Automated Annotation of Corals in Natural Scene Images Using Multiple Texture Representations

Jean-Nicola Blanchet[1], Sébastien Déry[2], Jacques-André Landry[3], Kate Osborne[4]

[1]Laboratory for Imagery, Vision and Artificial Intelligence, École de technologie supérieure, Montreal, Quebec, Canada
[2]Faculty of Medicine, McGill University, Montreal, Quebec, Canada
[3]Department of Automation Engineering, École de technologie supérieure, Montreal, Quebec, Canada
[4]Australian Institute of Marine Science, Cape Cleveland, Queensland, Australia

ABSTRACT

Current coral reef health monitoring programs rely on biodiversity data obtained through the acquisition and annotation of underwater photographs. Manual annotation of these photographs is a necessary step, but has become problematic due to the high volume of images and the high cost of human resources. While automated and reliable multi-spectral annotation methods exist, coral reef images are often limited to visible light, which makes automation difficult. Much of the previous work has focused on popular texture recognition methods, but the results remain unsatisfactory when compared to human performance for the same task. In this work, we present an improved automatic method for coral image annotation that yields consistent accuracy improvements over existing methods. Our method builds on previous work by combining multiple feature representations. We demonstrate that the aggregation of multiple methods outperforms any single method. Furthermore, our proposed system requires virtually no parameter tuning, and supports rejection for improved results. Firstly, the complex texture diversity of corals is handled by combining multiple feature representations: local binary patterns, hue and opponent angle histograms, textons, and deep convolutional activation feature. Secondly, these multiple representations are aggregated using a score-level fusion of multiple support vector machines. Thirdly, rejection can optionally be applied to enhance classification results, and allows efficient semi-supervised image annotation in collaboration with human experts.

Keywords: Coral reef, Annotation, Texture features, Multi-classifier, Rejection

INTRODUCTION

Coral reef across the globe are endangered. In time, this will have a significant economic impact on many societies (Hoegh-Guldberg et al. 2007). To provide a scientific basis for reef preservation, protection and monitoring programs have been established. These however require information about the marine substrate coverage, which is typically obtained by manual labeling of images acquired underwater. Recently, it was shown that acquisition methods based on stereo-vision or multi-spectral imagery can be used to perform reliable automatic image annotation (Gleason et al. 2007; Johnson-Roberson et al. 2006; Sasano et al. 2013). Unfortunately, these re-

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2026v2 | CC-BY 4.0 Open Access | rec: 5 May 2016, publ: 5 May 2016

quire expensive equipment, and are not applicable to the large volume of unlabeled images gathered in the last decades acquired using simple digital or analog cameras. Furthermore, they were only shown to discriminate a few classes such as live coral, dead coral, sand and algae. Visible light imaging at a close range from the substrate has the advantage of containing detailed information on individual species. This leads to a much better understanding of coral reef ecosystems.

To extract data from these RGB photographs, various manual annotation protocols based on image content sampling have been adopted such as that proposed by Jonker et al. (2008). These provide specific directives on the software to use, the image and point sampling methodologies, the labeling categories (often called codes), the label decision process as well as other software parameters. While manual expert annotation can be used for biodiversity data extraction it is a time consuming task and cannot be applied to large datasets of benthic images given the available resources.

Challenges

Underwater natural scene images present multiple challenges around which our method is developed. While acquisition-related challenges may vary significantly from one dataset to another, the following list presents common difficulties encountered with underwater benthic images.

1. **Scale, orientation** and **illumination** varies. This is expected from organic objects in natural scene images.
2. **Red channel information loss** is a frequent artifact caused by red wavelength attenuation when traveling through water.
3. **Imbalanced data** is a common problem, as some coral species are extremely rare, while algae and sand samples are abundant for examples.
4. **Incorrect expert labeling** occurs because of the difficulty of the task. This affects the correctness of the ground truth and impacts the machine learning process. A study by Ninio et al. (2003) presented quantitative data on the disagreement frequency between multiples experts for the task of coral annotation in analog video frames. The reported overall disagreement rate was between 10% and 20% depending on the taxonomic ranking used which affects the difficulty of the task. While most classes were labeled with an estimated error rate well below 10%, error rates of up to 60% were reported for some of the rarest classes, mostly at the life form ranking. Nonetheless, these results are important, as they provide an approximation of the satisfactory accuracy threshold, which is considered to be between 80% and 90%.
5. The **sampling method** used for manual annotation is typically random or systematic (grid or uniform sampling) which causes considerable ambiguity: points are often close to the boundary between two or more observable classes. This ambiguity is responsible for much of the incorrect expert labeling, and poses a challenge for training in automation, because these points are used as a ground truth. Furthermore, sampled points are usually not the center of a homogenous region, which adds considerable background noise to the local texture.
6. The **complex environment** of underwater natural scene images contains many irrelevant, occluding objects, such as fishes, markers, acquisition equipment (*e.g.* quadrat). When performing manual annotation, these are either ignored, or simply flagged as part of the

class "others". The large diversity of objects and the few samples make the task of modeling these objects difficult.

7. Many classes are **difficult to model.** Firstly, a high intra-class variance caused by many environmental and geographical factors result in significant variations in coral appearances. The shape, hue, and texture of a single coral or algae class varies. Regardless of the taxonomic ranking, all classes should be considered multi-modal. Secondly, low inter-class variance can cause two coral classes to appear identical to a non-expert eye.

Previous work

A typical recognition problem is often broken down into a set of sub-problems: preprocessing, segmentation, feature extraction and classification. In this section, we briefly survey previous work on each sub problem, and discuss how previous efforts have addressed some of the challenges.

