Identifying sea scallops from benthic camera images

Prasanna Kannappan [*]

Justin H. Walker [†]

Art Trembanis [‡]

Herbert G. Tanner [§]

[*]Department of Mechanical Engineering, University of Delaware Newark, DE 19716

[†]Department of Geological Sciences, University of Delaware, Newark, DE 19716

[‡]Department of Geological Sciences, University of Delaware, Newark, DE 19716

[§]Department of Mechanical Engineering, University of Delaware, Newark, DE 19716
Corresponding author. Email:btanner@udel.edu

Automated Scallop Counting from Images

Acknowledgments

The authors extend thanks to the captain and crew of the *f/v Christian and Alexa*. This work grew out of a project with Dr. Bill Phoel (deceased) and his guiding vision and enthusiasm has carried this work forward. Special thanks are also extended to Scott Gallager and Amber York of Woods Hole Oceanographic Institution for making available to us datasets on which Matthew Dawkings and Charles Stewart of RPI's Computer Science Department worked on. The autonomous underwater vehicle field effort was funded by the NOAA Research Set-Aside Program under Award Number: NA11NMF4540011, and the automated image analysis work was supported by the National Science Foundation under IIS grant #0913015.

Abstract

The paper presents an algorithmic framework for the automated analysis of benthic imagery data. The data are collected by an autonomous underwater vehicle for the purpose of population assessment of epibenthic organisms, such as scallops. The architecture consists of three layers of processing: visual attention, graph-cut segmentation methods, and template matching. The visual attention layer filters the imagery input, focusing subsequent processing only on regions in the images that are likely to contain target objects. The segmentation layer prepares for subsequent template matching. Finally, template matching classifies filtered objects into targets and distractors. The significance of the proposed approach is in its modular nature and its ability to process imagery datasets of low resolution, brightness, and contrast.

Introduction

Objectives

The sea scallop *(Placopecten magellanicus)* fishery in the US EEZ (Exclusive Economic Zone) of the northwest Atlantic Ocean has been, and still is, one of the most valuable fisheries in the United States. Historically, the inshore sea scallop fishing grounds in the New York Bight, i.e., Montauk Point, New York to Cape May, New Jersey, have provided a substantial amount of scallops (Caddy 1975; Serchuk et al. 1979; Hart and Rago 2006; Naidu and Robert 2006; Fisheries of the United States 2012). These mid-Atlantic Bight "open access" grounds are especially important, not only for vessels fishing in the day boat category, which are usually smaller vessels with limited range opportunities, but also all the vessels that want to fish in near-shore "open access" areas to save fuel. These areas offer high fish densities, but are at times rapidly depleted due to overfishing (Rosenberg 2003).

Dredge-based surveys have been extensively used for scallop population density assessment (National Marine Fisheries Service Northeast Fisheries Science Center (NEFSC) 2010). The process involves dredging part of the ocean floor, and manually counting the animals of interest found in the collected material. In addition to being invasive and detrimental to the creatures' habitat (Jenkins et al. 2001), these methods have accuracy limitations and can only generalize population numbers up to a certain extent. The goal of this paper is to demonstrate: (a) the efficacy of non-invasive techniques of monitoring and assessing such populations through the use of an Autonomous Underwater Vehicle (AUV) (Trembanis et al. 2011), and (b) the potential for automated methods of detection and enumeration of scallops.

The paper thus reports on efforts to accomplish these goals through a combination of underwater robotic image surveys and the development of an novel automated scallop recognition system. The automated recognition process workflow includes visual attention methods, which mark possible scallop regions, followed by segmentation and classification algorithms.

Related Literature

Robotic Marine Surveys

Optical based surveys of benthic habitats, either from towed camera sleds or underwater robots, constitute a leap forward in terms of increasing data density for habitat studies. However, the abundance (thousands to millions) of seabed images is both a boon and a challenge for researchers and managers. So far, the development of image acquisition strategies and platforms have outstripped the development of image processing techniques. This mismatch provides the motivation behind efforts to automate the detection of images containing scallops.

One of the earliest video based surveys of scallops (Rosenkranz et al. 2008) reports that it took from 4 to 10 hours of tedious manual analysis in order to review and process one hour of collected seabed imagery. The report suggests that an automated computer technique for processing of the benthic images would be a great leap forward; to this time, however, no such system is available. There is anecdotal evidence of in-house development efforts by the HabCam group (Gallager et al. 2005) towards an automated system but as yet no such system has emerged to the community of researchers and managers. A recent manual count of our AUV-based imagery dataset indicated that it took an hour to process 2080 images, whereas expanding the analysis to include all benthic macro-organisms reduced the rate down to 600 images/hr (Walker 2013). Another manual counting effort (Oremland et al. 2008) reports a processing time of 1 to 10 hours per person to process each image tow transect (the exact image number per tow was not reported). The same report indicates that the processing time was reduced to 1–2 hours per tow by subsampling 1 % of the images.

Vision-based Detection of Marine Creatures

There have been attempts to count marine species using stationary underwater cameras (Edgington et al. 2006; Spampinato et al. 2008). Background subtraction and shape detection (Williams et al. 2006) has been used to count salmon. However, background subtraction becomes inherently challenging when counting sedentary and sea-floor inhabiting animals like scallops. Other marine survey applications, such as zooplankton assessment (Stelzer 2009; McGavigan 2012), can require specialized imaging and sampling apparatus that cannot be easily re-tasked for other applications.

Marine survey cases that admit non-specialized imaging equipment may be amenable to automation, and form a natural application range for underwater robotics. AUVs with mounted cameras have been used for identification of creatures like clam and algae (Forrest et al. 2012). In such cases, very simple processing techniques like thresholding and color filtering are used. Yet, these techniques can be ineffective when low-resolution, depth, and deposited sediment can deprive scallops of unique color and texture.

Scallops, especially when viewed in low resolution, do not provide features that would clearly distinguish them from their natural environment. This presents a major challenge in automating the identification process based on visual data. To compound this problem, visual data collected from the species' natural habitat contain a significant amount of speckle noise. Some scallops are also partially or almost completely covered by sediment, obscuring the scallop shell features. A highly robust detection mechanism is required to overcome these impediments.

Existing approaches to automated scallop counting in artificial environments (Enomoto et al. 2009, 2010) employ a detection mechanism based on intricate distinguishing features like fluted patterns in scallop shells and exposed shell rim of scallops, respectively. Imaging these intricate scallop shell features might be possible in artificial scallop beds with stationary cameras and minimal sensor noise, but this level of detail is difficult to obtain from images of scallops in their natural environment. A major factor that contributes to this loss in detail is the poor image resolution obtained when the image of the target is captured several meters away from it. Overcoming this problem by operating an underwater vehicle too close to the ocean floor will adversely impact the image footprint (i.e. area covered by an image) and increase the risk of damaging the vehicle.

Furthermore, existing work on scallop detection (Dawkins 2011; Einar Óli Guòmundsson 2012) in their natural environment is limited to small datasets (often less than 100 images). From these studies alone, it is not clear if such methods can be used effectively in cases of large datasets comprising several thousand seabed images. An interesting example of machine-learning methods applied to the problem of scallop detection (Fearn et al. 2007) utilizes the concept of Bottom-Up Visual Attention (BUVA). The approach is promising but it does not use any ground truth for validation. As with several machine learning and image processing algorithms, porting the method from the original application set-up to another may not necessarily yield the anticipated results, and the process has to be tested and assessed.

