired me and made this possible.

How fortunate I was, first of all, to be born in a time and place which enabled me to spend 11 years at University in pursuit of knowledge and understanding of the natural world. How extraordinary lycky I was to be born in Sweden which offers free higher-education for all, and to then complete my education in the U.S. which welcomed me with open arms and full financial support.

How fortunate I was, also, to be born in to a loving, caring, and nurturing family. How well I remember our dinner-conversations about technology and science where we, as I fondly recall, concluded that the “smack-effect” was indeed the only reason one would want to let go of a ladder should it start falling backwards; and how my father, Peter, took the time to fax his friend to help me sort out a logarithmic equation. I have tried to take the words of my mother, Elisabeth, to heart when she encouraged me to stop spending all my time studying and instead “go out and play”, and her reminders that at the end of the day, it's the relationships we build in life that matter. And no matter how far I am from my brother, Carl, we are always close and his values, ethics, wisdom, and stamina continues to impress and inspire me.

Mallory Pickett, how fortunate I am to have a partner like you to console and encourage tirelessly as the ups and downs of the PhD sometimes brought me to my knees. I love you.

I owe huge gratitude towards my PhD advisor, David Kriegman. David, thank you for taking me under your wings and for your constant and tireless mentorship. Thank you for being a solid wall against which to bounce ideas against, and for giving me freedom to find my own academic path. I think few computer science Professors would, like David did, not only encourage me to become a scientific diver and embark on several field-trips, but also take the time to do so themselves. David is an extraordinary advisor who always keeps the best of his students in mind. He sets exceedingly high academic standards, and has an uncanny way of finding any holes or weaknesses in an argument. I have, through David, learned to push my thinking and to not leave any stones unturned. It has been an honor and a pleasure.

I have been fortunate to work closely with my whole PhD committee, and I'd like to thank the other members: Serge Belongie, for co-advising my early thesis-work and for including me in the exciting work of his research group; Nuno Vasconcelos, for inspiring discussions on machine learning research, and for your excellent course in generative probabilistic models; Sanjoy Dasgupta, for equally inspiring discussions on sampling theory and statistics, for always taking the time to meet and engage, and for your extraordinary teaching skills; and Greg Mitchell for inviting me into the world of ocean exploration and natural science.

I would also not be here today were it not for my MS advisors at Lund University, Karl Åström, Sven Hedlund, and Andres P. Eriksson, as well as countless other teachers and mentors throughout my schooling. Thank you all.

My PhD work was seeded through a NSF grant: Computer Vision Coral Ecology (CVCE), and I owe much gratitude towards the CVCE team. I'd like, in particular to thank Greg Mitchell and David I. Kline for conceiving the project idea so many years ago and for securing the funding. I'd also like to thank the other team members: Ben Neal, for inspiring conversations and for always being there to lend a helping hand; Tali Treibitz for excellent mentorship throughout my PhD and for showing me how to make scientific figures; and Paul Roberts for valuable mentorship and advice. I'd also like to thank other collaborators from the coral ecology community: Pete Edmunds and Vincent Moriarty at California State University Northridge; Matthew Dunlap and Benjamin Richards at NOAA; Ove Hoegh-Guldberg and Manuel González-Rivero at University of Queensland; and Jennifer Smith at Scripps Institution of Oceanography.

In addition, I am thankful to the past and present members of the UCSD Computer Vision Lab, whose advice, friendship, and engaging discussions have been critical to my achievements: Sam Kwak, Tsung-Yi Lin, Catherine Wah, Florian Schroff, Tali Treibitz, Boris Babenko, Steven Branson, Carolina Galleguillos, Piotr Dollár, Manmohan Chandraker, Ana Cristina Murillo, Hani Altwaijry, Mohammad Moghimi, Zachary Murez, Phuc Nguyen, Arturo Flores, Eric Christiansen, Jan Jakeš, Tomas Matera, and Michael Wilber.

This dissertation is based on the following published papers and manuscripts, which, for the most-part, were co-authored with others:

- Chapter 1 is based, in part, on *Towards automated annotation of benthic survey images: variability of human experts and operational modes of automation*, Oscar Beijbom, Peter J. Edmunds, Chris Roelfsema, Jennifer Smith, David I. Kline, Benjamin Neal, Matthew J. Dunlap, Vincent Moriarty, Tung-Yung Fan, Chih-Jui Tan, Stephen Chan, Tali Treibitz, A. Gamst, B. Greg Mitchell, David Kriegman. PLoS ONE, 2015 [9]. The dissertation author was the primary investigator and author of this paper.
- Chapter 2 is a reformatted version of "*Automated Annotation of Coral Reef Survey Images*". Oscar Beijbom, Peter. J. Edmunds, David. I. Kline, B. Greg Mitchell and David Kriegman. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), June 2012. The dissertation author was the primary investigator and author of this paper. © 2012 IEEE.
- Chapter 3 is a reformatted version of "*Towards automated annotation of benthic survey images: variability of human experts and operational modes of automation*", Oscar Beijbom, Peter J. Edmunds, Chris Roelfsema, Jennifer Smith, David I. Kline, Benjamin Neal, Matthew J. Dunlap, Vincent Moriarty, Tung-Yung Fan, Chih-Jui Tan, Stephen Chan, Tali Treibitz, A. Gamst, B. Greg Mitchell, David Kriegman. PLoS ONE, May 2015. The dissertation author was the primary investigator and author of this paper.
- Chapter 4 is a reformatted version of "*Random Sampling in an Age of Automation: ?Minimizing Expenditures through Balanced Collection and Annotation*" by Oscar Beijbom, which is in submission.
- Chapter 5 is a reformatted version of "*Improving Automated Annotation of Coral Reef Images Using Wide-Field-Of-View Fluorescence Imaging*" by Oscar Beijbom, Tali Treibitz, David I. Kline, Gal Eyal, Adi Khen, Ben Neal, Yossi Loya, B. Greg Mitchell and David Kriegman, which is in submission. The dissertation author was the primary investigator and author of this paper.

VITA

2005	Bachelor of Science in Engineering Physics, Lund University
2007	Master of Science in Engineering Physics, Lund University
2012	Master of Science in Computer Science, University of California, San Diego
2015	Doctor of Philosophy, University of California, San Diego

PUBLICATIONS

O. Beijbom, P. J. Edmunds, C. Roelfsema, J. Smith, D. I. Kline, B. Neal, M. J. Dunlap, V. Moriarty, T-Y. Fan, C-J. Tan, S. Chan, T. Treibitz, A. Gamst, B. G. Mitchell, D. Kriegman. *"Towards automated annotation of benthic survey images: variability of human experts and operational modes of automation"*. PLoS ONE, 2015.

O. Beijbom, D. Morris, S. Saponas, N. Joshi. *"Menu-Match: Restaurant-Specific Food Logging from Images"*, Winter Conference on Applications of Computer Vision, 2015.

T. Treibitz, B. P. Neal, D. I. Kline, O. Beijbom, P. L. D. Roberts, B. G. Mitchell, D. Kriegman. *"Wide Field-of-View Fluorescence Imaging of Coral Reefs"*, Scientific reports, 2015.

M. Gonzalez-Rivero, P. Bongaerts, O. Beijbom, O. Pizarro, A. Friedman, A. Rodriguez-Ramirez, B. Upcroft, D. Laffoley, D. Kline, R. Ververs, O. Hoegh-Guldberg. *"The Catlin Seaview Survey - kilometre-scale seascape assessment, and monitoring of coral reef ecosystems"*. Aquatic Conservation: Marine and Freshwater Ecosystems, 2014.

O. Beijbom, M. Saberian, N. Vasconcelos, D. Kriegman. *"Guess Averse Loss Functions For Cost-Sensitive Multiclass Boosting"*. International Conference on Machine Learning (ICML), Beijing, China, 2014.

T. Treibitz, B. P. Neal, D. I. Kline, O. Beijbom, P. L. D. Roberts, B. G. Mitchell, D. Kriegman *"Wide Field-of-View Daytime Fluorescence Imaging of Coral Reefs"*. Journal of the Marine Technology Society, 2013.

S. Branson, O. Beijbom, S. Belongie. *"Efficient Large-Scale Structured Learning"*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, Oregon, 2013.

O. Beijbom, P.J. Edmunds, D.I. Kline, G.B. Mitchell, D. Kriegman. *"Automated Annotation of Coral Reef Survey Images"*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, Rhode Island, 2012.

O. Beijbom. *"Domain Adaptations for Computer Vision Applications"*. Research exam: University of California San Diego, arXiv:1211.4860, April 2012.

O. Beijbom. *"A method, device, and computer program product for event detection while preventing misclassifications"* (W83960002). International filing date: 20.07.2009.

S. Hedlund, O. Beijbom, M. Almers. *"A method for determining an in-focus position and a vision inspection system"* (PCT/EP20008/051993). International filing date: 19.02.2008.

O. Beijbom. *"Single Image Focus Level Assessment Using Support Vector Machines"*. Master's Thesis Report, Lund University, 2007.

O. Beijbom. *"Random Sampling in an Age of Automation: ?Minimizing Expenditures through Balanced Collection and Annotation"*. In submission.

O. Beijbom, T. Treibitz, D. I. Kline, G. Eyal, A. Khen, B. Neal, Y. Loya, B. G. Mitchell, D. Kriegman. *"Improving Automated Annotation of Coral Reef Images Using Wide-Field-Of-View Fluorescence Imaging"*. In submission.

ABSTRACT OF THE DISSERTATION

Automated Annotation of Coral Reef Survey Images

by

Oscar Beijbom

Doctor of Philosophy in Computer Science

University of California, San Diego, 2015

Professor David Kriegman, Chair

The digital age is profoundly changing how data is collected, stored and analyzed in the ecological sciences. Where scientists would previously resort to *in-situ* observations of nature, they are now increasingly relying on digital recording devices to collect large data-sets. Collected data typically requires annotation in order to yield the desired data-products, but unlike collection, annotation often requires expert skills, which are in short supply. As a consequence, a manual-annotation bottleneck has developed in many disciplines between collected data and desired data-products.

Concurrently, significant advances have been made in automated identification of common objects & scenes using machine learning and computer vision methods. This thesis builds on

these advances in order to address the manual-annotation bottleneck in coral reef surveys. The focus on reef surveys was established due to the severity of the manual-annotation bottleneck, and the critical importance of large-scale reef surveys in understanding the contemporary global coral reef crisis.

Towards this goal, we investigate several aspects of the data collection and annotation work. First, methods for automated annotation of coral reef survey images are developed and compared to human expert annotators. Random sampling designs and semi-automated frameworks are proposed which enable deployment of automated annotation without sacrificing the estimation quality of derived data-products (e.g. percent coral cover). Such methods are appealing as they fuse the high accuracy of expert annotators with the speed and low cost of automated annotators. Further, a multi-channel annotation algorithm, which utilizes broad-spectrum fluorescence images in conjunction with standard reflectance images, is proposed and shown to significantly reduce the error-rate compared to using reflectance images only. Finally, the developed methods are implemented in a web-portal, CoralNet (coralnet.ucsd.edu), and made publicly available. At the time of writing, over 80,000 images have been uploaded and annotated by over 200 users, including research labs and government agencies.

We believe automated annotation methods will have a transformative effect on benthic surveys, allowing a better understanding of global reef health and ecology.

Introduction

The digital age is having a profound impact on how scientific data is collected, stored and analyzed. This is true, in particular, for the ecological sciences. Where scientists would previously resort to *in-situ* observations of nature, they are increasingly relying on digital recording devices, such as cameras or microphones [124, 48, 145, 61, 110, 4]. This has several critical benefits: (1) it creates a permanent record, (2) it allows for more effective field deployment, and (3) collection can be done by a lay-person or a device. As a result, scientists now have access to increasingly large data-streams. However, the increasing data-stream often commands a significant data annotation effort. For example, “how much of a coral reef survey image is covered in the coral genera *Acropora*?”, or “how many dolphin individuals are present in the underwater recording?”. Contrary to data collection, this annotation effort often requires expert skills, which is in short supply [80, 33, 128]. As a consequence, there is a “manual-annotation bottleneck” between collected data and the relevant data-products needed for analysis. The goal of this thesis work is, in very general terms, to investigate methods for alleviating this bottleneck.

I focus on one data collection task in particular, namely surveys of coral reefs. This focus was established due to the critical importance of automated annotation for coral reef survey as well as the challenges associated with underwater image analysis.

Due to global and local stressors, coral reefs are in unprecedented global decline [58, 47, 17]. This is problematic since coral reefs provide refuge and habitat for countless marine organisms. A coral reef collapse is also a collapse of the local fishery, and in some cases, a collapse of the local economy [30, 24]. While some global stressors, such as global warming and ocean acidification, are hugely challenging to address, local efforts, such as establishing marine protected areas, and managing fisheries, can go a long way to prevent reef collapse [12]. But

such efforts need to be put in place, monitored, and enforced. This all requires effective survey methods.

Underwater surveys, however, are challenging. Light attenuation in the water column means that aerial and satellite surveys are not always effective for measuring ecosystem changes. Instead, most monitoring efforts rely on *in-situ* surveys [56, 90]. Advances in underwater robotics such as Autonomous Underwater Vehicles [95], Remotely Operated Vehicles [52], and underwater scooters [51], have enabled collection of vast amounts of reef-survey imagery. However, as with other sciences, annotation is hugely time-consuming and expensive. For example, our collaborators at the Moorea Coral Reef Long Term Ecosystem Research program spend three person-months annotating images that are collected in two person-weeks [1]. Other survey efforts, such as the Catlin Seaview Survey collect hundreds of thousands of images, for which manual annotation would require tens of years [51]. The need for automated annotation methods is thus dire. This thesis takes a broad perspective on the problem of automating the annotation work of coral reef survey images, and considers several aspects of the data-collection pipeline.

Chapter 2 discusses the core problem of semantic interpretation of textured images. In this work we developed image descriptors based on color and texture and show how to build an end-to-end recognition pipeline. Our method achieved an accuracy of 70-80% on a dataset of survey image from Moorea, in French Polynesia [8].

The results of Chapter 2 prompted several questions related to whether the 70-80% accuracy is high enough to be of practical use. First: since our method learns from expert annotation, it can not be expected to out-perform the experts themselves. So how accurate are the experts? Second: even if the automated annotation accuracy is lower than that of a human annotator, it may still provide useful data-products, such as the percent coral cover. In other words: what level accuracy is acceptable for the underlying ecological needs? Third: how can we balance the trade-off between an imperfect but cheap algorithm and an accurate but expensive expert annotator. These questions are all discussed in **Chapter 3** for coral reef surveys, where we show that semi-automated annotation performed in the proposed “Alleviate” annotation mode allows for a 50% reduction of the annotation work without losing the estimation quality of the

derived data-products, e.g. percent coral cover [9].

Chapter 4 examines the design of random sampling surveys in the presence of an automated annotation algorithm. Specifically: can the additional noise introduced by using a machine learning algorithm be ameliorated by collecting and annotating more samples? If so, what are the annotation and collection costs under which this is beneficial? We provide an analysis of these questions and define a “hybrid” sampling design, which can reduce total sampling expenditures under precisely-defined requirements on annotation accuracies and sampling costs. As exemplified on coral and plankton surveys, the hybrid sampling design reduced the total sampling costs by 50 - 79% compared to a conventional sampling design [7].

Regardless of whether the automated annotator is deployed in a hybrid sampling design of Chapter 4, or during a semi-automated mode of Chapter 3, the accuracy of the underlying method is critical to maximize the utility. In **Chapter 5** we examine whether fluorescence information can be utilized in conjunction with standard RGB images in order to improve the accuracy of the automated annotator. Our results indicate an error-rate reduction of 22% when utilizing both sources of information, compared to using standard RGB images only, suggesting that it may be beneficial to deploy a fluorescence camera during reef surveys.

Finally, in **Chapter 6** we present CoralNet (coralnet.ucsd.edu), a web-portal which implements the developed methods and makes them available to the coral ecology community. CoralNet enables users to upload their survey images and annotate them directly in the web-browser. At the time of writing this thesis, over 80,000 images has been annotated in CoralNet with over three million expert point annotations. CoralNet users include the Coral Reef Ecosystem Division of the National Oceanic and Atmospheric Administration (NOAA-CRED), the Australian Institute of Marine Science (AIMS), University of Washington, Scripps Institution of Oceanography, University of North Carolina, Chapel-Hill, and many others.

Chapter 1

Background

1.1 Data Collection

Coral reef surveys have undergone several important transitions over the last half decade (Fig. 1.1). Reef surveys were traditionally performed through *in situ* counting by scuba divers skilled in marine ecology. *In situ* counting enable accurate observations, but they are time-consuming and allow only small areas of the reef to be surveyed. Additionally, this method is dependent on the skill of the experts conducting the surveys, and the data cannot be reliably returned to for reanalysis. *In situ* counting of corals was largely replaced by image-based surveys when high quality underwater cameras became available at an economic price in the early 1960's [62], coincident with the proliferation of scuba as a tool for underwater research. Images-based techniques are advantageous for underwater applications, because they maximize data collection in the limited underwater time possible with scuba-based field deployments [39], and they provide permanent records that can be analyzed for organism abundance, size, and percentage cover. However, benthic photographs pose significant challenges to quantification because the organisms are seldom well defined in color, outline, or texture, and the process of underwater photography is prone to the vagaries of underwater conditions (e.g., surge, turbidity, light penetration, spectral truncation, etc.), equipment limitations in the corrosive seawater environment, and photographers with variable experience. Thus, underwater images often are dominated by blue colors, inadequate illumination, and deep shadows, and together these can make the analysis of complex images such as those found on tropical coral reefs difficult.

Since the turn of the new millennium, the application of image-based tools to underwater research has experienced rapid and dramatic changes in three domains. First, improvements in digital photography enable gathering large numbers of images at increasingly high resolution, thus vastly increasing the capacity to census marine environments. Second, the capabilities of computers and software to manipulate, store, and analyze images has increased by orders of magnitude. Third, advances in robotics and control theory have enabled construction of autonomous underwater vehicles (AUVs), remotely operated vehicles (ROVs), and imaging sleds that can capture thousands of survey images in a single deployment [39, 124, 95].

However, the capacity to analyze the collected images has not advanced at a pace commensurate with the capacity to collect them. This has created a ‘manual-annotation’ bottleneck between the rapid image acquisition and the quantitative data needed for ecological analysis.

1.2 Point Annotations

Generating quantitative ecological information from underwater images typically involves random point annotation to estimate the percent cover of the substrata of interest. To generate these data, substratum types are identified, usually manually by an expert, for a number of randomly selected locations in each image (Fig. 1.2). Point annotations are typically preformed using manual, off-line annotation software like Coral Point Count with Excel Extensions (CPCe) [70] or photoQuad [135] which facilitate the annotation process by providing user interfaces and tools for export of relevant data. More recently, web-based annotation tools such as Biigle (biigle.de) and Catami (catami.org) have been developed, however these also require manual annotation, with consequent extensive time requirements for analysis. Another tool, Squidle (squidle.acfr.usyd.edu.au), which aims to use machine learning methods to facilitate annotation, is under development at the University of Sydney.

While this thesis focus on the automated annotation of point annotations, there are two other principal annotation paradigms: full image classification and image segmentation. Full image classification is commonly used for benthic habitat mapping from AUV and ROV surveys [121], and as ground truth for satellite imagery [113]. The purpose of such surveys is

Figure 1.1. Evolution of methods for benthic cover estimation. The y-axis represents a qualitative comparison of required time per collected piece of information, for various improvements in survey technology. (A) *in-situ* inspection of the benthos is time-consuming during deployment but the desired data-product is available without much post-processing in lab. (B) photographic surveys enable more effective field-deployment, however significant time is required to annotated the image with desired meta-data, such as coral cover. (C) Software improvements such as Coral Point Count, help facilitate the logistics of the manual annotation process. (D) Improvements in automated collections technologies, by e.g. autonomous underwater vehicles, have further reduced the collection time, but the collected images still require a time-consuming manual annotation step. (E) Automated annotation methods, which is the focus of this thesis, can address the manual-annotation bottleneck between collected images and desired data-product.

to determine the large-scale structures of the reef-scape, and create spatial maps. Full image segmentation, on the other hand, requires each pixel to be mapped to a semantic label. This is the most detailed type of annotation, and is used when detailed population statistics are desired (e.g. a histogram of coral colony sizes). This, in turn, typically require high-resolution image mosaics to be collected [40, 123], since single photo-quadrants generally are too small to fully contain coral colonies.

Random point annotations can thus be considered a middle-ground which is utilized when cover statistics of main benthic categories are required, but not the full population-statistics. It is the most commonly used annotation paradigm, and the sole focus of this thesis. We recommend

the excellent thesis by Ariell Friedman [46], and Timm Schoening [120] for automation of the other annotation paradigms.

Figure 1.2. Sample point-annotations of an image from the Moorea Coral Reef Long Term Ecological Research monitoring program in French Polynesia. The point locations are randomly scattered on the image plane, and then annotated by an expert annotator. Between 10 and 200 points are typically used per image.

1.3 Computer Vision Methods For Automated Annotation

The timing of this thesis coincided with a time of great excitement in the field of computer vision, in particular with respect to semantic interpretation of images. During my early research the bag-of-visual-words (BOVW) paradigm [125] was the dominant recognition pipeline, and was used for tasks such as object recognition [44], object detection [42], and textures [35]. In the BOVW paradigm, one or many base-features are first extracted: color [69], histogram of oriented gradients (HOG) [34], scale-invariant feature transforms (SIFT) [79], local binary patterns (LBP) [98], and filter responses from the MR8 filter bank [141], each capturing some aspect of the visual space. The base-features were then encoded to “visual words” using some vector quantization step, such as k-means clustering. Finally, image descriptors were extracted by pooling the visual words across some region in the image. One popular pooling scheme is spatial pyramid matching [74], in which the visual words are pooled in multiple overlapping image regions to create a loose encoding of the large-scale image-structures. We utilized the BOVW

paradigm in the automated annotation method of Chapter 2, and introduced a rotational-invariant pooling scheme.

More recently, the BOVW paradigm has been made all-but-obsolete by deep convolutional neural networks (CNNs). The superiority of CNNs was established in 2012 by Krizhevsky et al. [71], in their seminal paper that vastly improved the state of the art performance on a large-scale image classification dataset, ImageNet (www.image-net.org). CNNs have received considerable attention the last years, and continue to show great promise on a wide array of semantic image recognition tasks [63, 50]. We utilized CNNs in Chapter 5 for classification of registered fluorescence and reflectance image-pairs, and I believe these methods hold great promise for improving methods for automated annotation of coral reef survey images in general.

1.4 Computer Vision Coral Ecology Data-sets

As this thesis work has progressed, several data-sets have been gathered and packaged to enabled bench-marking and development of new analytical methods. The data-sets are listed in Table 1.1.

Acknowledgments

Chapter 1 is based, in part, on *Towards automated annotation of benthic survey images: variability of human experts and operational modes of automation*, Oscar Beijbom, Peter J. Edmunds, Chris Roelfsema, Jennifer Smith, David I. Kline, Benjamin Neal, Matthew J. Dunlap, Vincent Moriarty, Tung-Yung Fan, Chih-Jui Tan, Stephen Chan, Tali Treibitz, A. Gamst, B. Greg Mitchell, David Kriegman. PLoS ONE, 2015 [9]. The dissertation author was the primary investigator and author of this paper.

Table 1.1. Thesis data-sets overview. Data-set collected, annotated, and analyzed during this thesis-work. All images are captured < 1 meter from the ocean floor, covers $0.12\text{--}0.38\text{ m}^2$, and are annotated at 10-200 point locations by expert annotators. All data-sets can be accessed at <http://vision.ucsd.edu/data>.

	Moorea Labeled Corals	Pacific Labeled Corals	Eilat Fluorescence Corals
Characteristic	Multi-year.	Multi-reef & multi annotator.	Multi-modal imagery.
Nbr. images	2,055	5,090	212
Nbr. total point annotations	393,299	251,988	42,400
Nbr. sets of expert annotations	1	training-set:1, test-set:7	1
Spatial layout	Six sites around Moorea, French Polynesia.	Four Pacific coral reefs: Moorea, French Polynesia; Heron Reef, Australia; Nanwan Bay, Taiwan & northern Line-Islands.	One site in the shallow reefs of Eilat, Israel.
Temporal layout	Three replicates between 2008 and 2010.	One replicate for each reef captured between 2005 and 2012.	One replicate captured during 2015.
Camera setup	Digital SLR with two strobes.	Digital SLF with mixed artificial lighting.	Registered reflectance and fluorescence image-pairs.
Used in	Chapter 2, Chapter 4	Chapter 3	Chapter 5, Chapter 7

Chapter 2

Texture and color descriptors

Chapter 2 is a reformatted version of "*Automated Annotation of Coral Reef Survey Images*" [8]. Oscar Beijbom, Peter. J. Edmunds, David. I. Kline, B. Greg Mitchell and David Kriegman. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), June 2012.

This chapter describes our work on developing color and texture descriptors for automated annotation of coral reef survey images. The descriptors are a corner stone of this thesis work, and they are utilized in all subsequent chapters, as well as in CoralNet (coralnet.ucsd.edu).

Figure 2.1. Example images from the Moorea Labeled Corals dataset. Top row images are from the *Outer 10m reef* habitat, bottom from the *fringing reef*. Superimposed on each image is a subset of the ground truth annotations (smaller symbols) and the estimated classifications by the proposed algorithm (larger symbols). Circles represent coral genera *Acropora*, *Pavona*, *Montipora*, *Pocillopora*, *Porites* and triangles are non-coral substrates, *Crustose Coralline Algae*, *Turf algae*, *Macroalgae* and *Sand*. Note the organic borders between the substrates and complex class morphologies. The white transect line going through the images is part of the sampling methodology, as is the metal frame seen along the edges. This figure is best viewed in color.

Abstract

With the proliferation of digital cameras and automatic acquisition systems, scientists can acquire vast numbers of images for quantitative analysis. However, much image analysis is conducted manually, which is both time consuming and prone to error. As a result, valuable scientific data from many domains sit dormant in image libraries awaiting annotation. This work addresses one such domain: coral reef coverage estimation. In this setting, the goal, as defined by coral reef ecologists, is to determine the percentage of the reef surface covered by rock, sand, algae, and corals; it is often desirable to resolve these taxa at the genus level or below. This is challenging since the data exhibit significant within class variation, the borders between classes are complex, and the viewpoints and image quality vary. We introduce Moorea Labeled Corals, a large multiyear dataset with 400,000 expert annotations, to the computer vision community, and argue that this type of ecological data provides an excellent opportunity for performance benchmarking. We also propose a novel algorithm using texture and color descriptors over multiple scales that outperforms commonly used techniques from the texture classification literature. We show that the proposed algorithm accurately estimates coral coverage across locations and years, thereby taking a significant step towards reliable automated coral reef image annotation.

2.1 Introduction

In many scientific disciplines experts routinely analyze large quantities of image data. However, not only has the capacity for acquiring digital images vastly outpaced the resources to manually annotate those images, but there are also issues with lack of consistency and objectivity in human labeling [80, 33]. One such domain is coral reef ecology, which is particularly important given the crucial ecological roles of coral reefs, and their current state of decline in health and abundance [114]. To understand this development, ecologists need accurate and large-scale coral reef coverage data. As satellite images are ineffective for this purpose [57] and low altitude photography suffers from problems such as surface effects, coral ecologists commonly do *in situ* studies. Recent innovations in image acquisition techniques, such as autonomous underwater vehicles [106] and towed diver sleds [67] have greatly increased the number of images available and researchers, reef managers, and government agencies such as the National Oceanic and Atmospheric Administration (NOAA), routinely collect millions of images per year as part of reef surveys. However, the lack of automated annotation methods effectively caps the monitoring effort since manual annotation is slow, tedious, and expert time is limited; anecdotally, as little as 1-2 % of captured images gets annotated.

Concurrently, there have been great advances in recognition of specific objects, object classes, and scenes on widely used benchmark data sets (e.g., Caltech 101 [44], Pascal [42]), and there is intense competition within the research community to develop new methods. Yet recognition in much of the biological image data introduces new challenges. As noted by Adelson [2], rather than trying to recognize *things* as found in most of the object recognition datasets, biological datasets often require identifying *stuff*. In the case of coral reef image analysis the size, color, shape, and texture of each of the classes may vary significantly and the boundaries are often organic and ambiguous. Further, water turbidity can change dramatically between sites and years due to currents, plankton, algal blooms, etc., affecting ambient light and image colors. We argue that this type of data pose challenges quite different from those posed by the above mentioned datasets, and also from benchmarks commonly used for texture based recognition, Brodatz [109] and CURET [35].

Due to the above mentioned challenges, standard annotation techniques such as image labels, bounding boxes or full segmentations are inappropriate. A full manual segmentation would be too time consuming to collect and neither bounding boxes nor gross image labels provide the level of detail needed. Instead, coral ecologists commonly rely on random point sampling. In this process images are annotated by asking an expert to identify at a specific point (pixel) in the image, which class is present (most prevalent) at that single location. The process is repeated for a fixed number of randomly chosen locations in the image (typically 10-200) and often facilitated by software such as Coral Point Count [70]. Once annotated, coverage statistics can be calculated over sets of images. A typical survey might include 1000 images with 200,000 annotations requiring >400 person hours to label. These expert annotated images are a bonanza for computer vision researchers – a rich source of labeled and relevant data.