Preprocessing of underwater images was largely studied and methods have been developed to correct common underwater image artifacts (Bazeille et al. 2006; Carlevaris-Bianco et al. 2010; Prabhakar & Kumar 2012). However, the resulting image quality is subjective to the observer, and because acquisition conditions can be so different, these methods may not apply well to all datasets. The current trend seems to be to apply image correction and enhancement steps according to the dataset based on empirical results (Beijbom et al. 2012; Shihavuddin et al. 2013), i.e. finding the best preprocessing method for a given dataset. This simple approach targets dataset-specific acquisition artifacts. Most importantly, preprocessing can address the red channel information loss challenge, which is necessary for extracting useful color features.

Segmentation finds a region of interest in the surrounding of a labeled point. Promising work was done by Tusa et al. (2014) using a three class supervised pixel classification method based on Gabor wavelet response. Other semi-supervised methods for similar problems have been proposed by Costa & Battista (2013) as well as Neal et al. (2015) to help experts label entire images. Despite these efforts, unsupervised segmentation on a full scale dataset remains an open problem and is beyond the scope of this work. We therefore settle for fixed-size patches around target points. Segmentation however has two expected benefits: it may improve classification accuracy by ignoring irrelevant background information, and would yield additional surface coverage data. Full image segmentation of contiguous regions of homogenous textures may overcome the challenge related to the ambiguous sampling method. A few reasons explain why unsupervised segmentation remains difficult to apply. Firstly, the wide range of textures found in benthic images make segmentation methods difficult to parametrize. Secondly, ground truths often consists of single points which cannot be used to measure the accuracy of a segmentation method on a large scale dataset. Furthermore, quality metrics used for segmentation such as Dice index and Jaccard index do not consider that only a fraction of the coral's texture is sufficient to perform texture based recognition. Thirdly, given a point, it is difficult to segment a surrounding region in accordance with the expertly labeled ground truth: labels found on coral boundaries where two objects meet are ambiguous and the resulting region may not reflect the expert’s intentions.

Feature extraction has been the focus of much work. The dominant approach consists in a feature level fusion (i.e. concatenation of multiple feature vectors) of statistical features and global descriptors invariant to scale, orientation and illumination. These include intensity histogram statistics, gray-level co-occurrence matrix (GLCM) statistics, Gabor wavelet response statistics, local binary patterns (LBP), hue histograms, etc. Such methods are well established for the texture recognition problem, and were applied several times to automated coral images annotation (Bewley et al. 2012; Bouchard 2011; Marcos et al. 2005; Prévost 2015; Shihavuddin et al. 2013). Though these feature representations are popular, none was shown to perform at an acceptable level on a full-scale natural scene image dataset. Recently, a powerful dictionary-based texture descriptor, textons, was proposed as a feature representation by Beijbom et al. (2012). The method was shown to achieve between 67% and 83% accuracy for a nine-class dataset of natural images with over one hundred thousand labeled points. Dictionary-based methods were further investigated by Bewley et al. (2015) using small patches represented with principal component analysis dimensionality-reduced intensity values. Their results, however, suggest that a simple LBP representation remains competitive with such methods.

Classification was attempted using several classifiers such as nearest neighbors and neural networks (Marcos et al. 2005; Shihavuddin et al. 2013). However, the radial basis function (RBF) kernel support vector machines (SVM) has yielded much more promising results (Beijbom et al. 2012; Bouchard 2011). These have been widely used for various texture recognition problems. SVMs have the distinct advantage of being very flexible: they can be trained for regression to produce a likelihood estimation, and they are known to perform well when assembled into a broader multi-classifier system. Score-level fusion of multiple SVM has been applied to complex data in many fields to improve accuracy, including remote sensing (Waske & Benediktsson 2007) and biometrics (He et al. 2010). While this remains unexplored by previous work, a multi-classifier modeling approach would be more appropriate given the challenge of modeling the high diversity of complex textures.

Data

The dataset was provided by the Australian Institute of Marine Science (AIMS) (Sweatman et al. 2001). It contains 15,165 images of the Great Barrier Reef acquired between 2006 and 2012 over hundreds of unique transects. Image acquisition was performed underwater with a 6 mm lens at a distance of approximately 50 cm from the substrate, resulting in 25 x 34 cm ground coverage per image. Two different resolutions are used: 3,264 x 2,448 pixels for images from 2006 to 2010 and 2,112 x 2,816 pixels for 2011 and 2012. No artificial light source was used, resulting in images of variable illumination and sharpness.

Each image was expertly hand labeled at five distinct points located at the following relative coordinates $(x, y) = \left(\frac{1}{4}, \frac{1}{4}\right), \left(\frac{1}{4}, \frac{3}{4}\right), \left(\frac{1}{2}, \frac{1}{2}\right), \left(\frac{3}{4}, \frac{1}{4}\right), \left(\frac{3}{4}, \frac{3}{4}\right)$. This task was performed at five taxonomic rankings, from the highest to the lowest: group description, benthos description, family, genus, species description. To handle this multi-level annotation, we initially use the lowest species ranking, and then perform simple mapping to higher levels. This allows us to present results at all levels, while focusing our analysis to the broader group description level. The only exception is the benthos description rank, which does not follow a clear one-to-many tree-like mapping structure. While there are multiple possible solutions, overcoming this limitation is beyond the scope of this work. We have therefore ignored the benthos description level.

Across all taxonomic rankings, the number of samples per class is subject to an extreme variance. Many classes contain too few examples for practical machine learning. To overcome the imbalanced data challenge, a reasonable number of samples is required. Consequently, classes with less than 50 samples at the species rank were eliminated, therefore filtering out 0.83% of the data. This was done in previous work: Beijbom et al. (2012) discarded 4% of the Moorea labeled corals dataset for the same reason. Table 1 presents statistical data on the number of classes as well as the number of samples per class used in the filtered dataset at all four rankings of interest when performing mapping. Figure 1 shows examples of the 300x300 pixel patches used for twenty classes at the species level.

Table 1. Statistics on the classes and the number of samples per class (spc) at every taxonomic rank in the filtered AIMS dataset. Classes with less than 50 samples at the species level were eliminated.