Visual Attention

Visual attention is a neuro-physiologically inspired machine learning method (Koch and Ullman 1985) that attempts to mimic the human brain function in its ability to rapidly single out objects that are different from their surroundings within imagery data. The method is based on the hypothesis that the human visual system first isolates points of interest in an image, and then sequentially processes these points based on the degree of interest associated with each point. The degree of interest associated with a pixel is called *salience*, and points with the highest salience values are processed first. The method is used to pinpoint regions in an image where the value of some pixel attributes may be an indicator to its uniqueness relative to the rest of the image.

According to the visual attention hypothesis (Koch and Ullman 1985), in the human visual system the input video feed is split into several feature streams. Locations in these feature streams that are different from others in their neighborhood would generate peaks in the *center-surround* feature maps (explained later in more detail; see (1) for an example). The different center-surround feature maps can be combined to obtain a saliency *map*. Peaks in these resulting saliency maps, otherwise known as *fixations*, become points of interest, processed sequentially in descending order of their salience values.

Itti et al. (1998) proposed a computational model for visual attention. According to this model, an image is first processed along three feature streams (color, intensity, and orientation). The color stream is further divided into two sub-streams (red-green and blue-yellow) and the orientation stream into four sub-streams ($\theta \in \{0°, 45°, 90°, 135°\}$). The image information in each sub-stream is further processes in 9 different scales. In each scale, the image is scaled down using a factor $\frac{1}{2^k}$ (where $k = 0, \ldots, 8$), resulting in some loss of information as scale increases. The resulting image data for each scale factor constitutes the *spatial scale* for the particular sub-stream.

The sub-stream feature maps are compared across different scales to expose differences in them. Through the spatial scales in each sub-stream feature map, the scaling factors change the information contained. Resizing these spatial scales to a common scale through interpolation, and then comparing them, brings out the mismatch between the scales. Let $\ominus$ be an pixel operator that takes pixel-wise differences between resized sub-streams. This function is called the *center-surround* operator, and codifies the mismatches in the differently scaled sub-streams in the form of another map: the center-surround feature map. In the case of the intensity stream, with $c \in \{2, 3, 4\}$ and $s = c + \delta$ for $\delta \in \{3, 4\}$ denoting the indices of two different spatial scales, the center-surround feature map is given by

$$I(c, s) = |I(c) \ominus I(s)| \quad . \tag{1}$$

Similarly center-surround feature maps are computed for each sub-stream in color and orientation streams.

In this way, the seven sub-streams (two in color, one in intensity and four in orientation), yield a total of 42 center-surround feature maps. Now all center-surround feature maps in an original stream (color, intensity, and orientation) are then combined into a *conspicuity map* (CM): one for color $\bar{C}$, one for intensity $\bar{I}$, and one for orientation $\bar{O}$. Define the cross-scale operator $\oplus$ that adds up pixel values in different maps. Let w_{cs} be scalar weights associated with how much the combination of two different spatial scales c and s contributes to the resulting conspicuity map. If M is the global maximum over the map resulting from the $\oplus$ operation, and $\bar{m}$ is the mean over all local maxima present in the map, let $\mathcal{N}(\cdot)$ be a normalization operator that scales that map by a factor of $(M - \bar{m})^2$. For the case of intensity, this combined operation produces a conspicuity map based on the formula

$$\bar{I} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} w_{cs} \mathcal{N}(I(c,s)) \ . \tag{2}$$

The three conspicuity maps—for intensity, color and orientation—are combined to produce the *saliency map*. If scalar weights for each data stream are selected, say $w_{\bar{I}}$ for intensity, $w_{\bar{C}}$ for color, and $w_{\bar{O}}$ for orientation, the saliency map can be expressed mathematically as

$$S = w_{\bar{I}} \mathcal{N}(\bar{I}) + w_{\bar{C}} \mathcal{N}(\bar{C}) + w_{\bar{O}} \mathcal{N}(\bar{O}) \ . \tag{3}$$

In a methodological variant of visual attention known as BUVA, all streams are weighted equally: w_{cs} is constant for all $c \in \{2,3,4\}$, $s = c + \delta$ ($\delta \in \{3,4\}$) and $w_{\bar{I}} = w_{\bar{C}} = w_{\bar{O}}$. A winner-takes-all neural network is typically used (Itti et al. 1998; Walther and Koch 2006) to compute the maxima, or fixations, on this map—other discrete optimization methods are of course possible. In the context of visual attention, fixations are the local maxima of the saliency map. These fixations lead to shifts in *focus of attention*, or in other words, enables the human vision processing system to preferentially process regions around fixations in an image.

In a different variant of visual attention referred to as Top-Down Visual Attention (TDVA) (Navalpakkam and Itti 2006), the weights in (2) and (3) are selected judiciously to bias fixations toward particular attributes. One method to select these weights in the general case when N_m maps are to be combined with those weights, is discussed in Navalpakkam and Itti (2006). Let N be the number of images in the learning set, and N_{iT} and N_{iD} be the number of targets—in this case, scallops—and distractors (similar objects) in image i within the learning set. For image i, let P_{ijT_k} denote the local maximum of the numerical values of the map for feature j in the neighborhood of the target indexed k;

similarly, let P_{ijD_r} be the local maximum of the numerical values of the map for feature j in the neighborhood of distractor indexed r. The weights for a combination of maps are determined by

$$\begin{aligned} w'_j &= \frac{\sum_{i=1}^{N} N_{iT}^{-1} \sum_{k=1}^{N_{iT}} P_{ijT_k}}{\sum_{i=1}^{N} N_{iD}^{-1} \sum_{r=1}^{N_{iD}} P_{ijD_r}} \\ w_j &= \frac{w'_j}{\frac{1}{N_m} \sum_{j=1}^{N_m} w'_j} \end{aligned} \tag{4}$$

Where $j \in \{1, \ldots, N_m\}$ is the index set of the different maps to be combined. Equations (4) are used for the selection of weights w_{cs} in (2), and $w_{\bar{I}}$, $w_{\bar{O}}$, $w_{\bar{C}}$ in (3).

Contributions

The paper describes a combination of robotic-imaging marine survey methods, with automated image processing and detection algorithms. The automated scallop detection algorithm workflow involves 3 processing layers: customized TDVA pre-processing, robust image segmentation, and object recognition methods. The value of the proposed approach is primarily in providing a novel engineering solution to a real-world problem with economic and societal significance, which goes beyond the particular domain of scallop population assessment and can possibly extend to other problems of environmental monitoring, or even defense (e.g. mine detection). Given the general unavailability of similar automation tools, the proposed one can have potential impact in the area of underwater automation. The multi-layered approach not only introduces several minor technical innovations at the implementation level, but also provides a specialized package for benthic habitat assessment. At a processing level it provides the flexibility to re-task individual data processing layers for different detection applications. When viewed as a complete package, the proposed approach offers an efficient tool to benthic habitat specialists for processing large image datasets.

