Towards Generalized Benthic Species Recognition and Quantification using Computer Vision

Adam F. Gobi
Faculty of Engineering and Applied Science
Memorial University of Newfoundland
St. John's, NL, Canada
Email: afgobi@mun.ca

***Abstract*—Seabed resource exploitation and conservation efforts are extending to offshore areas where the distribution of benthic epifauna (animals living on the seafloor) is unknown. There is a need to survey these areas to determine how biodiversity is distributed spatially and to evaluate and monitor ecosystem states. Seafloor imagery, collected by underwater vehicles, offer a means for large-scale characterization of benthic communities. A single submersible dive can image thousands of square metres of seabed using video and digital still cameras. As manual, human-based analysis lacks large-scale feasibility, there is a need to develop efficient and rapid techniques for automatically extracting biological information from this raw imagery. To meet this need, underwater computer vision algorithms are being developed for the automatic recognition and quantification of benthic organisms. Focusing on intelligent analysis of distinct local image features, the work has the potential to overcome the unique challenges associated with visually interpreting benthic communities. The current incarnation of the system is a significant step towards generalized benthic species mapping, and its feature-based nature offers several advantages over existing technology.**

I. Introduction

Ocean-related activities are extending to offshore areas where little is known about the distribution of benthic epifauna (animals living on the seafloor). There is a need to survey these areas to determine such distributions and to evaluate and monitor ecosystem states. Seafloor imagery, collected by underwater vehicles, offer a means for large-scale characterization of benthic communities. A single submersible dive can image thousands of square metres of seabed using video and digital still cameras [1]. Visual and acoustic seafloor imaging by autonomous underwater vehicles has augmented data gathering capabilities to square kilometre scales [2]. The benthic information contained within this optical data has wide application in harvesting industries (e.g., stock assessment), the oil and gas industry (e.g., population monitoring for environmental impact assessment), as well as oceanographic research (e.g., population studies, habitat analysis).

It is evident that such large amounts of data quickly become unmanageable. In order to use these imaging tools effectively, there is a need to develop efficient means of extracting pertinent benthic information from the raw imagery. With few exceptions, this step is presently performed manually; researchers examine individual images and video sequences to, for example, identify and count organisms and study their relationship to the habitat. From both scientific and commercial points of view, it would be costly and simply impractical to scale up this labour-intensive process, as there are literally millions of images collected annually during submersible missions around the world. Clearly, a means of *automatically* extracting the desired knowledge from the collected imagery is a much more preferable solution.

This study is part of an ongoing research program that is developing a system capable of performing automated underwater image analysis that has wide, general application. The primary focus is the extraction of knowledge pertaining to the abundance and distribution of benthic epifauna. As such, the main goal is for the vision system is to find, identify, and count organisms appearing on the seafloor. Their diversity presents a highly varied array of visual challenges, making the achievement of generality the principal challenge.

II. Challenges Faced

Automating the task of benthic species recognition and quantification is non-trivial. Fig. 1 and Fig. 2 demonstrate some of the key challenges faced. Fig. 1 shows how organisms exist amongst varying background clutter, providing a potential source of confusion, as well as often causing partial occlusion. The organisms (in this case, sea cucumbers) are also in close proximity to each other, sometimes touching and obscuring their shape; this can be problematic as it could be difficult for a system to tell them apart, potentially mistaking their combined shape for something else. Fig. 2 shows an extreme case where starfish are heavily overlapped. Other significant challenges include problems such as variation in lighting, poor quality images with blurring or very low contrast, and particles in the water column causing reflectance noise.

The principal issue here is the high level of complexity of the imagery. It is not comparable to, for instance, a simple case of segmenting an outdoor image into sky, ground and trees; even human visual systems need to spend a little more time analyzing before they can extract anything meaningful. In the development of a solution, the system must have the capability to effectively seek out a target organism, and be able to quantify instances of it. The first is necessary when dealing with large amounts of clutter and other artifacts obscuring a