2.1.1 Related work

In a recent effort to automate coral classification, Stokes and Deane [129] devised a method that utilizes the discrete cosine transform and a k-nearest neighbor classifier to do patch-wise classification of benthic substrates. Their work shows promising results, but their small dataset (16 images) makes it hard to draw conclusions on scalability. Other work includes [84, 122], that use texture and color descriptors, but presume the availability of patches hand cropped from the image data.

There is a rich literature on object recognition and texture-based classification in the computer vision community. Dana et al. created the CURET database in 1999, and this became a popular benchmark for texture recognition [35]. Malik et al. [81] made an important contribution by introducing the texton as a cluster center of filter responses, and later Cula and Dana [31], were able to do robust classification from a single image. Varma and Zisserman further improved the CURET benchmarks by introducing the rotational invariant Maximum Response (MR) filter bank and reported classification accuracies of up to 98.61 % [139]. Some years later Varma and Zisserman proposed the use of patch exemplars instead of filter responses [140]. They report a minor improvement on the CURET database, but in general the classification scores seem

Figure 2.2. Sample patches from the Moorea Labeled Corals dataset. Each class is represented by two example patches. First row: *Acropora*, *Porites*, *Montipora*. Second row: *Pocillopora*, *Pavona*, Macroalgae. Third row: Sand, Turf algae, CCA. Each class exhibits great variability with respect to growth form, color, scale and viewing angle. For example, note the two different growth forms of *Acropora* and *Porites*, and the difference in color and scale of the *Pocillopora* patches. Macroalgae varies tremendously in shape and color, and often projects from underneath the corals. Both CCA and Turf algae tend to overgrow dead coral, which pose a challenge for automated analysis since the coral skeleton retains its shape, but now classifies as the the algae that overgrows it. Also, CCA and Turf are similar and are hard to discriminate from an RGB photography. Finally, we point out that the images in this grid has been hand picked to clearly exemplify each class. A random selection of patches would include a large amount of patches that straddles class boundaries, and where class designation is less clear.

to have saturated in the high nineties. One of the limitations of CURET is that all objects are cropped and each image contains only one type of texture. Winn et al. [146] introduced the MSRC dataset in 2005 as an alternative benchmark for texture based methods, but this suffers from other problems as pointed out by Torralba and Efros [132]. In contrast to CURET, MSRC, and most other datasets used for performance benchmarks, the Moorea Labeled Corals (MLC) dataset was *not* created explicitly for the purpose of developing computer vision methods. We believe this provides a better opportunity for benchmarking, and indicates a direction for future computer vision challenges.

2.1.2 Contributions

The contribution of this paper is twofold: (1) We introduce the analysis of coral reef survey images as a relevant and important application domain to the computer vision community and propose a benchmark dataset. (2) We propose a method based on texture and color descriptors over multiple scales that outperforms standard texture classification methods and establishes a strong baseline on the MLC dataset. This is, to the best of our knowledge, the first paper to address automated annotations of coral reef survey image on a large scale.

2.2 Dataset

The Moorea Coral Reef Long Term Ecological Research (MCR LTER) project has been collecting image data from the island of Moorea (French Polynesia) since 2005. This project monitors six sites around the island, and four habitats at each site. Each year, transect lines are attached between points permanently anchored in the reef. Then, divers move along the transect line, acquiring images using a digital SLR equipped with underwater strobes. The camera is attached to a frame (50 cm by 50 cm) to keep the distance from the bottom constant and the camera orientation parallel to the reef surface. Sample images are provided in Fig. 2.1 where the frame is seen along the image edges, and the white transect lines pass through the middle. The orange lines are attached to the frame to help the diver align the center of the frame with the target. Even with this careful method of acquisition there are large differences in image appearance as mentioned in Section 2.1. Across years things become even more challenging as the ecosystem changes. On this reef an outbreak of Crown of Thorns Sea Star (a predatory sea star) in 2006, followed by Cyclone Oli in February 2010, together destroyed the majority of the corals. Data from 2011 show signs of rapid regrowth of the reef through recruitment. All images are labeled using the random point sampling method mentioned in Section 2.1. In this dataset 200 points were laid over on each image at random locations and then labeled by an expert according to an established taxonomy (~ 30 categories) for the site.

2.2.1 Moorea Labeled Corals

The MLC dataset is a subset of the MCR LTER packaged for computer vision research. It contains 2055 images from three habitats: *fringing reef*, *outer 10m* and *outer 17m*, from 2008, 2009 and 2010. It also contains random point annotation (row, col, label) for the nine most abundant labels, four non coral labels: (1) Crustose Coralline Algae (CCA), (2) Turf algae, (3) Macroalgae and (4) Sand, and five coral genera: (5) *Acropora*, (6) *Pavona*, (7) *Montipora*, (8) *Pocillopora*, and (9) *Porites*. These nine classes account for 96% of the annotations and total to almost 400,000 points. There is a large variation in the number of samples from each class, Fig. 2.5. The MLC dataset is available at (<http://vision.ucsd.edu/data>).

2.2.2 Labeling Ambiguity

The annotations in this settings are not ‘ground truth’ in the common sense, meaning an absolute truth. Culverhouse et al. studied this phenomena for plankton classification and reported surprisingly weak inter and intra operator consistency rates [33]. The authors list several reasons for this: short term human memory, fatigue and boredom, recency effects (bias towards labels recently used) and positivity bias (bias towards what one expects to see) [33]. Further, the labels may be noisy for reasons outside the control of the human operator. For example, in the present case, there might be insufficient signal in an RGB image to discriminate Turf algae and CCA. Ninio et al. investigated some of these effects for underwater video images of the Great Barrier Reef. They report accuracies of 96% for hard corals, 92.5% for soft corals and 80.6% for algae when compared to *in situ* observations. They also report on inter- and intra-operator variabilities but only on aggregated coverage statistics [96]. Carleton & Done observed a “far inferior” ability to discriminate among corals using video transects compared to *in situ* observations [21]. In general there is an acknowledgement within the community that errors exists in distinguishing among benthic groups and that it can be severe in a few cases, for example discerning between Turf algae and CCA. This notion is supported by a recent (unpublished) NOAA study comparing the annotations of a benthic expert to those of moderately trained operators. Their data indicate 97% agreement between the operators and the expert on the binary task of discriminating between

corals and non-corals, 85% agreement on the task of discriminating between the 5 most abundant hard corals, and 79% agreement when discriminating between Macroalgae, Turf algae, CCA, and Sand [49]. We are currently studying these effects for four Pacific datasets including the MLC dataset.

2.3 Method

In this section we describe the proposed method. As the objects we aim to classify are small, lack clear boundaries and a clear sense of shape, we represent them using texture descriptors. Varma and Zisserman [139] proposed a method where textures are modeled as histograms of textons [77], and introduce a rotational invariant filter bank, Maximum Response (MR). Their method is straight forward and achieves accuracy of $>95\%$ on the CURET database [35]. Driven by difficulties in the problem domain, the proposed method extends [139] in two ways: the use of multiple scales and the incorporation of color information. As shown in Fig. 2.3, and further discussed in Section 2.3.3, there is no straightforward way to choose a fixed patch size over which to integrate image information, which led us to a data-driven, multiple patch approach. The color extension was introduced to encode important color information. However, this is not trivial as colors are corrupted under water, e.g., the red colors are attenuated as a function of depth [23]. We discuss this further in Section 2.3.2.

2.3.1 Preprocessing

All images are first downsized by a factor of 2. We then apply a method, *ColorChannelStretch*, which, for each color channel individually, finds the 1% and 99% intensity percentile, subtracts the lower from all intensities in that channel (and sets any negative pixel values to 0), and then divides by the upper. This method slightly modifies the image colors, and was found to be empirically superior to *IntensityStretch* which stretches the global image intensities across all channels, Table 2.1.

Table 2.1. Comparison of color spaces and preprocessing methods. In each cell, the percent of correctly predicted points are shown for: first row, the 2008 $\Rightarrow$ 2008, experiment; second row, the 2008 $\Rightarrow$ 2009 experiment; third row, the 2008 + 2009 $\Rightarrow$ 2010 experiment; fourth row, the average across all experiment (see Section 2.4 for experimental details). The best score for each experiment is represented in bold, and the lower right cell represent base-line performance. While the differences are subtle and noisy, we see that the L*a*b* color space and *ColorChannelStretch* works best, in particular for the across year experiments. By stretching the image histogram we compensate for differences in water turbidity and lighting. The comparison shows that color information is useful even though absolute color references are unavailable.

	L*a*b*	HSV	RGB	Gray
Intensity - Stretch	74.6%	65.3%	71.8%	69.5%
	63.9%	58.9%	64.3%	63.6%
	80.1%	79.1%	78.2%	76.4%
	[72.9%]	[67.8%]	[71.4%]	[69.8%]
Color - Channel - Stretch	74.7%	72.6%	72.5%	70.1%
	67.4%	64.9%	66.0%	64.5%
	80.1%	80.6%	80.8%	79.8%
	[74.1%]	[72.2%]	[73.1%]	[71.1%]
None	75.1%	73.2%	72.7%	70.4%
	64.1%	59.3%	65.8%	63.9%
	81.3%	82.3%	80.8%	79.2%
	[73.5%]	[71.6%]	[73.1%]	[71.2%]

2.3.2 Texture and Color

We use the Maximum Response (MR) filter bank introduced in [139] which encodes rotational invariance by first filtering with bar and edge filters at different orientations and then outputting the maximum over the orientations. It also contains a circular Gaussian and Laplacian filter. By cross validating over different sizes we arrive at bar and edge filters with standard deviations of 1, 3 and 8 pixels along the short dimension, and circular filters standard deviation of 6 pixels, thus producing an 8 dimensional filter output vector. The filter sizes correspond to 0.5, 1.5, 4 and 3 mm actual distance on the reef surface which is small enough to capture the polyps characterizing the different coral types, along with the finer coral structures. After filtering, following [81, 85], we apply a contrast normalization,

$$\mathbf{F}(\mathbf{x}) \leftarrow \mathbf{F}(\mathbf{x})[\log(1 + L(x)/0.03)]/L(x) \quad (2.1)$$

Figure 2.3. Illustration of the patch size selection problem. Four different patches from the dataset, each with a region of 61 by 61 pixels indicated with a red box. In top left, the signal from the small algal patch might be overwhelmed by the surroundings if the patch is too big. In top right, on the other hand, a small patch would not capture the larger textures. Also, in bottom left, the histogram of visual words would be more stable if a larger area is used. Bottom right show a case where the point of interest is on the edge between classes, in which case any patch size will be problematic.

where $L(x) = \|\mathbf{F}(\mathbf{x})\|_2$ is the magnitude of the filter response vector, $\mathbf{F}(\mathbf{x})$ at pixel x .

Color

The method as presented in [139] does not give special care to color. While colors are not generally a robust source of information [66], particularly underwater with artificial light sources [134], we show that color information is useful for this task. We encode color information by applying the filters to each color channel in the $L^*a^*b^*$ color space and then stack the filter response vectors. Ideally, when using color descriptors for discrimination, color correction should be performed. We were however unable to do this as: (1) there was no reliable color references in the images, (2) there is a deficiency in the literature of methods for underwater color correction. This issue was left to future work, and we settle for intensity normalization by histogram stretching as discussed in Section 2.3.1. An empirical comparison on the end-to-end system shows that *ColorChannelStretch* and $L^*a^*b^*$ yield the most consistent results across years (Table 2.1). It seems that by stretching the image histogram we are able to compensate for some differences in water turbidity and lighting.

2.3.3 Descriptors

To create the texton maps, a dictionary of textons is required. For this purpose we select a subset of the 2008 images that were not used in any test sets. Filter responses from each

Figure 2.4. Importance of using appropriate patch-size for recognition. End to end performance result using different patch sizes (ps). The $ps = multi$ results are generated using all four integration areas as described in Section 2.3.3. For size reference, note that the patches in Fig. 2.2 are all 221 by 221 pixels. All runs are evaluated on the same test set comprising 200 images from 2008. The training data comprises subsets of the remaining images from 2008. Note how the single patch size performance levels out at $ps = 61$. From this experiment it is evident that by concatenating the histogram counts over multiple scales, a more discriminative descriptor is generated.

of the nine classes were separately aggregated across the images, and k-means [55] clustering with 15 cluster centers was applied to each set of filter responses. Finally the cluster centers, or *textons*, from the different classes were merged to create the dictionary with 135 words, each 24 dimensions.

To extract the texture descriptor, we first apply the filters over the whole image. This yields a 24-dimensional feature vector for each pixel in the image. The filter responses are then mapped to the texton with smallest 2-norm distance, creating an integer valued textonmap. The actual feature vector, or descriptor, is set to be the normalized histogram of textons around a patch of interest. As illustrated in Fig. 2.3, the selection of an appropriate patch size is not trivial. We evaluate the performance across four different patch sizes (21, 61, 121 and 221) pixels, and also using a concatenation of the descriptors (bin counts) across all scales. The evaluation shows that the multi-scale approach is superior to any of the single patch sizes. This is illustrated in Fig. 2.4 where we sweep over training set sizes. The complete image processing pipeline is summarized in Algorithm 1 where I is the input image, MR the filter bank, D the dictionary of textons, A the row and column index of the annotated image points (with a total of n points), and T the texton map.

2.3.4 Machine Learning

The LIBSVM implementation of Support Vector Machines (SVM), with Radial Basis Function kernel were used throughout this work [29, 25]. For each experiment the RBF-SVM training step is preceded by 4-fold cross validation on the training data, where parameters γ and C are optimized by a logarithmic grid search over values $\log(\gamma, C) \in \{-5, -4, \dots, 4, 5\}$. The training data was subsampled so that a maximum of 15000 samples from each class is represented; the subsampled classes get assigned a weight inversely proportional to the subsample rate.

Algorithm 1. Extract Descriptors(I, MR, D, A)

```

1:  $I \leftarrow \text{COLORCHANNELSTRETCH}(I)$ 
2:  $I \leftarrow \text{CONVERTTOLAB}(I)$ 
3:  $F \leftarrow \text{APPLYFILTERS}(I, MR)$ 
4: for each pixel,  $p \in I$  do
5:    $T_p = \arg \min_w ||F_p - D_w||_2$ 
6: end for
7: for  $i = 1$  to  $n$  do
8:    $D_i \leftarrow \text{MULTISCALEHISTOGRAMS}(T, A_n)$ 
9: end for

```

2.4 Experiments

We evaluate end-to-end system performance by running three experiments. In 2008 $\Rightarrow$ 2008, we trained on two thirds of the images from 2008 and tested on the last third. In 2008 $\Rightarrow$ 2009 and 2008 + 2009 $\Rightarrow$ 2010 we evaluate across-year performance. The last two experiments are of particular interest for this application as they explore the opportunity to let the machine annotate all data from a certain year. They are also challenging since the appearance of the reef changes between years as discussed in Section 2.2. The training and testing data are created by the procedure described in Algorithm 1, and then run through the SVM. Precision is determined as the ratio of correctly classified patches in relation to the expert annotation. We also considered using the f-score, or harmonic mean which is often used for multi class classification performance evaluation, but this was too noisy due to the small number of samples in some of the classes. In addition we evaluate performance by comparing coverage statistics extracted from the estimated

annotations to those extracted from the expert annotations. Percent cover is calculated as the ratio of annotations of a certain type compared to the total number of annotations.

2.4.1 Patchwise Results

Table 2.2 shows results for the full experiment, and three important sub classification tasks. It also shows random classification (based on the prior probabilities) scores as these get quite large in some cases due to the abundance of CCA. Best performance on the full labelset is achieved on the within year experiment (74%). For the second experiment, $2008 \Rightarrow 2009$, performance drops to 67%. This illustrates the difficulties encountered when classifying across years, and is present in all across-year classification tasks. One fundamental problem is that the data does not comply with the IID assumption of the learning algorithm due to differences in reef appearance. The third experiment, $2008 + 2009 \Rightarrow 2010$, shows an apparent gain in accuracy (83%) but much of this can be attributed to the increase in the random classification baseline. On the sub-classification tasks we note the excellent ability of this method to discriminate corals vs. non-corals: 95% (random assignment achieve up to 75%). Even stronger results are achieved on the coral discrimination task (given that we know it's a coral), with up to 97% accuracy. The non-corals discrimination task is more challenging and hence shows more confusion, in particular among the algae classes. Figure 2.5 shows full confusion matrices for each experiment. Note how the main source of confusion is between CCA and each of the other classes respectively. Part of this confusion is likely because CCA overgrows dead coral; while the coral structure retains its shape it now classifies as CCA.

2.4.2 Coverage Results

Keeping the application domain in mind we also consider how well the coverage statistics can be estimated. For this purpose we generated scatter plots showing, for each category, the coverage based on the computer generated annotations compared to the coverage based on the expert annotations for the $2008 + 2009 \Rightarrow 2010$ experiment (Fig. 2.6). The first sub-plot shows coral coverage, where we consider the binary coral vs. non-coral scenario. The slope coefficient, $p1 = 1.01$, indicates a very robust ability to estimate coral coverage even for this challenging

Figure 2.5. Confusion matrices between automated and human annotations. The rows correspond to the true class assignments, the columns to the predicted assignments. A perfect performance would generate a matrix with ones on the diagonal, and zeros elsewhere. The numbers on the right side show how many samples are in the test set for each category. Note that some of the coral classes are rare when compared to the larger CCA class. The main source of confusion is between CCA and each of the other classes respectively, while there is little confusion between the coral classes

across-years experiment. This is important since many surveys, in particular those carried out over large spatio-temporal scales by reef managers and government agencies, often focus on this one metric (coral coverage) as an indicator of reef health. Coverage of individual coral genera *Porites*, *Pocillopora*, *Acropora* and *Pavona* are also accurately estimated while *Montipora* is underestimated. This might be related to the encrusting morphology of this genus which makes it less visually distinctive than the branching corals. The algal plots exhibit more variation.

Table 2.2. Result summary table. For each experiment we report the accuracy on the full 9-class experiment, as well as three key sub problems. The first ‘coral vs. non-corals’, considers the binary classification task where all corals are grouped together against non-corals. The second, ‘within corals’, considers the situation where we know a certain point is a coral, and we need to assign a genus level label. The third, ‘within non-corals’, consider the situation where we know a certain point is a non-coral and we need to assign the correct label (CCA, Turf algae, Macroalgae, Sand). The numbers in parenthesis show the accuracy by random assignment of labels according to the priors.

	2008 ⇒ 2008	2008 ⇒ 2009	2008, 2009 ⇒ 2010
Full	0.74 (0.28)	0.67 (0.29)	0.83 (0.39)
Coral, non-corals	0.92 (0.66)	0.93 (0.70)	0.95 (0.75)
Within corals	0.97 (0.35)	0.91 (0.37)	0.97 (0.43)
Within non-corals	0.78 (0.43)	0.70 (0.42)	0.87 (0.52)

2.5 Discussion

As discussed in Section 2.2, the method accuracy must be evaluated with respect to the inconsistent nature of the ‘ground truth’ annotations. Compared to the CRED report discussed in Section 2.2 the results in Table 2.2 seem very encouraging indeed, but since the CRED effort analyzed a different reef, used a different label set and used non-expert annotators, one must regard this comparison with caution. Again, we emphasize the importance of a large inter-operator variability study to better understand ‘how good is good’ for this task. Other future work will include a more careful evaluation of the challenges posed by color correction under water, as well as an appropriate treatment of the non-IID data situation.

2.6 Conclusion

We introduce coral reef image annotation as a relevant and important application domain to the computer vision community and propose a benchmark dataset. We argue that an effective treatment of this problem could have an impact on many disciplines such as landscape surveying and remote sensing and that our data set provides an excellent opportunity for the evaluation of computer vision methods. We also propose a novel method for object recognition in this setting that offers a strong baseline on the MLC dataset. This method handles ambiguous and ‘organic’ object boundaries by a multiple scale approach, which is superior to using any single patch size, and we show that careful treatment of color information further boosts performance. The proposed method yields 83.1% accuracy on the MLC 2008 + 2009 $\Rightarrow$ 2010 experiment. We also show that the proposed method accurately estimates coral coverage across reef sites and multiple years, which offers exciting potential for large scale coral reef analysis.

Acknowledgments

The authors like to thanks Vincent Moriarty at California State University Northridge, for the annotation work that made this study possible, and Tali Treibitz at University of California, San Diego for valuable advice on the treatment of color information. This work was partly

Figure 2.6. Scatter plots of benthic coverage estimation. for the $2008 + 2009 \Rightarrow 2010$ experiment. Each dot indicates the ground truth and the estimated coverages respectively per image. For each label type we performed linear interpolation (constrained to zero intersect) and report slope $p1$, and the Pearson product-moment correlation coefficient, ρ . A perfect estimation would generate numbers $p1 = 1, \rho = 1$. The first plot is an aggregated plot showing the coverage estimation for the binary coral vs. non-coral classification task.

supported by NSF Computer Vision Coral Ecology grant #ATM-0941760. The image data was collected, curated, and analyzed with funding from grants OCE 04-17413 and OCE 10-26851 from the National Science Foundation and gifts from the Gordon and Betty Moore Foundation. This is a contribution of the Moorea Coral Reef (MCR) LTER, and is contribution number: 179 of the CSUN Marine Biology Program.

Chapter 2 is a reformatted version of "*Automated Annotation of Coral Reef Survey Images*". Oscar Beijbom, Peter. J. Edmunds, David. I. Kline, B. Greg Mitchell and David Kriegman. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), June 2012. The dissertation author was the primary investigator and author of this paper. © 2012 IEEE.

Chapter 3

Variability of Human Experts and Operational Modes of Automation

Chapter 3 is a reformatted version of "*Towards automated annotation of benthic survey images: variability of human experts and operational modes of automation*", Oscar Beijbom, Peter J. Edmunds, Chris Roelfsema, Jennifer Smith, David I. Kline, Benjamin Neal, Matthew J. Dunlap, Vincent Moriarty, Tung-Yung Fan, Chih-Jui Tan, Stephen Chan, Tali Treibitz, A. Gamst, B. Greg Mitchell, David Kriegman. PLoS ONE, May 2015.

This chapter investigates two questions rising from the conclusion of Chapter 2. First, it measures the variability of human annotators in order to determine how well we can expect an automated annotator to perform. Second, it proposes three ways in which a cheap but less-than-perfect automated annotator can be used in conjunction with a human annotator. Such hybrid approaches can reduce the manual annotation burden while providing high quality annotations.

Abstract

Global climate change and anthropogenic stressors have sharpened the need to characterize ecological changes in marine benthic communities across large scales of space and time. Digital photography and robotics enable rapid collection of digital survey images to meet this need, but the subsequent image annotation remains largely a time consuming, manual task. We investigated the feasibility of using automated point-annotation to expedite cover estimation of key taxa from images of coral reefs in four Pacific locations. Intra- and inter- annotator variability among six human experts was quantified and compared to two proposed operational modes of automated annotation. Our results show that cover estimation errors using 50% automation were similar to those arising from intra-annotator errors of the human experts, but lower than inter-annotator errors. Further, fully automated annotations yielded unbiased cover estimates but with increased standard errors. These results show that automated annotation can effectively be utilized for image-based reef surveys to increase spatial coverage and decrease time and financial outlay. Our method is made publicly available at coralnet.ucsd.edu.

3.1 Introduction

Coral reefs provide habitat to a wide diversity of organisms, and substantial economic and cultural benefits to coastal communities [30, 24]. These functions are threatened by global declines in coral cover caused by a wide diversity of natural and anthropogenic disturbances including global climate change and ocean acidification [58]. The decline in coral cover has been dramatic, with $> 80\%$ decrease in the Caribbean over the last five decades [47], and $\sim 1\%$ lost each year in the Indo-Pacific since 1997 [17]. These rates of decline are forecasted to increase [137], and extensive surveys are urgently needed to better understand the coral reef crises [60] and contextualize effective ecosystem-based management for coral reefs [12].

Reef surveys were traditionally performed through *in situ* counting by scuba divers skilled in marine ecology. *In situ* counting surveys enable accurate observations, but they are time-consuming and allow only small areas of the reef to be surveyed. Additionally, this method is dependent on the skill of the experts conducting the surveys, and the data cannot be reliably

returned to for reanalysis. *In situ* counting of corals was largely replaced by image-based surveys when high quality underwater cameras became available at an economic price in the early 1960's [62], coincident with the proliferation of scuba as a tool for underwater research. Images-based techniques are advantageous for underwater applications, because they maximize data collection in the limited underwater time possible with scuba-based field deployments [39], and they provide permanent records that can be analyzed for organism abundance, size, and percentage cover. However, benthic photographs pose significant challenges to quantification because the organisms are seldom well defined in color, outline, or texture, and the process of underwater photography is prone to the vagaries of underwater conditions (e.g., surge, turbidity, light penetration, spectral truncation, etc.), equipment limitations in the corrosive seawater environment, and photographers with variable experience. Thus, underwater images often are dominated by blue colors, inadequate illumination, and deep shadows, and together these can make the analysis of complex images such as those found on tropical coral reefs difficult. Although some efforts have been made to quantify accuracy and inter annotator variability in surveys conducted using underwater video 'manta-tows' versus belt transect, line intercept transects, and other methods [21, 89, 96, 56, 68], there is little information describing the variability associated with annotations of coral reef survey images.

Since the turn of the new millennium, the application of image-based tools to underwater research has experienced rapid and dramatic changes in three domains. First, improvements in digital photography enable gathering large numbers of images at increasingly high resolution, thus vastly increasing the capacity to census marine environments. Second, the capabilities of computers and software to manipulate, store, and analyze images has increased several orders of magnitude in just 15 years. Third, advances in robotics and control theory have enabled the construction of autonomous underwater vehicles (AUVs), remotely operated vehicles (ROVs), and imaging sleds that can capture thousands of survey images in a single deployment [39, 124, 95]. However, the capacity to analyze the collected images has not advanced at a pace commensurate with the capacity to collect them. This has created a 'manual-annotation' bottleneck between the rapid image acquisition and the quantitative data needed for ecological analysis.

Generating quantitative ecological information from underwater images typically involves random point annotation to estimate the percent cover of the substrata of interest. To generate these data, substratum types are identified, usually manually by an expert in that ecosystem, for a number of randomly selected locations in each image. Statistically valid sampling of benthic habitats with image- and point- based tools requires careful attention to the primary purpose of the study, choice of the experimental and statistical approaches necessary to answer the questions being addressed, and the statistical power (i.e., a function of sample size, variance, and desired difference to be detected) required to test the relevant hypotheses. These experimental design components are not the focus of this paper, and interested readers are referred to the many excellent texts on these subjects [104, 90]. Instead, we focus on the means and accuracy by which quantitative information can be manually and automatically extracted from underwater images using random point annotation, specifically for near-shore tropical marine environments.

Point annotations are typically preformed using manual, off-line annotation software like Coral Point Count with Excel Extensions (CPCe) [70], photoQuad [135], or pointCount99 which facilitate the annotation process by providing user interfaces and tools for export of relevant data. More recently, web-based annotation tools such as Biigle (biigle.de) and Catami (catami.org) have been developed, however these also require manual annotation, with consequent extensive time requirements for analysis. Recently there have been several efforts to automate point annotation of benthic survey images [84, 129, 8, 15, 123], but the implementation of these tools has been hindered by procedural gaps. Critically, while human experts have been annotating benthic images manually for decades, the replication accuracy of this procedure remains largely unknown (i.e., intra-annotator error), as does the extent to which results can vary between two annotators (i.e., inter-annotator error). This information is critical for evaluating the efficacy of new analytical tools, as it provides the baseline by which performance of these new tools can be assessed. Additionally, there is scope to develop “hybrid” analytical systems in which automated procedures annotate a significant portion of the data, but defer the most uncertain identification decisions to human annotators. The value of this approach will depend on the quantitative means by which “uncertain” is evaluated, and the trade-off between accuracy and speed (i.e., efficiency)

that is acceptable for the question at hand.

In this paper we focus on codifying the criteria for implementing an existing automated image analysis tool [8] for coral reef survey images. Specifically, we sought to answer two questions: (1) what is the baseline of variation among human experts in the analysis of benthic communities that we would hope to equal or improve upon through automated procedures, and (2) what is the quantitative analytical framework that defines the trade-off between accuracy and efficiency, and how can this inform user-defined choices for operational modes of an automated system? To answer the first question, we designed a study in which we quantified the variance among multiple experts in the analysis of benthic images from coral reefs in Moorea (French Polynesia), Taiwan, the northern Line Islands, and Heron Reef (Great Barrier Reef), and compared their results to those obtained through computer-based automated analysis. To address the second question, we then altered the requirements for more (or less) human supervision in the modes of automated operation, in order to quantify the trade-off between accuracy and efficiency. Finally, a key contribution of this work is that the methods developed for automated annotations are incorporated in the random point annotation tool of CoralNet (coralnet.ucsd.edu), which is publically available.

3.2 Materials & Methods

Digital photoquadrats from coral reefs in four locations throughout the Pacific, with 671 - 3472 images per location, were used to provide a diverse model system in which we tested our analytical tools (Table 3.1, Fig. 3.1). These images were originally annotated by local experts at 24–200 random point locations superimposed on each image. From each image set at each location, 200 images were randomly selected and designated as the Evaluation set; the remainder were designated the Reference set. The four Evaluation sets were then annotated again independently, and without prior knowledge, of the initial annotations by each of six human experts at 10 points on each image, sub-sampled at random from the original point locations. Following the human annotation, an implementation of a computer vision algorithm [8] was used to automatically annotate the same 10 points per image. The four Reference sets were used to

train the automated annotator, and were made available as training material for the human experts. The multiple sets of annotations were used to evaluate intra- and inter-annotator variability and to evaluate the proposed operational modes of automated annotation. All images and data used in this study are made available as a benchmark dataset¹.