	Group	**Family**	**Genus**	**Species**
Number of classes eliminated	1	12	63	148
Number of classes used	6	30	54	76
Mean spc	12,532	2,507	1,393	989
Median spc	4,987	456	296	190

METHOD

We have presented the general challenges concerning benthic images as well as previously proposed solutions at the preprocessing, feature extraction or classification levels. The errors caused by the ambiguous point sampling methods will be investigated in future work on segmentation, and we accept the error for now. Amongst the more difficult aspects of the coral classification task are the high intra-class variance and low inter-class variance. The data is complex and difficult to model using a single feature representation. Each class should be thought of as following a multi-modal distribution, where each mode may be significantly different than the previous. We hypothesize that, given the complexity of the textures, different modes within a single class require different features to be properly characterized. Previous work on feature extraction has shown that integrating new feature representations using feature-level fusion tends to increase accuracy. However, feature-level fusion has two main limitations. Firstly, the computational cost of training and testing a classifier increases dramatically with the size of the feature vector, thus making large vectors impractical. Secondly, classification accuracy does not necessarily consistently improve as more features are added, even if these are uncorrelated and discriminant. This effect is known as the "curse of dimensionality", which we can observe here when combining too many features.

Alternatively, we propose using a score-level fusion to aggregate the complex information from multiple feature representations. Our proposed method uses three state-of-the-art feature representations: Local binary patterns combined with color information, textons, and a convolutional neural network-based feature. While we only use these three representations, the method can be easily extended to support additional ones.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2026v2 | CC-BY 4.0 Open Access | rec: 5 May 2016, publ: 5 May 2016

A score-level fusion uses an aggregation function to combine the class-wise prediction likelihood estimation from multiple classifiers. The fusion process creates a fusion score as a by-product which can then be used to implement rejection. Because all class predictions come with their own score, a threshold can be set to reject lower score predictions, which are more likely to be errors than high score predictions.

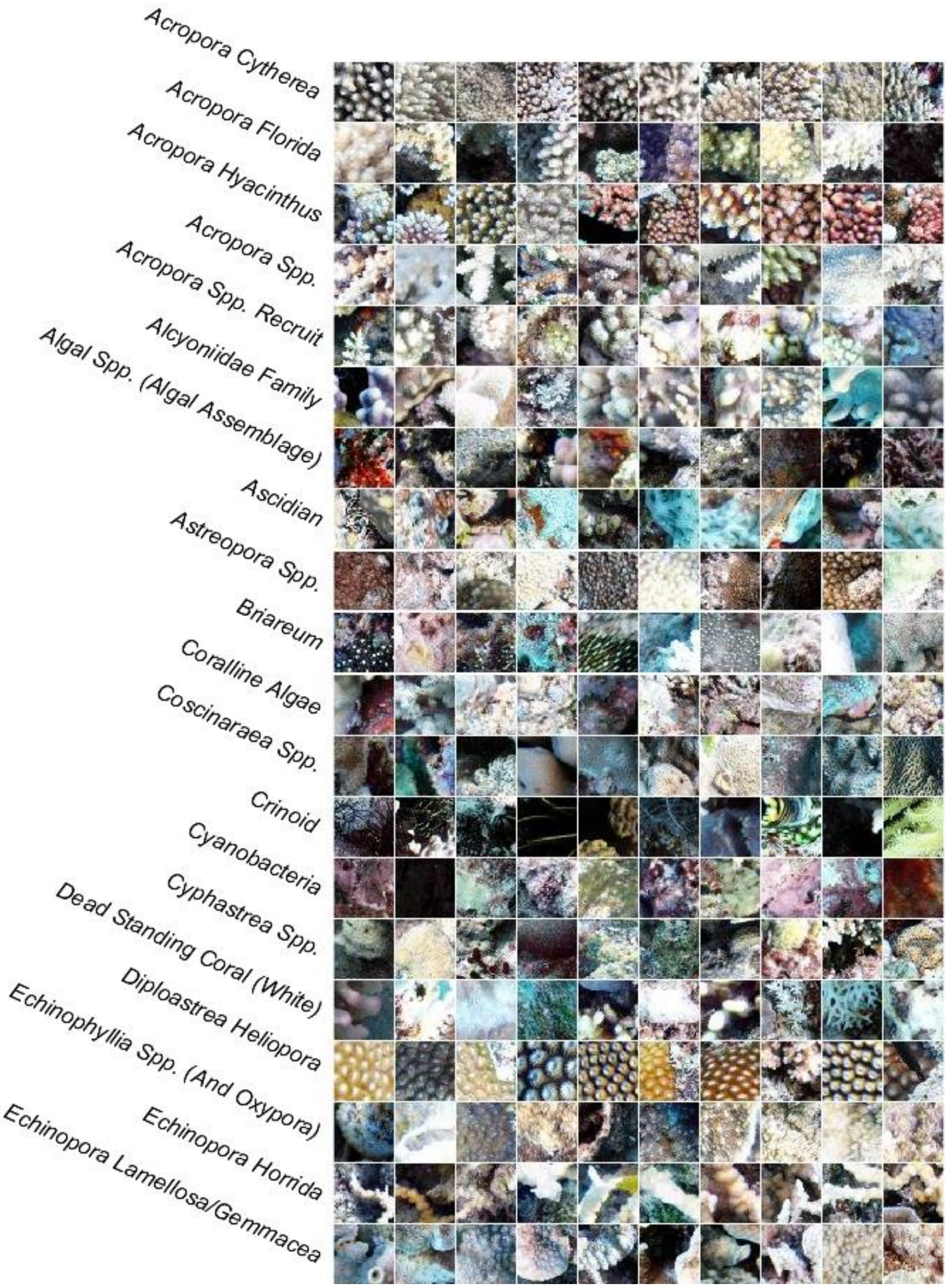

Figure 1. Ten samples per class for the first 20 classes from the AIMS dataset at the species description ranking. Histogram equalization was applied to enhance visualization.