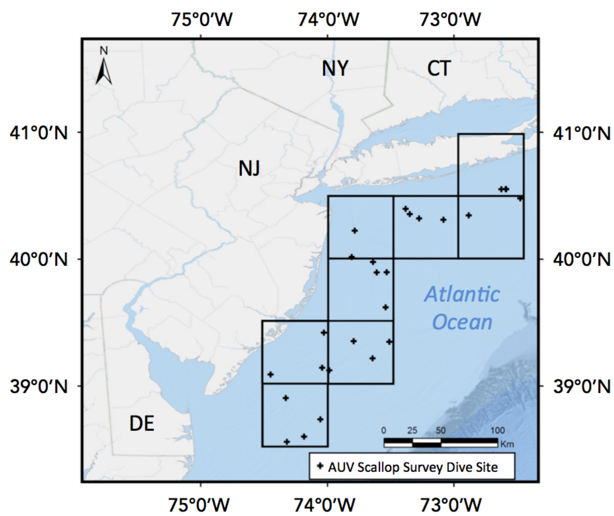


Figure 1: Map of the survey region from Shinnecock, New York to Cape May, New Jersey, divided into eight blocks or strata

Materials and Procedure

The 2011 Research Set-Aside (RSA) project (Titled: "A Demonstration Sea Scallop Survey of the Federal Inshore Areas of the New York Bight using a Camera Mounted Autonomous Underwater Vehicle") was a proof-of-concept project that successfully used a digital, rapid-fire camera integrated to a Gavia AUV, to collect a continuous record of photographs for mosaicking, and subsequent scallop enumeration. In July 2011, transects were completed in the northwestern waters of the mid-Atlantic Bight at depths of 25-50 m. The AUV continuously photographed the seafloor along each transect at a constant distance of 2 m above the seafloor. Parallel sets of transects were spaced as close as 4 m. Georeferenced images were manually analyzed for the presence of sea scallops using position data logged (using Doppler Velocity Log (DVL) and Inertial Navigation System (INS)) with each image.

Field Survey Process

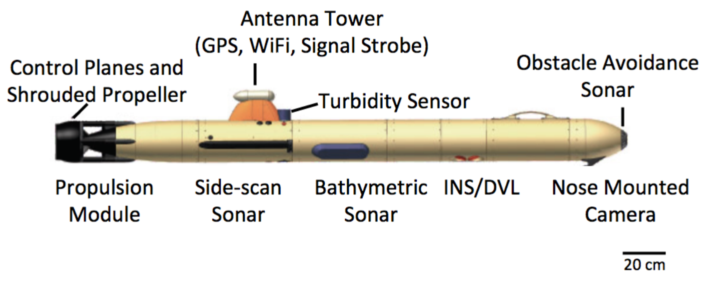
In the 2011 demonstration survey, the federal inshore scallop grounds from Shinnecock, New York to Ocean View, Delaware, was divided into eight blocks or strata (as shown in Figure 1). The *f/v Christian and Alexa* served as the surface support platform from which a Gavia AUV (see Figure 2) was deployed and recovered. The AUV conducted photographic surveys of the seabed for a continuous duration of approximately 3 hours during each dive, repeated 3–4 times in each stratum, with each stratum involving roughly 10 hours of imaging and an area of about 45 000 m^2. The AUV collected altitude (height above the seabed) and attitude (heading, pitch, roll) data, allowing the georectification of each image into scaled images for size and counting measurements. During the 2011 pilot study survey season, over 250 000 images of the seabed were collected. These images were analyzed in the University of Delaware's laboratory for estimates of scallop abundance and size distribution. The *f/v Christian and Alexa* provided surface support, and made tows along the AUV transect to ground-truth the presence of scallops and provide calibration for the size distribution. Abundance and sizing estimates were computed manually for each image using a GUI-based digital sizing software. Each image included embedded metadata that allowed it to be incorporated into existing benthic image classification systems (HabCam mip (Dawkins et al. 2013)).

During this proof of concept study, in each stratum the f/v *Christian and Alexa* made one 15-minute dredge tow along the AUV transect to ground-truth the presence of scallops and other fauna, and provide calibration for the size distribution. The vessel was maintained on the dredge track by using Differential GPS. The tows were made with the starboard 15 ft (4.572 m) wide New Bedford style commercial dredge at the commercial dredge speed of 4.5–5.0 knots. The dredge was equipped with 4 inch (10.16 m) interlocking rings, an 11 inch (27.94 cm) twine mesh top, and turtle chains. After dredging, the catch was sorted, identified, and weighed. Length-frequency data were obtained for the caught scallops. This information was recorded onto data logs and then entered into a laptop computer database aboard ship for comparison to the camera image estimates.

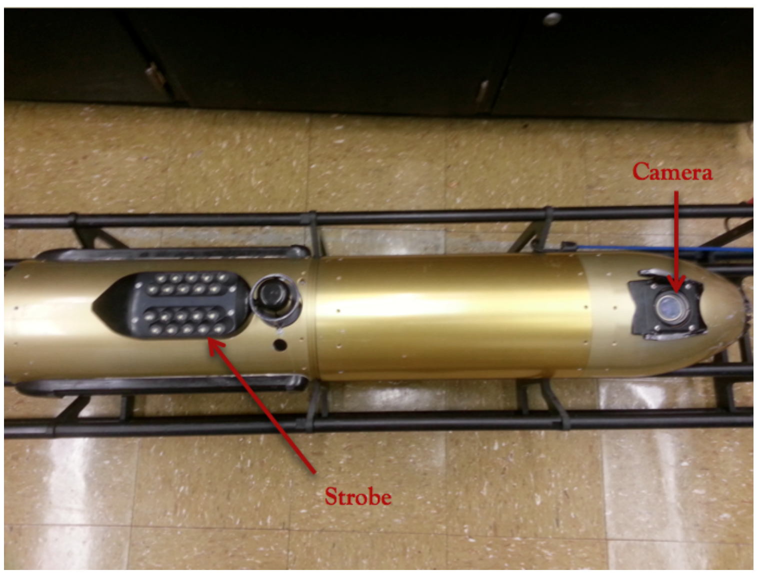
The mobile platform of the AUV provided a more expansive and continuous coverage of the seabed compared to traditional fixed drop camera systems or towed camera systems. In a given day, the AUV surveys covered about 60 000 m^2 of seabed from an altitude of 2 m above the bed, simultaneously producing broad sonar swath coverage and measuring the salinity, temperature, dissolved oxygen, and chlorophyll-a in the water.

Sensors and Hardware

The University of Delaware AUV (Figure 2) was used to collect continuous images of the benthos, and simultaneously map the texture and topography of the seabed. Sensor systems associated with this vehicle include: (1) a 500 kHz GeoAcoustics GeoSwath Plus phase measuring bathymetric sonar; (2) a 900/1800 kHz Marine Sonic dual-frequency high-resolution side-scan sonar; (3) a Teledyne Rd Instruments 1200 kHz acoustic doppler velocity log (DVL)/Acoustic doppler current profiler (ADCP); (4) a Kearfott T-24 inertial navigation system; (5) an Ecopuck flntu combination fluorometer / turbidity sensor; (6) a Point Grey Scorpion model 20SO digital camera and LED strobe array; (7) an Aanderaa Optode dissolved oxygen sensor; (8) a temperature and density sensor; and, (9) an altimeter. Each sensor separately records time and spatially stamped data with frequency and spacing. The AUV is capable of very precise dynamic positioning, adjusting to the variable topography of the seabed while maintaining a constant commanded altitude offset.



(a)



(b)

Figure 2: Schematics and image of Gavia AUV

Data Collection

The data was collected over two separate five-day cruises in July 2011. In total, 27 missions were run using the AUV to photograph the seafloor (For list of missions see Table 1). Mission lengths were constrained by the 2.5 to 3.5 hour battery life of the AUV. During each mission, the AUV was instructed to follow a constant height of 2 m above the seafloor. In addition to the 250 000 images that were collected, the AUV also gathered data about water temperature, salinity, dissolved oxygen, geoswath bathymetry, and side-scan sonar of the seafloor.