978-1-4244-5222-4/10/$26.00

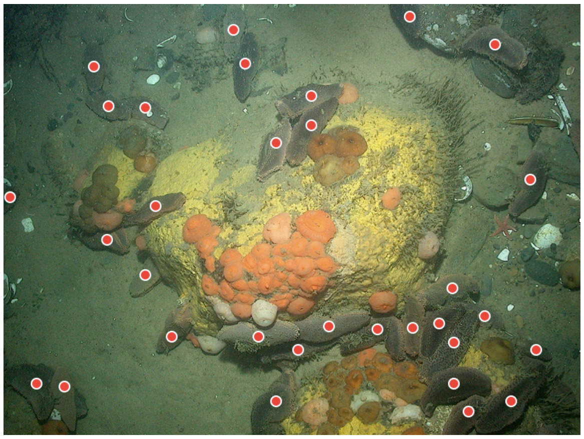


Fig. 1. Sea cucumbers in close quarters, manually annotated with red dots. Photo courtesy of John Anderson, Northwest Atlantic Fisheries Centre.

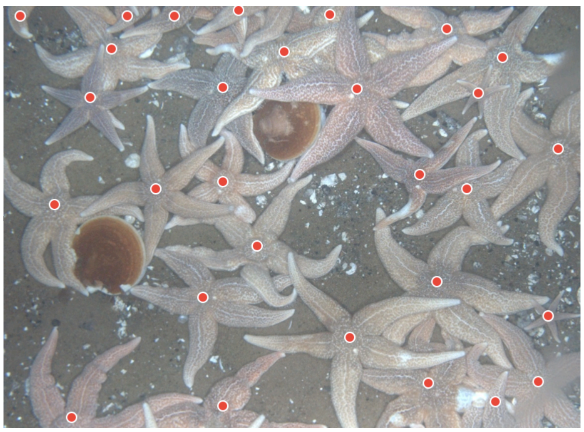


Fig. 2. Starfish piling on top of each other (manually annotated with red dots), along with a few scallops. Photo courtesy of Scott Gallager, Woods Hold Oceanographic Institution.

target, and the second is important when instances of the same target obscure each other.

While it is challenging to create a system that specifically solves just one species counting problem, the difficulty is significantly amplified if the desire is to generalize to a variety of sea life. The nature of each of the above-mentioned problems will change when considering different types of sea animals and their variations in shape, size, colour, texture, orientation, visibility, interaction, and habitat.

III. Existing Technology

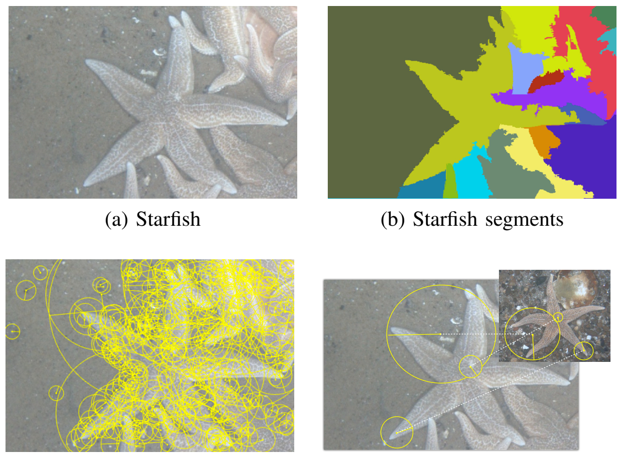
Research involving benthic organism location and quantification has predominantly focused on one type of organism. Few of the challenges outlined in the previous section have been addressed, and there is not much to draw upon for solving the generalized species problem. Gesu *et al.* [3] describe a system for tracking and counting starfish in a video sequence. Powell *et al.* [4] created a system for counting sardine eggs. Texture-based identification of Crown-of-Thorns Starfish using Local Binary Patterns is described by Clement *et al.* [5]. Somewhat more general approaches have been taken as well, including the detection of circular objects [6], a hierarchical classification system for object recognition [7], a protocol for the design and performance estimation of underwater video classification systems [8], and a system for detecting man-made objects [9]. Unfortunately, none of these approaches provide useful methods for the problem at hand. They are either very specific (i.e., intended to count one type of animal), or attempt to recognize only simple objects.