Feature Extraction

A combination of *completed local binary patterns* (CLPB) and *color information* inspired by previous work (Shihavuddin et al. 2013) is used as the first texture representation. We reduced computational complexity and vector dimensionality by using simpler parameters, and eliminating most of the descriptors. These choices were made heuristically, as we found that fine optimization had a significant effect on the classification accuracy when using the vector by itself, but had almost none on the multi-classifier fusion results. We emphasize that our approach does not require fine optimization of any parameter to achieve reasonable performance. The following descriptors were combined in this first representation:

1. **CLBP** were initially proposed by Guo et al. (2010) and use non-linear mapping functions to describe the local pattern around each pixel using binary codes. We used a sampling of eight neighbors at a distance of one pixel, and applied the uniform rotation-invariant mapping to reduce the bin count while achieving orientation invariance. The three histograms (sign, center and magnitude) are aggregated by concatenating a 20 bin 2d center-magnitude joint histogram with a 10 bin 1d sign histogram. This results in 30 CLBP bins.
2. **Hue** and **opponent angle** histograms are powerful color descriptors tolerant to geometric and photometric variations proposed by Van De Weijer & Schmid (2006). Both hue and opponent angle values are averaged over blocs of 20 by 20 pixels, which are then quantized into 16-bin histograms, resulting in 32 color feature bins. We found that applying the comprehensive image normalization (Finlayson et al. 1998) yielded better color features, as inconsistent red channel attenuation causes inaccurate representations biased towards the cyan color.

Textons were applied to coral reef image annotation by Beijbom et al. (2012). A texton is a quantized pixel response to a Gabor filter bank. Initially, a dictionary of textons is learned for quantization using clustering. Then, a texton feature vector can be computed in the form of a normalized histogram of the textons found on the texture patch. We used the same 135-texton dictionary (Beijbom et al. 2012) trained using the Maximum Response (MR) Gabor filter bank, applied the same channel stretch image enhancement for color consistency, and used the Lab color space. We also extracted textons using four square patch sizes: 300, 165, 80 and 30 pixels similar to how it was done in the original work. This reportedly addresses part of the challenges related to the lack of segmentation and variable texture scale, and results in a 540-value feature representation. These patch sizes are free parameters in our system. However, we've observed that their optimization has very little impact. This was also the conclusion of the work by Prévost (2015).

Deep convolution activation feature (DeCAF) is a transfer learning description method based on activation weights of the last convolution layer in a convolutional neural network (CNN). While transfer learning using CNNs is an active field a research, DeCAF was applied to texture recognition by Cimpoi et al. (2014), which demonstrated that a CNN trained for object recognition provided features able to describe textures that are statistically different, yet semantically alike. We applied this technique here as our third feature representation, using the CNN trained by Simonyan & Zisserman (2014) on the ImageNet object recognition dataset, made available by the MatConvNet library (Vedaldi & Lenc 2014). The 4,096 activation weights are extracted directly from the last convolution layer, and normalized using the L1 norm. The extraction process requires no additional training or parameter optimization.

Classification

To aggregate these multiple representations, we use a score-level fusion, as presented in figure 2. Each feature representation trains a one-against-one regression SVM that provides a probability estimate, or a score, for each possible class. We used the LIBSVM library to perform this task (Chang & Lin 2011). Grid search is performed on at most 60 samples per class from the training set to approximate the cost parameter as well as the gamma kernel parameter. The score outputs are then normalized and aggregated using a product fusion function and the maximum aggregated score value indicates the predicted class. We initially applied several fusion functions including sum, mean, product, maximum and vote, and found that the product function was consistently better on all tested datasets, even on benchmark texture datasets. We therefore settle for product fusion, which eliminates the need for costly adaptive fusion function selection. While product is not a commonly used fusion function, we explain its performance here by the high frequency of low score scores due to the high class count as well as the low inter-class variance. The resulting score from fusion can optionally be used for rejection thresholding, thereby further enhancing the prediction accuracy.

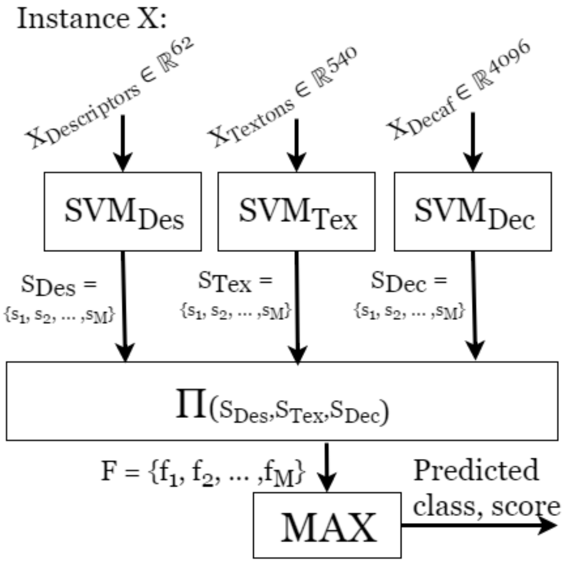
Instance X:
$X_{Descriptors} \in \mathbb{R}^{62}$
$X_{Textons} \in \mathbb{R}^{540}$
$X_{Decaf} \in \mathbb{R}^{4096}$
SVM_{Des}
SVM_{Tex}
SVM_{Dec}
$S_{Des} = \{s_1, s_2, ... ,s_M\}$
$S_{Tex} = \{s_1, s_2, ... ,s_M\}$
$S_{Dec} = \{s_1, s_2, ... ,s_M\}$
$\Pi(S_{Des},S_{Tex},S_{Dec})$
$F = \{f_1, f_2, ... ,f_M\}$
MAX
Predicted class, score

Figure 2. Our classification architecture using a score-level fusion of three texture representations: CLBP and color global descriptors (Des), Textons (Tex), and CNN activation weights (Dec). The other variables are the number of classes (M), the SVM scores (S), and the fusion score (F).