The camera on the AUV, a Point Grey Scorpion model 20SO (for camera specifications see Table 2), was mounted inside the nose module of the vehicle. It was focused at 2 m, and captured images at a resolution of 800×600. The camera lens had a horizontal viewing angle of 44.65 degrees. Given the viewing angle and distance from the seafloor, the image footprint can be calculated as 1.86×1.40 m^2. Each image was saved in jpeg format, with metadata that included position information (including latitude, longitude, depth, altitude, pitch, heading and roll) and the near-seafloor environmental conditions analyzed in this study. This information is stored in the header file, making the images readily comparable and able to be incorporated into existing RSA image databases, such as the HabCam database. A manual count of the number of scallops in each image was performed and used to obtain overall scallop abundance assessment. Scallops counted were articulated shells in life position (left valve up) (Walker 2013).

Layer I: Top-Down Visual Attention

Counting scallops manually, through observation and tagging of the AUV-based imagery dataset, is a tedious process that typically proceeds at a rate of 600 images/hr (Walker 2013). The outcome usually includes an error in the order of 5 to 10 percent. An automated system

Table 1: List of missions and number of images collected

Mission	Number of images	Mission	Number of images
LI1[1]	12 775	NYB6	9 281
LI2	2 387	NYB7	12 068
LI3	8 065	NYB8	9 527
LI4	9 992	NYB9	10 950
LI5	8 338	NYB10	9 170
LI6	11 329	NYB11	10 391
LI7	10 163	NYB12	7 345
LI8	9 780	NYB13	6 285
LI9	2 686	NYB14	9 437
NYB1[2]	9 141	NYB15	11 097
NYB2	9 523	ET1[3]	9 255
NYB3	9 544	ET2	12 035
NYB4	9 074	ET3	10 474
NYB5	9 425		

[1] LI–Long Island
[2] NYB–New York Bight
[3] ET–Elephant Trunk

Table 2: Camera specifications

Attribute	Specs
Name	Point Grey Scorpion 20SO Low Light Research Camera
Image Sensor	8.923 mm Sony ccd
Horizontal Viewing Angle	44.65 degrees (underwater)
Mass	125 g
Frame rate	3.75 fps
Memory	Computer housed in AUV nose cone
Image Resolution	800 × 600
Georeferenced metadata	Latitude, longitude, altitude, depth
Image Format	jpeg

Figure 3: Seabed image with scallops shown in circles

that would merely match this performance would still be preferable to a manual process.

Classification methods exploit characteristic features in the objects of interest. What features can be chosen for scallops can be debated, and different options may be available depending on the observation setup. The dataset which the reported algorithm applied on, (see Figure 3 for a representative sample) did not offer any unequivocal feature choices, but there were still some identifiable recurring visual patterns.

One example is a dark crescent on the upper perimeter of the scallop shell, which is the shadow cast by the upper open scallop shell produced from the AUV strobe light (Figure 4(a)). Another pattern that could serve as a feature is a frequently occurring bright crescent on the periphery of the scallop, generally being the visible inside of the right (bottom) valve when the scallop shell is partly open (Figure 4(b)). A third pattern is a yellowish tinge associated with the composition of the scallop image (Figure 4(b)).

Learning

A customized TDVA algorithm can be designed to sift automatically through the body of imagery data, and focus on regions of interest that are more likely to contain scallops. The process of designing the TDVA algorithm is described below.

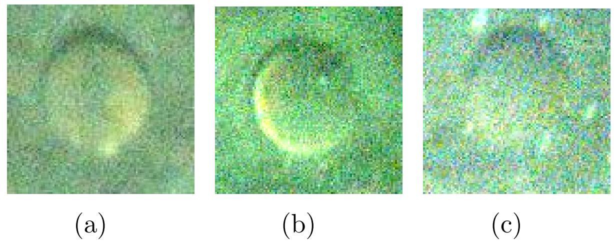


Figure 4: (a) Scallop with yellowish tinge and dark crescent; (b) Scallop with yellowish tinge and bright shell rim crescent; (c) Scallop with no prominent crescents and texturally identical to the background

The first step is a small-scale, bottom-up (BUVA) saliency computation. The saliency computation is performed on a collection of randomly selected 243 annotated images, collectively containing 300 scallops. This collection constitutes the *learning set*. Figure 5 represents graphically the flow of computation and shows the type of information in a typical image that visual attention tends to highlight.

A process of extremum seeking on the saliency map of each image identifies fixations in the associated image. If a 100×100 pixel window—corresponding to an approximately $23 \times 23\ \text{cm}^2$ area on the seafloor—centered around a fixation point contained the center of a scallop, the corresponding fixation was labeled a *target*; otherwise, it is considered a *distractor*.

The target and distractor regions are determined in all the feature and conspicuity maps for each one of these processed images in the learning set. This is done by adaptively thresholding and locally segmenting the points around the fixations with similar salience values in each map. Then the mean numerical value in neighborhoods around these target and distractor regions in the feature maps and conspicuity maps are computed. These values are used to populate the P_{ijT_k} and P_{ijD_r} variables in (4), and determine the top-down weights for feature maps and conspicuity maps.

For the conspicuity maps, the center-surround scale weights w_{cs} computed through (4) and consequently used in (2), are shown in Table 3. For the saliency map computation, the