3.2.1 Study Locations

Digital photoquadrats were obtained from projects monitoring coral reef community structure in the outer and fringing reefs of Moorea (French Polynesia), Kingman, Palmyra, Tabuaeran and Kiritimati atolls (northern Line Islands), Nanwan Bay (Taiwan), and the platform reefs at Heron Reef (Great Barrier Reef, Australia). The study locations were selected because they offered legacy data involving large numbers of images annotated with equivalent random point methodologies by experts with extensive experience in identifying benthic organisms from photographs at their respective locations. In each location, multiple species of scleractinian corals, macroalgae, crustose coralline algae, and various non-coral invertebrates densely populate benthic surfaces, and photoquadrats are characterized by complex shapes, diverse surface textures, and intricate boundaries between dissimilar taxa. Additionally, in varying degrees for the four image sets, water turbidity and light attenuation degrade colors and image clarity, presenting a challenging task for either human or automated annotation. It should be noted however, that these are all typical conditions, and these image sets represent typical survey images taken for purposes of coral reef ecology.

The four locations also represent the large variation commonly found within and among photographic surveys of coral reefs in shallow water (< 20 m depth). This variation includes affects attributed to differences among locations in the coral species and the colony morphologies they produce, variation in camera equipment (e.g., angle of view, and resolution), distance between camera and benthos (and whether the distance was constant among photographs), and the mechanism employed to compensate for the depth-dependent attenuation of sunlight (i.e., through the use of strobes and/or manual white balance adjustment of camera exposures). The

---

¹All images and data used in this work are available at the University of California Computer Vision Laboratory's data repository: vision.ucsd.edu/content/pacific-labeled-corals

photographs from Moorea, the Line Islands and Nanwan Bay were captured using framers to hold the camera perpendicular to, and at a constant distance from, the sea floor. Underwater strobes were used in Moorea to restore surface color and remove shadows from images, and in the Line Islands, image-colors were adjusted through manual adjustment of the white balance for each series of images. Neither strobes nor color correction were used to record photoquadrats in Nanwan Bay. Finally, at Heron Island, the reef was photographed using a camera (without strobes or white balance correction) that was hand-held above the reef using a weighted line suspended below the camera to maintain an approximately fixed distance to the sea floor [108]. Refer to Fig. 3.1 for sample images from the locations, to Table 3.1 for a data summary, and to Appendix S1 for more details of the survey locations.

Figure 3.1. Sample photoquadrats acquired and annotated as part of four long-term monitoring projects. Photoquadrats acquired on shallow coral reefs (< 17 m depth) in: (A) Moorea, French Polynesia, acquired with strobes and framer (50 x 50 cm); (B) Palmyra, northern Line Islands, acquired using a manual white balance and framer (65 x 90 cm); (C) Nanwan Bay, Taiwan acquired with framer but neither strobes nor white balance; (D) Heron Reef, GBR, acquired without framer, strobes or white balance.

Table 3.1. Summary of image resources.

	Moorea	Line Islands	Nanwan Bay	Heron Reef
Investigator	P.J. Edmunds	J. Smith	T-Y Fan	C. Roelfsema
Sampling year	2008	2005	2012	2007
Reef type	Fringing reef	Fringing reef	Fringing reef	Platform reef
Geomorphic zone (Photo depth)	Reef flat / lagoon (2-5 m), Fore reef (10 m & 17 m)	Reef flat / lagoon (2-5 m)	Fore reef (2-5 m)	Fore reef (5 m) Reef flat (1 m)
# images	671	756	890	3472
# anns. image ⁻¹	200	100	50	24
# total anns. in Reference set	94,200	55,600	34,500	78,528
Image cover (cm)	50 × 50	65 × 90	35 × 35	50 x 65
Image size (px.)	6.24 M	7.1 M	9.98 M	6.2 M
Spatial res. px. (mm ²)	24.96	12	81.46	19.08
Camera	Nikon D70	Olympus 7070	Canon G12	Canon A540
Lighting	Dual Nikonos SB 105 strobes	No lighting, manual w.b.	No lighting	No lighting

= number of, M = million, res. = resolution, px. = pixel, w.b. = white balance

3.2.2 Label-set

When this study began, the photoquadrats from each of the four locations already had been analyzed manually by the local coral reef experts using four different label-sets defined by the respective experts. In order to make comparisons of annotator accuracies between the locations, a consensus label-set was created to which the four original label-sets were mapped (Table 3.2). The consensus label-set consisted of 8 scleractinian genera and 1 ‘other scleractinians’ label; 3 algal functional groups (macroalgae, crustose coralline algae (CCA), and turf algae); and 9 other labels that included sponges, sand, and the hydrozoan *Millepora*. In our analysis we also considered all coral labels together as a coral functional groups. The labels of the consensus label-set were chosen to enable a one-to-one or many-to-one mapping from the original label-sets in Moorea, Line Islands, and Nanwan Bay. The coral substrates were not resolved to genus level

in the original Heron Reef label-set and all Archived coral annotations were therefore mapped to the generic ‘other scleractinians’ label for this location.

Table 3.2. Label-set used in this work, and definition provided to human annotators.

Label	Definition
<i>Acropora</i>	Coral genus
<i>Favia</i>	Coral genus
<i>Favites</i>	Coral genus
<i>Montastraea</i>	Coral genus
<i>Pavona</i>	Coral genus
<i>Platygyra</i>	Coral genus
<i>Pocillopora</i>	Coral genus
<i>Porites</i>	Coral genus
Other scleractinians	Other hard corals
<i>Millepora</i>	Genus of hydrozoan coral
Sponges	All types of sponges
Soft Coral	All soft corals
Crustose coralline algae	All genera
Turf algae	Here defined as multi-specific assemblages: 1 cm or less in height
Macroalgae	All genera - defined as larger algae > 1 cm in height
Sand	Sand, silt or other fine-grained, soft substratum
Bare space	Rock, Basalt, Limestone, Dead coral, Rubble, or other hard substratum. Note, use this label only if <i>not</i> overgrown by algae
Transect hardware	Transect line, wand, and framer. Anything that is part of the sampling methodology
Unclear	Dots falling in shadowy, dark or blurry areas, where a class designation is not possible
All other labels	Any substratum not covered by the other labels, e.g., seagrass, other invertebrates, terrestrial trash

3.2.3 Manual Annotations

As part of the present analysis, manual annotations were performed by six human experts for the photoquadrats from each of the four locations. The local expert familiar with each location were designated as the ‘Host’, and the other five experts with less familiarity with the specific locations were designated ‘Visiting’. All 6 experts were experts in identify corals and benthic taxa in coral reefs of the tropical Pacific Ocean. The annotations completed by the Hosts prior

to the present study as part of the original ecological analyses were denoted ‘Archived’. One to six years had passed since the original annotations were conducted, and thus we reasoned that the Hosts would not be biased by their original annotations. All of the present annotations were performed using the random point annotation tool of CoralNet (Section 2.8; Figure S1A). To assist the experts in the new annotations, the images and archived annotations of the Reference sets from each location were stored in CoralNet and used as a virtual learning tool to improve identification of the benthic taxa (Figure S1B). All images were scored using 10 points per image, and therefore with 200 images per location, 4 locations, and annotations conducted by 6 experts, this study generated a set of 48,000 manual annotations.

3.2.4 Automated Annotations

We previously developed an automated annotation system for coral reef survey images [8]. In this system, the texture and color of a local image patch around a location of interest (i.e., a random point location) is encoded as a count of ‘visual words’, which is a quantification of the space of visual appearance of the image. The encoded visual information is then used, together with a set of labels, to train the automated annotator, using the Support Vector Machines (SVM) algorithm[43].

The method of [8] was modified in two ways to increase its accuracy and reduce the runtime of the method. First, the vector quantization step was replaced by Fisher Encoding [107] which was recently shown to improve classification accuracy for various image classification tasks [26]. Second, the kernel-based SVM was replaced by a linear SVMs which have significantly lower runtime [43, 26, 107]. In our experiments training a linear SVM required 5-20 minutes, while the kernel-based SVM required several hours. The final algorithm used in this work required 20 seconds to pre-process each image, the training of the SVM required 5–20 minutes for each location depending on the size of the Reference set (i.e., number of expert annotations), and automated annotations of an unknown image from the Evaluation set required < 1 second. The aforementioned times were measured on a single computational core.

The color and texture information of the images in the Evaluation and Reference sets

was encoded using Fisher encoding to generate a ‘feature vector’ for each annotated point in each image. Using a dictionary of 40 visual words, the dimensionality of each feature vector was 1920. The feature vectors from the Reference set were paired with labels encoding the Archived annotations to form a training-set, which was used to train a one-versus-rest linear SVM (Appendix S1). This training was performed separately for each location so that the training-set from each location was used to train a SVM, which was then used to automatically annotate the images in the Evaluation set from the same location.

Since the images in the Evaluation set were randomly selected from the set of available images, the mean percent cover of each label was similar in the Reference and Evaluation sets. This means that approximately correct cover estimates of the Evaluation set could be generated by a naive strategy in which labels were assigned to point locations in the Evaluation sets in the same proportions (base-rates) as in the Reference set. However, in other situations where the automated annotator is trained on data from alternative data, for example from a previous year or another location, the base-rate will likely differ between training (here the Reference set) and test data (here the Evaluation set). The following pre-processing step was applied to ensure that any conclusions from this study, with respect to the efficacy of the automated annotator to estimate percent covers, would be valid in situations where the base-rate differs between the training and test data. In this pre-processing step, the base-rates of the Reference sets were changed by discarding a randomly selected subset of the non-coral annotations, so that the coral base-rate increased by 10%. For example, there were 19,566 coral, and 74,634 non-coral annotations in the Moorea Reference set, for a coral base-rate of 20.77%. A random subset of 8,566 non-coral annotations was discarded so that 66,068 non-coral annotations remained. This resulted in a modified coral base-rate of 22.85%, which was a 10% increase from 20.77%.

3.2.5 Post-processing of Annotations

Two inconsistencies in the annotations of the Hosts and Visitors were noted and corrected. First, some expert annotators mapped points on the images that fell on transect hardware to the ‘Transect hardware’ label, while others mapped them to substratum types that were inferred to

occur beneath the hardware based on close proximity adjacent to the hardware. Such discrepancies were due to incomplete instructions to the experts rather than identification difficulties. Therefore, any point locations that two or more of the experts mapped to ‘Transect Hardware’ were considered to be transect hardware, and all annotations for that point location were changed to ‘Transect hardware’ in post-processing.

The second inconsistency arose with the scoring of images from Moorea because neither the original label-set in this location, nor the Archived annotations, contained the ‘Bare space’ label of the consensus label-set. Instead, in this location all “bare space” was effectively covered with CCA and therefore was mapped to a ‘CCA’ label. However, in the annotation of the present study, some experts mapped these areas to ‘Bare space’ as the best match to this state in the consensus label-set. To facilitate a comparison between the multiple sets of annotations, all ‘Bare space’ annotations for the Moorea location were changed to ‘CCA’ in post-processing.

3.2.6 Estimating Annotator Variability

Using the Archived annotations as baseline, the accuracy of annotator a was estimated using the Cohen’s kappa statistic, and denoted κ^a [22]. Additionally, κ_Ψ^a denoted the Cohen’s kappa of the binary classification task between a set of labels, Ψ (e.g., corals, macroalgae, or *Acropora*), and the labels not in Ψ . For example, κ_{coral}^a denoted the accuracy of annotator a , for the task of discriminating between coral and non-coral. For notational brevity, the super-script is henceforth dropped when it is clear from the context which annotator is considered. Differences in κ (calculated using the full label-set) between the Hosts and Visitors were assessed using non-parametric Mann-Whitney U tests, and difference between the four functional groups (coral, macroalgae, CCA, and turf algae) was assessed using a Kruskal-Wallis test.

Additionally, using again the Archived annotations as baseline; a confusion matrix, Q was estimated for each location and annotator. Each confusion matrix Q was 20 rows by 20 columns and values at row r , column c in the matrices indicate the ratio of annotations originally labeled by Archived as label r now classified by the Host, Visiting and Machine annotations respectively as label c .

3.2.7 Modes of Automated Annotation

In this section, the proposed operational modes for semi- and fully- automated annotation are detailed.

The Alleviate Operational Mode

A one-versus-rest automated classification procedure generates a vector of classification scores corresponding to each classification decision [112]. The score vector can be used to estimate the certainty of the automated classifier for an unseen example [78]. For example, if all scores are low and similar to each-other, the classifier is more likely to assign the incorrect label [78]. In such situation, an alternative is to let the classifier abstain from making an automated annotation decision and instead defer to a human expert. Such procedure enables a trade-off between the amount of human effort and the accuracy of the final annotations.

The classification scores were used in an operational mode, which we denote ‘Alleviate’ because it alleviates the workload for the human annotator. The scores were denoted $s_{i,k}(m)$ where subscript i indicate the image, subscript k the point location, and m the class label. For a given point in an image the vector or 20 scores, one for each class, was denoted $[s_{i,k}]$. Using Alleviate, an automated annotation was only assigned when, for a given point location, there existed a score $s_{i,k}(m)$ such that $s_{i,k}(m) > \epsilon$, for some threshold, ϵ . More formally, the Alleviate annotations, $y_{i,k}^{\text{Alleviate}}$ were defined as:

$$y_{i,k}^{\text{Alleviate}} = \begin{cases} y_{i,k}^{\text{Automated}} & \text{if } \max[s_{i,k}] > \epsilon \\ y_{i,k}^{\text{Host}} & \text{otherwise} \end{cases}$$

where $y_{i,k}^{\text{Automated}}$ was the automated annotation of point k in image i , and $y_{i,k}^{\text{Host}}$ the corresponding Host annotation. We denoted by $\lambda(\epsilon)$ the ‘level of alleviation’, i.e. the fraction of samples classified by the automated annotator for a certain threshold, ϵ . A level of alleviation of $\lambda(\epsilon) = 50\%$ was used in the final analysis of Alleviate.

The Abundance Operational Mode

While the automated annotations are, in general, less accurate than those provided by human experts [129, 8, 123], they have the benefit of being, in contrast to a human expert, deterministic, meaning that the automated annotator will always make the same classification decision for the same image. If the matrix of confusion is known for the automated annotator, it can be used to generate annotations that are accurate in aggregate (such as abundance of taxa) even though individual decisions might be erroneous [127, 59]. The corrected abundances will by design be unbiased but with larger variation than those based on the automated annotations [127].

Using the abundance correction method of [127, 59], a second fully automated deployment mode, Abundance was defined. In this mode, cover estimates for each image i were calculated as:

$$c_i^{\text{Abundance}}(m) = (Q')^{-T} c_i^{\text{Automated}}(m) \quad (3.1)$$

where $c_i^{\text{Automated}}(m)$ denoted the cover in image i of class m based on the automated annotations, Q' was a confusion matrix estimated through a 20 fold cross-validation on the Reference set, and the $-T$ superscript denotes matrix transpose followed by inverse. Note the difference between Q' , which was estimated for the automated annotator from the Reference set and used in Abundance, and Q , which was estimated for all annotators from the Evaluations sets and introduced in Section 3.2.6.

Performance evaluation based on cover estimates

To assess the efficacy of the proposed operational modes, their estimates of the ecological composition of the four studied locations were contrasted with the estimates based on the Archived, Host's, and Visitors' annotations. The cover of label m , for image i , was denoted $c_i^a(m)$ for each set of annotations:

$$a \in \{\text{Archived}, \text{Host}, \text{Visiting1}, \dots, \text{Visiting5}, \text{Alleviate}\}$$

In addition, $c_i^{\text{Abundance}}$ was derived from the Machine annotations using Eq. (3.1). The pairwise differences from the Archived cover estimates for each image were denoted:

$$d_i^a(m) = c_i^a(m) - c_i^{\text{Archived}}(m)$$

The average estimation error (bias) for annotator a , label m for the 200 images in a certain location was denoted:

$$e^a(m) = \frac{1}{200} \sum_{i=1}^{200} d_i^a(m)$$

The Mean Absolute Error (MAE) was denoted as the mean absolute value of $e^o(m)$ calculated for a particular substratum and group of annotators across the four locations (e.g., the macroalgae cover as estimated by the Hosts).

The differences $d_i^a(m)$ were also used to test the null hypothesis that, for each label, annotator and dataset, the mean of $d_i^a(m)$ is zero. A one-sample Kolmogorov-Smirnov test was first performed to determine normality of $d_i^a(m)$. This test yielded p-values < 0.001 for all label (or label groups), annotations, and locations, indicating non-normality and the consequent need for a non-parametric test. A permutation t-test was therefore performed at the 95% significance level with a Bonferroni correction for eight comparisons (Host, Alleviate, Abundance, and 5 Visitors) [100]. Test results are given in Fig. 3.5 (and Table S2) where a yellow dot indicates that the null hypothesis cannot be rejected. The differences $d_i^a(m)$ were also used to estimate a 95% confidence interval for all contrasts using the percentile-t bootstrap procedure (Fig. 3.6) [100].

3.2.8 CoralNet

As a key contribution of this work, we make implementations of the described modes of operation (Alleviate, Abundance, and Refine [Appendix S1]) publicly available on CoralNet (coralnet.ucsd.edu). CoralNet is a software module that has been designed as a repository and online annotation tool for benthic survey images. CoralNet allows users to upload and annotate survey images with a user-defined label-set. Annotations are performed using a random point annotation interface similar to that of CPCe [70]. CoralNet also offers tools for browsing

images and annotations as well as for viewing and exporting estimated cover statistics. The implementation of Alleviate allows users to determine the level of alleviation based on the accuracy of the machine annotator, and Refine is implemented with 5 suggested labels per point location.

3.3 Results

3.3.1 Annotator Accuracy

Substratum Identification Accuracies

Annotator accuracy was measured as Cohan's kappa, κ of the annotators, compared to the archived annotations. The Hosts annotation accuracies for the functional groups differed among the functional groups ($H = 10.05$, $df = 3$, $p = 0.015$). Specifically, the accuracy for each functional group was: $\kappa_{coral} = 89.7 \pm 1.2\%$, $\kappa_{macroalgae} = 71.0 \pm 4.0\%$, $\kappa_{CCA} = 51.0 \pm 9.3\%$, and $\kappa_{turf\ algae} = 61.6 \pm 6.1\%$ (mean $\pm$ SE, $n = 4$ locations, [Fig. 3.2]), with the main source of confusion occurring among algal substrata (Figure S3). The average annotation accuracy of the Hosts for common coral genera (i.e., with > 10 archived annotations) was $\kappa_{coral\ genera} = 79.4 \pm 4.2\%$ ($n = 19$, [Table S1]). Coral genera were most commonly confused with other coral genera, in particular with the general label of 'other scleractinian'. However, there was also notable confusion between the turf algae and CCA (Figure S3).

Relationship between Abundances and Identification Accuracies

The Host accuracies were not significantly correlated with the benthic covers ($r = 0.089$, $df = 33$, $p = 0.61$), although there was a few outliers recorded a low ($< 3\%$) cover (Fig. 3.3 A). The accuracies of corals (when considered as a group) was $> 80\%$ in all four locations. Annotation accuracy of algal groups was $> 60\%$ except for three samples, which were recorded at $< 12\%$ cover. Similarly, accuracy for identifying coral genera generally was $> 80\%$, except six samples recorded at $< 3\%$ cover.

Figure 3.2. Classification accuracy. The accuracy is measured by Cohen's kappa for the Host annotations with various level of alleviation, and for the Visiting (mean $\pm$ SD, $n = 5$) annotations. The left edge of the alleviation curve corresponds to Machine annotations only, and the right edge corresponds to Host annotations only. First column shows accuracy calculated from the full label-set. Column two to five show accuracy for the task of discriminating functional groups: coral, macroalgae, crustose coralline algae (CCA), and turf algae, respectively, versus the rest. Column six and seven show accuracy for the three dominant coral genera. Genus-level identification for Heron Reef is not provided, as Archived annotations were not available to this resolution.

Figure 3.3. Relationship between mean percent cover and annotation accuracies. Relationships displayed as (A) Cohen’s kappa for the Hosts, and (B) Difference in Cohen’s kappa between Alleviate and the Hosts, plotted against the cover of coral genera and functional groups. Data is drawn from all four locations except for the coral genera where data was not available for Heron Reef.

Intra-annotator Variability

The annotation accuracies of the Visiting experts were lower than for the Hosts ($U = 26$, $p = 0.029$). Specifically, the accuracy for each functional group was: $\kappa_{coral} = 84.0 \pm 1.0\%$, $\kappa_{macroalgae} = 58.7 \pm 2.5\%$, $\kappa_{CCA} = 35.5 \pm 3.7\%$, and $\kappa_{turf\ algae} = 43.3 \pm 3.6\%$ (mean $\pm$ SE, $n = 20$, [Fig. 3.2]). The average accuracy of the Visitors for the common coral genera (i.e., with > 10 archived annotations) was: $\kappa_{coral\ genera} = 58.6 \pm 2.9\%$ ($n = 95$, [Table S1]). As with the Hosts, the principal confusion occurred among the algal groups, and the principal confusion among coral genera occurred with the ‘other scleractinian’ label (Figure S3). The lower accuracy of the Visitors was evident in the lesser weight across the confusion matrix diagonals compared to the matrix diagonals of the Hosts (Figure S3).

3.3.2 Modes of Automated Annotation

The accuracy of the machine annotation algorithm was: $\kappa_{coral} = 63.5 \pm 4.3\%$, $\kappa_{macroalgae} = 48.3 \pm 11.7\%$, $\kappa_{CCA} = 28.3 \pm 16.3\%$, and $\kappa_{turf\ algae} = 43.8 \pm 4.2\%$ (mean $\pm$ SE, $n = 4$, [Fig. 3.2]).

)). This drop in accuracy compared to the Hosts (e.g., over 20% for κ_{coral}) suggested that a direct application of the automated annotation algorithm of [8] would not yield reliable cover estimates of benthic substrata. Instead, performance of the automated annotation algorithm was evaluated within the context of the proposed operational modes.

Hybrid Annotation Using Alleviate

Alleviate is a semi-automated annotation mode, in which the automated annotator has the option of deferring difficult decisions to a human expert annotator. The ratio of points that are automatically classified was denoted ‘level of alleviation’, λ . Because the expert annotators were more accurate than the automated annotator, the accuracy of Alleviate increased as λ decreased, and the highest accuracies were observed when $\lambda = 0\%$, i.e. when all annotations were done by the Hosts (Fig. 3.2). Specifically, using the full training-set of available expert-annotated images in the present study, $\lambda = 40 - 60\%$ incurred $< 5\%$ decrease in κ_{coral} compared to the Host annotations (Fig. 3.4). Moreover, a supplementary analysis established that $\lambda = 32 - 45\%$ incurred $< 5\%$ decrease in accuracy when only 5,000 manual point-annotations were used for training, which corresponds to $\sim 25 - 200$ training images for each of the four locations (Appendix S1).

The Mean Average Error (MAE) of cover estimates obtained using Alleviate at $\lambda = 50\%$, compared to cover estimates from the archived annotations were $1.3 \pm 0.4\%$ for coral, $2.0 \pm 0.8\%$ for macroalgae, $3.6 \pm 1.0\%$ for CCA, and $5.6 \pm 1.7\%$ for turf algae (mean $\pm$ SE, $n = 4$ locations, [Fig. 3.6]). The 1.3% coral cover MAE corresponds to a relative error of $\sim 4 - 6\%$ at the 22-31% coral cover recorded across the four locations (Fig. 3.5). These results should be viewed in the context of the Host annotator MAEs, which were $1.3 \pm 0.6\%$ for coral, $1.5 \pm 0.4\%$ for macroalgae, $3.4 \pm 2.0\%$ for CCA, and $5.0 \pm 1.6\%$ for turf algae for the Hosts (mean $\pm$ SE, $n = 4$, [Fig. 3.6]). For the Visitors, the MAEs were $2.2 \pm 0.4\%$ for coral, $4.2 \pm 1.2\%$ for macroalgae, $7.6 \pm 2.1\%$ for CCA, and $11.9 \pm 2.0\%$ for turf algae (mean $\pm$ SE, $n = 20$ [Fig. 3.6]). The Alleviate MAE for the three dominant coral genera found in Moorea, the Line Islands, and Nanwan Bay was $0.5 \pm 0.2\%$ ($n = 9$ [Fig. 3.6]). Again, this was similar to the Host MAE of $0.5 \pm 0.2\%$ ($n = 9$, [Fig. 3.6]), and lower than the Visitor MAE of 1.3 ± 0.23 ($n = 45$ [Fig. 3.6]).

Figure 3.4. Accuracy as measured by Cohen's kappa for the task of discriminating corals from other labels, κ_{coral} at various levels of Alleviation for the four studied locations. Accuracy was calculated by comparing the Archived annotations with the combined annotations of Host and automated annotations for various levels of alleviation. The black x on each curve indicates the point where κ_{coral} was 5% lower than its maximum value (i.e. a 5% drop compared to the κ_{coral} of the Hosts).

The difference in accuracy $\kappa^{\text{Alleviate}} - \kappa^{\text{Host}}$ (at $\lambda = 50\%$) was not significantly correlated with benthic cover ($r = 0.25$, $df = 33$, $p = 0.14$ [Fig. 3.3 B]). The differences for coral as a group were around -5% for the four locations, and between -9% and +3% for the coral genera (a positive difference means that $\kappa^{\text{Alleviate}}$ was higher than κ^{Host}). The only exception was *Platygyra*, where the difference was -28.0% at 1.6% cover and -15.3% at 0.5% cover in Nanwan Bay and Line Islands respectively. The differences for algal functional groups were $> -10\%$, except for macroalgae and CCA in Nanwan Bay where the differences were -19.8% at 6.5% cover, and -15.6% at 3.9% cover, respectively.

Fully Automated Annotation Using Abundance

Using the abundance correction method of [127], the Abundance operational mode can be deployed without any manual annotations. The MAE of cover estimates obtained in the Abundance mode was $1.8 \pm 1.5\%$ for coral, $3.4 \pm 2.3\%$ for macroalgae, $4.1 \pm 2.1\%$ for CCA, and $7.2 \pm 4.5\%$ for turf algae (mean $\pm$ SE, $n = 4$ [Fig. 3.6]). The Abundance cover MAE of the three dominant coral genera of Moorea, the Line Islands, and Nanwan Bay was $1.5 \pm 0.5\%$ ($n = 9$

Figure 3.5. Percentage cover estimates of functional groups (coral, macroalgae, CCA, and turf algae) and the three most abundant coral genera in each location as determined by different annotation methods. Values displayed as mean $\pm$ SE ($n = 200$ images) for each location: (A) Moorea, (B) the northern Line Islands, (C) Nanwan Bay, and (D) Heron Reef. Archived describes reference analyses, performed by a local expert, against which the other results were evaluated. Abundance and Alleviate refer to fully and semi-automated annotation modes. Host refers to re-annotation by the same local experts. Visiting refers to analyses completed by five coral biology experts who do not regularly work in those locations. Yellow dots indicate significant differences at 95% confidence between cover estimates from Archived versus other annotations for that label and dataset. Genus-level identification for Heron Reef is not provided, as Archived annotations were not available to this resolution.

Figure 3.6. Differences in percent cover estimates between the Host and Visiting experts as well as the Abundance and Alleviate operational modes and the Archived annotations. Differences are displayed as mean with 95% confidence interval ($n = 200$ images) for functional groups (coral, macroalgae, CCA, and turf) and the three most abundant coral genera in each location: (A) Moorea, (B) the northern Line Islands, (C) Nanwan Bay, and (D) Heron Reef. Archived describes reference analyses, performed by a local expert, against which the other results were evaluated. Abundance and Alleviate refer to fully and semi-automated annotation modes. Host refers to re-annotation by the same local experts. Visiting refers to analyses completed by five coral biology experts who do not regularly work in those locations. Genus-level identification for Heron Reef is not provided, as Archived annotations were not available to this resolution.

[Fig. 3.6]). As with Alleviate, these errors should be viewed in the context of the human errors.

3.4 Discussion

3.4.1 Annotator Accuracy

We have estimated the inter- and intra- annotator variability of human experts for annotating coral reef survey images. Quantifying this variability is critical to contextualize the performance of automated annotation methods. For example, if the accuracy of human experts is low for a certain substratum, we should expect the automated annotation accuracy to be equally low (since automated annotation methods are commonly trained on archived, manually annotated data). Conversely, for labels where the human annotator accuracy is high (e.g., for coral genera), our results establish baselines against which newer generations of automated annotation systems can be compared.

Intra-annotator Variability

Annotator accuracy was measured using Cohen’s kappa (κ) for intra- and inter- annotator agreement. While κ is widely used to assess annotator agreement for categorical data [3, 82, 91], there is no absolute scale against which values of κ can be gauged. For example, [72] characterized $\kappa < 0$ as indicating no agreement, 0–20% as slight agreement, 21–40% as fair agreement, 41–60% as moderate agreement, 61–80% as substantial agreement, and $> 81\%$ as almost perfect agreement. In contrast, [45] characterized $\kappa < 40\%$ as poor agreement, 40–75% as fair-to-good agreement, and $> 75\%$ as excellent agreement. Using these interpretations, the Host κ_{coral} of $89.7 \pm 1.2\%$ should be considered an “excellent” or “almost perfect” agreement between the Host and Archived annotations. Similar high accuracies have previously been noted for self consistency of human annotations of corals versus other substratum [96, 89]. Ninio et al. investigated accuracy (as defined by agreement with *in-situ* observations) for video transects from coral reefs, and observed 96% accuracy for identifying hard corals [96].