RESULTS AND ANALYSIS

Classification Accuracy

We ran a tenfold analysis on the filtered dataset of 75,195 patches at the species ranking with a RBF kernel SVM using each feature representation individually. Simple mapping to higher ranks is performed for other levels. Results are compared with the ones obtained using the proposed fusion method, and are shown in table 2. Both the overall and average accuracies observed using three aggregated feature representations are consistently higher at all ranking levels. These results suggest that not all textures can be described using a single feature representation, and that multiple representations are complimentary.

Table 2. Results before rejection: accuracy (Acc) and class-wise average accuracy (Avg Acc) for each feature representations individually, as well as for the multi-classifier fusion. The number of classes at each ranking is specified in parenthesis. The standard deviation across all 10 folds is also reported.

	Group Description		**Family**		**Genus**		**Species Description**	
	(6)		**(30)**		**(54)**		**(76)**	
	Acc	Avg Acc	Acc	Avg Acc	Acc	Avg Acc	Acc	Avg Acc
CLBP+hue+OA	54.9±0.6	47.8±1.1	39.6±0.6	27.2±0.7	36.6±0.5	21.3±0.5	35.0±0.6	19.4±0.6
Textons	61.5±0.9	55.1±1.5	45.8±0.9	35.7±1.0	42.9±0.8	29.9±0.9	40.6±0.9	26.9±1.2
DeCAF	62.5±0.9	57.1±1.3	44.0±1.2	40.0±1.5	41.1±1.1	37.2±1.8	37.7±1.1	35.5±1.6
Score-level fusion	**71.7±0.5**	**66.2±1.2**	**59.4±0.9**	**46.8±1.4**	**56.7±0.9**	**41.0±1.6**	**54.5±0.9**	**37.8±1.5**

Despite this significant accuracy improvement, the reported error before rejection is higher than the expected human performance for the same task. Much of the error can be explained by two of the challenges that have been ignored within the scope of this work.

1. Ambiguous patches due the sampling methods cause background information to appear in patches, therefore increasing the variance for the distribution of all classes in feature space. This further increases the difficulty of separating classes.
2. The ground truth used to measure the accuracy of our system is subject to expert errors, which is partially caused by patch ambiguity. The error should be seen as the disagreement rate between a single expert and our automated system, which are both prone to errors. Any level of performance above that of human would be insignificant, meaning that human performance is an upper bound to the measured accuracy of our automated system. A better accuracy metric could be obtained by querying the expert on the correctness of predicted labels over a sample. In addition, there are many cases of very dark or light saturated regions presenting no significant texture information where the expert is able to infer the correct class heuristically based on adjacent or previously encountered areas. These cases are difficult to model using machine learning.

In this work, we attempt to eliminate some of these errors through rejection by thresholding the post-fusion score. However, future work will investigate segmentation as a complimentary, but more appropriate solution.

Figure 3a presents the confusion matrix of the combined ten folds obtained at the coarsest group description ranking. The majority of the error comes from most classes being confused with hard corals, which is a complex class observed with many different textures and colors, and that occupies large portion of the feature space. Consequently, most other classes tend to overlap with hard coral in feature space. The sponge class has the highest error and is mostly confused with algae and hard coral. This is partially explained by the low representation of the sponge class, which remains difficult to model. Moreover, sponges in this particular data are mostly encrusting forms and therefore harder to identify than in many other benthic datasets. We propose combining the algae, other and sponge classes into a single “other” class considering they are confused and their separation is of little interest. For research purposes, it is most important to monitor, and therefore identify, the hard coral class above all, as ecologically it is of greatest interest and as a reef builder is the main contributor to coral reef structure. The resulting confusing matrix is shown in figure 3b, and yields an improved accuracy of 78.7% over 10 folds.

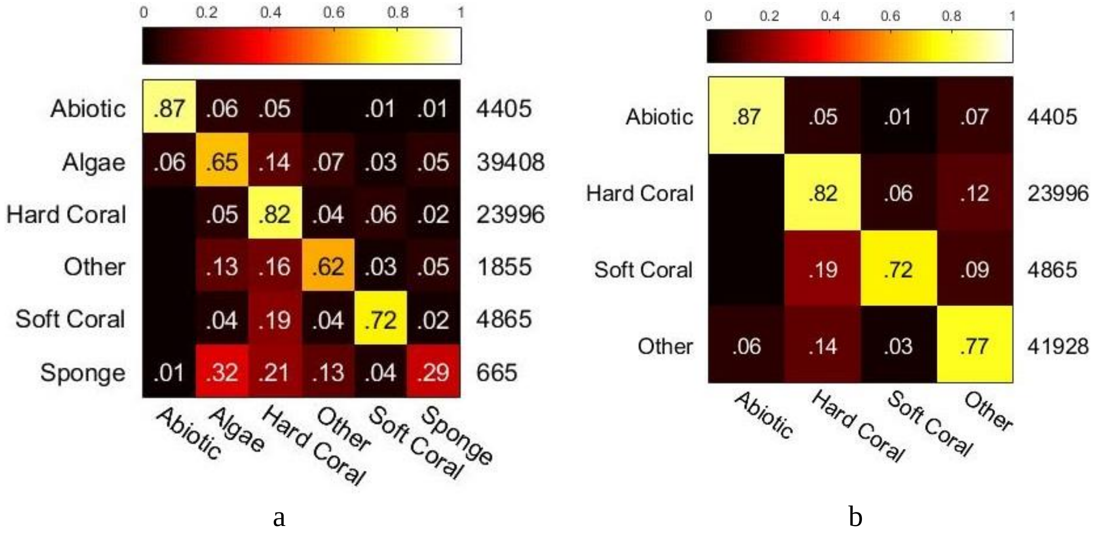

Figure 3. (a) Confusion matrix (before rejection) of the combined 10 folds for our proposed method at the group description ranking. (b) Confusion matrix (before rejection) after merging the algae, other and sponge classes into a single “other” class. Both matrices present the normalized prediction frequencies for all combination of real class (vertically) and predicted class (horizontally). The class frequency is displayed on the right side. Each cell represents the frequency at which a sample of the real class is classified as the predicted class.