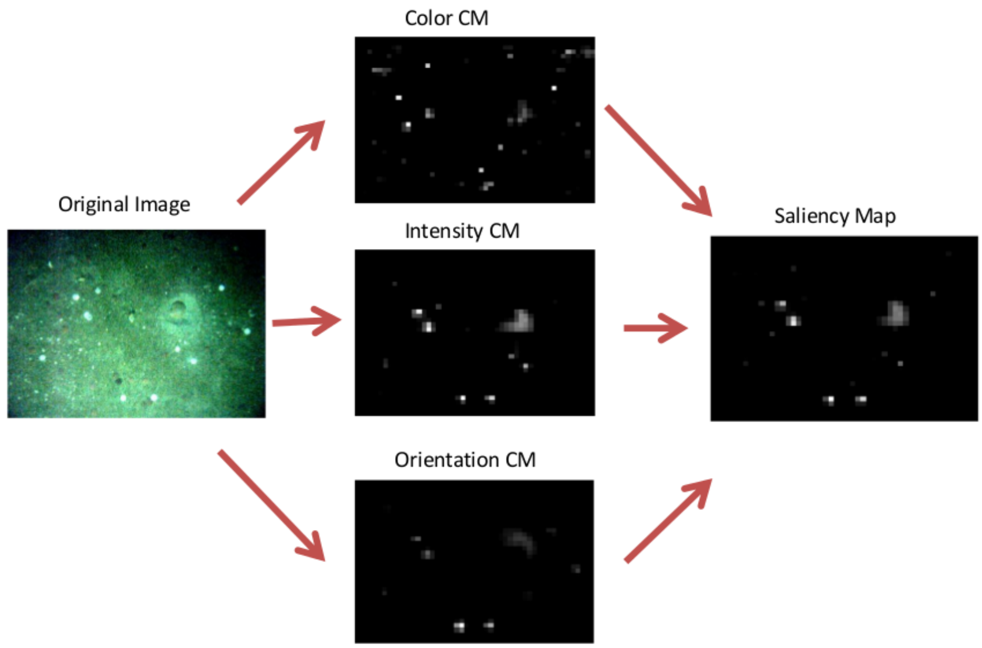


Figure 5: Illustration of computation flow for the construction of saliency maps

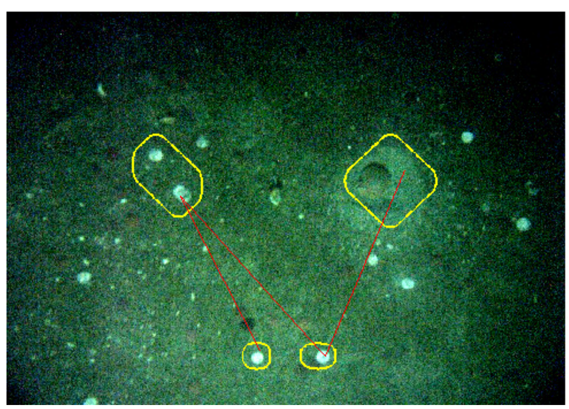


Figure 6: Illustration of fixations (marked by yellow boundaries): red lines indicate the order in which the fixations were detected with the lower-left fixation being the first.

Table 3: Top-down weights for feature maps

		Center Surround Feature Scales					
		1	2	3	4	5	6
Color	red-green	0.8191	0.8031	0.9184	0.8213	0.8696	0.7076
	blue-yellow	1.1312	1.1369	1.3266	1.2030	1.2833	0.9799
Intensity	intensity	0.7485	0.8009	0.9063	1.0765	1.3111	1.1567
Orientation	0°	0.7408	0.2448	0.2410	0.2788	0.3767	2.6826
	45°	0.7379	0.4046	0.4767	0.3910	0.7125	2.2325
	90°	0.6184	0.5957	0.5406	1.2027	2.0312	2.1879
	135°	0.8041	0.6036	0.7420	1.5624	1.1956	2.3958

weights resulting from the application of (4) on the conspicuity maps are $w_{\bar{I}} = 1.1644$, $w\bar{C} = 1.4354$ and $w_{\bar{O}} = 0.4001$. The symmetry of the scallop shell in our low-resolution dataset justifies the relatively small value of the orientation weight.

Implementation and Testing

To test the performance of the customized TDVA algorithm, it is applied on two image datasets, the size of which is shown in Table 4. In this application, the saliency maps are computed via the formulae (3) and (2), using the weights listed in Table 3. Convergence time of the winner-takes-all neural network that finds fixations in the saliency map of each image in the datasets of Table 4, is controlled using dynamic thresholding: It is highly unlikely that a fixation that contains an object of interest requires more than 10 000 iterations. If convergence to some fixation takes more than this number of iterations, then the search is terminated and no more fixations are sought in the image.

Given that an image in datasets of Table 4 contains two scallops on average, no more than ten fixations are sought in each image (The percentage of images in the datasets that contained more than 10 scallops was 0.002%). Since in the testing phase the whole

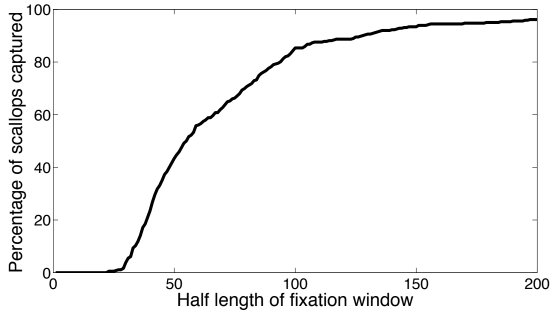


Figure 7: Percentage of scallops enclosed in the fixation window as a function of window half length (in pixels)

(a) (b) (c) (d) (e)

Figure 8: (a) Fixation window from layer I; (b) Edge segmented image; (c) graph-cut segmented image; (d) Region boundaries obtained when the edge segmented image is used as a mask over the graph-cut segmented image boundaries; (e) circle fitted on the extracted region boundaries.

scallop—not just the center—needs to be included in the fixation window, the size of this window is set at 270×270 pixels; more than 91% of the scallops are accommodated inside the window (Figure 7).

Layer II: Segmentation and shape extraction

This processing layer consists of three separate sub-layers: edge based segmentation (involves basic morphological operations like smoothing, adaptive thresholding and edge detection), graph-cut segmentation, and shape fitting. The flow of the segmentation process for a typical fixation window containing scallop is illustrated in Figure 8. Figure 8(a) shows a fixation window. Edge-based segmentation on this window yields the edge segmented image of Figure 8(b). At the same time, graph-cut segmentation process (Shi and Malik 2000) is applied on the fixation window to decompose it into 10 separate regions as seen in Figure 8(c). The boundaries of these segments are matched with the edges in the edge segmented image. This leads to further filtering of the edges, and eventually leads to the region boundaries on Figure 8(d). This is followed by fitting of a circle to each of the contours in the filtered region boundaries (Figure 8(d)). Only circles with dimensions close to that of a scallop (diameter $20 - 70$ pixels) are retained (Figure 8(e)), which in turn helps in rejection of other non-scallop round objects.

The choice of the shape to be fitted is suggested by the geometry of the scallop's shell. Finding the circle that fits best to a given set of points is formulated as an optimization problem along the lines of Taubin (1991) and Chernov (2010).

Given a set of n points on a connected contour each with coordinates (x_i, y_i) ($i \in \{1, 2, \ldots, n\}$), define a function of four parameters A, B, C, and D:

$$F_2(A,B,C,D) = \frac{\sum_{i=1}^{n}[A(x_i^2+y_i^2)+Bx_i+Cy_i+D]^2}{n^{-1}\sum_{i=1}^{n}[4A^2(x_i^2+y_i^2)+4ABx_i+4ACy_i+B^2+C^2]} \ . \tag{5}$$

It is shown (Taubin 1991) that minimizing (5) over these parameters yields the circle that fits best around the contour. The center (a, b) and the radius of this best-fit circle are given as a function of the parameters as follows:

$$a = -\frac{B}{2A} \ , \qquad b = -\frac{C}{2A} \ , \qquad R = \sqrt{\frac{B^2+C^2-4AD}{4A^2}} \ . \tag{6}$$

For all annotated scallops in the testing image dataset, the quality of the fit is quantified by means of two scalar measures: the center error e_c, and the percent radius error e_r. An annotated scallop would be associated with a triple (a_g, b_g, R_g)—the coordinates of its center (a_g, b_g) and its radius R_g. Using the parameters of the fit in (6), the error measures are evaluated as follows, and are required to be below the thresholds specified on the right hand side in order for the scallop to be considered detected.

$$e_c = \sqrt{(a_g - a)^2 + (b_g - b)^2} \leq 12 \text{ (pixels)} \qquad\qquad e_r = \frac{|R_g - R|}{R_g} \leq 0.3 \ .$$

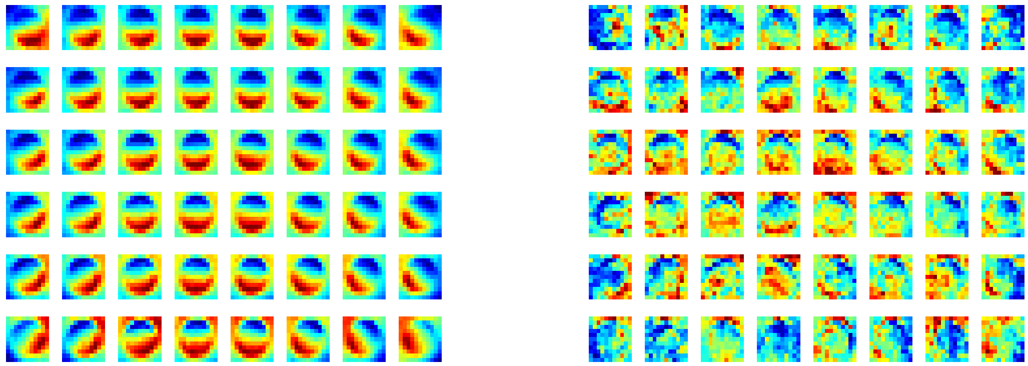
These thresholds were set empirically, taking into account that radius measurements in manual counts used as ground truth (Walker 2013) have a measurement error of 5–10%.