The intra-annotator accuracy of identifying algal substrata was lower than for corals. In particular, for CCA and turf algae Cohen’s kappa were $51.0 \pm 9.3\%$ and $61.1 \pm 6.1\%$, respectively,

although this agreement can still be considered ‘moderate’ or ‘fair to good’ [72, 45]. The lower accuracy of algal classification may be due to methodological limitations in discriminating algae in planar RGB photo-quadrats at common resolutions as used in the present study (Table 3.1), but also due to the label definitions of the consensus label-set. For example, the consensus label-set defined turf-algae as a “multi-specific assemblage < 1 cm in height”, while macroalgae were “larger algae > 1 cm in height” (Table 3.2). Considering the planar nature of the photo quadrats, such distinctions can be hard to consistently apply across photographs. Ninio et al. observed 80% annotation accuracy of algae (as a ‘main benthic’ group) [96], which is higher than our estimates of algal functional groups annotation accuracy. However, comparison of accuracy of algae as a ‘main benthic’ group, with algal functional groups (macroalgae, turf algae, and CCA) is misleading, since the majority of the algal classification errors occurred among the algal groups (Figure S3). Merging the algal functional groups used in this study, the Hosts annotation accuracy of discriminating algae (macroalgae, turf algae and CCA combined) versus everything else, was $\kappa_{turf\ algae} = 42.9 \pm 3.4\%$. This is similar to the 80% accuracy recorded in [96].

The Host accuracies were not correlated with the abundance of the substratum of interest, and were > 60% except for five rare substrata occupying < 5% of the benthos (Fig. 3.3). The only exception was turf algae in Moorea, with 43% accuracy at 11.3% cover. These results agree with the findings of Ninio et al. who noted highly variable precision when mean covers were < 3%, but “markedly decreased” variability at higher covers (i.e., > 3%) [96].

Inter-annotator Variability

Our results indicate the inter-annotator accuracy was lower than intra-annotator accuracy. Differences were large in particular for CCA and turf algae, where the accuracies were $35.5 \pm 3.7\%$ and $43.3 \pm 3.6\%$ respectively, which can be considered “fair” [72, 45]. This lower accuracy resulted in cover estimation errors by the Visitors that commonly exceeded 10% (Fig. 3.6). These results are in contrast with the results of Ninio et al., where the intra-annotator variability only contributed $\pm 0.5\%$ to the confidence intervals for benthic group ‘Algae’ [96]. We believe this contrast has two principal causes. First, the multiple human annotators of Ninio et al. were all

familiar and trained in the ecology of the study location, while the Visiting experts of the present study were less familiar with the local ecology and had not (with a few exceptions) visited the respective locations. Second, as noted above, Ninio et al. do not report cover estimation on algal functional groups but only in aggregate. Still, it is clear from our results that rigorous training of expert annotators is critical to achieve reliable manual annotation of turf and CCA algal groups, in particular where there are multiple experts involved in the annotation process. Studies on annotator variability of plankton classification have reached similar conclusions, namely that human annotator variability can be significant both across and within labs [32, 80, 33].

3.4.2 Modes of Automated Annotation

We have shown that the hybrid annotation mode, Alleviate, can reduce manual annotation time by 50% while generating cover estimates of comparable accuracy to those obtained by human annotators (Fig. 3.6), and that this method can be deployed when limited data are available for training the automated annotator (Appendix S1). We have also shown that differences in accuracy between Alleviate and the Hosts were generally small ($< 10\%$), and not correlated with the substratum abundances (Fig. 3.3). Two outliers were noted. First, the accuracy of Alleviate for *Platygyra* was 15-29% lower than for the Hosts (recorded at $< 2\%$ cover). This may be due to the limited amount of available training data for *Platygyra* (due to the low cover), and the visual similarity to other massive or encrusting corals with meandroid or submeandroid corallum morphologies. Second, the accuracy of Alleviate was 16% and 20% lower for macroalgae and CCA in Nanwan Bay respectively (both recorded at $< 7\%$ cover). This may be due to a combination of limited amount of available training data, and limitations in the photographic methodology indicated by the low accuracy of the Host: 60.4% and 30.3% respectively for macroalgae and CCA (Fig. 3.2).

Alleviate uses the classification scores to decide when to make an automated annotation and when to defer to a human expert. To the best of our knowledge, this approach has not been utilized for annotation of coral reef survey images. However, a similar approach was used for classification of plankton images acquired using imaging in-flow cytometry, where the specificity

of the classifier increased for all classes when only considering classification decisions where the classifier was $> 65\%$ confident of their identification [128, 99].

We have also evaluated the efficacy of a fully automated annotation mode Abundance. Our results indicates that Abundance generates unbiased cover estimates, but with larger standard errors (Fig. 3.6), which is to be expected by the nature of the abundance correction method [127]. Large standard errors are undesirable, but can be reduced by collecting and automatically analyzing more images. In fact, as discussed in Chapter 4, the introduction of fully automated annotation can inform the underlying random sampling survey design, where for example, more images of coral reefs could be collected in the field (which is often relatively easy and cost effective) in order to compensate for the larger variance of fully automated annotation modes such as Abundance.

As with Alleviate, the underlying principle of Abundance has been utilized in other applications. Indeed, it was originally proposed the method for automated plankton classification [127], and it has been commonly utilized for that purpose [128]. It has also been utilized for sentiment analysis from text corpora [59].

In order for Abundance to generate unbiased cover estimates, the confusion matrix must be known for the data that is being sampled [127]. In our experiments, the training data (i.e., the Reference Sets) were drawn from the same underlying probability density as the sampled data (i.e., the Evaluation Sets), and the confusion matrices could thus be estimated from the training data. If, however, the training data and the sampled data are drawn from e.g. different locations, sites, or years, the probability densities may be different [6, 101]. In such situations a subset of the sampled data must be annotated in order to estimate the confusion matrix, and Abundance can no longer be considered fully automated. However, it may still offer significant savings compared to Alleviate or fully manual annotation.

3.5 Conclusions

We have established baseline accuracies of human expert annotations of coral reef survey images, and compared this to a recent method for automated annotation. Our results indicate that the accuracy of human experts varies with the type of benthic substrata. Annotations of

coral genera have low inter- and intra- annotator variability, while annotations of algal groups, in particular turf and CCA algae, from those same survey images, have much larger intra- and inter- annotator variability. This suggests the need for development of photographic methodology for visualizing turf and CCA algal groups (e.g., through fluorescence [133, 148]) and rigorous training of expert annotators in order to achieve reliable manual annotation of these groups.

We have proposed two modes of operations in which methods for automated annotation can be deployed semi- or fully- automated. Our results indicate that cover estimates from the semi- automated annotation mode, *Alleviate* are of similar quality to cover estimates of manual annotations, while the cover estimates of the fully- automated mode, *Abundance*, are unbiased but with higher errors. The appropriate deployment mode will depend on the specific application. A reef manager for example, might utilize *Abundance* for rapid assessment of reef health, while a benthic ecologist might prefer *Alleviate*, which requires greater manual effort but is more accurate. We expect the reduction of annotation time to continue as improved methods for automated annotation become available. Implementations of the proposed modes of operation (*Alleviate*, *Abundance*, and *Refine* [Appendix S1]) are available to the public on CoralNet (coralnet.ucsd.edu).

3.6 Supporting Figures & Tables

Figure 3.7. Screenshots from CoralNet. A) Graphical user interface (GUI) used to create the Host and Visiting annotations. B) Browse tool used by the Visiting annotators to learn about the images and the label-set from previous annotations. Here a user has searched for all *Pocillopora* annotations from the Line Islands dataset.

Figure 3.8. Confusion matrices for Moorea, Line Islands, Nanwan Bay and Heron Reef. Values at row r , column c in the matrices indicate the ratio of annotations originally labeled by Archived as label r now classified by the Host, Visiting and Machine annotations respectively as label c . The numbers on the right margin indicate the total count of each row. For brevity, all annotations of the Visiting experts merged into a single confusion matrix for brevity and only labels for which more than 10 annotations were assigned by any of the annotators were included.

Table 3.3. Tabulated accuracies. Accuracy is measured by Cohen’s kappa for the Hosts, Visitors and two levels of Alleviate (ALL.), where 100% alleviation corresponds to using the automated annotations only. All accuracies are measured as compared to the Archived annotations. The ‘All classes’ row is the accuracy of the full confusion matrix, while the other rows indicate the accuracy of binary classification between the indicated label or label group and the other labels.

	Label	ALL.	ALL.	Host	Visitors					mean	std
		100%	50%		V1	V2	V3	V4	V5		
Moorea	All classes	58.82	72.66	71.76	57.63	61.41	27.43	52.61	48.19	49.46	13.29
	Coral	70.79	87.69	91.05	90.34	81.88	78.25	86.24	86.67	84.68	4.68
	Macro	57.62	72.1	69.34	57.63	65.25	59.4	65.8	54.43	60.5	4.92
	CCA	58.96	72.35	70.2	45.22	63.13	20.66	43.82	33.13	41.19	15.74
	Turf	31.36	44.41	43.81	33.9	48.81	18.06	24.51	19.15	28.89	12.78
	Acrop	43.32	92.23	92.61	78.75	64.54	60.42	64.26	72.51	68.1	7.4
	Monti	65.16	89.66	91.1	83.74	28.45	11.07	33.52	67.73	44.9	29.89
	Pavon	66.26	84.98	90.78	80.63	73.14	-2.01	79.75	85.54	63.41	36.84
	Pocill	74.85	88.89	91.13	90.16	80.11	25.87	85.98	88.89	74.2	27.3
	Porit	76.13	90.23	92.53	90.85	75.95	34.23	88.17	90.85	76.01	24.16
Line-Islands	All classes	52.72	69.35	72.01	52.65	49.06	54.65	50.3	59.25	53.18	4.02
	Coral	68.77	88.33	91.83	88.83	85.62	83.66	85.46	88.95	86.5	2.31
	Macro	59	71.6	76.79	57.42	61.11	64.11	63.24	57.29	60.63	3.19
	CCA	53.89	64.33	62.93	51.21	42.18	58.62	48.23	63.15	52.68	8.32
	Turf	49.2	63	62.51	24.15	24.42	51.83	36.11	49.13	37.13	13.14
	Acrop	77.8	91.4	95.6	96.06	95.14	94.43	93.53	94.43	94.72	0.94
	Favia	22.71	47.52	56.58	54.15	65.06	13.45	71.16	61.6	53.08	22.99
	Monti	41.03	84.15	88.15	78.12	67.41	24.16	68.3	70.89	61.78	21.44
	Pavon	8.53	59.6	64.74	8.09	9	-1.25	25.23	23.74	12.96	11.27
	Plat	39.82	62.37	77.68	95.21	79.9	78.14	0	89.95	68.64	39.01
	Pocill	69.27	91.09	93.68	89.86	87.72	70.69	85.92	85.01	83.84	7.58
	Porit	47.61	76.22	84.75	78.64	65.67	50.78	44.09	60.75	59.99	13.4
	Mille	15.31	95.63	100	69.03	0	0	28.2	62.56	31.96	33.04
Nanwan Bay	All classes	41.96	63.35	71.39	48.97	51.49	59.99	56.45	63.34	56.05	5.91
	Coral	51.55	81.7	89.37	87.09	76.27	83.23	87.98	90.5	85.01	5.54
	Macro	13.52	40.54	60.38	26.57	53.23	59.77	38.11	63.76	48.29	15.58
	CCA	-0.2	14.67	30.3	34.89	20.51	19.84	23.89	29.05	25.63	6.33
	Turf	47.53	63.55	71.4	30.26	57.53	64.47	47.05	66.92	53.25	14.99
	Acrop	71.97	71.14	74.76	51.16	37.28	54.4	63.26	45.5	50.32	9.72
	Favia	13.91	77.78	85.43	69.66	56.48	41.71	37.75	26.58	46.44	16.82
	Fav	29.99	76.52	83.27	18.48	62.38	37.82	-0.19	41.21	31.94	23.77
	Monti	57.77	82.19	90.34	86.77	49.32	84.39	86.51	87.34	78.86	16.55
	Plat	26.96	54.53	82.51	81.9	79.67	75.16	76.31	53.56	73.32	11.36
	Pocill	34.49	34.24	35.86	31.16	43.29	40.58	84.21	33.54	46.56	21.62
	Porit	-0.09	35.27	37	26.95	10.95	31.81	11.68	35.58	23.4	11.45
	Mille	82.9	95.21	95.54	91.02	82.07	92.17	92.7	95.9	90.77	5.19
Heron Reef	All classes	52.77	70.95	73.92	57.56	58.18	67.11	66.29	69.58	63.74	5.5
	Coral	62.73	81.35	86.39	77.92	79.89	79.28	78.97	82.35	79.68	1.65
	Macro	63.86	78.25	77.68	47.59	67.28	74.09	63.84	74.93	65.54	11.06
	CCA	-0.28	36.1	40.55	27.64	0	24	27.2	34.34	22.64	13.2
	Turf	47.02	65.25	66.82	42.87	44.92	61.66	58.53	61.63	53.92	9.27

Table 3.4. Tabulated p-values. Probabilities that the estimated cover from a set of annotations and for a certain label and dataset is the same as the cover estimated from the Archived annotations. Probabilities (p-values as estimated from the permutation t-test) in red italics are below $0.05 / 8 = 0.00625$, which means that the null hypothesis can be rejected at a 95% confidence level with a Bonferroni correction for eight repeated measurements. This implies that the particular set of annotations is unreliable for cover estimation. Rows 1–4 are the functional groups: coral, macroalgae, crustose coralline algae (CCA), and turf algae; rows 5–6 are the dominant coral genera, except for Heron Reef where annotations were not resolved to genus level in the original study. Columns ABU and ALL are the Abundance and Alleviate annotation modes respectively, and V1 to V5 are the Visiting annotators.

	Label	ABU.	ALL.	Host	V1	V2	V3	V4	V5
Moorea	Coral	0.6010	0.5089	0.0623	0.6841	0.0949	0.4529	0.4504	0.1466
	Macroalgae	0.8592	0.0157	<i><0.001</i>	<i>0.0053</i>	<i><0.001</i>	0.0086	<i><0.001</i>	0.7351
	Turf algae	0.0345	<i><0.001</i>	0.2911	<i><0.001</i>	<i><0.001</i>	<i><0.001</i>	<i><0.001</i>	<i><0.001</i>
	CCA	0.4933	<i><0.001</i>	<i><0.001</i>	<i><0.001</i>	<i>0.0050</i>	<i><0.001</i>	<i><0.001</i>	<i><0.001</i>
	<i>Porites</i>	0.6613	0.0120	0.0088	0.0834	0.2567	<i><0.001</i>	0.0574	0.1348
	<i>Pocillopora</i>	0.7716	0.2366	0.5804	0.0378	0.0249	<i><0.001</i>	0.0218	0.3802
	<i>Montipora</i>	0.8605	0.0476	0.0627	0.0534	0.8678	<i><0.001</i>	0.0835	0.2951
Line - Islands	Coral	0.9228	0.0703	0.9935	<i>0.0049</i>	0.0866	<i>0.0017</i>	0.0150	0.2427
	Macroalgae	0.3423	<i><0.001</i>	0.0156	<i>0.0023</i>	<i><0.001</i>	<i><0.001</i>	<i><0.001</i>	<i>0.0021</i>
	Turf algae	0.6993	<i><0.001</i>	<i><0.001</i>	<i><0.001</i>	<i><0.001</i>	0.8518	0.0436	<i>0.0015</i>
	CCA	0.7255	0.5407	<i><0.001</i>	<i><0.001</i>	<i><0.001</i>	0.3079	<i><0.001</i>	<i><0.001</i>
	<i>Acropora</i>	0.6352	0.6156	0.2216	0.1789	0.0671	0.7429	0.7618	0.3731
	<i>Pocillopora</i>	0.1856	0.6752	0.4232	0.1678	0.5183	<i><0.001</i>	0.2320	0.0153
	<i>Porites</i>	0.0083	0.5623	0.9389	0.6850	0.4432	0.0077	0.1414	<i><0.001</i>
Nanwan Bay	Coral	0.9520	0.0165	<i><0.001</i>	<i><0.001</i>	<i><0.001</i>	0.0073	0.0715	<i><0.001</i>
	Macroalgae	0.1168	<i><0.001</i>	0.0500	<i><0.001</i>	0.4438	0.3934	<i><0.001</i>	0.6704
	Turf algae	<i><0.001</i>	<i><0.001</i>	<i><0.001</i>	<i><0.001</i>	<i><0.001</i>	<i><0.001</i>	<i><0.001</i>	<i><0.001</i>
	CCA	0.3291	<i><0.001</i>	0.0557	<i><0.001</i>	0.2894	<i><0.001</i>	<i><0.001</i>	0.0066
	<i>Montipora</i>	0.0344	0.2066	0.0187	0.3763	<i><0.001</i>	0.2128	0.2564	0.9279
	<i>Pocillopora</i>	<i><0.001</i>	<i><0.001</i>	<i><0.001</i>	<i><0.001</i>	<i><0.001</i>	0.8761	0.0398	<i><0.001</i>
	<i>Favites</i>	0.2577	0.3682	0.2311	<i><0.001</i>	0.0555	0.0475	<i><0.001</i>	0.0122
Heron Reef	Coral	<i>0.0040</i>	0.0287	0.0332	<i><0.001</i>	<i><0.001</i>	<i>0.0022</i>	<i><0.001</i>	<i><0.001</i>
	Macroalgae	0.0628	0.3281	0.3338	<i><0.001</i>	0.1397	0.4804	0.1386	<i><0.001</i>
	Turf algae	0.9951	<i>0.0028</i>	0.1720	0.0108	<i><0.001</i>	0.8682	<i>0.0005</i>	0.0198
	CCA	<i><0.001</i>	<i><0.001</i>	0.1384	<i><0.001</i>	<i>0.0017</i>	<i>0.0043</i>	<i>0.0061</i>	<i><0.001</i>

Appendix

3.A Details of Coral Reef Survey Locations

The acquisition and analysis methods used in the coral surveys, which provided images and Archived annotations for the present study, are detailed in this section.

3.A.1 Moorea

In Moorea, sampling took place in fringing (2–5 m depth) and outer (10-m and 17-m depth) reef habitats, with two sites on each of the three shores of the island [1]. In each habitat, images (each 0.25 m²) were recorded at ~ 40 points scattered randomly along a 40 m transect, and were recorded with a Nikon D70 camera (6.24 megapixels) attached to two strobes and mounted to a frame holding the camera perpendicular to the reef. With this sampling scheme, ~ 720 images are recorded annually, and the resolution of the camera allows objects as small as ~ 1 cm to be resolved. Over the last 8 years, images have been analyzed using Coral Point Count with Excel Extensions (CPCe) software [70], in which a grid of 200 randomly-located dots is superimposed on each image, and a human operator identified the taxon beneath the dots. The organisms covering the coral reefs are resolved to four functional groups: scleractinian corals, macroalgae [algae = 1 cm high], algal turf [algae < 1 cm high], and crustose coralline algae (CCA), with the scleractinians further resolved to genus.

3.A.2 Northern Line Islands

Images from the Line Islands were recorded on outer reef communities (10-12 m depth) at Kingman, Palmyra, Tabuaeran and Kiritimati atolls in 2005. On each island, ten sites were surveyed with ~ 200 images per site, and at each site two 25 m transects were deployed with 10 randomly placed photoquadrats (0.54 m²) along each transect. No strobes were used, and the white balance of the camera was set prior to shooting using a white reference. Images were collected with an Olympus C-7070 camera (7.1 megapixels) in an underwater housing, and were edited for brightness and contrast in Photoshop. Benthic analysis was performed using PhotoGrid 1.0 with 100 points per photograph. Organisms were identified to genus when possible, but

genus-level annotations were pooled to six groups: scleractinian coral, soft coral, macroalgae, CCA, turf algae, and “other”.

3.A.3 Nanwan Bay, Taiwan

A total of 890 images were recorded in southern Taiwan at two sites in Nanwan Bay (Houbihu and Outlet) on 3 occasions during 2011 and 2012, and at 5 sites around Liuchiu Island during 2007 and 2011. At each site (all 2–5 m depth), ~ 30 photoquadrats (0.123 m^2) were recorded at random locations along three 10 m transects [131]. Images were recorded with a Canon G12 camera (9.98 megapixels) in an underwater housing mounted to a frame holding the camera perpendicular to the reef. The percentage cover of each benthic category was determined using CPCe software with 50 points distributed randomly on each photo and then assigned to 154 benthic cover types [70].

3.A.4 Heron Reef, Great Barrier Reef

Heron Reef images were recorded coinciding with satellite imagery as part of an annual mapping survey in 2007 [108]. Geo-referenced photoquadrats (1 m^2) were recorded at $\sim 3 \text{ m}$ intervals along transects using a Canon A540 camera (6.2 megapixels) held 0.5 m above the benthos. The position of each photoquadrat was geo-referenced using a GPS floating at the surface and towed behind the operator. Thirty-two transects were censused, each between 200 and 1000 m in length for a total of $\sim 13 \text{ km}$ (3,500 photos). The benthic cover category for each photo was determined by randomly distributing 24 points on each photograph, then manually assigning each point to one of 92 benthic cover types using CPCe [70]. The coral categories were resolved to functional group (e.g. branching coral) instead of genus as in the other three locations.

3.B Classification Using Linear Support Vector Machines

A binary linear Support Vector Machine [29] solves the following unconstrained optimization problem,

$$\min_{\omega} \left[\frac{1}{2} \omega^T \omega + C \sum_{i=1}^n \xi(\omega; x_i, z_i) \right]$$

Figure 3.A.1. Sample photoquadrats drawn from the long-term coral reef projects that served as test data for the present analysis. (First row) 4 photoquadrats (50 × 50 cm) from Moorea; (second row) 3 photoquadrats (65 × 90 cm) from the Line Islands; (third row) 4 photoquadrats (35 × 35 cm) from Nanwan Bay (Taiwan); (bottom row) 3 photoquadrats (50 × 65 cm) from Heron Reef (GBR).

where $z_i \in \{-1, 1\}$ is a binary label, x_i is a data sample, $C > 0$ is a regularization parameter and

$$\xi(\omega; x_i, y_i) = \max(1 - z_i \omega^T x_i, 0) \quad (3.2)$$

is a loss function. The weight vector $\omega \in \mathbb{R}^d$ is the output of the learning algorithm and is used to classify new, unseen sample, x_u as

$$z_u = \text{sign}(\omega^T x_u). \quad (3.3)$$

Multiclass classification can be mapped to binary classification through a one-versus-rest classification scheme. In this scheme a weight vector ω_m is trained separately for each class m , by converting the multiclass labels $y_i \in \{1, \dots, 20\}$ to binary labels $z_i = \mathbb{S}(y_i, m)$, $i \in 1, \dots, n$ where $\mathbb{S}$ is the signed indicator function

$$\mathbb{S}(a, b) = \begin{cases} 1 & \text{if } a=b \\ -1 & \text{if } a \neq b \end{cases} \quad (3.4)$$

Classification is then done by selecting the class for which the score $s_u(m) = \omega_m^T x_u$, has the highest value

$$y_u = \text{argmax}_m [s_u(m)]$$

3.C Importance of Training Size for Alleviate

In this analysis the accuracy of Alleviate, in terms of κ_{coral} , was measured as a function of training data size. (As defined previously, κ_{coral} is the Cohen's Kappa [22] score of classifying coral versus non-corals.) As detailed in the paper, after randomly selecting the 200 images in the Evaluation Set, the remaining data was referred to as the Reference Set, and these data was used to train the machine-learning algorithm. The total number of training samples available for each location is detailed in Table 1, and varied between 34,260 and 94,200. In 30 repeated experiments, a random subset of increasing size was selected from this training data, so that *e.g.* $94,200 \times \frac{i}{30}$

samples were used in iteration i for Moorea. This data was used to train an automated annotator, and generate Machine annotations on the Evaluation Set. These annotations, together with the Host annotations were used to calculate κ_{coral} for 30 levels of alleviation ($\lambda = 0, \dots, 1$). This procedure was repeated 30 times for each location, and the mean results for each training set size and level of alleviation is shown in Fig. 3.C.1. The results are normalized so that κ_{coral} for the Host annotator of each location is 1. The level curves thus indicate what proportion of points can be classified by the Machine while maintaining a certain accuracy ratio compared to the Host. For example, with a 5% decrease in κ_{coral} , $\sim 10,000$ training samples are required in Moorea for 40% alleviation, and $\sim 80,000$ for 50% alleviation. The left side of Fig. 3.C.1 shows that the accuracy of the Machine annotations increases as more training samples become available, but that the increase-rate decreases as additional training samples are added. On the right side of Fig. 3.C.1 are the 95%-level curves for each location plotted on the same axis for comparison. The machine performance on Nanwan Bay and Heron Reef is significantly lower than for Line Islands and Moorea probably because of the difference in photographic quality between the locations. This analysis indicates that machine classification performance saturates around 50,000 training samples. For this amount of training data, 40–60% alleviation results in 95% of max κ_{coral} , and 50–80% alleviation result in 90% of max κ_{coral} .

3.D Refine: a Supplementary Operation Mode

Label-sets of benthic surveys can include over 100 labels when fine taxonomic resolution is required. In such situations, point annotation tools such as CPCe suffer due to limited visual resolution and tedious manual distinction among categories. We define an interactive annotation mode, Refine, that addresses these situations. As denoted in the paper, $s_{i,k}(m)$ is the score given to class m by the automated classification algorithm for a certain image i and point k ; these scores are used also by Refine. A set of t labels (where t is smaller than the total number of labels) with the highest scores is retrieved from the full label-set and displayed to the human operator. The human operator selects the correct label from the retrieved set, or asks to see more labels. This allows the operator to quickly identify the correct label, and avoids the problems associated with

Figure 3.C.1. Training set size analysis. Left: level curves indicating the accuracy of Alleviate at various levels of alleviation and training set sizes. Accuracy is expressed as the normalized κ_{coral} score (so that κ_{coral} of the Hosts is 1). Right: 95% level curves for the four locations drawn on the same axis. Note how the better quality photographs of Moorea and Line Islands enable stronger scores for the same amount of training data compared to Nanwan Bay and Heron Reef. Also note how the main gain in alleviation level is achieved, in general, by the first 50k training samples.

limited screen-space. In the case of $t = 1$, the operator can rapidly verify if the sole label is correct, or else select from the full label-set.

We conducted an experiment to measure to what extent the retrieved set of Refine contains the correct label. For each point in each image, the t labels with highest scores were retrieved. If the Host annotation for that point was among the retrieved labels, it was assigned to that point; otherwise the retrieved label with the highest score was assigned. The full set of annotations thus derived were compared to the Archived annotation to calculate Cohen's kappa (κ) [22]. This procedure was repeated for each t until all labels were in the retrieved set, at which point $\kappa \approx 70\%$. The κ scores are plotted against t in Fig. 3.D.1. The retrieval set sizes required for a 5% drop in κ compared to the Hosts were 5 Moorea, 7 for the Line Islands, and 16 for Nanwan Bay. This indicates this strong ability of Refine include the correct labels in the retrieved set. Since the label-set used for Heron Reef only contain 5 labels, it was not included in this experiment.

Figure 3.D.1. Performance evaluation of the Refine operational mode. Cohen's kappa (κ) for various size retrieval sets of Refine for Moorea, Line Islands and Nanwan Bay. The κ was calculated by comparing the joint set of Host and Machine annotations to the Archived annotations. The black x on each curve indicates the point where κ is 5% lower than its maximum value (i.e. a 5% drop compared to the κ of the Hosts). Note how the better Machine annotation accuracy for Moorea and Line Islands enable smaller sets of labels to be retrieved compared to Nanwan Bay.

Acknowledgments

This work was supported by the US National Science Foundation (NSF) Division of Ocean Sciences (OCE) 0941760 to BGM. We gratefully acknowledge the support of the U.S. National Science Foundation's Long Term Ecological Research (LTER) program and the Moorea Coral Reef (MCR) LTER Site (through OCE 0417412, 1026851 and 1236905).

Chapter 3 is a reformatted version of "*Towards automated annotation of benthic survey images: variability of human experts and operational modes of automation*", Oscar Beijbom, Peter J. Edmunds, Chris Roelfsema, Jennifer Smith, David I. Kline, Benjamin Neal, Matthew J. Dunlap, Vincent Moriarty, Tung-Yung Fan, Chih-Jui Tan, Stephen Chan, Tali Treibitz, A. Gamst, B. Greg Mitchell, David Kriegman. PLoS ONE, May 2015. The dissertation author was the primary investigator and author of this paper.

Chapter 4

Cost-Effective Sampling Designs for Automated Annotation

Chapter 4 is a reformatted version of "*Random Sampling in an Age of Automation: Minimizing Expenditures through Balanced Collection and Annotation*" by Oscar Beijbom, which is in submission.