Rejection

We propose using the fusion score to implement rejection in our system. We focused most of our analysis on the modified four-class group description ranking, as it offers close to satisfactory performance, but rejection can be applied at all taxonomic rankings. Figure 4 presents the ROC curves obtained for each class when attempting to separate correct predictions from errors using a score threshold. Ideally, we would like to eliminate all errors (false acceptance, or FA) and retain all correct predictions (true acceptance, or TA). While errors cannot be perfectly eliminated, good threshold candidates for rejection become apparent. For instance, at least 50 % of the prediction errors can be eliminated for the abiotic and soft coral classes, while losing no more than 15 % of correct predictions. The ROC curve however does not provide details on the TA and FA frequency, which are important to select good rejection thresholds. Figure 5 presents the class wise score probability density functions (PDF) for FAs and TAs. Rejection aims to find a score threshold that best separates the two populations. The four PDFs suggest that rejection is not equally beneficial for all classes, and that class-specific thresholds are important to optimally eliminate errors. To find these class-specific thresholds, we applied two methods:

1. Chow's rule (Chow 1970) is a rejection rule based on best error-reject trade-off from the PDF. While this method minimizes the absolute error, its main disadvantage is the lack of control over the number of rejected samples, which could be problematic for biodiversity statistic.
2. A greedy search algorithm is applied to minimize the absolute error with the desired rejection percentage as an exit condition. We used with method for two thresholds of 5% and 10% of the total data.

Both of these methods require knowledge of the score PDF, which we estimate for each fold using the scores from the training data. Table 2 reports the accuracies obtained using various thresholds. Moreover, we explore an alternative rejection application: a semi-supervised mode where most of the samples are automatically labeled, and the small rejected fraction is manually reviewed and corrected by an expert. This is done by setting the rejected samples to their ground truth label. The rationale is that blindly rejecting too many samples may lead to biased biodiversity statistics. There are also many advantages to a semi-supervised mode in an operational setting, such as online learning for improved and adaptive recognition. This is however beyond the scope of this work.

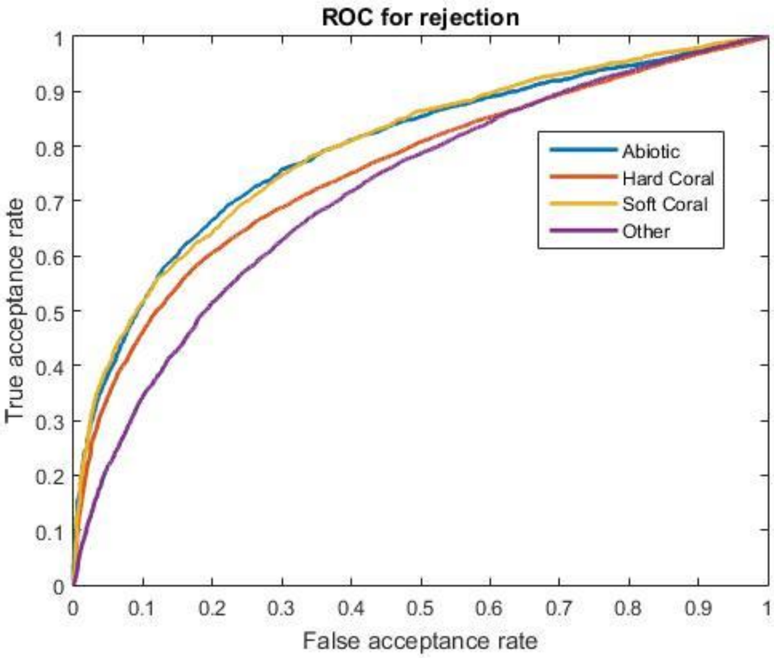

Figure 4. Rejection ROC curves. Every point represents the impact of rejection (true acceptance rate, false acceptance rate) for a unique rejection threshold.

We have established that rejection can be used to enhance results. However, we do not draw conclusions on the “optimal” rejection threshold selection method. Rejection is flexible and has many practical applications. The task at hand should be considered when applying it. For instance, for locating examples of a specific rare species in a large dataset, 80 % of the data can be rejected for good results. However, for practical surface estimation, rejection needs to be tuned carefully based on the error for important classes such as hard coral to avoid a bias towards classes that are easier to recognize. In a semi-supervised collaborative annotation mode between our automated system and an expert user, bias is no longer an issue, but the rejection percentage needs to consider the availability of human resources. The confusion matrices obtained with and without expert correction are presented in figure 6.

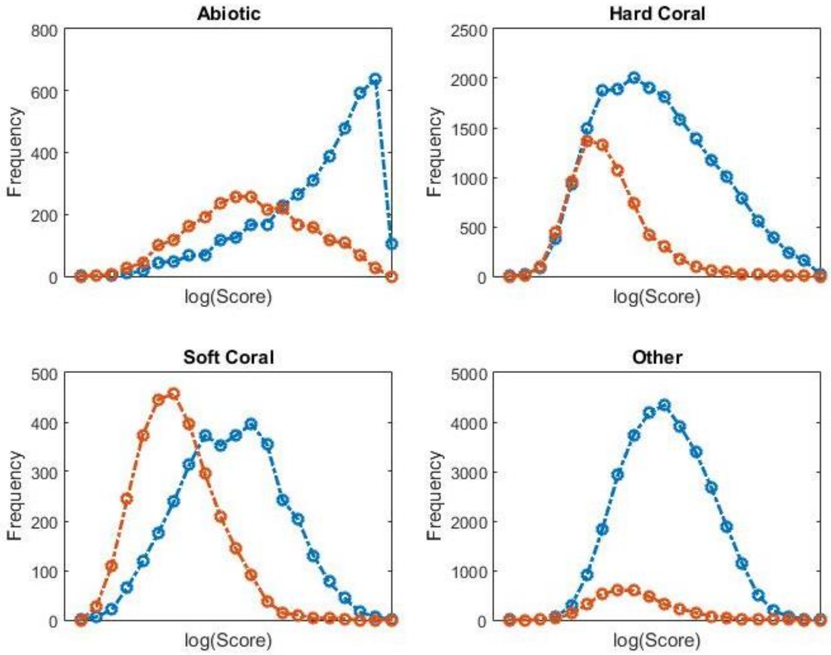

Figure 5. Class-wise score distribution of true acceptance (blue) and false acceptance (orange). The curves were smoothed by quantizing data to twenty points. The log scale for the score is used for visualization purposes to account for the product fusion function.