Some of the more successful work in benthic organism quantification is being done at the Woods Hole Oceanographic Institution (WHOI), lead by Gallager *et al.* [10]. They have developed algorithms for counting scallops and a few other species from the images collected by their camera system. They use edge flow segmentation [11], a powerful but computationally expensive algorithm, to extract disjoint regions of interest (ROIs), followed by support vector machine (SVM) based classification. The methods are general and could be extended to other species. Like much of the above-cited literature, however, they still take a conventional approach of extracting ROIs, and proceed to classify them using features (colour, shape, texture, etc.) extracted from these regions. Unfortunately, such an approach would likely have difficulty achieving success in complex images such as Fig. 1 and Fig. 2. Often seabed clutter is so confusing, or organisms are so close together and overlapping, that it is not possible to select singular targets for eventual classification. While this may work fine in many cases where regions of interest are distinct, sometimes an entire image is a region of interest that must be analyzed more closely than these methods are capable of doing. It is apparent that a different approach is needed.

IV. Feature-Based Recognition

Looking at terrestrial computer vision, a much larger research area, much recent work has focused on local invariant features [12], with applications including object recognition [13], robotic mapping and navigation [14], image stitching [15], and many others. In the context of local invariant features, a feature represents a specific structure in the image itself, ranging from simple structures such as points or edges to more complex structures such as objects.

Using a feature detector, a seabed image can be abstracted into a collection of these features, where a feature could be as large as an entire animal, e.g., a crab, or as small as the fine textural pattern on the leg of a starfish. This is done without the need for segmentation. Unlike segmented regions, features are separate but not necessarily disjoint; smaller features can be found within an area covered by a larger feature, representing finer details. Once features are detected, they can be described via the computation of feature descriptors. A feature descriptor is an array of numbers that describe the nature of the feature in a way that is more meaningful than raw pixel data; it gives the feature a distinctive description that allows it to be matched with similar looking features in other images. Features can also be designed to be invariant to transformations such as position, scale, and orientation, and it is possible for them to



(a) Starfish (b) Starfish segments

(c) Starfish features (d) Starfish feature matching

Fig. 3. Using local features instead of segmenting an image. Details in text.

be robust to changes in illumination, noise, blur, discretization, and changes in viewpoint.

Fig. 3 illustrates the advantages of extracting local features over more traditional image segmentation. Using the original image of starfish shown in Fig. 3a as input, a typical segmentation algorithm is applied in Fig. 3b. It becomes difficult to distinguish between starfish, and their shape is obscured among multiple segments. Compare this with Fig. 3c, which illustrates hundreds of local features detected in the original starfish image (features are located at the centre of every yellow circle, with the contents of the circle representing the feature). These features provide much more information than segmentation, as many features can be generated, even for small objects. Here they are based on the appearance of the starfish at particular interest points; note how the features capture large details about the shape of the starfish, as well as smaller details like the texture of their skin. As well, because the features are both distinct and invariant, they can be matched to new images, allowing the detection of other starfish. This matching is conceptually illustrated in Fig. 3d, where a few example starfish features are matched to similar features in a different image, despite differences in position, scale, rotation, illumination, etc.

It is evident local image features could be very useful in solving the problems with generalized benthic species recognition. Their fundamental components are not aimed at any specific application in computer vision, making them powerful tools for more general problems. If the organism is complex, a number of distinctive features can provide an accurate representation of what its species looks like. If the organism is simple in shape, a single feature could be used to describe what it, and others of its species, looks like. Taking advantage of invariance to position, scale, and rotation enables the identification of similar organisms in other images, no matter where they are, what their orientation is, or what scale they are seen at. As well, robustness to changes in illumination, noise, blurring, and viewpoint becomes useful in overcoming imperfections with the imaging system.

The challenges involved with partially hidden animals can be overcome as long as one or two distinct features are visible. For large numbers of organisms that obscure each other, it is reasonable that learning feature-based compositions of these organisms will allow a system to reduce the complexity of such a task. Organisms could be broken down into parts in order to perform a quantification estimate based on which parts belong to each individual animal.

As a final note, due to the popularity of feature detection algorithms, there is also the potential to take advantage of hardware acceleration [16]. This would aid greatly in achieving real-time recognition rates.

V. The First Attempt

Conceptually, feature-based recognition is an interesting and plausible idea. However, before building a system based around local invariant features, there is an important practical question to ask: *What features should be used?* To realize all of the advantages mentioned in the last section, the features must be both reliably detected, and distinctively described. Perhaps the most popular algorithm available today is the Scale-Invariant Feature Transform (SIFT) [13].