Chapter 4 considers automated annotation in the context of random sampling surveys. The goal of ecological surveys is often some aggregated statistic of the data, i.e. the percent coral cover of a certain region. In such situations, one can allow some annotation errors as long as the final ecological estimate is correct. This chapter investigates whether the noise of an automated annotator can be ameliorated by collecting and annotating additional samples, and if so; under what cost-structures it is beneficial to do so. Chapter 4 is similar in spirit to the second part of Chapter 3, in that it considers a blend of expert and automated annotators, but it is more general and a more rigorous theoretical analysis. It is also different in that it models sample collection in conjuncture with annotation while Chapter 3 strictly models annotation.

Abstract

Methods for automated collection and annotation are changing the cost-structures of sampling surveys for a wide range of applications. Digital samples in the form of images, audio recordings or electronic documents can be collected cheaply, and computer programs or crowd workers can be utilized to provide cheap annotations of collected samples. We consider the problem of estimating a population mean using random sampling under these new cost-structures and propose a novel ‘hybrid’ sampling design. This design utilizes a pair of annotators: a primary, which is accurate but costly (e.g. a human expert) and an auxiliary which is noisy but cheap (e.g. a computer program), in order to minimize the total cost of collection and annotation. We show that hybrid sampling is applicable under a key condition: that the noise of the auxiliary annotator is smaller than the variance of the sampled data. Under this condition, hybrid sampling can reduce the amount of primary annotations needed and minimize total expenditures. Experiments were conducted on two marine ecology data mining applications, where computer programs utilized in hybrid sampling designs reduced the total cost by 50 – 79% compared to sampling designs that relied only on human annotators. In addition we define an ‘auxiliary’ sampling design which can be very cost-effective, but requires *a priori* knowledge of the auxiliary annotator misclassification rates.

4.1 Introduction

Using random sampling to estimate the mean of a population is a fundamentally important method to science and society at large, and has been studied extensively [130, 37]. Deployment of any random sampling design requires collection of some number of observations sampled randomly from nature. This was traditionally done *in situ* by an expert. Recently, advances in robotics, sensor technology, digital storage, and information technology have enabled rapid collection of samples in digital format, such as images [73, 110, 61], audio [64], or electronic documents [103]. The popularity of digital sample collection can be attributed to three key factors: it creates a permanent record; it can be done cheaply using automated sampling vehicles or non-expert personnel; and it is generally fast. However, such samples (e.g. a photoquadrat of the

forest floor) typically require annotation by an expert in order to reveal the desired quantity of interest (e.g. a count of insects). Such annotation work can be slow, tedious, expensive, and prone to error [80, 96].

Concurrent with the development of automated collection methods, advances in computer-vision [8], computer-audition [5], and document-analysis [83] have enabled automation of said annotation work. Such methods often rely on machine learning where expert annotated archived data sets are utilized to train automated annotators. Automation is a compelling low-cost alternative to expert annotations, but it's generally less reliable and may be biased [8, 54]. This is particularly problematic if the probability density of the archived data differs from the density of the data to be sampled [101].

Crowdsourcing offers another low-cost alternative to expert annotation for e.g. document or image annotation [126, 97]. Crowdsourced annotations can be noisy, and much work has been devoted to improving the quality of such annotations. This is generally done either by carefully designing the tasks given to the crowd workers [97], or by collecting multiple crowd annotations for the same sample and then modeling, and compensating for, the annotation errors [143, 105].

We consider the problem of estimating a population mean under these new cost-structures of data collection and annotation. This is formalized as follows: Given is a procedure for collecting random samples, $x_i \in \mathcal{X}$ (e.g. images, audio recordings, or documents), each with an associated quantity of interest (value), $y_i \in \mathcal{Y} \subseteq \mathbb{R}$; and two annotators: a primary, $f_a(x_i) \in \mathcal{Y}$ which is accurate but costly (e.g. a human expert), and an auxiliary, $f_b(x_i) \in \mathcal{Y}$ which is cheap but noisy (e.g. a computer program, lay-person, or crowd worker). Our goal is to derive a sampling designs that achieves unbiased estimates of the population mean ($E[y_i]$) at a target error and confidence, while minimizing the Total Sampling Cost, which is the total cost of collection and annotation. In particular, we ask what is the optimal balance between the number of samples annotated by the primary and auxiliary annotators? This work is, to the best of our knowledge, the first to consider this problem.

A key challenge is to define a procedure that can correct for the potential bias of the auxiliary annotator. This is difficult as we cannot assume any prior knowledge of the underlying

probability density from which the samples are drawn. Indeed, if this density was known, the population mean could be estimated directly, making the sampling work unnecessary. If the auxiliary annotator is based on machine learning and trained on archived data with a different probability density, the problem of transfer learning arises [101] for which the generalization bounds of statistical learning theory generally do not apply [138]. Methods for bias-correction have been proposed independently by Solow et. al. [127] and by Hopkins and King [59], that do not require knowledge of the full underlying probability density (of the sampled data) but only of the conditional probabilities of a label given a sample. However, as we shall see, this information may not always be at hand, particularly if the auxiliary annotator is based on machine learning.

A key insight of our work is that an offset estimator can be utilized that directly estimates the offset (bias) of the auxiliary annotator. The offset estimator requires a subset of the samples to be annotated by both annotators, and is unbiased if the samples are independent and identically distributed, which they are by construction under random sampling. We incorporate the offset estimator in a sampling design which we denote hybrid sampling. Population mean estimates based on hybrid sampling are unbiased, and we show that the cost of hybrid sampling can be significantly lower than for a conventional design that utilizes the primary annotator only. These savings are possible under general, and precisely defined, conditions on the auxiliary annotator.

After introducing notation and formalizing the problem, we recall the conventional sampling design where all collected samples are annotated by the primary annotator. We then introduce and discuss the hybrid sampling design as well as an auxiliary sampling design based on the methods of [127, 59]. We finally compare the utility of the designs on two marine ecology data mining applications: coral reef and plankton surveys.

4.1.1 Related Work

Our work is related to the literature on survey sampling with auxiliary data [130, 115]. In that context, “auxiliary data” is typically not a direct estimate of the variable of interest but some other, related quantity. For example, if the variable of interest, y_i is the number of animals per plot, the auxiliary data can be the plot area, vegetation type, or plot elevation. In this literature,

auxiliary data is incorporated using a ratio estimator which can reduce the sample size if there is an approximately linear relationship between the auxiliary data and the variable of interest [130]. However, the ratio estimator is design-biased, and analysis of the estimator variance typically assumes that the auxiliary data is available for the whole population [130, 115]. In this work, in contrast, we do not assume that the auxiliary data is available for all samples, but that this data can be acquired at a cost by collecting and annotating additional samples (using the auxiliary annotator). This additional cost is then taken into account when deriving optimal sampling sizes. Another key difference is that we do not use the ratio estimator but an offset estimator which directly estimates the bias of the auxiliary annotator. The offset estimator is design-unbiased and allows for a more straight-forward analysis.

Another line of related work utilize stratified random sampling [13], importance sampling [119] or generative models of the classifier score distribution [53, 144], to achieve cost-effective estimates of classifier performance on new data. However, these methods typically consider classifier performance on a fixed set of data. In contrast, we include the data collection in our model which enables e.g. noisy classifier annotations to be ameliorated by processing more data.

Our work is also related to active [27], and transfer learning [101]. It is related, in particular, to recent work on active transfer learning where labels are queried to optimize classifier performance in a target domain [142]. The key difference between this work and ours is that active learning methods optimize the labeling effort to create the best *classifier* (which then can, presumably, be used to label more data and in order to estimate the desired data-products). In contrast, we directly optimize the labeling effort to derive the desired *data-product* (i.e. the population mean).

4.2 Problem Setup

We denote by $\mu_p \equiv E[y_i]$ and $\sigma_p^2 \equiv \text{var}[y_i]$ the first and second moments of the (unknown) probability density function of the values. We further denote by $f_a : \mathcal{X} \rightarrow \mathcal{Y}$ the primary, and by $f_b : \mathcal{X} \rightarrow \mathcal{Y}$ the auxiliary annotator, $\epsilon_{a,i} \equiv f_a(x_i) - y_i$ the error of f_a on sample i , $\mu_a \equiv E[\epsilon_{a,i}]$

and $\sigma_a^2 \equiv \text{var}[\varepsilon_{a,i}]$. Similarly, $\mu_b \equiv E[\varepsilon_{b,i}]$ and $\sigma_b^2 \equiv \text{var}[\varepsilon_{b,i}]$.

We denote by n_a and n_b the number of samples annotated by f_a and f_b , respectively. The number of collected samples is given by $\max(n_a, n_b)$ since samples needs to be annotated to provide any information, and conversely, needs to be collected in order to be annotated. We denote by c_c, c_a and c_b the cost per sample for collection, annotation by f_a and annotation by f_b , respectively. The ‘accurate and expensive’ characteristics of the primary annotator are operationalized by letting $\sigma_a^2 < \sigma_b^2$ and $c_a > c_b$.

We can now precisely state the goal of this work: Given costs c_c, c_a and c_b , and two annotators, f_a and f_b , derive a sampling design that estimates the population mean, μ_p by defining the number of annotated samples (n_a and n_b), so that $E[\hat{\mu}_p] = \mu_p$ and $\Pr(|\hat{\mu}_p - \mu_p| > d) < \delta$, for a target error, d and confidence, δ . The utility of the sampling design is evaluated by the Total Sampling Cost (TSC)

$$t(n_a, n_b) = c_a n_a + c_b n_b + \max(n_a, n_b) c_c. \quad (4.1)$$

We make three assumptions. First, we assume that the number of collected samples is small in comparison with the total size of the population. This allows us to omit the finite-population correction factor [130]. Second, we assume that the primary annotator, f_a is unbiased, i.e. $\mu_a = 0$. Third, we assume that the annotator errors $\varepsilon_{a,i}$ and $\varepsilon_{b,i}$ are uncorrelated with y_i .

4.3 Conventional Sampling

We denote by ‘conventional’, a sampling design where all collected samples, x_i are annotated by the primary annotator f_a , i.e. $n_b = 0$. In such design, an unbiased estimator of μ_p is given by

$$\check{\mu}_p = \frac{1}{n_a} \sum_{i=1}^{n_a} f_a(x_i), \quad (4.2)$$

with variance

$$\text{var}[\check{\mu}_p] = \frac{1}{n_a}(\sigma_p^2 + \sigma_a^2). \quad (4.3)$$

The variance $(\sigma_p^2 + \sigma_a^2)$ is often unknown, but can be estimated from $f_a(x_1), \dots, f_a(x_{n_a})$. The sample size, n_a needs to be large enough to ensure that $\Pr(|\check{\mu}_p - \mu_p| > d) < \delta$, for a target error d and confidence δ . By standard results, this is satisfied when

$$\zeta_\delta \sqrt{\text{var}[\check{\mu}_p]} \leq d, \quad (4.4)$$

where ζ_δ is the upper $1 - \delta/2$ point on the standard normal distribution curve [130]. The target sample size is given by inserting (4.3) into (4.4) yielding

$$\check{n}_a = \frac{\zeta_\delta^2}{d^2}(\sigma_p^2 + \sigma_a^2), \quad (4.5)$$

for a TSC of

$$t(n_a, n_b) = (c_c + c_a)\check{n}_a. \quad (4.6)$$

4.4 Hybrid Sampling

Now consider a design where $n_b \geq \check{n}_a$ samples are collected and annotated by the auxiliary annotator f_b , and where a subset $n_a \leq n_b$ is also annotated by the primary annotator f_a . An offset estimator of μ_p under this design is given by

$$\hat{\mu}_p = \frac{1}{n_b} \left[\sum_{i=1}^{n_a} f_a(x_i) + \sum_{i=n_a+1}^{n_b} (f_b(x_i) - \hat{\mu}_b) \right], \quad (4.7)$$

which can be rewritten on the simpler form

$$\hat{\mu}_p = \frac{1}{n_b} \sum_{i=1}^{n_b} f_b(x_i) - \hat{\mu}_b. \quad (4.8)$$

The offset estimator is unbiased and an unbiased estimate of μ_b is given by

$$\hat{\mu}_b = \frac{1}{n_a} \sum_{i=1}^{n_a} f_b(x_i) - f_a(x_i). \quad (4.9)$$

The variance of $\hat{\mu}_p$ can be calculated from first principles as

$$\text{var}[\hat{\mu}_p] = \frac{1}{n_b} (\sigma_p^2 - \sigma_b^2) + \frac{1}{n_a} (\sigma_a^2 + \sigma_b^2). \quad (4.10)$$

If the costs are such that a large number of samples, n_b can be collected and annotated by the auxiliary annotator, the magnitude of the first term in (4.10) becomes small and the sampling error depends mainly on the auxiliary annotation error, σ_b^2 and the number of samples annotated by the primary annotator, n_a . In contrast, the conventional sampling design depends mainly on the data variance, σ_p^2 and n_a . Thus, if the auxiliary annotation error is smaller than the data variance, the hybrid sampling error can be smaller than the conventional sampling error. This is formalized in the following theorem:

Theorem 4.1. *For any fixed $n_b > n_a$, $\text{var}[\hat{\mu}_p] < \text{var}[\check{\mu}_p]$ if and only if $\sigma_b^2 < \sigma_p^2$.*

Proof. It follows from (4.3) and (4.10) that if $\sigma_b = \sigma_p \Rightarrow \text{var}[\hat{\mu}_p] = \text{var}[\check{\mu}_p]$. Since $\text{var}[\hat{\mu}_p]$ increases monotonically with increasing σ_b (since $n_b > n_a$) the theorem follows trivially. $\square$

Theorem 4.1 implies that the uncertainty introduced by f_b can be compensated for by using more samples if and only if $\sigma_b^2 < \sigma_p^2$. However, the additional collection of samples is only economical for certain cost functions, and should in the general case be determined by comparing the TSC of the two designs.

To determine the TSC of hybrid sampling we begin by deriving optimal sample sizes $\hat{n}_b$ and $\hat{n}_a$. By combining (4.10) and (4.4), and solving for equality, the following trade-off between

Figure 4.1. Amount of auxiliary (n_b), and primary (n_a) annotation required to achieve error $d \leq 0.02$ at 95% confidence for $\sigma_p = 0.2$, $\sigma_a = 0.02$, and $\sigma_b = \{0.25\sigma_p, 0.5\sigma_p, \sigma_p\}$ under the hybrid sampling design. Solid gray line indicates $n_b = n_a$. Optimal operating points for when the relative cost of primary annotation ($k = \frac{c_a}{c_c + c_b}$) is $k = 10$ and $k = 100$ are marked with X on the $\sigma_b = 0.5\sigma_p$ curve.

n_b and n_a is derived

$$n_a = \frac{\sigma_b^2 + \sigma_a^2}{\frac{d^2}{\xi_\delta^2} - \frac{1}{n_b}(\sigma_p^2 - \sigma_b^2)}. \quad (4.11)$$

Example trade-off curves (Fig. 4.1) demonstrate how the amount of primary annotations can be reduced by increasing the amount of auxiliary annotations. Note that if $n_b = \check{n}_a$, it follows from (4.11) that $n_a = \check{n}_a$, and the hybrid design reduces to the conventional design.

The optimal operating point along the n_b, n_a trade-off curve can be derived by minimizing the TSC. Using (4.11) to eliminate n_a , the TSC becomes

$$t(n_a, n_b) = (c_c + c_b) \left(n_b + \frac{k(\sigma_b^2 + \sigma_a^2)}{\frac{d^2}{\xi_\delta^2} - \frac{1}{n_b}(\sigma_p^2 - \sigma_b^2)} \right) \quad (4.12)$$

where

$$k = \frac{c_a}{c_c + c_b} \quad (4.13)$$

is the relative cost of f_a . The optimal sample size, $\hat{n}_b$ is given by minimizing (4.12) under the constraint that $n_b \geq \check{n}_a$. This yields the following theorem (proof in Appendix 4.A).

Theorem 4.2. *Under a hybrid sampling design (4.8), the optimal auxiliary annotation size is*

$$\hat{n}_b = \max \left[\frac{\zeta_{\delta}^2}{d^2} \left(\sigma_p^2 - \sigma_b^2 + \sqrt{k(\sigma_b^2 + \sigma_a^2)(\sigma_p^2 - \sigma_b^2)} \right), \check{n}_a \right] \quad (4.14)$$

The corresponding primary annotation size is given by (4.11).

Using the optimal sample sizes, the TSC of hybrid sampling can be calculated from (4.12) and compared to the TSC of conventional sampling in order to determine the most cost-effective design. For the important special case where $c_b = 0$ (which can occur e.g. if f_b is a computer algorithm) the following theorem applies (proof in Appendix 4.B).

Theorem 4.3. *If $c_b = 0$, the TSC of hybrid sampling is smaller than the TSC of conventional sampling if and only if $k > k_0$, where k is given by (4.13) and*

$$k_0 = \frac{\sigma_a^2 + \sigma_b^2}{\sigma_p^2 - \sigma_b^2}. \quad (4.15)$$

For example, if the primary annotation errors are zero ($\sigma_a^2 = 0$), and the auxiliary errors are half as large as the data variance ($\sigma_b^2 = \sigma_p^2/2$), then $k_0 = 1$. This means that the hybrid sampling is cheaper than conventional sampling if $k > 1$, which occurs if the cost of collection is smaller than the cost of primary annotation. The difference in TSC between the two sampling designs is shown in Fig. 4.2 for various values of k , σ_p and σ_b .

4.5 Auxiliary Sampling

A third sampling design can be defined under one critical additional assumption. This assumption, which we will denote the ‘population drift assumption’, is that the performance of f_b can be defined in terms of a matrix of confusion, Q (which excludes real-valued output spaces $\mathcal{Y}$), and that Q is known *a priori* for the data to be sampled.

For these derivations we will let $\mathcal{Y} = \{0, 1\}$, meaning that the samples y are drawn from $\text{Ber}(\mu_p)$, a Bernoulli distribution with mean μ_p . This corresponds to annotating each sample x_i as containing or not containing the quantity of interest. In Appendix 4.C we derive the statistics of

Figure 4.2. Difference in Total Sampling Cost ($TSC_{\text{conventional}} - TSC_{\text{hybrid}}$) for $\sigma_p = 0.2$, $\sigma_a = 0.02$, $\sigma_b = \{0.25\sigma_p, 0.5\sigma_p, 0.75\sigma_p\}$, $c_b = 0$ and $c_c = 1$, as a function of the relative cost of primary annotation: $k = c_a/(c_c + c_b)$. Both sampling designs are calibrated for error $d \leq 0.02$ at 95% confidence. The threshold costs, k_0 (4.15) are marked with black stumps. If $k < k_0$ the sampling designs and costs are identical

auxiliary sampling in a two-stage sampling design [130, 37], where each sample x_i is annotated by second stage sampling, from which the corresponding y_i is obtained.

A matrix of confusion, Q characterizes the misclassification rates of f_b . In the binary case, Q is a two by two matrix

$$Q = \begin{bmatrix} \alpha & 1 - \beta \\ 1 - \alpha & \beta \end{bmatrix},$$

where α is the sensitivity and β the specificity. As noted independently by [127, 59] Q can be used to create an unbiased estimate of y , and we begin by recalling this procedure. The expected value of the auxiliary annotation f_b of a sample x with value y is

$$E \begin{bmatrix} f_b(x) \\ 1 - f_b(x) \end{bmatrix} = Q \begin{bmatrix} y \\ 1 - y \end{bmatrix},$$

and an unbiased estimator of y is given by inverting Q

$$\tilde{y} = \frac{f_b(x) + \beta - 1}{\alpha + \beta - 1}. \quad (4.16)$$

We refer to this as the ‘abundance corrected’ value, and derive a sampling procedure based on

this correction. The variance of $\tilde{y}_i$, given the true value y_i is

$$\text{var}[\tilde{y}_i|y_i] = \frac{\text{var}[f_b(x_i)|y_i]}{(\alpha + \beta - 1)^2}, \quad (4.17)$$

which follows directly from (4.16). We also note that

$$\text{var}[f_b(x_i)|y_i] = y_i\alpha(1 - \alpha) + (1 - y_i)(1 - \beta)\beta, \quad (4.18)$$

since if $y_i = 1$, $f_b(x_i) \sim \text{Ber}(\alpha)$, and if $y_i = 0$, $f_b(x_i) \sim \text{Ber}(1 - \beta)$. Combining (4.17) and (4.18), yields

$$\text{var}[\tilde{y}_i|y_i] = \frac{y_i\alpha(1 - \alpha) + (1 - y_i)(1 - \beta)\beta}{(\alpha + \beta - 1)^2}. \quad (4.19)$$

Finally, $\text{var}[\tilde{y}_i]$ is given by the law of total variance

$$\text{var}[\tilde{y}_i] = E[\text{var}(\tilde{y}_i|y_i)] + \text{var}[E(\tilde{y}_i|y_i)] \quad (4.20)$$

$$= \sigma_s^2 + \sigma_p^2, \quad (4.21)$$

where σ_p^2 is the data variance and

$$\sigma_s^2 = \frac{\mu_p\alpha(1 - \alpha) + (1 - \mu_p)(1 - \beta)\beta}{(\alpha + \beta - 1)^2}, \quad (4.22)$$

the variance introduced by the abundance correction. If the classifier is balanced, i.e. $\alpha = \beta$, (4.22) simplifies to $\sigma_s^2 = \frac{\alpha(1-\alpha)}{(2\alpha-1)^2}$. Since $\tilde{y}_i$ is an unbiased estimator of y_i , we can achieve an unbiased estimation of μ_p as

$$\tilde{\mu}_p = \frac{1}{n_b} \sum_{i=1}^{n_b} \tilde{y}_i, \quad (4.23)$$

with variance, assuming that σ_s^2 and σ_p^2 are uncorrelated

$$\text{var}[\tilde{\mu}_p] = \frac{1}{n_b}(\sigma_s^2 + \sigma_p^2), \quad (4.24)$$

and sample size

$$\tilde{n}_b = \frac{\zeta_\delta^2}{d^2}(\sigma_s^2 + \sigma_p^2). \quad (4.25)$$

Finally, the TSC, since $n_a = 0$, is given by

$$t(n_a, n_b) = (c_c + c_b)\tilde{n}_b. \quad (4.26)$$

The auxiliary sampling design requires annotation of $\tilde{n}_b$ samples by the auxiliary annotator, but as it does not require any annotations by the primary annotator, the TSC can be low. The following theorem is given directly from (4.26) and (4.6).

Theorem 4.4. *For binary output spaces, $\mathcal{Y} = \{0, 1\}$, the TSC of auxiliary sampling is smaller than the TSC of conventional sampling if and only if*

$$k' > \frac{\sigma_p^2 + \sigma_s^2}{\sigma_p^2 + \sigma_a^2} \quad (4.27)$$

where $k' = (c_c + c_a)/(c_c + c_b)$.

If f_b is accurate then σ_s^2 is small and auxiliary sampling is cheaper than conventional sampling even for low primary annotation costs. For example, if $\alpha = \beta = 0.9 \Rightarrow \sigma_s^2 \approx 0.14$, $\sigma_p^2 = 0.04$, and $\sigma_a^2 = 0$, it suffices that k' is larger than 4.5, which is satisfied e.g. if $c_b = 0$ and $c_a > 3.5c_c$. If, on the other hand, $\alpha = \beta = 0.7 \Rightarrow \sigma_s^2 \approx 1.3$, k' must be larger than 33.5.

4.6 Experimental Verification

We performed two sets of experiments. First, the cost-structures, data variances, and performance of commonly used machine annotation algorithms were investigated for two marine

Table 4.1. Empirical evaluation of hybrid sampling for two marine ecology data-mining tasks. All results calculated for error, $d < 0.02$, and confidence $\delta = 0.05$. All costs are measured in time required by a human operator (s:seconds, m:minutes, h:hours).

Details		Variances			Costs (time)			Conv. sampl.		Hybrid. sampl.		
Task	$\mathcal{Y}$	σ_p	σ_a	σ_b	c_c	c_a	c_b	n_a	TSC	n_a	n_b	TSC
Coral cover	$[0, 1]$	0.27	0.01	0.12	1m	10m	0s	707	130h	229	1458	62h
copepod abund.	$\{0, 1\}$	0.5	0.0	0.22	0	5s	0s	2401	200m	501	32071	42m

data mining applications to determine the applicability of hybrid sampling. Second, simulations were carried out to verify our general assumptions and the applicability of auxiliary sampling.

4.6.1 Applicability of Hybrid Sampling

Coral Cover Estimation

Quantifying the benthic composition of the ocean floor is a cornerstone of understanding and monitoring marine ecosystems [47, 58]. Millions of benthic survey images are collected annually by research labs, reef managers, government agencies and non-profit organizations. We used data from the Moorea Coral Reef Long Term Ecological Research program which surveys six sites around the island of Moorea in French Polynesia each year. Approximately 700 images are collected annually by a research diver, and annotated by a coral reef ecologist. Annotations are done using random point sampling, in which the substratum is identified at 200 point locations randomly scattered on each image [70]. The point annotations are then used to approximate the percent coral cover, y_i in each image, x_i . This procedure of using random sampling to annotate each collected sample is commonly referred to as two-stage sampling [130, 37].

We consider the task of designing a survey to estimate the coral cover of Moorea during 2009. We use the data from 2008 to train a machine annotator, and to make *a priori* estimates of the variances (Table 4.1). The data variance, σ_p^2 was directly estimated as the sample variance of the coral cover from the 2008 data, and the error variances σ_a^2 and σ_b^2 were estimated through cross validation. The cost of collection, c_c was set to 1 minute as a diver typically collects 50 – 75 images per hour and the cost of expert annotation, c_a was set to 10 minutes as annotation typically

requires 5 – 30 minutes per image.

We used the method of [8] to encode color and texture at the 200 point locations in each of the images. A linear binary Support Vector Machine [43], f'_b was used to identify coral versus other substrates in each point location and the machine annotator coral cover estimate, $f_b(x_i)$, was the normalized count of the 200 point annotations by f'_b .

As shown in Table 4.1, a hybrid sampling design can reduce the required manual effort by 50% which is equivalent to 60h per year. The actual reduction, however, may be larger. Sample sizes derived from Theorem 4.2 rely on a linear approximation of the cost per collected photograph. In reality, however, the cost per photograph is likely to decrease as more photographs are captured (due to a relatively lower amount of time spend in transition). Since hybrid sampling requires more photographs to be captured, the estimated reduction is a lower bound on the actual reduction. Additionally, as the collection process of reef surveys becomes increasingly automated, and as new methods for automated annotation becomes available [124], the cost-savings of a hybrid approach will increase.

Plankton Abundance

Monitoring plankton abundances is another important data mining task in the marine ecology community [99, 128]. Images are commonly captured using an imaging-in-flow cytometer [99], and annotated by an expert or by an automated annotation system. Here we consider data from the Scripps Phytoplankton Camera¹, which collects tens of thousands of *in-situ* plankton images each day. We consider the task of estimating the daily abundance of a plankton subclass, *copepod* from the collected images. Using the method of [128] a machine classifier is trained using 5500 manually annotated images selected at random, and the machine error variance, σ_b was estimated through cross-validation. Manual annotation requires approximately 5s per image and we assume that $\sigma_a = 0$. Since the costs of collection and automated annotation are zero, we set $k = 1000$ in order to calculate the hybrid sampling sizes (4.14). The annotation task is binary in this case, so $y_i \sim \text{Ber}(\mu_p)$, and since the *copepod* abundance varies between days, we let $\sigma_p^2 = 0.25$ since this is an upper bound on $\text{var}[y_i]$.

---

¹<http://jaffeweb.ucsd.edu:5001/>

As shown in Table 4.1, the hybrid design reduced the required manual effort by 79% which is equivalent to 2.5 hours per day. This reduction is an upper bound since the highest possible data variance, $\sigma_p^2 = 0.25$ was used in the derivations. However, even with a smaller data variance, $\sigma_p^2 = 0.09$, which corresponds to $\mu_p = 90\%$ (or $\mu_p = 10\%$), the savings of using hybrid sampling are still 43% (32 minutes) per day. Since the expert-annotated samples are selected randomly these can readily be used to improve the accuracy of the machine annotator for the next day, which can further reduce the manual annotation effort.

Figure 4.1. Sample coral survey images. Coral survey images from the Moorea Coral Reef Long Term Ecological Research program.

4.6.2 Validity of Assumptions

A Priori Estimates of Sample Size

To determine the sample size of any random sampling survey the variance of the data, σ_p^2 needs to be estimated *a priori*. This prior knowledge could be based on data collected at a previous occasion or at a different location. If σ_p^2 is erroneously estimated, the sample size will either be too small (yielding larger errors than desired) or too large (resulting in a unnecessarily high cost). However, the conventional sampling design will still yield an unbiased estimate of μ_p . For the hybrid sampling design, the error variance, σ_b^2 also needs to be estimated. If f_b is based on machine learning this can be difficult, since subtle differences between the underlying probability densities of the archived and the sampled data may affect the machine performance and thus σ_b^2 . However, it is important to note that even though an erroneous *a priori* estimate of σ_b^2 may result in too few or too many samples being collected and annotated, the hybrid estimator,

like the conventional estimator, is unbiased.

In order to validate the sampling designs suggested in Section 4.6.1, simulations were carried out on 695 coral survey images collected and annotated during 2009. Using the *a priori* estimates (from the 2008 data) given in Table 4.1, sample sizes were determined for twenty budgets between 10 and 100 man-hours. For each budget, and for 1000 iterations, the required number of images were drawn randomly with replacement², and the mean estimates of μ_p were calculated. These 1000 mean estimates were used to calculate sample variances $\widehat{\text{var}}[\check{\mu}_p]$ and $\widehat{\text{var}}[\hat{\mu}_p]$ as well as sample means $\widehat{E}[\check{\mu}_p]$ and $\widehat{E}[\hat{\mu}_p]$.