Table 2. Accuracies and Average Accuracies using different rejection thresholds at the group description level using 4 classes. Samples are either rejected and ignored (R) or rejected and manually corrected (R+C). The standard deviation across all 10 folds is also reported.

		Greedy selection				**Chow’s rule**	
	No Rejection	5% R	5% R+C	10% R	10% R+C	19.3% R	19.3% R+C
Accuracy	78.7±0.5	81.3±0.6	82.3±0.5	83.3±0.5	84.9±0.5	85.6±0.6	88.3±0.7
Average Acc.	79.5±0.9	80.1±1.0	81.7±0.9	81.3±1.0	83.8±0.9	82.5±1.6	86.9±1.4

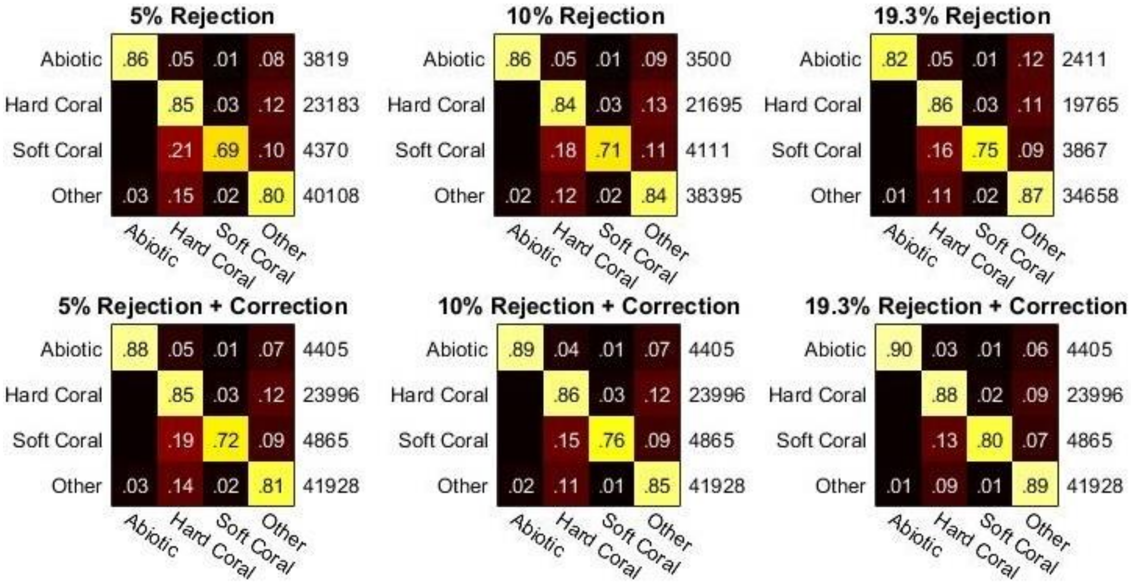

Figure 6. Confusion matrices at the group ranking using the three proposed thresholds (from left to right: 5%, 10%, and 19.3% using Chow's rule) for rejection without (top row) and with (bottom row) manual correction of the rejected samples. See figure 3 for details.

DISCUSSION

Comparison between state-of-the-art methods

While our results have demonstrated that combining multiple method through a score-level fusion yields consistently better results than using any single method, individual results from previously published state-of-the-art methods cannot be compared with each other. For a fair comparison, parameters of each method should be finely optimized for a given dataset in order to achieve maximum performance. For instance, we used a texton dictionary trained for the Moorea Labeled Coral dataset, which is unlikely to be the optimal dictionary for our dataset despite still performing reasonably well.

Feature representations

The three feature representations used in this work were selected based on the most promising state-of-the-art work on texture. This selection is however somewhat arbitrary. A better way of selecting feature representations could be to study the confusion matrix, and design features that are specifically good at discriminating classes that are currently confused. This could be done using unsupervised feature learning for example.

Parameters

Because large scale analysis is time consuming on regular hardware with limited resources, parameters selection is often an impractical process. We attempted to either eliminate or approximate heuristically every parameter selection step. The only exception is the SVM grid search model selection, which is efficiently approximated on about 6 % of the data. While fine optimization, in general, does improve results slightly, we empirically determined that its impact was negligible when applying score-level fusion. The lack of free parameter gives our system an out-of-the-box applicability to new datasets, which is a highly desirable aspect in any operational setting, and makes it a good candidate for a practical solution to the large scale coral annotation problem.

Cost of error

Our work did not cover extensively the cost of error. While we considered hard corals to be particularly important, the exact cost of error for each class was not discussed. This aspect is beyond the scope of this work, but it is incorrect to assume that confusing two types of algae has the same weight as confusing sand with soft coral in a biodiversity study. Nonetheless, even without rejection, our results have shown that a hard coral texture patch has a high chance of being classified correctly.

CONCLUSION

We demonstrated that pooling multiple texture representations at the score level using multiple support vector machines yields more accurate results for automated coral reef annotation when compared to using a single method. We have combined three texture representations: a mix of global descriptors (CLBP, hue and opponent angle histograms), textons, and activation weights from a deep convolutional neural network trained for object recognition. Rejection was applied on the resulting fusion scores to eliminate ambiguous points, and improve accuracy. Our multi-representation pooling system does not require fine tuning of any parameter to achieve reasonable performance, and can be extended to support additional representation designed specifically for a given dataset.