SIFT detects and describes local features in images, and possesses the particular characteristics of invariance and robustness mentioned in the previous section. For feature detection, the SIFT algorithm works by first convolving the image with Gaussian filters at different scales, and then taking the difference of successive Gaussian-blurred images. Feature points are then taken as maxima/minima of the Difference of Gaussians (DoG) that occur at multiple scales, achieving scale-invariance.

Once features are detected, rotational invariance is achieved by assigning one or more orientations based on local image gradient directions. SIFT then computes a descriptor for each feature such that the descriptor is highly distinctive and partially invariant to the remaining variations such as illumination, 3D viewpoint, etc. The descriptor, inspired by the human visual system, is computed by first splitting the image region defining the feature into sub-regions. Histograms are then formed by computing gradient magnitude and orientation values in each sub-region, and combining them into a descriptor of 128 elements.

Using SIFT features as-is, a preliminary system was designed for the recognition and quantification of benthic organisms. It has been successfully tested on a case study of quantifying deep-sea crabs in seabed imagery collected off the West coast of Canada by the Remotely Operated Platform for Ocean Sciences (ROPOS) [17]. In order to describe the system, its function will be illustrated using this particular case study. Further detail about the data used is given in Section VI.

A. Target Annotation and Extraction

In order to create a system that detects a target, the system must be presented with examples of what it looks like, so it can be trained or calibrated. With crabs as the target, a

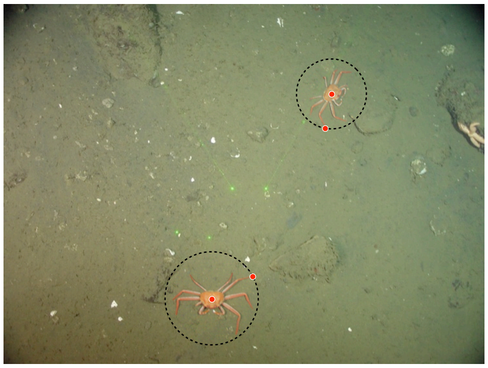


Fig. 4. Target annotation and extraction. A human annotator uses two points (red dots) to define a circle containing a target, which can be extracted from the image.

Fig. 5. Target feature detection and description of an example crab (left). The SIFT algorithm is used to detect feature locations (middle), and compute feature descriptors, representing the visual information contained in each circle (right).

human expert is required to analyze a quantity of training images, annotating them with the locations of any targets. This is accomplished by marking a point at the centre of the target, and marking another somewhere near the edge of the target. As can be seen in Fig. 4, those two points define a circle that encompasses the target, including all visual information used to describe it. When the expert finishes their annotations, the targets are extracted from the training images by simply isolating the circular image regions describing each target.

B. Target Feature Detection and Description

Once a set of targets is extracted from the training set, SIFT is used to detect and describe the target features. The standard SIFT algorithm for local feature detection/description is employed [13] without any modifications or enhancements. As seen in Fig. 5, each circular target image is subjected to SIFT, resulting in local features that attempt to describe important details about the target. While there are many features within the circular target regions that describe the crab shapes well, others are less useful. Only target features that effectively describe target examples are desired, and thus a means of discarding the non-useful features is needed. This reduction of features is taken care of in the next step.

C. Target Feature Reduction

Ultimately, the goal is to build a concise collection of target features that can be used to match against features detected in unseen (non-training) images, in order to find new instances of the target. As noted in the previous section, many of the

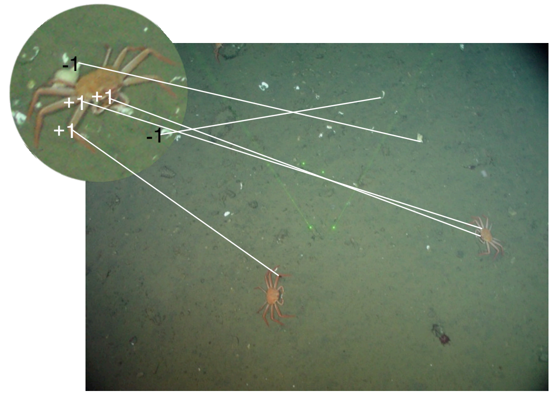


Fig. 6. Target feature reduction. Target features are scored based on their ability to match up to other target features in the training set.

features detected within the extracted circular target regions are not relevant to the targets themselves, and so they would not be desired for the final collection. Taking advantage of training images, the system employs a training or optimization procedure that automatically determines which features are useful as target features.