Our simulations show that the hybrid mean estimator is unbiased, and that the cost of hybrid sampling is lower (Fig. 4.3 A-B).

Figure 4.2. Sample plankton images. Image of *copepod* (left) and other (right) from the Scripps Institution of Oceanography Phytoplankton Camera.

Auxiliary Sampling

Cross validation on the coral images from 2008 was used to estimate σ_s : $\alpha = 0.72, \beta = 0.94, \mu_p = 0.31 \Rightarrow \sigma_s \approx 0.48$ (4.22). Because of the two stage sampling design used in the coral survey, the variance of auxiliary sampling is given by

$$\text{var}[\check{\mu}_p] = \frac{1}{n_b} \left(\frac{1}{s} [\sigma_s^2 + \mu_p(1 - \mu_p)] + \sigma_p^2 \right), \quad (4.28)$$

²Sampling with replacement was used to avoid finite-population artifacts due to the limited pool of 695 image. In an actual application of this method, there would be a very large number of locations from where to capture the images and finite-population correction would not be needed.

Figure 4.3. Simulations on the coral reef image dataset. Mean coral cover estimation (A) and estimation standard errors (B) as estimated from 1000 repeated experiments on the coral survey data from the island of Moorea in French Polynesia. (C) Scatter plot of true percent coral cover (y_i) versus machine estimation errors ($\epsilon_{i,b}$) ($r^2 = 0.15$).

where $s = 200$ is the number of secondary samples (SI Appendix 4.C). Using costs and σ_p^2 as estimated in Section 4.6.1, gives a sampling size $\tilde{n}_b \approx 720$ samples, and a TSC of 12 hours. This is a remarkably low cost, but it can only be realized if the population drift assumption holds. To investigate this, the auxiliary sampling design was added to the simulations detailed in the previous section. These simulations indicate that the estimated Q was not valid for the coral images from 2009, and the coral cover was underestimated by 2.4 pp, a 17.4% difference from the ground truth (Fig. 4.3 A).

The population drift assumption is that the matrix of confusion, Q for the data to be sampled (in this case the data from 2009) is known *a priori*. In absence of better options, Q was estimated by cross validation using the data from 2008. Since the same camera-equipment was used and the same sites were visited both years, there are reasons to believe that the population drift assumption would be satisfied. However, there are also several reasons why it would not be. First, the ecology changes over time which affects the substrate appearance. Second, water turbidity varies due to currents, plankton, or algal blooms etc. which affects image colors and contrasts. Third, the settings of the camera and artificial lightning (strobes) may have differed.

We believe that similar difficulties can be expected in other applications where f_b is based on machine learning. If the underlying probability densities differ between the archived data (on which the machine is trained and Q estimated) and the data to be sampled, the matrix of

confusion may differ significantly resulting in biased estimates [127]. However, this effect may be less severe if f_b is a layperson or a crowd-worker. Human perception is generally more robust to shifts in appearances than a machine learning algorithm, and the matrix of confusion estimated on the archived data may be valid for the data to be sampled.

Correlation of Errors

We have assumed that the errors $\epsilon_{a,i}$ and $\epsilon_{b,i}$ of the annotators are not correlated with y_i . These correlations are related to the accuracy of the annotators. If the accuracy is 100%, the errors are all zero and there is no correlation. One would therefore expect the errors of the primary annotator, f_a to be uncorrelated with the sample values. If on the other hand the accuracy is low, there may be a correlation, and this needs to be verified experimentally. Note that such correlation does not render the hybrid sampling design invalid, but the covariance between $\epsilon_{b,i}$ and y_i must be taken into account when calculating the error of $\hat{\mu}_p$.

To investigate the correlation between errors and values, the machine errors, $\epsilon_{i,b}$ were plotted against the true coral covers, y_i for each of the coral survey images collected during 2009 (Fig. 4.3 C). According to these simulations, the r^2 correlation using a linear model was 0.15 between $\epsilon_{b,i}$ and $y_i, i = 1, \dots, n_a$. This indicates a low correlation and thus the validity of (4.10) for this application.

4.7 Discussion

Before concluding, we discuss two aspects of the proposed method.

4.7.1 Sampling under Domain Shift

As mentioned in the introduction, the underlying probability density of the sampled data is likely to differ between the sampled (target) data and the archived data. Such differences are commonly referred to as domain shifts and, as mentioned previously, may cause difficulties in particular if the auxiliary annotator is based on machine learning. We briefly discuss the implications of these difficulties.

If the population drift assumption holds, i.e. the domain shift is solely due to shifting base-rates, $\Pr(y_i)$, then the abundance correction methods of [127, 59] can be used and the sampling can be designed to use the auxiliary annotator only. However, as ours and others experiments have indicated [132, 128], this is unlikely to occur, which renders the auxiliary sampling design biased and unpredictable. A domain shift will also affect the hybrid sampling design. Specifically, the *a priori* estimate of σ_b may not be correct for the target data. However, as our results indicate, this effect is less severe for the offset estimator at the heart of the hybrid sampling design, than for the abundance correction of [127, 59]. Also, the hybrid mean estimator, $\hat{p}$ is unbiased regardless of the domain shift.

If the domain shift between the archived and target data is sufficiently large, it may be necessary to first determine the error rates on the target data. This can be achieved e.g. by stratified sampling of [13], or by fitting a parametric model to [53, 144], the classifier output scores. After the target error rates have been determined, appropriate sampling sizes can be calculated using Theorem 4.2. However, this strategy is viable only in situations where additional samples can readily be collected (such as the plankton application discussed in section 4.6.1).

4.7.2 Estimation of Multiple Quantities

In this work we have focused on the estimation of a single quantity of interest (e.g. coral cover from survey images). However, sampling is often used to simultaneously estimate multiple quantities of interest (e.g. coral, algae *and* sand cover). The proposed hybrid sampling design extends to such situations by utilizing a separate function $f_b^{(j)}(x)$ for each quantity of interest, j . Machine bias estimation can then be performed independently for each quantity. Auxiliary sampling, on the other hand, requires the full matrix of confusion to be estimated and inverted [127], which may be unstable if the classification accuracy of any class is low (since the matrix needs to be full rank).

4.8 Conclusion

We have investigated the implications of modeling and incorporating the cost and accuracy of two annotators in random sampling designs for population mean estimation. We have derived a novel hybrid sampling design which utilizes both annotators to minimize the total cost of the sampling procedure, and we have provided sufficient and necessary conditions for when such procedure is preferred over conventional sampling. The hybrid design was evaluated on two marine ecology data mining applications, where the total cost of the sampling work was reduced by 50% – 79%. Other applications include other image-based marine applications, such as fish surveys [145], image-based terrestrial applications including crop [61], forest [48], rangeland [110] and desert [4] surveys; audio-based surveys of e.g. marine mammal or bird populations [64]; and sentiment analysis surveys from text corpora [103]. It can also include medical surveys, where the two annotators represent medical tests or treatments of varying quality and cost. Additionally, we have shown how to utilize the abundance correction of [127, 59] to rely solely on the cheap annotator in an auxiliary sampling design. This design is applicable under additional conditions which are defined and discussed.

To the best of our knowledge, this is the first work to model random sampling from this perspective and we believe that there are several interesting directions of future work, notably with respect to stratified and sequential sampling procedures.

Appendix

4.A Proof of Theorem 2

We start with the following lemma:

Lemma 1. *The cost function of hybrid sampling (4.12) is convex for $\sigma_b^2 < \sigma_p^2$, $n_b \geq \check{n}_a$ where $\check{n}_a$ is given by (4.5).*

Proof. We begin by deriving the first and second derivatives of (4.12):

$$\frac{\partial t(n_a, n_b)}{\partial n_b} = (c_c + c_b) \left(1 - \frac{k(\sigma_b^2 + \sigma_a^2)(\sigma_p^2 - \sigma_b^2)}{(n_b \frac{d^2}{\zeta_\delta^2} - (\sigma_p^2 - \sigma_b^2))^2} \right) \quad (4.29)$$

$$\frac{\partial^2 t(n_a, n_b)}{\partial n_b^2} = (c_c + c_b) \left(\frac{2k(\sigma_b^2 + \sigma_a^2)(\sigma_p^2 - \sigma_b^2)(\frac{d^2}{\zeta_\delta^2})}{(n_b \frac{d^2}{\zeta_\delta^2} - (\sigma_p^2 - \sigma_b^2))^3} \right). \quad (4.30)$$

Since by assumption $\sigma_b^2 < \sigma_p^2$, the second derivative is positive for

$$n_b \geq \check{n}_a > \frac{\zeta_\delta^2}{d^2}(\sigma_p^2 - \sigma_b^2), \quad (4.31)$$

where the strict inequality requires either σ_a^2 or σ_b^2 to be non-zero. This concludes the proof. $\square$

Since, according to Theorem 4.1, hybrid sampling should only be considered if $\sigma_b^2 < \sigma_p^2$, and since by design, $n_b \geq \check{n}_a$, (4.12) is convex, and Theorem 4.2 follows by setting the first derivative to zero and solving for n_b .

4.B Proof of Theorem 3

Proof. The TSC of conventional sampling, under the assumption that $c_b = 0$, is given by $n_a(c_c + c_a)$ and the TSC of hybrid sampling is given by $n_b c_c + n_a c_a$. As noted previously, if $\hat{n}_b = \hat{n}_a = \check{n}_a$, hybrid sampling reduces to conventional sampling and the TSCs are equal. Since $\hat{n}_b$ minimize the hybrid TSC, which is convex for $n_b \geq \check{n}_a$ (Appendix 4.A), it follows that the TSC of hybrid sampling is smaller than the TSC of conventional sampling if and only if $\hat{n}_b > \check{n}_a$. The threshold, k_0 for when this occurs can be calculated by equating the two arguments inside the max operator of (4.14) and solving for k :

$$\frac{\zeta_{\delta}^2}{d^2}(\sigma_p^2 + \sigma_a^2) = \frac{\zeta_{\delta}^2}{d^2} \left(\sigma_p^2 - \sigma_b^2 + \sqrt{k_0(\sigma_b^2 + \sigma_a^2)(\sigma_p^2 - \sigma_b^2)} \right) \Rightarrow \quad (4.32)$$

$$\sigma_a^2 + \sigma_b^2 = \sqrt{k_0(\sigma_b^2 + \sigma_a^2)(\sigma_p^2 - \sigma_b^2)} \Rightarrow \quad (4.33)$$

$$k_0 = \frac{\sigma_a^2 + \sigma_b^2}{\sigma_p^2 - \sigma_b^2}. \quad (4.34)$$

□

4.C Two-Stage Auxiliary Sampling

In two-stage sampling designs each first stage sample, x_i is again sampled randomly using some number, s of second-stage samples [130]. An analysis of the errors using such designs is provided by Deming [37]. Second stage sampling is commonly used e.g. in benthic surveys where each collected photoquadrat is annotated using random point sampling [70]. This protocol requires s points to be overlaid on each image at locations selected randomly with replacement. The substrate under each point is then annotated by an expert as pertaining to one of some number of classes. An unbiased estimator of the abundance (benthic cover) of each class for a certain sample can be derived by counting how many of the s annotations that were annotated as that class.

We derive the statistics of two-stage sampling under the population drift assumption, namely that each decision is made by some noisy annotator f_b , with known matrix of confusion. We will denote by $x_{i1}, \dots, x_{is}$ the s locations to be annotated in each sample x_i , and $u_{ij} \in \{0, 1\}$ the true value associated with each location. The value of each first stage sample is approximated by $y_i = \sum_{j=1}^s u_{ij}$. We do not make any assumptions on the probability density from which the first stage samples are drawn, but as previously let μ_p denote the expected value and σ_p^2 the variance. Given a classifier f_b with known matrix of confusion, an unbiased estimator of u_{ij} is given as

previously by

$$\tilde{u}_{ij} = \frac{f_b(x_{ij}) + \beta - 1}{\alpha + \beta - 1}. \quad (4.35)$$

From this an unbiased estimator of $\tilde{y}_i$ is given by

$$\tilde{y}_i = \frac{1}{s} \sum_{j=1}^s \tilde{u}_{ij}. \quad (4.36)$$

We have derived the variance of $\tilde{y}_i$ for the special case where $s = 1$ in the main paper. Next, we show how to derive the variance of $\tilde{y}_i$ for a general s by applying the law of total variation twice. We begin by noting that

$$\text{var}(\tilde{y}_i) = E[\text{var}(\tilde{y}_i|y_i)] + \text{var}[E(\tilde{y}_i|y_i)]. \quad (4.37)$$

The second term is simply $\text{var}[E(\tilde{y}_i|y_i)] = \text{var}[y_i] = \sigma_p^2$, and the first term can be expressed in terms of $\tilde{u}_{ij}$

$$\text{var}(\tilde{y}_i|y_i) = \frac{1}{s^2} \sum_{i=1}^s \text{var}(\tilde{u}_{ij}|y_i), \quad (4.38)$$

which can be expressed, by again using the law of total variation, as

$$\text{var}(\tilde{u}_{ij}|y_i) = E[\text{var}(\tilde{u}_{ij}|u_{ij}, y_i)] + \text{var}[E(\tilde{u}_{ij}|u_{ij}, y_i)]. \quad (4.39)$$

The second term of (4.39) is simply given by $\text{var}[E(\tilde{u}_{ij}|u_{ij}, y_i)] = \text{var}[u_{ij}|y_i] = y_i(1 - y_i)$, but the first term is less obvious. Following Solow et al. [127], we first note that

$$\text{var}(f_b(x_{ij})|u_{ij}, y_i) = u_{ij}\alpha(1 - \alpha) + (1 - u_{ij})\beta(1 - \beta), \quad (4.40)$$

since if $u_{ij} = 1$, $f_b(x_{ij}) \sim \text{Ber}(\alpha)$, and if $u_{ij} = 0$, $f_b(x_{ij}) \sim \text{Ber}(1 - \beta)$. We then note that

$$\text{var}[\tilde{u}_{ij}|u_{ij}, y_i] = \frac{\text{var}[f_b(x_{ij})|u_{ij}, y_i]}{(\alpha + \beta - 1)^2}, \quad (4.41)$$

which follows directly from (4.35), and also that $E[u_{ij}] = y_i$. Putting this together yields the following expression for the first term of (4.39):

$$E[\text{var}(\tilde{u}_{ij}|u_{ij}, y_i)] = \frac{y_i \alpha (1 - \alpha) + (1 - y_i) \beta (1 - \beta)}{(\alpha + \beta - 1)^2}. \quad (4.42)$$

Putting this all together yields

$$\text{var}(\tilde{y}_i) = E[\text{var}(\tilde{y}_i|y_i)] + \text{var}[E(\tilde{y}_i|y_i)] \quad (4.43)$$

$$= E\left[\frac{1}{s^2} \sum_{i=1}^s \text{var}(\tilde{u}_{ij}|y_i)\right] + \sigma_p^2 \quad (4.44)$$

$$= \frac{1}{s^2} \sum_{i=1}^s E[\text{var}(\tilde{u}_{ij}|y_i)] + \sigma_p^2 \quad (4.45)$$

$$= \frac{1}{s^2} \sum_{i=1}^s E\left[\frac{y_i \alpha (1 - \alpha) + (1 - y_i) \beta (1 - \beta)}{(\alpha + \beta - 1)^2} + y_i(1 - y_i)\right] + \sigma_p^2 \quad (4.46)$$

$$= \frac{1}{s^2} \sum_{i=1}^s \left[\frac{\mu_p \alpha (1 - \alpha) + (1 - \mu_p) \beta (1 - \beta)}{(\alpha + \beta - 1)^2} + \mu_p(1 - \mu_p)\right] + \sigma_p^2 \quad (4.47)$$

$$= \frac{1}{s} [\sigma_s^2 + \mu_p(1 - \mu_p)] + \sigma_p^2, \quad (4.48)$$

where σ_s^2 is given by (4.22). Interestingly, the variance of $\tilde{y}$ approach σ_p for large number of secondary stage samples, s . This is to be expected under the assumption that f_b is perfectly modeled by a known matrix of confusion Q . Since $\tilde{y}_i$ is an average across s decisions, the variance introduced by the abundance correction cancels out with large values of s .

Finally, the total variance of $\tilde{\mu}_p$ is given by

$$\text{var}[\tilde{\mu}_p] = \frac{1}{n_b} \left(\frac{1}{s} [\sigma_s^2 + \mu_p(1 - \mu_p)] + \sigma_p^2 \right), \quad (4.49)$$

and the sample size by

$$n_b = \frac{\zeta_\delta^2}{d^2} \left[\frac{1}{s} [\sigma_s^2 + \mu_p(1 - \mu_p)] + \sigma_p^2 \right]. \quad (4.50)$$

Acknowledgments

The author would like to thank S. Dasgupta, C. Elkan and D. Kriegman at the Department of Computer Science & Engineering at the University of California, San Diego for invaluable advice and discussions. The author also gratefully acknowledge the support of the U.S. NSF Long Term Ecological Research (LTER) program and the Moorea Coral Reef LTER Site and support by the National Science Foundation (NSF) Division of Ocean Sciences grant #0941760.

Chapter 4 is a reformatted version of "*Random Sampling in an Age of Automation: Minimizing Expenditures through Balanced Collection and Annotation*" by Oscar Beijbom, which is in submission.

Chapter 5

Fluorescence imaging

Chapter 5 is, in full, a reformatted version of "*Improving Automated Annotation of Coral Reef Images Using Wide-Field-Of-View Fluorescence Imaging*" by Oscar Beijbom, Tali Treibitz, David I. Kline, Gal Eyal, Adi Khen, Ben Neal, Yossi Loya, B. Greg Mitchell and David Kriegman, which is under review for *Methods in Ecology and Evolution*.

One key conclusion of Chapters 3 & 4 is that a less-than-perfect automated annotator can be deployed to reduce the need for fully manual annotation and thus sampling expenditures. This is encouraging since machine-learning based methods rarely achieve human performance levels for visual recognition tasks. However, Chapters 3 & 4 also conclude that the cost savings increase with the accuracy of the automated annotation method. In this chapter we consider augmenting the photographic methodology in order to achieve higher automated annotation accuracy. More precisely, we deploy a fluorescence camera in conjuncture with a standard reflectance camera and show how to learn from the joint information in both images. This enables a trade-off where manual annotation can be reduced by more sophisticated image collection. As a simple example in Chapter 7 demonstrates, this trade-off can be used to significantly reduce sampling expenditures.

Abstract

With the increasing use of large-scale imaging technologies such as autonomous aerial or underwater vehicles, vast numbers of digital images are collected for ecological applications. However, the time and cost associated with manual image annotation can be prohibitive, which creates a bottleneck between collected images and desired data-products. This bottleneck is particularly severe for underwater surveys, in which millions of benthic marine digital images are obtained each year by researchers, government agencies, and reef managers. Recent automated annotation methods have been shown to provide a potential solution, but standard white light reflectance photographs alone do not always contain enough discriminative information due to turbidity, color changes underwater, and morphological similarity for targets that may be quite different. In this work we hypothesize that wide-band, wide-field-of-view fluorescence imagery acquired with a low cost modified consumer camera can be incorporated with reflectance imagery to improve automated annotation accuracy of underwater benthic surveys. To evaluate this hypothesis we acquired registered reflectance and fluorescence images during a coral reef survey in Eilat, Israel. A novel automated annotation method based on convolutional neural networks was developed, and the performance was compared for (a) reflectance images only, (b) fluorescence images only, and (c) registered reflectance and fluorescence images. Our results show that the classification accuracy increased from 87.8% to 90.5% when using both images types compared to using reflectance images only. This is equivalent to a error-rate reduction by 22%, meaning that roughly one fifth of the annotation decisions were corrected when incorporating fluorescence information. The improvements were particularly large for coral genera *Platygyra* and *Acropora*, as well as for the hydrozoan *Millepora*, where the classification recall increased by 33 - 41%. We conclude that wide-field-of view, low cost fluorescence cameras can be deployed in coral reef surveys in order to improve automated annotation accuracy. Such improvements reduce the need for manual annotation, which enables larger and more frequent surveys. The developed methods can readily be applied to other terrestrial and marine surveys where multi-spectral and multi-modal images are collected and analyzed.

5.1 Introduction

Advances in robotics, control theory, and digital imaging technology have enabled the collection of large-scale digital photographic data-sets throughout the ecological sciences [51, 65, 28, 99, 123]. However, obtaining the relevant scientific data from the collected images typically requires time-consuming and expensive manual image annotation. Although recent automated annotation methods offer a compelling alternative to manual annotation [128, 8], the accuracy is rarely as high as that of human experts [54, 33]. In this work we investigate whether multi-modal image channels can contribute to improved automated annotation accuracy, thereby reducing the need for manual image annotation and verification. Specifically, we focus on using fluorescence information, as a mode distinct from reflectance, for automated annotation of coral reef survey images.

Coral reefs are essential to coastal societies throughout the world, providing food, resources and income to over 500 million people [19]. In the last three decades up to 80% of coral coverage has been lost in the Caribbean [47] and up to 50% in the Indo-Pacific [17, 102] largely due to anthropogenic stressors including over-fishing, pollution, sedimentation, habitat destruction and climate change [20, 111, 117]. This accelerated rate of decline creates a need for rapid assessment of reef health in order to develop more effective management and conservation strategies [92].

Currently, the most prevalent method for reef assessment is *in-situ* digital photographic surveys. However, obtaining ecological data, such as percent cover of key benthic groups, from the collected images requires time-consuming and expensive manual image analysis, often in the form of point annotation [104].

Recent advances in computer vision have enabled automated annotation of coral reef survey images [129, 136, 123, 8], offering a compelling alternative to manual annotation. Unfortunately, automated annotation in this context is challenging for several reasons: degradation of colors underwater, image distortions due to water turbidity, large variability of within-class appearance, and complex boundaries and juxtaposition of different classes. As shown in this work, however, these difficulties can be alleviated by incorporating fluorescence information in

the automated annotation method.

While reef surveys are commonly conducted using underwater digital consumer reflectance cameras, several other instruments have been utilized to study coral reef health. Multi-spectral cameras, underwater spectrometers, underwater radiometers and underwater fluorometers (diving PAMs) have been previously used to quantify physiological parameters of coral reef organisms [133]. Another option is to measure fluorescence signatures which have been shown to contain ecologically relevant information such as level of bleaching [93], recruitment levels [133], and physiological state [149]. However, until recently, fluorescence imaging systems were limited either by resolution [149], spatial coverage of the measurements [87, 147] and/or ease of operation [87, 149, 88].

We have recently developed the FluorIS (Fluorescence Imaging System) [133], a modified consumer camera which can be used for underwater fluorescence imaging of both the green (520-630 nm), and red (630-800 nm) wide-band components of the fluorescence spectra which correspond to the emission spectra of green fluorescence proteins (GFP-like) and chlorophyll-a, respectively. Its low price, ease of operation, and wide-field-of-view (0.5m^2) makes FluorIS a compelling tool for reef surveys. In the present work we hypothesized that the information captured by FluorIS could be used as auxiliary information to improve automated image annotation accuracy.

To test this hypothesis we used a custom-designed framer and both a standard SLR and FluorIS camera to capture registered fluorescence and reflectance image pairs during a reef survey in Eilat, Red Sea (Israel). We then developed a supervised automated annotation algorithm, based on Convolutional Neural Networks (CNN) [76], which utilized the joint information in the image pairs. While CNNs have recently shown remarkable progress on several important visual recognition tasks, such as image classification [71], and segmentation [50], it has not previously, to the best of our knowledge, been utilized for automated annotation of underwater images. Here we show that our CNN-based methods were able to utilize the additional information captured by FluorIS to significantly improve the annotation accuracy compared to methods which only utilized the reflectance images [8]. This demonstrates that (1) fluorescence signatures contain information

that can be directly related to the taxonomic identity of the benthos, and that (2) such signatures can be captured during reef surveys in conjunction with reflectance information. In addition, improved automated annotation accuracy can enable more extensive spatial and temporal studies by reducing the need for costly manual annotation and verification.

5.2 Materials & Methods

5.2.1 Image Collection

Co-located image-pairs were captured during nighttime in the shallow reefs adjacent to the Interuniversity Institute for Marine Sciences in Eilat. Nighttime deployments were used due to the higher effectiveness of the fluorescence imaging system [133] and the logistical ease of night diving in Eilat. Image locations were chosen randomly along a 3-15 m depth gradient. A custom-made framer was used that enabled rapid attachment and release of the imaging modules holding the cameras and strobes (Fig. 5.1). At each location, reflectance and fluorescent images were taken, each covering 50 x 70cm of the benthos. Using this methodology, 212 image-pairs were collected during three dives, each approximately one hour long. The quality of the high-resolution images was sufficient for identifying most of the corals to the genus level, with some identification to the species level.

For both reflectance and fluorescence imaging, the camera system comprised a Canon 5D Mark II professional grade off-the-shelf camera with a Sigma 20mm wide-angle lens, and a Sea&Sea underwater housing with the Fisheye Dome Port 240, fitted with a 40mm extension ring for better alignment of the dome port with the lens to reduce distortions. For fluorescence imaging, we used the Fluorescence Imaging System (FluorIS) which we developed in previous work [133]. In the FluorIS, the infrared filter on the camera sensor was removed for increased sensitivity to red fluorescence (chlorophyll-a emission). In addition, a yellow Tiffen #12 barrier filter was mounted on the camera lens, and blue NightSea excitation filters were mounted on the strobes as previously described [133]. As the modified FluorIS camera had an expanded spectral range in the long wavelengths, an additional filter (Schott BG39) was used for the strobes to block IR wavelengths that pass through the primary excitation filter. The reflectance camera system was

deployed with two Ikelite DS161 (guide number 24) strobes while the FluorIS system used two Sea&Sea YS250s (guide number 32) and two Inon Z240s (guide number 24) to maximize the fluorescence signal.

Figure 5.1. Co-located images captured using a custom framer. Two research divers positioned the framer (A). The FluorIS image was captured (B). The FluorIS camera with its four strobes was un-docked and removed from the framer (C). Finally, the reflectance camera was docked and a co-located reflectance image captured.

5.2.2 Post-processing

Using the custom framer, co-registered images were captured (Fig. 5.2). However, there were some minor registration issues due to slight framer movements. To ensure a high-quality registration, we therefore applied a post-processing registration step. In this step, four to ten corresponding points were hand-clicked in all image-pairs, and an affine image warp was applied using the `imtransform.m` command from the MATLAB Image Processing Toolbox (The MathWorks, Inc).

Image annotation of the reflectance images was performed by trained coral ecology experts (co-authors: AK, DE, and DK) with knowledge of the local ecosystem. In this procedure, the substrate under 200 random point locations in each image was assigned one of ten pre-defined labels using the point-annotation tool of CoralNet (coralnet.ucsd.edu). The label-set included the five dominant coral genera: *Faviidae*, *Stylophora*, *Platygyra*, *Acropora* and *Pocillopora*; and non-coral labels for bare-substrate and colonial hydrozoans *Millepora*. The label-set also included ‘unknown’, which was used when identification was not possible (e.g. for dark image

Figure 5.2.

Three registered image pairs. (Left) Reflectance images, (Center) FluorIS images. (Right) Pixel-wise average of the reflectance and fluorescence images, visually demonstrating the advantage of combining the two information sources. The fluorescence increases the contrast of the corals, and the reflectance gives context and information on the non-fluorescing substrates. The registration quality is evident from this average.

areas), as well as 'other hard-coral' and 'other invertebrates'. A 'macroalgae' label was used during annotation, but was excluded from the subsequent analysis since it was exceedingly rare ($< 0.03\%$). The bare-substrate label included crustose coralline algae (CCA) and turf-algae since the distinction between these substrates was difficult, and sometimes not possible. A histogram of the total annotation counts is shown in Fig. 5.3, and example cropped patches from each of the labels are shown in Fig. 5.4. Since our goal was to evaluate the utility of fluorescence information for *automated* annotation, the "ground truth" was the human expert annotations of reflectance images, which is the standard used in previous studies [8, 123]. Therefore, fluorescence images were *not* used during the manual annotation process. To verify the accuracy of the human expert,

1500 point annotations across 30 images were repeated by a second expert. A comparison of the two sets of manual annotations indicated excellent agreement (97.8%) between the two experts.

Figure 5.3. Histogram of point annotations. Histogram shows counts across the 212 image-pairs. The total number of annotated points is 42,400.

5.2.3 Automated Annotation

For the purpose of method evaluation, 212 annotated image-pairs were divided into two sets. The training-set comprised 142 randomly selected image-pairs, and was used to train the automated annotation methods. The test-set comprised the remaining 70 image-pairs and was used to evaluate the annotation methods. With 200 point annotations per image, the training-set contained 28,400 point annotations, and the test-set contained 14,000.

Discriminatory Information in Reflectance and Fluorescence Imagery

Before seeking to incorporate fluorescence information into an automated annotation system, we first wanted to verify that fluorescence image intensities indeed contain discriminatory information. Since benthic substrates lack distinct outlines, and since colors often change too much with the underwater conditions to be relied upon for classification [8, 123], automated annotation methods chiefly rely on textural cues [8, 123]. On the other hand, fluorescent spectra have been shown to contain information that distinguishes between benthic substrates, [148], which suggests that the colors (i.e. image intensities) of the fluorescence images can be used directly as predictive information in an automated annotation method. To determine whether fluorescence intensity provides useful discriminatory information, we conducted the following

Figure 5.4. Cropped patch-pairs from the data set. Six randomly selected pairs of patches from each class are cropped out from the annotated point locations. The left member of each pair is from the conventional reflectance camera and right member is from FluorIS. All patches are 256×256 pixels. The dotted blue line separates the hard corals from other substrates.

experiment.