ACKNOWLEDGEMENTS

This research was made possible through the financial support of the Natural Sciences and Engineering Research Council of Canada and École de Technologie Supérieure. Thanks to the Australian Institute of Marine Science for providing high quality data.

REFERENCES

Bazeille S, Quidu I, Jaulin L, and Malkasse J-P. 2006. Automatic underwater image pre-processing. CMM'06. p xx.

Beijbom O, Edmunds PJ, Kline D, Mitchell BG, and Kriegman D. 2012. Automated annotation of coral reef survey images. Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on: IEEE. p 1170-1177.

Bewley M, Nourani-Vatani N, Rao D, Douillard B, Pizarro O, and Williams SB. 2015. Hierarchical classification in AUV imagery. Field and Service Robotics: Springer. p 3-16.

Bewley MS, Douillard B, Nourani-Vatani N, Friedman A, Pizarro O, and Williams SB. 2012. Automated species detection: An experimental approach to kelp detection from sea-floor AUV images. Proceedings of Australasian Conference on Robotics and Automation, Victoria University of Wellington, New Zealand.

Bouchard J. 2011. Méthodes de vision et d'intelligence artificielles pour la reconnaissance de spécimens coralliens. École de technologie supérieure (Master thesis). École de Techonologie Supérieure, Montreal, Quebec, Canada.

Carlevaris-Bianco N, Mohan A, and Eustice RM. 2010. Initial results in underwater single image dehazing. OCEANS 2010: IEEE. p 1-8.

Chang C-C, and Lin C-J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2:27.

Chow CK. 1970. On optimum recognition error and reject tradeoff. *Information Theory, IEEE Transactions on* 16:41-46.

Cimpoi M, Maji S, Kokkinos I, Mohamed S, and Vedaldi A. 2014. Describing textures in the wild. Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on: IEEE. p 3606-3613.

Costa B, and Battista T. 2013. The semi-automated classification of acoustic imagery for characterizing coral reef ecosystems. *International journal of remote sensing* 34:6389-6422.

Finlayson GD, Schiele B, and Crowley JL. 1998. Comprehensive colour image normalization. *Computer Vision—ECCV'98*: Springer, 475-490.

Gleason A, Reid R, and Voss K. 2007. Automated classification of underwater multispectral imagery for coral reef monitoring. OCEANS 2007: IEEE. p 1-8.

Guo Z, Zhang L, and Zhang D. 2010. A completed modeling of local binary pattern operator for texture classification. *Image Processing, IEEE Transactions on* 19:1657-1663.

He M, Horng S-J, Fan P, Run R-S, Chen R-J, Lai J-L, Khan MK, and Sentosa KO. 2010. Performance evaluation of score level fusion in multimodal biometric systems. *Pattern Recognition* 43:1789-1800.

Hoegh-Guldberg O, Mumby P, Hooten A, Steneck R, Greenfield P, Gomez E, Harvell C, Sale P, Edwards A, and Caldeira K. 2007. Coral reefs under rapid climate change and ocean acidification. *science* 318:1737-1742.

Johnson-Roberson M, Kumar S, Pizarro O, and Willams S. 2006. Stereoscopic imaging for coral segmentation and classification. OCEANS 2006: IEEE. p 1-6.

Jonker M, Johns K, and Osborne K. 2008. Surveys of benthic reef communities using underwater digital photography and counts of juvenile corals. Long-term Monitoring of the Great Barrier Reef. Standard Operational Procedure.

Marcos MSA, Soriano M, and Saloma C. 2005. Classification of coral reef images from underwater video using neural networks. *Optics express* 13:8766-8771.

Neal BP, Lin T-H, Winter RN, Treibitz T, Beijbom O, Kriegman D, Kline DI, and Mitchell BG. 2015. Methods and measurement variance for field estimations of coral colony planar area using underwater photographs and semi-automated image segmentation. *Environmental monitoring and assessment* 187:1-11.

Ninio R, Delean J, Osborne K, and Sweatman H. 2003. Estimating cover of benthic organisms from underwater video images: variability associated with multiple observers. *Marine Ecology-Progress Series* 265:107-116.

Prabhakar C, and Kumar P. 2012. An image based technique for enhancement of underwater images. *arXiv preprint arXiv:12120291*.

Prévost I. 2015. Application de la vision artificielle à l'identification de groupes benthiques dans une optique de suivi environnemental des récifs coralliens (Master thesis). École de Techonologie Supérieure, Montreal, Quebec, Canada.

Sasano M, Imasato M, Yamano H, and Oguma H. 2013. Optical Design & Engineering Monitoring the viability of coral reefs.

Shihavuddin A, Gracias N, Garcia R, Gleason AC, and Gintert B. 2013. Image-based coral reef classification and thematic mapping. *Remote Sensing* 5:1809-1841.

Simonyan K, and Zisserman A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:14091556*.

Sweatman HH, Cheal AA, Coleman GG, Delean SS, Fitzpatrick BB, Miller II, Ninio RR, Osborne KK, Page CC, and Thompson AA. 2001. Long-Term Monitoring of the Great Barrier Reef, Status Report Number 5.

Tusa E, Reynolds A, Lane DM, Robertson NM, Villegas H, and Bosnjak A. 2014. Implementation of a fast coral detector using a supervised machine learning and Gabor Wavelet feature descriptors. Sensor Systems for a Changing Ocean (SSCO), 2014 IEEE: IEEE. p 1-6.

Van De Weijer J, and Schmid C. 2006. Coloring local feature extraction. *Computer Vision–ECCV 2006*: Springer, 334-348.

Vedaldi A, and Lenc K. 2014. MatConvNet-convolutional neural networks for MATLAB. *arXiv preprint arXiv:14124564*.

Waske B, and Benediktsson JA. 2007. Fusion of support vector machines for classification of multisensor data. *Geoscience and Remote Sensing, IEEE Transactions on* 45:3858-3866.