As shown in Fig. 6, each candidate target feature is scored on its ability to match to other target features in the training images. This is performed by taking each extracted target region, and matching its detected features to all of the features in every training image. When a candidate target feature matches to another target feature, it gets a score of +1. When it matches to a feature that is outside a target region, it gets a score of -1. After scoring is completed, any candidate target feature that has accumulated a negative score is discarded. This procedure is effective since the discarded features often represent details of the seabed that are common throughout the entire image set. Ideally, all features representing target information would remain, and features representing irrelevant background information would be discarded.

D. Target Finding

Once the system has produced a useful collection of target features, they may be used to find new targets in unseen images. Features are detected in the new image, and matched to the target features. Some of these new features match to one target feature, some match to many, and some match to none. Each feature is assigned a weight equal to how many target features it matches, and this aids in determining the locations of new targets. This process is illustrated in Fig. 7, where the features in an unseen image are first matched to target features (the red lines indicate matches to training targets); weights are then assigned to the features, shown as transparent blue circles with diameters directly proportional to the weights.

The system selects clusters of high-weighted features. If a cluster has a total weight above a threshold, a mark is made at the centre designating a detected target. This threshold is calibrated experimentally by measuring quantification performance on the training images.

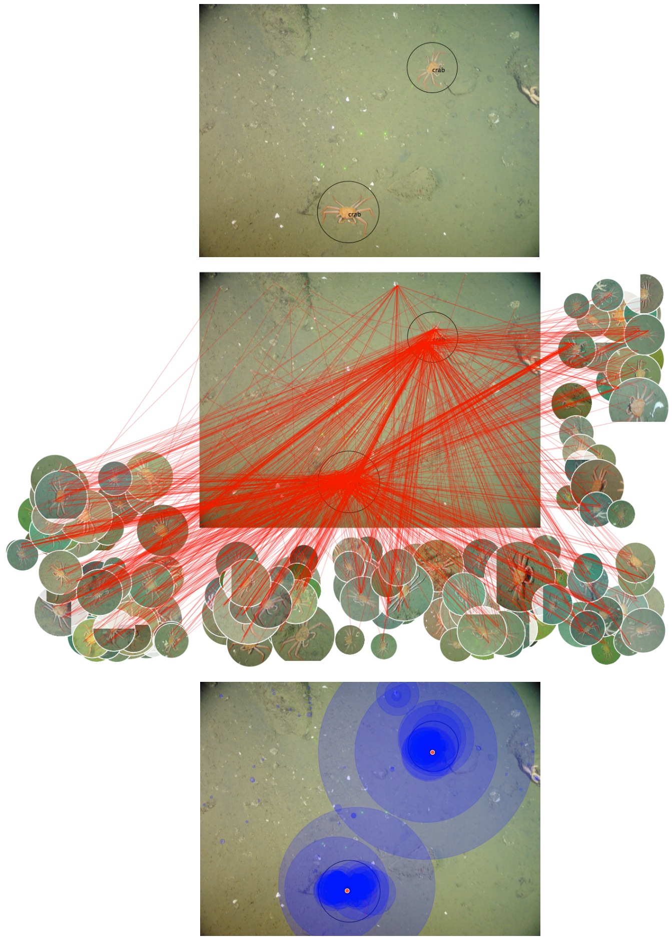


Fig. 7. Target finding. New targets are detected based on interpretation of new features matching to target features.

VI. Experimental Results

As mentioned in the previous section, the system has been successfully tested in an application of deep-sea crab quantification. A set of 776 images containing 428 crabs was used for the experiments. In each experiment, a portion of the set is used for training images, with the rest used for testing the system's performance. The 5 megapixel images are generally of good quality, though the altitude does vary significantly. At least 10 percent of the images were taken too high above the seabed, resulting in poor image quality.