Layer III: Classification

The binary classification problem solved in this layer consists of identifying specific features in the images which mark the presence of scallops. These images are obtained by a using a camera at the nose of the AUV, illuminated by a strobe light close to its tail (mounted to the hull of the control module at an oblique angle to the camera). Our hypothesis were that due to this camera-light configuration, scallops appear in the images with a bright crescent at the lower part of its perimeter and a dark crescent at the top—a shadow Though crescents appear in images of most scallops, their prominence and relative position with respect to the scallop varies considerably. The hypothesis regarding the origin of the light artifacts implies that the approximate profile and orientation of the crescents is a function of their location in the image.

Scallop Profile Hypothesis

A statistical analysis was performed on a dataset of 3 706 manually labeled scallops (each scallop is represented as (a, b, R) where a, b are the horizontal and vertical coordinates of the scallop center, and R is its radius). For this analysis, square windows of length $2.8 \times R$ centered on (a, b) were used to crop out regions from the images containing scallops. (Using a slightly larger window size ($> 2 \times R$, the size of the scallop) includes a neighborhood of



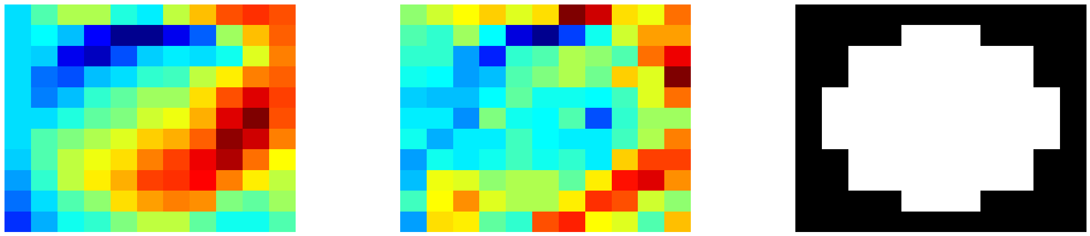
(a) (b)

Figure 9: (a) Mean map of scallops in each quadrant (b) Standard deviation map of scallops in each quadrant. Red corresponds to higher numeric values and blue correspond to lower numeric values.

pixels just outside the scallop which is where crescents are expected. This also improves the performance of local contrast enhancement, leading to better edge detection.) Each cropped region was filtered in grayscale, contrast stretched, and then normalized by resizing to 11×11 dimension or 121 bins. To show the positional dependence of the scallop profiles, the image plane is discretized into 48 regions (6×8 grid). Scallops whose centers lie within each grid square are segregated. The mean (Figure 9(a)) and standard deviation (Figure 9(b)) of the 11×11 scallop profiles of all scallops per grid square over the whole dataset of 3 706 images was recorded. The lower standard deviation found in the intensity maps of the crescents on the side of the scallop facing away from the camera reveal that these artifacts are more consistent as markers compared to the ones closer to the lens.

Scallop Profile Learning

The statistics of the dataset of 3 706 images used to produce Figure 9 form a look-up table that represents reference scallop profile (mean and standard deviation maps) as a function of scallop center pixel location. To obtain the reference profile for a pixel location, the statistics



(a) (b) (c)

Figure 10: Intensity statistics and mask for a region centered at a pixel with coordinates $470, 63$) in the image (a) Map of mean intensity; (b) Map of intensity standard deviation; (c) Mask applied to remove background points.

from all the scallops whose centers lie inside a 40×40 window centered on the pixel is used. This look-up table can be compressed; it turns out that not all of the 121 bins (11×11) within each map is equally informative, because bins close to the boundary are more likely to include a significant number of background pixels. For this reason, a circular mask with a radius covering 4 bins is applied to each map (Figure 10), thus reducing the number of bins that are candidates as features for identification to 61. Out of these 61 bins, an additional 15 bins having the highest standard deviation is ignored, leading to a final set of 46 bins. The value in the selected 46 bins from mean map forms a 46-dimensional feature vector associated with that region. The corresponding 46 bins from the standard deviation map are also recorded, and are used to weight the features (as seen later in (7)).

Scallop Template Matching

With this look-up table that codes the reference scallop profile for every scallop center pixel location, the resemblance of any segmented object to a scallop can now be assessed. The metric used for this comparison is a weighted distance function between the elements of the feature vector for the region corresponding to the segmented object, and that coming from the look-up table, depending on the location of the object in the image being processed. If

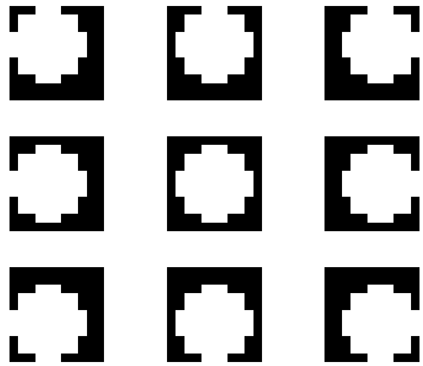


Figure 11: Nine different masks slightly offset from the center used to make the classification layer robust to errors in segmentation

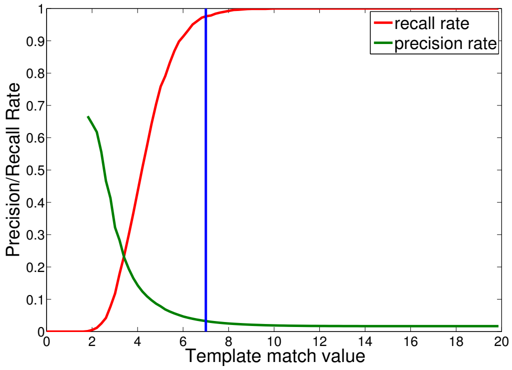
this distance metric is below a certain threshold D_{thresh}, the object is classified a scallop. Technically, let $X^o = (X_1^o, X_2^o, \ldots, X_{46}^o)$ denote the feature vector computed for the segmented object, and $X^s = (X_1^s, \ldots, X_{46}^s)$ the reference feature vector. Every component of the X^s vector is a reference mean intensity value for a particular bin, and is associated with a standard deviation σ_k from the reference standard deviation map. To compute the distance metric, first normalize X^o to produce vector $X^{\bar{o}}$ with components

$$X_p^{\bar{o}} = \min_k X_k^s + \left(\frac{\max_k X_k^s - \min_k X_k^s}{\max_k X_k^o - \min_k X_k^o} \right) \left[X_p^o - \min_k X_k^o \right] \text{ for } p = 1, \ldots, 48 \ ,$$