Image intensities at the labeled locations in the training-set were used to estimate the marginal probability (base-rate), $\Pr(y)$ of each class, and the class conditional probability, $\Pr(x_c|y)$ of each image channel c . Here, $y \in \{1, \dots, 10\}$ denotes the class label, and $x_c \in [0, 1]$ the image intensity of channel c . The class conditional probabilities, $\Pr(x_c|y)$ were approximated by fitting a Beta distribution to the intensities using maximum likelihood estimation. As shown in Fig. 5.5, the beta distribution models the image intensities well.

After estimating the base-rate and the class conditional probabilities, labeled image locations in the test-set were classified using a naïve Bayes classifier [116]. The naïve Bayes classifier used Bayes rule to approximate the posterior probability of each class given the intensities ($\Pr(y|x_c)$), by assuming that the class conditional probabilities were independent for each image channel. The classification decision was given by

$$\arg \max_y \Pr(y) \prod_{c \in \mathcal{C}} \Pr(x_c|y), \quad (5.1)$$

for various combinations of image channels, $\mathcal{C}$. For example, $\mathcal{C}$ could include the RGB color channels of the reflectance camera or the red channel of FluorIS. The predicted labels were then compared to the ground-truth labels provided by the human expert in order to determine the classification accuracy for each set of image channels.

Figure 5.5. Beta distribution estimation of image channel intensities. Estimates shown for the red FluorIS image channel for five selected classes. The close fit of the curves to the intensities indicates that the Beta distribution is an appropriate model for the data. Similar close fits were seen for all other image channels and classes.

Automated Annotation Using Convolutional Neural Networks

Our proposed automated annotation system used convolutional neural networks (CNNs) [76]. CNNs are commonly trained on color or gray-scale images, i.e., with one or three image channels [71]. The registered reflectance and FluorIS images collected in this work, however, have five channels of information (the FluorIS blue channel does not contain any information [133]). We investigated two approaches for utilizing CNNs on this data. The first was to train a CNN

directly using the five-channel information, and the second was to train two separate CNNs, one on each image type, and then train a second-stage classifier to consolidate the output of the two networks. In our experiments, the second approach was more effective and we focused our method development on this.

We used Caffe [63], an open-source framework for training the CNNs. Specifically, we used the publicly available “cifar10” network structure (caffe.berkeleyvision.org), which is designed to learn from 32 x 32 pixels images. The cifar10 network comprises three rounds of consecutive convolutions, pooling and non-linear rectification layers, and all network parameters are learned directly from the training-set through back-propagation [75]. For each of the 200 annotated point locations in the training-set images, a centered 128 x 128 pixel patch was cropped out, and re-sized to 32 x 32 pixels. Rotated (by 0, 90, 180 and 270 degrees) and mirrored versions of each cropped patch were included in the training data in order to prevent over-fitting (Fig. 5.6) [71, 50]. Using this data-preparation procedure, two networks were trained. Denoting $p \in \mathbb{R}^{32 \times 32 \times |\mathcal{C}|}$ as an image patch with channels $\mathcal{C}$, and by $s \in \mathbb{R}^{10}$ a vector of classification scores for each of the 10 classes in the label-set, the networks can be written as a mapping $f^{(\mathcal{C})} : p \rightarrow s$. Classification of a new patch, p^* , from the test-set was done by assigning the class, y , corresponding to the largest score:

$$s^* = f^{(\mathcal{C})}(p^*) \quad (5.2)$$

$$y = \arg \max[s_1^*, \dots, s_{10}^*] \quad (5.3)$$

f^{REF} represents the network trained on the reflectance images, and f^{FLR} , the network trained on the FluorIS photos. To consolidate the information from the two networks a second-stage classifier was used that mapped the concatenated network scores to class labels. To create this second-stage classifier, all image patches were propagated through the networks after training. This generated two, 10-dimensional score vectors for each patch (one for each network and image type). The score vectors were then concatenated to a 20-dimensional joint score vector

for each patch, which encoded information from both the reflectance and fluorescence images. The joint score vectors pertaining to the training-set were used to train a linear Support Vector Machine [43], which was then used to predict the labels of the patches in the test-set. This second stage classifier was denoted f^{JOINT} , and is described in more detail below. Using a NVIDIA Tesla K40 GPU, the full network was trained in approximately 5 hours, and prediction took less than 1 second per image.

Figure 5.6. Training-set preprocessing procedure for the proposed automated annotation method. Around each annotation point-location, a 128 x 128 pixel patch was cropped (left). The patch was then resized using bilinear interpolation to 32 x 32 pixels, rotated (by 90, 180, and 270 degrees), and mirrored. This generated eight training-patches from each annotated point location (right).

5.2.4 Statistical Analysis

A Wilcoxon signed-rank test was used to test the hypothesis that the annotation accuracy of f^{JOINT} was higher than the f^{REF} network. In other words, we examined whether the information from the FluorIS images improved annotation accuracy compared to using reflectance images alone. The accuracy was calculated for each image in the test-set as the ratio of correctly classified point-locations. Let $a^{\text{JOINT}}(i)$ be the accuracy of f^{JOINT} for image i , and $a^{(\text{REF})}(i)$ the accuracy of f^{REF} for the same image. The paired annotation accuracy differences were then calculated as $d(i) = a^{\text{JOINT}}(i) - a^{(\text{REF})}(i)$, and d was used to perform the Wilcoxon signed-rank test.

5.3 Results

5.3.1 Analysis of Image Intensities

Annotation accuracy of the naïve Bayes classifier was lower for the reflectance image channels than for the FluorIS image channels (Fig. 5.1). For example, using all three channels of the reflectance camera, the accuracy was $70.1 \pm 0.50\%$ (mean $\pm$ SE, $n = 70$) and the only class that was somewhat reliably distinguished from the dominant bare-substrate class was the ‘unknown’ class (Fig. 5.2). In contrast, the red and green channels of the FluorIS camera yielded $76.6 \pm 0.50\%$ accuracy. The annotation accuracy using all five channels ($74.3 \pm 0.42\%$) was slightly lower than the accuracy using the FluorIS channels alone (Fig. 5.1). The higher accuracy of the fluorescence channels was mostly due to a 57% recall for the dominant coral genera, *Faviidae* (Fig. 5.2).

Among the reflectance camera channels, the red channel contained slightly more information, achieving $70.1 \pm 0.50\%$ (mean $\pm$ SE, $n = 70$) annotation accuracy when used alone. In comparison, the green channel achieved $69.2 \pm 0.51\%$ and the blue channel $68.7 \pm 0.51\%$ (Fig. 5.1). Among the fluorescence channels the red fluorescence channel contained slightly more information than the green fluorescence channel, achieving $73.7 \pm 0.49\%$ vs. $72.3 \pm 0.52\%$ accuracy (Fig. 5.1).

Figure 5.1. Accuracy of the naïve Bayes classifier for different combinations of image channels. Accuracy displayed as mean $\pm$ SE. REF denotes the reflectance camera, FLR the fluorescence camera, and R, G, B denote the red, green and blue color channel respectively. The black rightmost bar indicate the accuracy of using FLR G+B together with REF R+G+B.

Figure 5.2. Confusion matrices for the naïve Bayes classifier. Numbers given on row r and column c are the probability of classifying class r as class c . The values along the diagonal indicate the recall of each class. REF denotes the reflectance camera, FLR the fluorescence camera, and R, G, B denote the red, green and blue color channel respectively. The dotted lines separate the hard-corals from the other substrates.

5.3.2 Automated Annotation Using Convolutional Neural Networks

While the results of the naïve Bayes classifier detailed in the previous section were useful to understand the discriminatory information associated with the raw image intensities, a reliable automated annotation method needs also to take into account texture and context in order to make more accurate annotation decisions [8].

The overall annotation accuracy of f^{REF} , the CNN which used the reflectance images only, was $87.8 \pm 0.61\%$ (mean $\pm$ SE, $n = 70$; Fig. 5.3). This was lower than the $90.5 \pm 0.47\%$ accuracy of f^{JOINT} , which used information from both images ($p < 0.0001$). This accuracy increase is equivalent to an error-rate reduction by 2.7%, from 12.2% for f^{REF} to 9.5% for f^{JOINT} . This is a 22% relative reduction in error-rate, meaning that approximately one in five annotation

errors were corrected by incorporating fluorescence information.

The increase in performance was particularly large for coral genera *Platygyra* and *Acropora*, and for the hydrozoan fire coral genus *Millepora*, where the classification recall increased 38%, 33%, and 41%, respectively, which is a relative increase of 512%, 100% and 86% respectively from the classification recalls of f^{REF} (Figures 5.3, 5.4).

The overall accuracy of f^{FLR} ($85.5 \pm 0.7\%$), which used the fluorescence images only, was lower than f^{REF} ($p = 0.0002$), but higher on several important benthic substrates: *Platygyra*, *Acropora*, and *Millepora* (Figures 5.3, 5.4). In addition, for *Pocillopora*, and *Faviidae*, the f^{FLR} network, while less accurate than f^{REF} , provided complementary information that the joint f^{JOINT} classifier utilized to outperform f^{REF} (Figures 5.3, 5.4). These findings support our main hypothesis that fluorescent information can improve the accuracy of automated annotation methods.

Figure 5.3. Confusion matrices for the proposed convolutional neural network method. Results displayed for networks using information from (left) the reflectance camera, (middle) the FluorIS fluorescence camera, and (right) the joint classifier. Numbers given on row r and column c are classification recalls, i.e. the probability of classifying class r as class c . The values along the diagonal indicate the recall of each label (a value of 1 along the diagonal indicate an ideal classifier). The dotted lines separate the hard-corals from the other substrates.

5.3.3 Training a Five-channel Network

While CNNs are commonly used to model color or gray scale images, a CNN can operate on any number of channels [63]. Since the reflectance and FluorIS images are registered, they can

be viewed as five-channel images. As noted in Section 5.2.3, this offers a compelling opportunity to train a five-channel network directly on all the information from the image-pairs.

We trained a CNN, f^{FIVE} directly on the annotated images of the training-set, and evaluated the accuracy, as described previously, on the test-set. This method directly incorporates both image types without the need for a subsequent second-stage classifier. The accuracy of f^{FIVE} was $88.9 \pm 1.1\%$ (mean $\pm$ SE, $n = 70$), which was higher than the $87.8 \pm 0.61\%$ accuracy of f^{REF} ($p < 0.0001$), but lower than the $90.5 \pm 0.47\%$ accuracy of f^{JOINT} ($p = 0.0029$).

Figure 5.4. Precision vs. recall curves for f^{REF} , f^{FLR} , and f^{JOINT} . For each class, precision is defined as the proportion of correct classification among all samples classified as that class, and recall is defined as the proportion of correct classification among all samples from that class. The curves were created by varying the classification threshold for each class.

5.3.4 Comparison to Traditional Automated Annotation Methods

This work is the first to utilize convolutional neural networks (CNNs) for automated annotation of benthic survey images. It was therefore critical to determine how the accuracy of CNNs relate to more traditional methods based on hand-tuned filter or color descriptors explicitly designed for automated annotation of reflectance images [129, 136, 123, 8]. Such methods commonly report on accuracy of around 70-80% [129, 136]; however, this is highly dependent on the data-set on which they are evaluated [123]. For this reason, we used the toolbox of

Beijbom et al. [8], which is publicly available (vision.ucsd.edu/content/moorea-labeled-corals), as a representative for the ‘traditional methods’, to compare the efficacy of such methods against the CNNs used in this work. The method of [8], here denoted $g^{\mathcal{C}}$, with the super-script indicating the image-type (REF or FLR), was trained and evaluated on the same training-set and test-set used throughout this work. The accuracy of g^{REF} was $87.7\% \pm 0.70\%$ (mean $\pm$ SE, $n = 70$), which was not different from the $87.8 \pm 0.47\%$ accuracy of f^{REF} ($p = 0.22$). When trained on the FluorIS images, the traditional method, g^{FLR} achieved $81.0 \pm 0.94\%$ accuracy, which was lower than the $85.5 \pm 0.70\%$ accuracy achieved by the proposed CNN-based method, f^{FLR} ($p = 1.3 * 10^{-5}$).

5.4 Discussion

We have demonstrated that fluorescence information can be captured during an image-based benthic reef survey and utilized to improve the annotation accuracy of automated annotation methods. Our work is novel from three key perspectives. First, we used fluorescence information to improve the classification accuracy of the dominant benthic classes by 2.7%, which is equivalent to a 22% relative reduction of error-rate. In contrast to previous work that investigated the use of fluorescence spectra to classify groups defined by their level of bleaching [93], fluorescence signatures [148, 88], or other physiological states [118, 149, 36], our work provides evidence that the fluorescence signatures can be directly related to benthic taxonomy. Second, most work on coral fluorescence has relied on relatively expensive spectrometers or custom multi-spectral cameras which have high spectral but limited spatial resolution [148, 88]. In contrast, we use an inexpensive, modified consumer camera, the FluorIS, which has high spatial resolution but only two spectral channels. The high spatial resolution of FluorIS enables deployment alongside a reflectance camera for regular benthic surveys. Third, we are the first, to the best of our knowledge, to use CNNs for automated annotation of joint fluorescence and reflectance image data. We believe this type of method can be applied to other ecological applications deploying multi-spectral or multi-modal cameras, e.g. vegetation monitoring [14], inter-tidal landscapes [18], or plant health for agronomic applications [94].

The utility of the fluorescent information depended on the substrate class. Different coral

species have different types of FPs and some species are more fluorescent than others [133, 86, 38]. In our study the FLR data was most helpful for improving the classification accuracy of the highly fluorescent coral genera *Platygyra* and *Acropora*, as well as for the hydrozoan *Millepora*. For these genera, the classification recall improved by 38%, 33% and 41% respectively (Fig. 5.3). We expect that including fluorescent information will increase the automated annotation accuracy of most coral reef surveys in areas with a large number of highly fluorescence species and in benthic surveys with a variety of fluorescent organisms (e.g. various algae species, diverse corals and other invertebrates). We also would hypothesize that fluorescent information might assist in distinguishing between different types of algae that can be challenging to distinguish using automated annotation methods such as turf algae and CCA [8]; but additional studies will need to be performed in order to verify this hypothesis. Salih et al. (2000) carried out a survey of the distribution of fluorescent corals on the Great Barrier Reef and found that 124 species from 56 genera contained fluorescent morphs, suggesting that fluorescence should be a highly successful factor for improving annotation accuracy on coral reefs. However, they also found that within a given species there were often fluorescent and non-fluorescent morphs growing side by side. This fluorescence polymorphism within a coral species could explain why the fluorescence information did not improve the annotation accuracy further [118], and raises interesting ecological questions not explored here. Salih et al. (2000) also found that the highest number of fluorescent morphs occurred at the shallowest sites, and suggested that fluorescent pigments could have photoprotective properties. It is therefore likely that fluorescence information will be most useful in shallow coral reef surveys where corals have the maximal fluorescent photoprotective pigments and less useful beyond the photic zone. Future work should explore the utility of fluorescence information for improving automated annotations at different depths, in different reef locations, and across different benthic habitats.

Furthermore, our experiments have indicated the importance of careful registration of the image pairs. This is supported by the stronger results of the joint classifier, f^{JOINT} , which merged the information on the patch-level, compared to the five-channel network, f^{FIVE} , which merged the information on the pixel-level. We believe that this difference in accuracy was an

effect of the registration quality. As shown in Figures 5.2 & 5.4, the registration was good, but not perfect. Had it been perfect, with every pixel of the two images corresponding, we believe the f^{FIVE} network would have done better, as it had direct access to all image information. Also, in our experiments, when the registration was omitted, the joint classifier failed to outperform the accuracy of the f^{REF} network. This emphasizes the importance of careful registration and the development of new imaging systems and protocols for registered image capture.

Higher accuracy of an automated annotation system can reduce the amount of expensive human effort required during image annotation. Our results show that a 22% reduction in error-rate can be achieved by incorporating fluorescence information, which means that one fifth of the automated annotation errors can be corrected if fluorescence information is available. This suggests that it may be cost-effective to deploy a wide-field-of-view fluorescence imaging system with a reflectance camera during surveys, in order to reduce the human annotation time. However, we emphasize that the image collection utilized in this survey should be considered a first attempt at obtaining co-registered images, and that additional work is required to facilitate robust co-located image capture of fluorescence and reflectance images on a large scale. Ideally, the image-pairs should be captured in a way that does not require a manual post-processing registration step. In addition, we previously showed that daytime fluorescence imaging is possible with the FlourIS using ambient light subtraction [133]. Future camera development that integrates reflectance and fluorescence channels in the same camera system together with automatic acquisition of an ambient light image would facilitate acquiring registered images at daytime.

We have also shown how to utilize CNNs for automated annotation of benthic survey images. To contextualize our results, we compared the efficacy of the CNNs against the method of Beijbom et al. (2012). This comparison showed that the CNNs performed on par with the method of [8] on the reflectance images, but that it more effectively utilized the fluorescence information. These results were not surprising. Traditional automated annotation methods, such as that of [8], were optimized on reflectance images, and could be expected to perform well on such images. However, the FlourIS images were of a different character both in terms of contrast and texture. For example, as we have shown, the FlourIS image intensities were more effective in

discriminating between coral and non-corals than the reflectance image intensities. Therefore, traditional automated annotation methods, which were not designed for fluorescence imagery, perform poorly. CNNs, in contrast, were able to adapt and learn directly from this data.

Acknowledgments

This work was supported in part by the US National Science Foundation (NSF) Division of Ocean Sciences (OCE) grant No. 09-41760 to BGM and DK, by the Israel Science Foundation (ISF) grant No. 341/12 to YL, and by the National Oceanic and Atmospheric Administration (NOAA) grant No. NA10OAR4320156 to OB. We thank Ben Herzberg and Tali Mass for help with dives, Charles H. Maze1 for valuable advice on the design of FluorIS, and gratefully acknowledge the support of NVIDIA Corporation for their donation of the Tesla K40 GPU used in this research.

Chapter 5 is a reformatted version of "*Improving Automated Annotation of Coral Reef Images Using Wide-Field-Of-View Fluorescence Imaging*" by Oscar Beijbom, Tali Treibitz, David I. Kline, Gal Eyal, Adi Khen, Ben Neal, Yossi Loya, B. Greg Mitchell and David Kriegman, which is under review for *Methods in Ecology and Evolution*. The dissertation author was the primary investigator and author of this paper.

Chapter 6

CoralNet

A website, CoralNet, has been developed that implements the core methods of this thesis. CoralNet was created by the thesis author together with several UCSD undergrad students, notably Stephen Chan, Devang Sampat and Jeff Sandvik and it is currently hosted at UCSD (coralnet.ucsd.edu). Because CoralNet is under constant development, this is not intended to be a user-manual, as this would quickly become outdated. Instead user instructions for CoralNet are integrated in the site. In addition, instruction videos are available at the the CoralNet vimeo channel (vimeo.com/channels/coralnet). Here we discuss the ideas underlying CoralNet and how these are implemented. In addition, we provide some statistics on the data that is currently on the site.

6.1 CoralNet Design Overview

The main goals of CoralNet are threefold. First and foremost, it provides coral reef researchers, government agencies, non-profits and others with direct access to sophisticated semi- and fully- automated as well as interactive point annotation tools. Second, the image data uploaded and annotated on CoralNet provide challenging, multi-year, multi-locations data-sets for research in computer vision and machine learning. Third, CoralNet serves as a collaborative platform and repository for benthic surveys from around the world. This will facilitate large scale meta-analysis in support of global hypothesis related to the health of the worlds coral reefs.

CoralNet can operate fully manually, without any computer vision integration. In this fully manual mode, users upload images and annotate them directly in the web-browser using a

point annotation tool similar to that of Coral Point Count (CPCe) [70]. Users select the label-set they desire for the annotation work from a list of pre-defined labels which are shared across the site, or create their own. While this manual annotation mode does not alleviate the actual annotation work, it offers some benefits over offline annotation programs. First, all images are stored on the CoralNet server which makes it easy to coordinate the annotation work for multiple annotators. Second, since the label-set is defined by the source administrator, it is guaranteed that all annotators use the same label-set. Third, all annotations are searchable, which offers easy verification of previous annotations (Fig. 3.7).

Each night at 3pm PST, the CoralNet computer vision backend is activated. For each source (annotation project), the backend learns from all manual annotations and creates a new version of the automated annotation robot, “Cora” *for each source*. Cora then annotates all remaining images in the source. All point annotations performed by Cora change status from “unannotated” to “unconfirmed”, meaning that they need confirmation by a human expert. The accuracy of Cora is estimated on a subset of the manually annotated images and displayed to the user. This accuracy information includes confusion matrices, and alleviation curves similar to Fig. 3.4.

Depending on the displayed accuracy, the users may then choose to (1) ignore the automated annotations and proceed manually, (2) trust the automated annotations, and export them for the remaining images along with the manual annotations, (3) keep annotating interactively by verifying Cora’s suggestions, and (4) trust a subset of Cora’s annotations, and manually verify the rest. The two latter options are detailed in Chapter 3 as Refine and Alleviate respectively.

For Refine, the top three suggestions by Cora are overlaid the image and facilitate rapid verification by the human expert. Preliminary (unpublished) results indicate that this save 7-12% of the annotation time without affecting the annotation accuracy. For Alleviate, the user chooses the desired level of alleviation, e.g. 40% which sets a threshold on the confidence required for an annotation decision to be alleviated. The actual alleviation happens when a user enters the annotation tool for an “unconfirmed” image. All point annotations for which Cora is sufficiently confident then change from “unconfirmed” to “confirmed”, and are thus treated on par with the

manually annotations. The remaining point annotations, for which Cora was not sufficiently certain, can then be verified with or without the Refine suggestions, according to what the user desires.

6.2 Snapshot of CoralNet Vision Backend Performance

In this section we provide a meta-analysis of the performance of the current CoralNet computer-vision backend algorithm by aggregating the performance across all sources. It is important to note that this performance overview is highly dependent on this particular algorithm used in the backend, which, at the time of writing this thesis, was the one described in Chapter 2. However, with next generation of vision algorithms, the performance can change drastically.

Before we proceed, a few words are needed on the metrics used. First, accuracy is measured on the functional group level to avoid artifacts associated with varying label-set size across the sources. The functional groups are, as defined in CoralNet: hard coral, algal, hard-substrate, soft-substrate, seagrass, other invertebrates, and all other. The functional-group accuracy is measured using Cohen’s Kappa (κ), which can be thought of as a normalized accuracy, and is particularly useful when there is large imbalance among the number of samples per class [22]. While κ is widely used to assess annotator agreement for categorical data [3, 82, 91], there is no absolute scale against which values of κ can be gauged. For example, [72] characterized $\kappa < 0$ as indicating no agreement, 0–20% as slight agreement, 21–40% as fair agreement, 41–60% as moderate agreement, 61–80% as substantial agreement, and $> 81\%$ as almost perfect agreement. In contrast, [45] characterized $\kappa < 40\%$ as poor agreement, 40–75% as fair-to-good agreement, and $> 75\%$ as excellent agreement. Second, “level of alleviation” (λ) is defined as the alleviation which generate a 5% decrease of Cohen’s Kappa of discriminating coral versus non-coral (Chapter 3). This was introduced to CoralNet in September 2014, so only a subset of the Cora versions have associated alleviation levels. All results are broken down by the maximum number of samples per class used during the training-phase, β , meaning that if, for any source, there were more than β samples from a certain class, only β samples from that class were used during training (selected randomly). This threshold is necessary to control the training time of the

learning algorithm, and was originally set to $\beta = 7500$. However, as more users uploaded larger and larger sources, some reaching over 400,000 samples, the vision backend started running to slow to cope with the demand. Hence, β was adjusted down to 4000 and then to 2500.

The median accuracy of Cora was 45% which can be considered ‘fair’ or ‘moderate’ (Fig. 6.1 A) [45, 72], and the median level of alleviation was 25%, which means that a quarter of the annotation work can be done automatically without affecting the quality of the benthic percent cover estimates (Fig. 6.1 B, Chapter 3). However, several sources show very high alleviation levels, of 80-100%. It is also clear that β is quite important. For example, the mean accuracy when $\beta = 7500$, increased to $\sim 60\%$ (Fig. 6.1 A). It is therefore important to find the appropriate balance between training accuracy and how quickly the backend can process all sources.

The accuracy was, not surprisingly, correlated to the total number of training samples available during training (Fig. 6.2 A). Again, the importance of β is evident, in particular for the larger sources. There was also a correlation between the number of labels in the label-set and the total training time (Fig. 6.2 B). In particular, with the exception for a peak around 20 pertaining to two distinct sources, the very large training times ($> 40h$) only occurred for larger label-sets (with > 80 labels). This may be an artifact of how we control the size of the training-set (namely using a threshold, β , *per class*). Since a particular class can only contribute β samples during training, Cora versions pertaining to sources with larger label-sets are more likely to utilize larger *total number* of samples during training, which in turn increase the training-time.

Figure 6.1. Accuracy and alleviation levels of CoralNet. Histogram of A) accuracy, κ and B) level of alleviation, λ , for all versions of Cora where alleviation was recorded. The results are resolved based on the threshold, β , for the number of samples per class used during training.

Figure 6.2. Scatterplots of CoralNet accuracy and training time. Scatterplots of A) accuracy, κ versus number of samples, and B) total training time versus label-set size. The results are resolved based on the threshold, β , for the number of samples per class used during training.

Chapter 7

Conclusion

7.1 Final Thoughts

The contributions of this thesis are interdisciplinary. Chapter 2 is primarily a computer vision contribution, Chapter 3 is an user-study and a machine learning contribution, Chapter 4 is a sampling theory and data-mining contribution, and Chapter 5 offers contributions in coral-physiology as well as multi-modal machine-learning. However, the contributions are all integral parts of the larger goal of this thesis, namely to reduce the manual-annotation bottleneck of coral reef surveys.

To emphasize this point we can consider deploying the photographic and algorithmic developments of the fluorescence imaging system of Chapter 5 in a hybrid sampling design of Chapter 4. Let's say that we want to estimate the coral cover for a particular site in the vicinity of where the Eilat-dataset (Table 1.1) was collected, and that we want the cover estimate to have $< 2\%$ error with 95% confidence. Using the Eilat data-set we can get an *a-priori* estimate of the standard deviation of the coral cover among the samples (images): $\sigma_p = 11.4\%$. This means that a conventional sampling-design would require the collection and annotation of 124 images to get an estimate of the desired precision (4.5). Using a collection-cost of 1 person-minute, and an annotation-cost of 10 person-minutes, this would result in a Total Sampling Cost (TSC) of 23 person-hours (4.6). Now we can consider utilizing the hybrid-sampling design proposed in Chapter 4. Let's first look at what happens if we collect standard reflectance images only, and use the f^{REF} network of Chapter 5 to annotate the images. The standard deviation of the errors

of f^{REF} , again estimated from the Eilat data-set, is $\sigma_b = 5.0\%$. Since $\sigma_b < \sigma_p$ we can expect a hybrid design to be more cost effective (Theorem 4.1). Indeed, a hybrid design would require collection of 255 images and manual annotation of 39 images (Theorem 4.2), for a TSC of 11 person-hours (4.12); this is a $\sim 50\%$ reduction compared to a conventional sampling design. Now assume that the FluorIS is deployed alongside the reflectance camera, and the joint classifier, f^{JOINT} of Chapter 5 is used. The standard deviation of the errors of f^{JOINT} is $\sigma_b = 2.1\%$, which is significantly lower than the annotation errors of f^{REF} . If the FluorIS deployment can be done without affecting the collection time, the hybrid design would require collection of 192 image-pairs and manual annotation of only 12 images, for a TSC of 5 person-hours; which is another $\sim 50\%$ reduction. It is however likely that the collection time may increase since two images are captured at each location. If the collection time per sample (now an image-pair) doubled, to 2 person-minutes per sample, it would still be beneficial to deploy the FluorIS; a hybrid design would in that case require collection of 171 images, and manual annotation of 15, for a TSC of 8 person-hours. This example illustrates the importance of a holistic approach including the photographic methodology, the machine learning approach, and the sampling design, in order to be able to make the correct design decisions.

Similarly, we can consider the developments of Chapter 5 in the context of the intra-operator study measured in Chapter 3. In Chapter 3, the intra-operator agreement of discriminating between coral and other, as measured by Cohen's kappa, was $\kappa_{\text{coral}} = 89.7 \pm 1.2\%$ (Fig. 3.2), suggesting that this is an upper bound on the accuracy of the human annotators. The accuracy of the f^{REF} network in Chapter 5 which utilized reflectance images only was $\kappa_{\text{coral}} = 84\%$, while the accuracy of the f^{JOINT} network which utilized both reflectance and fluorescence images was $\kappa_{\text{coral}} = 89.6\%$. This indicates that by adding fluorescence information, the automated annotation accuracy is approaching that of a human operator.