This particular case study is a challenging one, with crabs varying greatly in size, shape, and orientation. There are also many that are partially hidden or obscured, or displayed with very low contrast, to the point of where it becomes easy for even the human visual system to overlook them. Because the altitude of the ROV was not always constant, the crabs also appear at many different levels of scale.

Despite these challenges, the system is able to perform reasonably well, as seen in Table I. With a standard 60%/40% training/testing split, the system is able to find 74% of the test crabs, while yielding a low rate of false positives (5.3%). As well, even when using a very small amount of training data, the system manages to maintain the majority of its quantification performance.

TABLE I
Crab quantification performance (fpr = false positive rate)

Training Ratio	Testing Performance
60/40	74.0% (5.3% fpr)
30/70	70.0% (7.0% fpr)
10/90	66.4% (8.3% fpr)

The target feature reduction procedure proved instrumental throughout all experiments, as system performance was non-existent when the procedure was not applied. Target feature sets were often reduced considerably. In the case of using 60% of the data for training, target feature sets of over 30,000 features were reduced by 80% or more. This would indicate the not only were unwanted background features removed, but also features describing features of example targets that were too similar to other background features in the training data.

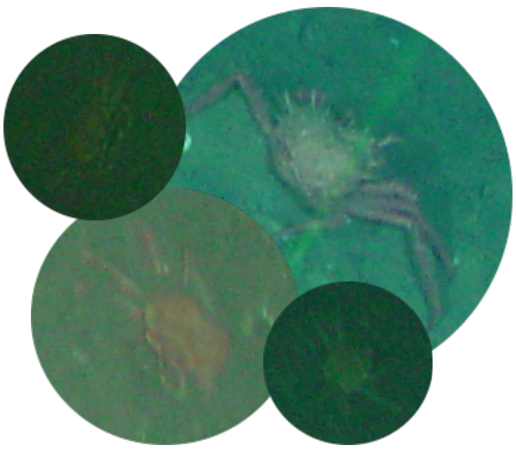
Fig. 8 shows a few examples of detected and undetected crabs. The successfully detected crabs were highly varied in shape and size, and many were partially hidden under rocks or other seabed clutter, or partially cut out of the photo. The system was also able to deal with very low-contrast crabs. As seen in Fig. 8b, however, some undetected crabs were simply of such poor contrast that they went undetected, which can be attributed to inadequate numbers of SIFT features found to describe them. There were also one or two examples of crabs that appeared too different from training examples to be detected. However, the large majority of undetected crabs were seen in images taken too far above the seabed, resulting in poor quality. If the imaging platform had been more consistent, higher detection rates would have been achieved.

The system was tested on a 2.8GHz Intel Core 2 Duo CPU with NVIDIA GeForce 9400M GPU. GPU acceleration is used for the SIFT algorithm via GPU-SIFT [16], and feature matching is carried out using a kd-tree with Best Bin First searching [13]. In its current state, from feature detection through to crab quantification, the system can process a 5 megapixel test image in under one second.

VII. Conclusions

This study demonstrated the use of local invariant features, specifically SIFT features, in the visual recognition and quantification of benthic organisms. An initial system was created and successfully tested on a case study of quantifying deep-sea crabs in seabed imagery collected off the West coast of Canada. While the system performed well, it is evident that a small percentage of images were of insufficient quality for reliable target recognition. It may be possible that some system components could be improved in order to better deal with this problem, such as applying contrast-enhancement to the images before feature detection. Another possibility would be to impose slightly stricter constraints during image collection, e.g., make sure the imaging platform stays within a certain distance above the seabed.

(a) Detected crabs



(b) Undetected crabs

Fig. 8. Some examples of detected and undetected crabs.

The ongoing research aims at generalized benthic species recognition, and the usefulness of SIFT features has been demonstrated here. They are powerful tools for more general computer vision, providing invariance to position, scale, rotation, and robustness to changes in illumination, noise, and viewpoint. They can provide accurate and information-rich representations of benthic organisms, despite the challenges encountered with variations in shape, size, colour, texture, orientation, visibility, interaction, and habitat.

The research will continue further refining the detection system while it is tested on other benthic species. Future improvements include the addition of colour-based features, automatic calibration of system parameters in the target feature reduction and target finding stages, machine learning of feature-based compositions of organisms, and extending the system to handle high-definition video streams.