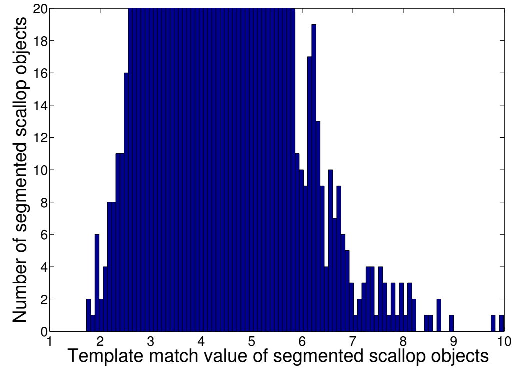
and then evaluate the distance metric D_t quantifying the dissimilarity between the normalized object vector $X^{\bar{o}}$ and the reference feature vector X^s as

$$D_t = \sqrt{\sum_{k=1}^{n} \frac{\|X_k^{\bar{o}} - X_k^s\|^2}{\sigma_k}} \ . \tag{7}$$

Small variations in segmentation can produce notable deviations in the computed distance metric (7). To alleviate this effect, the mask of Figure 10(c) was slightly shifted in



(a)



(b)

Figure 12: (a) Precision-Recall curve with D_{thresh} shown as a vertical line; (b) Histogram of template match of segmented scallop objects.

different directions and the best match in terms of the distance was identified. This process enhanced the robustness of the classification layer with respect to small segmentation errors. Specifically, nine slightly shifted masks were used (shown in Figure 11). Out of the nine resulting distance metrics $D_t^{o_1} \ldots D_t^{o_9}$, the smallest $D_{\mathsf{obj}} = \min_{p \in \{1,\ldots,9\}} D_t^{o_p}$ is found and used for classification. If $D_{\mathsf{obj}} < D_{\mathsf{thresh}}$, the corresponding object is classified as a scallop. Based on Figures 12(a)–12(b), the threshold value was chosen at $D_{\mathsf{thresh}} = 7$ to give a recall rate of 97%. (*Recall* refers to the fraction of relevant instances identified: fraction of scallops detected over all ground truth scallops; *precision* is the fraction of the instances returned that are really relevant compared to all instances returned: fraction of true scallops over all objects identified as scallops.) Evident in Figure 12(a) is the natural trade-off between increasing recall rates and keeping the number of false positives low.

Assessment

The reported multi-layered detection approach was tested on two separate datasets containing 1 299, and 8 049 images respectively. The results are shown in Table 4. Scallops

Table 4: Results of multi-layer scallop classification

	Dataset 1	Dataset 2
Number of images	1 299	8 049
Ground Truth Scallops	363	3 698
Valid Ground Truth Scallops	250	2 781
After Visual Attention Layer	231 (92.4%)	2 397 (86.2%)
After Segmentation Layer	185 (74%)	1 807 (64%)
After Classification Layer	183 (73%)	1 759 (63.2%)
False Positives	17 785	52 456

that were closer than 60 pixels vertically and 80 pixels horizontally to the image boundaries were excluded from the ground truth set as they were affected by severe vignetting effects caused by the strobe light on the AUV—boundaries become too dark (Figure 3) to correct with standard vignetting correction algorithms, and the characteristic crescents blend in the dark image boundaries (Figure 9(a)). Similar criteria were used by the manual annotation and counting process (Walker 2013).

For performance comparison purposes, a Support Vector Machine (SVM) classifier with a linear kernel was applied to the dataset of 8 049 images in Table 4. The percentage of scallops with respect to ground truth detected by the SVM was 48.5%, with three times fewer false positives—the trade-off of Figure 12(a) manifested here. The reported method leans toward maximizing true positives at the expense of a large number of false positives, by design, since some manual post-processing is deemed necessary anyway.

Discussion

The three-layer automated scallop detection approach discussed here works on feature-poor, low-light imagery and yields overall detection rates in the range of 60–75% . Related work on scallop detection using underwater imaging (Dawkins 2011; Dawkins et al. 2013), reported higher detection rates, but the quality of the images used was visibly better. Specifically, the datasets on which the algorithms of Dawkins et al. (2013) operated on, exhibit much more uniform lighting conditions, higher resolution, brightness, contrast, and color variance between scallops and background (see Figure 13). Evidence of this can be seen in Figure 13: the color variation between scallops and background data is reflected in the saturation histogram of Figure 13. While the histograms of scallop regions in the datasets of Table 4 is often identical to the global histogram of the image, the histograms of the Woods Hole data used by Dawkins et al. (2013) present a bimodal saturation histogram (Figure 13(c)), from which foreground and background are easily separable.

Compared to an alternative approach that uses a series of bounding boxes to cover the entire image (Einar Óli Guòmundsson 2012), the one reported here employs only ten windows per image, scanning the images at a much faster rate. Additionally, the detection rates of Einar Óli Guòmundsson (2012) were based on a dataset of just 20 images; statistically significant differences in performance rates between that approach and the one reported here would need much larger image samples.

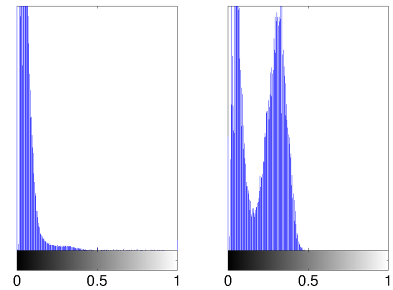
Comments and Recommendations

This work is a step toward the development of an automated procedure for scallop detection, classification and counting, based on low-resolution imagery data obtained in the organisms' natural environment. The reported method, in addition to being able to handle poor lighting and low-contrast imaging conditions, offers potential for computational time savings due to the targeted processing of image regions indicated by visual attention.

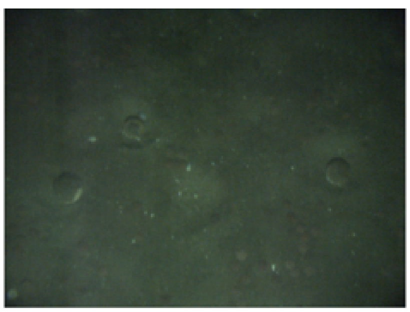
Significant improvements in terms of detection and classification accuracy can be expected in the context of pre-filtering and processing of raw image data. Another possibility for improvement could be in the direction of reducing the number of false positives. Given

(a)

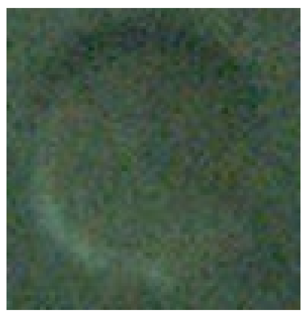
(b)



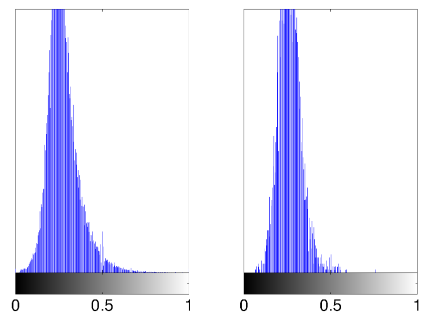
(c) Histogram of saturation values of background (left) and cropped scallop (right) from dataset in Dawkins et al. (2013)



(d)



(e)



(f) Histogram of saturation values of background (left) and cropped scallop (right) from our dataset