In conclusion, monitoring vast benthic areas is a daunting task, and compromises are imperative. When reef surveys shifted from *in-situ* observations to photographic surveys, there was much gained in terms of both making the underwater deployment more effective and in creating permanent visual records. But there was also disadvantages, in particular in that identification

from photograph is less accurate than *in-situ* identification. The same can be said for automated annotation, where the advantage of being able to process vastly larger amounts of images, comes at the cost of lower annotation accuracy. As better annotation methods are being developed, the costs will decrease, making the choice to rely on automation, partly or fully, easier.

In 2009, as this thesis work began, the thought of automated annotation of coral reef survey images, was, understandably, met with much hesitation from coral ecology collaborators. As methods have evolved, this hesitation is giving way and I believe that automated annotation will be an integral part of the data annotation cycle in the next few years.

7.2 Future Work

We outline three directions of future work.

- Hybrid annotation systems:** As in other sciences, there is a reluctance to defer annotation decisions to an automated system, and wide-spread belief in the experts' ability to provide consistent, high quality annotations. However, as shown in this work (Chapter 3), and in other fields [33], experts make mistakes too. Culverhouse et. al give four reasons for this (1) human short-term memory limit of 5 to 9 items; (2) fatigue and boredom; (3) recency effects whereby a new classification is biased toward those in the set of most recently used labels; (4) positivity bias, whereby labeling a specimen is biased by one's expectations of the species present in the sample [33, 41]. There is hence scope to not only match the human experts, but actually improve the annotation accuracy by relieving the human experts of the mind-numbing repetitive annotations of easy samples, allowing them to focus their attention on the more difficult samples. The semi-automated Alleviate operational modes introduced in Chapter 3 is one way to achieve this, but careful large-scale studies using a human-in-the loop needs to be conducted to establish the effects of overall accuracy.
- Annotation with guarantees:** Ecologists and scientists are used to handling uncertainty, with modelers taking into account noise from measurements, finite sample sizes, and other application-specific factors. The automated annotation noise can also be incorporated in

such models, but this requires that the error-levels can be quantified. This is deceptively difficult due to a problem known as domain-shift [6]. Under domain shift, the archived (source), data used to train the classifier, differs from the data of interest (target). Here, "differ" means that the joint probability over labels, Y , and data, X , are different between the source and target domains: $\Pr_{\text{source}}(X, Y) \neq \Pr_{\text{target}}(X, Y)$. This can, in turn, be broken down into differences in the class-conditional densities: $\Pr_{\text{source}}(X|Y) \neq \Pr_{\text{target}}(X|Y)$ and base-rates: $\Pr_{\text{source}}(Y) \neq \Pr_{\text{target}}(Y)$. In the ecological sciences, the archived data is typically data that was collected and annotated in a nearby location or at a previous occasion, and it is often the case that both the class-conditional densities and base-rates differ. As it turns out, predicting the classifier performance under domain-shift is exceedingly difficult [6, 101]. This is unfortunate, since, as mentioned above, scientific modeling require reliable estimates of error-levels. It seems, therefore, imperative to request a subset of the newly collected data to be manually annotated in order to model the automated annotation errors. Our work in Chapter 4 is one way to handle this, and there has been other [144, 13], but I believe there is ample room for improvement.

- Deep Learning:** The writing of this thesis is taking place during a period of immense excitement in the machine learning and computer vision communities, due to a family of learning methods, colloquially referred to as Deep Learning. Deep Learning is deep artificial neural networks (ANNs) trained on large datasets often using highly parallel General Processing Units (GPUs). Deep learning algorithms, in the form of large convolutional neural networks (CNNs) [76], have yielded immense improvements on important computer vision tasks such as image classification [71], and segmentation [50]. In Chapter 5 we have relied on CNNs to learn from fluorescence images, and preliminary, unpublished experiments indicate that significant improvements can be realized also on automated annotation of standard RGB reef survey images. I believe deep learning will be a game-changer for the analysis of benthic survey images, significantly increasing the accuracy of automated annotation methods.

Bibliography

- [1] T. C. Adam, R. J. Schmitt, S. J. Holbrook, A. J. Brooks, P. J. Edmunds, R. C. Carpenter, and G. Bernardi. Herbivory, connectivity, and ecosystem resilience: response of a coral reef to a large-scale perturbation. *PLoS One*, 2011.
- [2] E. H. Adelson. On seeing stuff: the perception of materials by humans and machines. In *Photonics West 2001-Electronic Imaging*, pages 1–12. International Society for Optics and Photonics, 2001.
- [3] O. Allouche, A. Tsoar, and R. Kadmon. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (tss). *Journal of applied ecology*, 43(6):1223–1232, 2006.
- [4] S. Archer, C. Scifres, C. R. Bassham, and R. Maggio. Autogenic succession in a subtropical savanna: conversion of grassland to thorn woodland. *Ecological Monographs*, 58(2):111–127, 1988.
- [5] M. Bahoura and Y. Simard. Serial combination of multiple classifiers for automatic blue whale calls recognition. *Expert Systems with Applications*, 39(11):9986–9993, 2012.
- [6] O. Beijbom. Domain adaptations for computer vision applications. *arXiv:1211.4860*, 2012.
- [7] O. Beijbom. Cost-effective sampling for pairs of annotators. *arXiv:1410.7074*, 2014.
- [8] O. Beijbom, P. J. Edmunds, D. I. Kline, B. G. Mitchell, and D. Kriegman. Automated annotation of coral reef survey images. In *Proc Comput Vision Pattern Recognit (CVPR)*, 2012.
- [9] O. Beijbom, P. J. Edmunds, C. Roelfsema, J. Smith, D. I. Kline, B. Neal, M. J. Dunlap, V. Moriarty, T-Y. Fan, C-J. Tan, S. Chan, T. Treibitz, A. Gamst, B. G. Mitchell, and D. Kriegman. Towards automated annotation of benthic survey images: variability of human experts and operational modes of automation. *PLoS ONE*, 2015.
- [10] O. Beijbom, N. Joshi, D. Morris, S. Saponas, and S. Khullar. Menu-match: Restaurant-specific food logging from images. In *Applications of Computer Vision (WACV)*, 2015

- IEEE Winter Conference on*, pages 844–851. IEEE, 2015.
- [11] O. Beijbom, M. Saberian, D. Kriegman, and N. Vasconcelos. Guess-averse loss functions for cost-sensitive multiclass boosting. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 586–594, 2014.
  - [12] D. R. Bellwood, T. P. Hughes, C. Folke, and M. Nyström. Confronting the coral reef crisis. *Nature*, 429(6994):827–833, 2004.
  - [13] P. N. Bennett and V. R. Carvalho. Online stratified sampling: evaluating classifiers at web-scale. In *International Conference on Information and Knowledge Management (CIKM)*, 2010.
  - [14] J. Berni, P. J. Zarco-Tejada, L. Suárez, and E. Fereres. Thermal and narrowband multispectral remote sensing for vegetation monitoring from an unmanned aerial vehicle. *Geoscience and Remote Sensing, IEEE Transactions on*, 47(3):722–738, 2009.
  - [15] M. S. Bewley, B. Douillard, N. Nourani-Vatani, A. Friedman, O. Pizarro, and S. B. Williams. Automated species detection: An experimental approach to kelp detection from sea-floor auv images. In *Australasian Conference on Robotics and Automation (ACRA) 2012*, 2012.
  - [16] S. Branson, O. Beijbom, and S. Belongie. Efficient large-scale structured learning. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1806–1813. IEEE, 2013.
  - [17] J. F. Bruno and E. R. Selig. Regional decline of coral cover in the indo-pacific: timing, extent, and subregional comparisons. *PLoS one*, 2(8):e711, 2007.
  - [18] M. Bryson, M. Johnson-Roberson, R. J. Murphy, and D. Bongiorno. Kite aerial photography for low-cost, ultra-high spatial resolution multi-spectral mapping of intertidal landscapes. *PLoS one*, 8(9):e73550, 2013.
  - [19] L. M. Burke, K. Reytar, M. Spalding, A. Perry, et al. *Reefs at risk revisited*. World Resources Institute Washington, DC, 2011.
  - [20] J. G. Canadell, C. Le Quéré, M. R. Raupach, C. B. Field, E. T. Buitenhuis, P. Ciais, T. J. Conway, N. P. Gillett, R. A. Houghton, and G. Marland. Contributions to accelerating atmospheric co₂ growth from economic activity, carbon intensity, and efficiency of natural sinks. *Proceedings of the National Academy of Sciences*, 104(47):18866–18870, 2007.
  - [21] J. H. Carleton and T. J. Done. Quantitative video sampling of coral reef benthos: large-scale application. *Coral Reefs*, 14(1):35–46, 1995.
  - [22] J. Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational*

linguistics, 22(2):249–254, 1996.

- [23] N. Carlevaris-Bianco, A. Mohan, and R.M. Eustice. Initial results in underwater single image dehazing. In *OCEANS*, 2010.
- [24] H. Cesar, L. Burke, and L. Pet-Soede. The economics of worldwide coral reef degradation. *Technical Report*, 2003.
- [25] C-C. Chang and C-J. Lin. LIBSVM: A library for support vector machines. *ACM TIST*, 2011.
- [26] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *Proc British Machine Vision Conference (BMVC)*, 2011.
- [27] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 1996.
- [28] P. B. Conn, J. M. Ver Hoef, B. T. McClintock, E. E. Moreland, J. M. London, M. F. Cameron, S. P. Dahle, and P. L. Boveng. Estimating multispecies abundance using automated detection systems: ice-associated seals in the bering sea. *Methods in Ecology and Evolution*, 5(12):1280–1293, 2014.
- [29] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 1995.
- [30] R. Costanza, R. d’Arge, R. Groot, S. Farber, M. Grasso, B. Hannon, K. Limburg, S. Naeem, R. V. O’Neill, J. Paruelo, et al. The value of the world’s ecosystem services and natural capital. *Nature*, 4:520–522, 1997.
- [31] O. G. Cula and K. J. Dana. 3d texture recognition using bidirectional feature histograms. *IJCV*, 59:2004, 2004.
- [32] P. F. Culverhouse, N. Macleod, R. Williams, M. C. Benfield, R. M. Lopes, and M. Picheral. An empirical assessment of the consistency of taxonomic identifications. *Marine Biology Research*, 10(1):73–84, 2014.
- [33] P. F. Culverhouse, R. Williams, B. Reguera, V. Herry, and S. Gonzales-Gil. Do experts make mistakes? a comparison of human and machine identification of dinoflagellates. *Marine Ecology Progress Series*, 247:17–25, 2003.
- [34] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.
- [35] K. J. Dana, B. Van Ginneken, S. K. Nayar, and J. J. Koenderink. Reflectance and texture of real-world surfaces. *ACM TOG*, 18, 1999.

- [36] C. D'Angelo, A. Denzel, A. Vogt, M. V. Matz, F. Oswald, A. Salih, G. U. Nienhaus, and J. Wiedenmann. Blue light regulation of host pigment in reef-building corals. *Marine Ecology Progress Series*, 364:97–106, 2008.
- [37] W. E. Deming. *Some theory of sampling*. Courier Dover Publications, 1966.
- [38] S. G. Dove, O. Hoegh-Guldberg, and S. Ranganathan. Major colour patterns of reef-building corals are due to a family of gfp-like proteins. *Coral reefs*, 19(3):197–204, 2001.
- [39] A. Eleftheriou. *Methods for the study of marine benthos*. John Wiley & Sons, 2013.
- [40] A. Elibol, N. Gracias, R. Garcia, A. Gleason, B. Gintert, D. Lirman, and P. R. Reid. Efficient autonomous image mosaicing with applications to coral reef monitoring. In *IROS 2011 Workshop on Robotics for Environmental Monitoring, San Francisco, CA*, 2011.
- [41] J. B. T. Evans. *Bias in human reasoning: Causes and consequences*. Lawrence Erlbaum Associates, Inc, 1989.
- [42] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.
- [43] R-E. Fan, K-W. Chang, C-J. Hsieh, X-R. Wang, and C-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.
- [44] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *CVIU*, 2007.
- [45] L. Fleiss, B. Levin, and M. C. Paik. The measurement of interrater agreement. In *Statistical Methods for Rates and Proportions*. Citeseer, 1981.
- [46] A. L. Friedman. *Automated interpretation of benthic stereo imagery*. PhD thesis, University of Sydney., 2013.
- [47] T. A. Gardner, I. M. Côté, J. A. Gill, A. Grant, and A. R. Watkinson. Long-term region-wide declines in caribbean corals. *Science*, 2003.
- [48] S. Getzin, K. Wiegand, and I. Schöning. Assessing biodiversity in forests using very high-resolution images and unmanned aerial vehicles. *Methods in Ecology and Evolution*, 2012.
- [49] Personal correspondence with M. Dunlap, NOAA Coral Reef Ecosystem Division, Honolulu, HI, 2012.

- [50] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 580–587. IEEE, 2014.
- [51] M. González-Rivero, P. Bongaerts, O. Beijbom, O. Pizarro, A. Friedman, A. Rodriguez-Ramirez, B. Upcroft, D. Laffoley, D. Kline, C. Bailhache, et al. The catlin seaview survey—kilometre-scale seascape assessment, and monitoring of coral reef ecosystems. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 24(S2):184–198, 2014.
- [52] L. E. Greene and W. S. Alevizon. Comparative accuracies of visual assessment methods for coral reef fishes. *Bulletin of Marine Science*, 44(2):899–912, 1989.
- [53] J. Gu, S. Ghosal, and A. Roy. Bayesian bootstrap estimation of roc curve. *Statistics in medicine*, 2008.
- [54] D. J. Hand. Classifier technology and the illusion of progress. *Statistical Science*, 2006.
- [55] J. A. Hartigan and M. A. Wong. A K-means clustering algorithm. *Applied Statistics*, 1979.
- [56] J. Hill and C. Wilkinson. Methods for ecological monitoring of coral reefs. *Australian Institute of Marine Science, Townsville*, 117, 2004.
- [57] E. J. Hochberg and M. J. Atkinson. Capabilities of remote sensors to classify coral, algae and sand as pure and mixed spectra. *Remote Sensing of Environment*, 85, 2003.
- [58] O. Hoegh-Guldberg, P. J. Mumby, A. J. Hooten, R. S. Steneck, P. Greenfield, E. Gomez, C. D. Harvell, P. F. Sale, A. J. Edwards, K. Caldeira, et al. Coral reefs under rapid climate change and ocean acidification. *Science*, 2007.
- [59] D. J. Hopkins and G. King. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247, 2010.
- [60] T. P. Hughes, A. H. Baird, D. R. Bellwood, M. Card, S. R. Connolly, C. Folke, R. Grosberg, O. Hoegh-Guldberg, J. B. C. Jackson, J. Kleypas, et al. Climate change, human impacts, and the resilience of coral reefs. *Science*, 301(5635):929–933, 2003.
- [61] E. R. Hunt, W. D. Hively, S. J. Fujikawa, D. S. Linden, C. S. T. Daughtry, and G. W. McCarty. Acquisition of nir-green-blue digital photographs from unmanned aircraft for crop monitoring. *Remote Sensing*, 2010.
- [62] J. S. Jaffe, K. D. Moore, J. McLean, and M. P. Strand. Underwater optical imaging: status and prospects. *Oceanography*, 14(3):66–76, 2001.
- [63] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of*

the ACM International Conference on Multimedia, pages 675–678. ACM, 2014.

- [64] M. P. Johnson and P. L. Tyack. A digital acoustic recording tag for measuring the response of wild marine mammals to sound. *Oceanic Engineering, IEEE Journal of*, 28(1):3–12, 2003.
- [65] G. P. Jones IV, L. G. Pearlstine, and H. F. Percival. An assessment of small unmanned aerial vehicles for wildlife research. *Wildlife Society Bulletin*, 34(3):750–758, 2006.
- [66] C. Kanan, A. Flores, and G. W. Cottrell. Color constancy algorithms for object and face recognition. In *Advances in Visual Computing*, pages 199–210. Springer, 2010.
- [67] J. C. Kenyon, R. E. Brainard, R. K. Hoeke, F. A. Parrish, and C. B. Wilkinson. Towed-diver surveys, a method for mesoscale spatial assessment of benthic reef habitat: A case study at midway atoll in the hawaiian archipelago. *Coastal Management*, 2006.
- [68] J. C. Kenyon, R. E. Brainard, R. K. Hoeke, F. A. Parrish, and C. B. Wilkinson. Towed-diver surveys, a method for mesoscale spatial assessment of benthic reef habitat: a case study at midway atoll in the hawaiian archipelago. *Coastal Management*, 34(3):339–349, 2006.
- [69] Rahat Khan, Joost Van de Weijer, Fahad Shahbaz Khan, Damien Muselet, Christophe Ducottet, and Cecile Barat. Discriminative color descriptors. *CVPR*, 2013.
- [70] 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*, 2006.
- [71] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [72] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174, 1977.
- [73] L. Lapierre and B. Jouvencel. Robust nonlinear path-following control of an auv. *Oceanic Engineering, IEEE Journal of*, 2008.
- [74] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. IEEE, 2006.
- [75] Y. Le Cun, D. Touresky, G. Hinton, and T. Sejnowski. A theoretical framework for back-propagation. In *The Connectionist Models Summer School*, volume 1, pages 21–28, 1988.

- [76] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361:310, 1995.
- [77] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision*, 43(1):29–44, 2001.
- [78] D. D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the eleventh international conference on machine learning*, pages 148–156, 1994.
- [79] David G Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.
- [80] N. MacLeod, M. Benfield, and P. Culverhouse. Time to automate identification. *Nature*, 2010.
- [81] J. Malik, S. Belongie, J. Shi, and T. Leung. Textons, contours and regions: Cue integration in image segmentation. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 918–925. IEEE, 1999.
- [82] S. Manel, H. C. Williams, and S. J. Ormerod. Evaluating presence–absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, 38(5):921–931, 2001.
- [83] L. M Manevitz and M. Yousef. One-class svms for document classification. *the Journal of machine Learning research*, 2:139–154, 2002.
- [84] 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.
- [85] D.R. Martin, C.C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(5):530–549, 2004.
- [86] M. V. Matz, K. A. Lukyanov, and S. A. Lukyanov. Family of the green fluorescent protein: journey to the end of the rainbow. *Bioessays*, 24(10):953–959, 2002.
- [87] C. H. Mazel. Diver-operated instrument for in situ measurement of spectral fluorescence and reflectance of benthic marine organisms and substrates. *Optical Engineering*, 36(9):2612–2617, 1997.
- [88] C. H. Mazel, M. P. Strand, M. P. Lesser, M. P. Crosby, B. Coles, and A. J. Nevis. High-resolution determination of coral reef bottom cover from multispectral fluorescence laser line scan imagery. *Limnology and oceanography*, 48:522–534, 2003.

- [89] I. Miller and R. Müller. Validity and reproducibility of benthic cover estimates made during broadscale surveys of coral reefs by manta tow. *Coral Reefs*, 18(4):353–356, 1999.
- [90] P. P. Molloy, M. Evanson, A. C. Nellas, J. L. Rist, J. E. Marcus, H. J. Koldewey, and A. C. J. Vincent. How much sampling does it take to detect trends in coral-reef habitat using photoquadrat surveys? *Aquatic Conservation: Marine and Freshwater Ecosystems*, 23(6):820–837, 2013.
- [91] R. A. Monserud and R. Leemans. Comparing global vegetation maps with the kappa statistic. *Ecological modelling*, 62(4):275–293, 1992.
- [92] P. J. Mumby and K. Anthony. Resilience metrics to inform ecosystem management under global change with application to coral reefs. *Methods in Ecology and Evolution*, 2015.
- [93] M. R. Myers, J. T. Hardy, C. H. Mazel, and P. Dustan. Optical spectra and pigmentation of caribbean reef corals and macroalgae. *Coral Reefs*, 18(2):179–186, 1999.
- [94] S. Nebiker, A. Annen, M. Scherrer, and D. Oesch. A light-weight multispectral sensor for micro uavopportunities for very high resolution airborne remote sensing. *The international archives of the photogrammetry, remote sensing and spatial information sciences*, 37:1193–1200, 2008.
- [95] J. W. Nicholson and A. J. Healey. The present state of autonomous underwater vehicle (auv) applications and technologies. *Marine Technology Society Journal*, 42(1):44–51, 2008.
- [96] R. Ninio, S. Delean, K. Osborne, and H. Sweatman. Estimating cover of benthic organisms from underwater video images: variability associated with multiple observers. *Marine Ecology Progress Series*, 265:107–116, 2003.
- [97] J. Noronha, E. Hysen, H. Zhang, and K. Z. Gajos. Platemate: crowdsourcing nutritional analysis from food photographs. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 1–12, 2011.
- [98] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *PAMI*, 2002.
- [99] R. J. Olson and H. M. Sosik. A submersible imaging-in-flow instrument to analyze nano-and microplankton: Imaging flowcytobot. *Limnology and Oceanography: Methods*, 5(6):195–203, 2007.
- [100] Good P. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer, 2005.
- [101] S. J. Pan and Q. Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010.

- [102] J. M. Pandolfi, J. B. C. Jackson, N. Baron, R. H. Bradbury, H. M. Guzman, T. P. Hughes, C. V. Kappel, F. Micheli, J. C. Ogden, H. P. Possingham, et al. Are U.S. coral reefs on the slippery slope to slime? *Science*, 307(5716):1725–1726, 2005.
- [103] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.
- [104] E. Pante and P. Dustan. Getting to the point: Accuracy of point count in monitoring ecosystem change. *Journal of Marine Biology*, 2012.
- [105] F. Parisi, F. Strino, B. Nadler, and Y. Kluger. Ranking and combining multiple predictors without labeled data. *Proceedings of the National Academy of Sciences*, 111(4):1253–1258, 2014.
- [106] M.R. Patterson and N.J. Relles. Autonomous underwater vehicles resurvey bonair: a new tool for coral reef management. In *Proceedings of the 11th International Coral Reef Symposium*, pages 539–543, 2008.
- [107] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *Proc Conf Computer Vision and Pattern Recognition (CVPR)*, 2007.
- [108] S. R. Phinn, C. M. Roelfsema, and P. J. Mumby. Multi-scale, object-based image analysis for mapping geomorphic and ecological zones on coral reefs. *International Journal of Remote Sensing*, 33(12):3768–3797, 2012.
- [109] T. Randen and J.H. Husoy. Filtering for texture classification: A comparative study. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(4):291–310, 1999.
- [110] A. Rango, A. Laliberte, J. E. Herrick, C. Winters, K. Havstad, C. Steele, and D. Browning. Unmanned aerial vehicle-based remote sensing for rangeland assessment, monitoring, and management. *Journal of Applied Remote Sensing*, 3(1):033542–033542, 2009.
- [111] M. R. Raupach, G. Marland, P. Ciais, C. Le Quéré, J. G. Canadell, G. Klepper, and C. B. Field. Global and regional drivers of accelerating co2 emissions. *Proceedings of the National Academy of Sciences*, 104(24):10288–10293, 2007.
- [112] R. Rifkin and A. Klautau. In defense of one-vs-all classification. *The Journal of Machine Learning Research*, 5:101–141, 2004.
- [113] C. Roelfsema, S. Phinn, and W. Dennison. Spatial distribution of benthic microalgae on coral reefs determined by remote sensing. *Coral Reefs*, 21(3):264–274, 2002.
- [114] F. Rohwer and M. Youle. *Coral Reefs in the Microbial Seas*, volume 1. Plaid Press, 2010.
- [115] R. M. Royall and W. G. Cumberland. An empirical study of the ratio estimator and

- estimators of its variance. *Journal of the American Statistical Association*, 76(373):66–77, 1981.
- [116] S. Russell and P. Norvig. Artificial intelligence: A modern approach. *Prentice-Hall, Englewood Cliffs*, 25, 1995.
  - [117] C. L. Sabine, R. A. Feely, N. Gruber, R. M. Key, K. Lee, J. L. Bullister, R. Wanninkhof, C. S. I. Wong, D. W. R. Wallace, B. Tilbrook, et al. The oceanic sink for anthropogenic CO₂. *Science*, 305(5682):367–371, 2004.
  - [118] A. Salih, A. Larkum, G. Cox, M. Kühl, and O. Hoegh-Guldberg. Fluorescent pigments in corals are photoprotective. *Nature*, 408(6814):850–853, 2000.
  - [119] C. Sawade, N. Landwehr, S. Bickel, and T. Scheffer. Active risk estimation. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 951–958, 2010.
  - [120] T. Schoening. *Automated detection in benthic images for megafauna classification and marine resource exploration: supervised and unsupervised methods for classification and regression tasks in benthic images with efficient integration of expert knowledge*. PhD thesis, University of Bielefeld, 2015.
  - [121] J. Seiler, A. Friedman, D. Steinberg, N. Barrett, A. Williams, and N. J. Holbrook. Image-based continental shelf habitat mapping using novel automated data extraction techniques. *Continental Shelf Research*, 45:87–97, 2012.
  - [122] M. Shiela, A. Marcos, L. David, E. Peñaflor, V. Ticzon, and M. Soriano. Automated benthic counting of living and non-living components in ngedarrak reef, palau via subsurface underwater video. *Environmental monitoring and assessment*, 145(1-3):177–184, 2008.
  - [123] A. S. M. Shihavuddin, N. Gracias, R. Garcia, A. C. R. Gleason, and B. Gintert. Image-based coral reef classification and thematic mapping. *Remote Sensing*, 5(4):1809–1841, 2013.
  - [124] H. Singh, R. Armstrong, F. Gilbes, R. Eustice, C. Roman, O. Pizarro, and J. Torres. Imaging coral i: imaging coral habitats with the seabed auv. *Subsurface Sensing Technologies and Applications*, 5(1):25–42, 2004.
  - [125] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477. IEEE, 2003.
  - [126] R. Snow, B. O’Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263, 2008.

- [127] A. Solow, C. Davis, and Q. Hu. Estimating the taxonomic composition of a sample when individuals are classified with error. *Marine ecology. Progress series*, 216:309–311, 2001.
- [128] H. M. Sosik and R. J. Olson. Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. *Limnology and Oceanography: Methods*, 5(6):204–216, 2007.
- [129] M.D. Stokes and G.B. Deane. Automated processing of coral reef benthic images. *Limnology and Oceanography: Methods*, 7(2):157–168, 2009.
- [130] S. K. Thompson. *Sampling*. John Wiley & Sons, Inc., 2012.
- [131] K. S. Tkachenko, B-J. Wu, L-S. Fang, and T-Y. Fan. Dynamics of a coral reef community after mass mortality of branching acropora corals and an outbreak of anemones. *Marine Biology*, 151(1):185–194, 2007.
- [132] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1521–1528. IEEE, 2011.
- [133] T. Treibitz, B. P. Neal, D. I. Kline, O. Beijbom, P. L. D. Roberts, B. G. Mitchell, and D. Kriegman. Wide field-of-view fluorescence imaging of coral reefs. *Scientific reports*, 5, 2015.
- [134] T. Treibitz and Y.Y. Schechner. Active polarization descattering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(3):385–399, 2009.
- [135] V. Trygonis and M. Sini. photoquad: A dedicated seabed image processing software, and a comparative error analysis of four photoquadrat methods. *Journal of Experimental Marine Biology and Ecology*, 424:99–108, 2012.
- [136] E. Tusa, A. Reynolds, D. M. Lane, N. M. Robertson, H. Villegas, and A. Bosnjak. Implementation of a fast coral detector using a supervised machine learning and gabor wavelet feature descriptors. In *Sensor Systems for a Changing Ocean (SSCO), 2014 IEEE*, pages 1–6. IEEE, 2014.
- [137] R. Van Hooidonk, J. A. Maynard, and S. Planes. Temporary refugia for coral reefs in a warming world. *Nature Climate Change*, 3(5):508–511, 2013.
- [138] V. N. Vapnik. An overview of statistical learning theory. *Neural Networks, IEEE Transactions on*, 10(5):988–999, 1999.
- [139] M. Varma and A. Zisserman. A statistical approach to texture classification from single images. *International Journal of Computer Vision*, 62(1-2):61–81, 2005.
- [140] M. Varma and A. Zisserman. A statistical approach to material classification using image

- patch exemplars. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(11):2032–2047, 2009.
- [141] Manik Varma and Andrew Zisserman. A statistical approach to texture classification from single images. *IJCV*, 2005.
  - [142] X. Wang, T-K. Huang, and J. Schneider. Active transfer learning under model shift. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1305–1313, 2014.
  - [143] P. Welinder, S. Branson, P. Perona, and S. Belongie. The multidimensional wisdom of crowds. In *Advances in neural information processing systems*, pages 2424–2432, 2010.
  - [144] P. Welinder, M. Welling, and P. Perona. A lazy man’s approach to benchmarking: Semisupervised classifier evaluation and recalibration. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3262–3269. IEEE, 2013.
  - [145] T. J. Willis and R. C. Babcock. A baited underwater video system for the determination of relative density of carnivorous reef fish. *Marine and Freshwater research*, 51(8):755–763, 2000.
  - [146] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1800–1807. IEEE, 2005.
  - [147] G. Winters, R. Holzman, A. Blekhman, S. Beer, and Y. Loya. Photographic assessment of coral chlorophyll contents: implications for ecophysiological studies and coral monitoring. *Journal of Experimental Marine Biology and Ecology*, 380(1):25–35, 2009.
  - [148] D. G. Zawada and C. H. Mazel. Fluorescence-based classification of caribbean coral reef organisms and substrates. *PLoS one*, 9(1), 2014.
  - [149] D.G. Zawada and J.S. Jaffe. Changes in the fluorescence of the caribbean coral *Montastraea faveolata* during heat-induced bleaching. *Limnology and oceanography*, 48(1):412–425, 2003.