Acknowledgments

The author would like to express sincere thanks to his advisors Ray Gosine, Ralf Bachmayer, and Neil Bose for their support and advice throughout the research. In addition, the author is very appreciative of Kim Juniper and his image analysis team at the University of Victoria for generously providing the Barkley Canyon image data, and of Corinne Pomroy for her invaluable motivation and support throughout the duration of this work. Finally, support from the NSERC CGS Program, the NSERC Canadian Healthy Oceans Network, and the NRC GSSS Program is gratefully acknowledged.

References

[1] D. C. Gordon, E. L. R. Kenchington, K. D. Gilkinson, D. L. McKeown, G. Steeves, M. Chin-Yee, W. P. Vass, K. Bentham, and P. R. Boudreau, "Canadian imaging and sampling technology for studying marine benthic habitat and biological communities," *ICES Annual Science Conference*, pp. 25–29, 2000.

[2] H. Singh, R. Eustice, R. Pizarro, and C. Roman, "Optical and acoustic habitat characterization with the Seabed AUV," *OCEANS 2003. Proceedings*, vol. 1, 2003.

[3] V. D. Gesu, F. Isgro, D. Tegolo, and E. Trucco, "Finding essential features for tracking starfish in a video sequence," in *Proc. Int. Conf. on Image Analysis and Processing*, 2003, pp. 504–509.

[4] J. R. Powell, S. Krotosky, B. Ochoa, D. Checkley, and P. Cosman, "Detection and identification of sardine eggs at sea using a machine vision system," *OCEANS 2003. Proceedings*, vol. 1, p. 175, 2003.

[5] R. Clement, M. Dunbabin, and G. Wyeth, "Toward Robust Image Detection of Crown-of-Thorns Starfish for Autonomous Population Monitoring," *Proceedings of the Australasian Conference on Robotics and Automation*, 2005.

[6] M. Gordan, O. Dancea, I. Stoian, A. Georgakis, and O. Tsatos, "A new SVM-based architecture for object recognition in color underwater images with classification refinement by shape descriptors," in *Automation, Quality and Testing, Robotics, 2006 IEEE International Conference on*, 2006, pp. 327–332.

[7] G. Foresti and S. Gentili, "A hierarchical classification system for object recognition in underwater environments," *Oceanic Engineering, IEEE Journal of*, vol. 27, no. 1, pp. 66–78, 2002.

[8] D. Redpath, K. Lebart, and C. Smith, "A methodology for underwater video classifier: design and comparison on limited datasets," *Oceans 2005-Europe*, vol. 2, 2005.

[9] C. Barat and M. Rendas, "A robust visual attention system for detecting manufactured objects in underwater video," *OCEANS 2006*, pp. 1–6, 2006.

[10] S. Gallager, S. Lerner, J. Howland, N. Vine, R. Taylor, D. Hart, and P. Rago, "New methods for surveying sea scallops, benthic organisms and habitat classification," presented at DFO National Workshop on Survey Design, Database Design and Analysis of Underwater Video/Photographic Surveys, 2007.

[11] W. Ma and B. Manjunath, "Edge flow: a framework of boundary detection and image segmentation," *Proc. of CVPR*, pp. 744–749, 1997.

[12] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: A survey," *Foundations and Trends® in Computer Graphics and Vision*, vol. 3, no. 3, pp. 177–280, 2008.

[13] D. Lowe, "Object recognition from local scale-invariant features," in *International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157.

[14] S. Se, D. Lowe, and J. Little, "Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks," *The International Journal of Robotics Research*, vol. 21, no. 8, p. 735, 2002.

[15] M. Brown and D. Lowe, "Automatic panoramic image stitching using invariant features," *International Journal of Computer Vision*, vol. 74, no. 1, pp. 59–73, 2007.

[16] S. Sinha, J. Frahm, M. Pollefeys, and Y. Genc, "GPU-based video feature tracking and matching," in *EDGE, Workshop on Edge Computing Using New Commodity Architectures*, vol. 278. Citeseer, 2006.

[17] K. Shepherd and S. Juniper, "ROPOS: Creating a Scientific tool from an industrial ROV," *Marine Technology Society Journal*, vol. 31, no. 3, pp. 48–54, 1997.