Figure 13: Representative samples of different imagery data on which scallop detection algorithms may be called to operate on. Figures 13(a) and 13(d), show an image containing a single scallop from the dataset used by Dawkins et al. (2013) (used with permission from the authors) and the datasets used in this paper respectively. A magnified view of a scallop cropped from Figure 13(a) and 13(d) can be seen in Figures 13(b) and 13(e) respectively. Figure 13(c) gives the saturation histogram of background or the complete image in Figure 13(a) to left and saturation histogram of Figure 13(b) to the right. Similarly, Figure 13(f) gives the saturation histogram of Figure 13(d) to the left and saturation histogram of Figure 13(e) to the right. The bimodal nature of the scallop histogram in Figure 13(c) derived from the dataset used in (Dawkins et al. 2013), clearly portrays the distinguishing appearance of the scallop pixels from the rest of the image, making it easily identifiable. The datasets we used did not exhibit any such characteristics (as seen in Figure 13(f)) to aid the identification of scallops.

the natural trade-off between the template matching threshold and the number of false positives, cross-referencing of the regions which include positives against the original, pre-filtered data may offer pathways to further false negative reduction.

Computationally, there is more to be gained in terms of performance by specialized and optimized code generation for segmentation and template matching. In the reported implementation, the graph-cut based image segmentation component is the most taxing in terms of computation time, and this area is where computational improvements are likely to yield the largest pay-off. On the other hand, the overall architecture is modular, in the sense that the segmentation and classification layers of the procedure could in principle be implemented using a method of choice, once appropriately interfaced with the neighboring layers and due to the fact that it allows retraining for other object detection problems with very different backgrounds or characteristic object features.

References

Caddy, J. (1975). Spatial model for an exploited shellfish population, and its application to the georges bank scallop fishery. *Journal of the Fisheries Board of Canada*, 32(8):1305–1328.

Chernov, N. (2010). *Circular and linear regression: Fitting circles and lines by least squares.* Taylor and Francis.

Dawkins, M. (2011). Scallop detection in multiple maritime environments. Master's thesis, Rensselaer Polytechnic Institute.

Dawkins, M., Stewart, C., Gallager, S., and York, A. (2013). Automatic scallop detection in benthic environments. In *IEEE Workshop on Applications of Computer Vision*, pages 160–167.

Edgington, D. R., Cline, D. E., Davis, D., Kerkez, I., and Mariette, J. (2006). Detecting, tracking and classifying animals in underwater video. In *OCEANS 2006*, pages 1–5.

Einar Óli Guòmundsson (2012). Detecting scallops in images from an auv. Master's thesis, University of Iceland.

Enomoto, K., Masashi, T., and Kuwahara, Y. (2010). Extraction method of scallop area in gravel seabed images for fishery investigation. *IEICE Transactions on Information and Systems*, 93(7):1754–1760.

Enomoto, K., Toda, M., and Kuwahara, Y. (2009). Scallop detection from sand-seabed images for fishery investigation. In *2nd International Congress on Image and Signal Processing*, pages 1–5.

Fearn, R., Williams, R., Cameron-Jones, M., Harrington, J., and Semmens, J. (2007). Automated intelligent abundance analysis of scallop survey video footage. *AI 2007: Advances in Artificial Intelligence*, pages 549–558.

Fisheries of the United States (2012). Fisheries of the United States, Silver Spring, MD. Technical report, National Marine Fisheries Service Office of Science and Technology.

Forrest, A., Wittmann, M., Schmidt, V., Raineault, N., Hamilton, A., Pike, W., Schladow, S., Reuter, J., Laval, B., and Trembanis, A. (2012). Quantitative assessment of invasive species in lacustrine environments through benthic imagery analysis. *Limnology and Oceanography: Methods*, 10:65–74.

Gallager, S., Singh, H., Tiwari, S., Howland, J., Rago, P., Overholtz, W., Taylor, R., and Vine, N. (2005). High resolution underwater imaging and image processing for identifying essential fish habitat. In Somerton, D. and Glentdill, C., editors, *Report of the National*

Marine Fisheries Service Workshop on Underwater Video analysis, NOAA Technical Memorandum NMFS-F/SPO-68, pages 44–54.

Hart, D. R. and Rago, P. J. (2006). Long-term dynamics of US Atlantic sea scallop Placopecten magellanicus populations. *North American Journal of Fisheries Management*, 26(2):490–501.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259.

Jenkins, S., Beukers-Stewart, B., and Brand, A. (2001). Impact of scallop dredging on benthic megafauna: a comparison of damage levels in captured and non-captured organisms. *Marine Ecology Progress Series*, 215:297–301.

Koch, C. and Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4(4):219–227.

McGavigan, C. (2012). A quantitative method for sampling littoral zooplankton in lakes: The active tube. *Limnology and Oceanography: Methods*, 10:289–295.

Naidu, K. and Robert, G. (2006). Fisheries sea scallop placopecten magellanicus. *Developments in Aquaculture and Fisheries Science*, 35:869–905.

National Marine Fisheries Service Northeast Fisheries Science Center (NEFSC) (2010). 50th northeast regional stock assessment workshop (50th saw) assessment report. Technical Report 10-17, p.844, US Dept Commerce, Northeast Fisheries Science Center.

Navalpakkam, V. and Itti, L. (2006). An integrated model of top-down and bottom-up attention for optimizing detection speed. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2049–2056.

Oremland, L., Hart, D., Jacobson, L., Gallager, S., York, A., Taylor, R., and Vine, N. (2008). Sea scallop surveys in the 21st century: Could advanced optical technologies ultimately replace the dredge-based survey? Presentation made to the NOAA Office of Science and Technology.

Rosenberg, A. A. (2003). Managing to the margins: the overexploitation of fisheries. *Frontiers in Ecology and the Environment*, 1(2):102–106.

Rosenkranz, G. E., Gallager, S. M., Shepard, R. W., and Blakeslee, M. (2008). Development of a high-speed, megapixel benthic imaging system for coastal fisheries research in alaska. *Fisheries Research*, 92(2):340–344.

Serchuk, F., Wood, P., Posgay, J., and Brown, B. (1979). Assessment and status of sea scallop (placopecten magellanicus) populations off the northeast coast of the united states. In *Proceedings of the National Shellfisheries Association*, volume 69, pages 161–191.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

Spampinato, C., Chen-Burger, Y.-H., Nadarajan, G., and Fisher, R. B. (2008). Detecting, tracking and counting fish in low quality unconstrained underwater videos. In *3rd International Conference on Computer Vision Theory and Applications*, pages 514–519.

Stelzer, C. P. (2009). Automated system for sampling, counting, and biological analysis of rotifer populations. *Limnology and oceanography: Methods*, 7:856.

Taubin, G. (1991). Estimation of planar curves, surfaces, and nonplanar space curves defined by implicit equations with applications to edge and range image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(11):1115–1138.

Trembanis, A. C., Phoel, W. C., Walker, J. H., Ochse, A., and Ochse, K. (2011). A demonstration sea scallop survey of the federal inshore areas of the new york bight using a camera mounted autonomous underwater vehicle. Technical report, National Oceanographic and Atmospheric Administration.

Walker, J. (2013). Abundance and size of the sea scallop population in the mid-atlantic bight. Master's thesis, University of Delaware.

Walther, D. and Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395–1407.

Williams, R., Lambert, T., Kelsall, A., and Pauly, T. (2006). Detecting marine animals in underwater video: Let's start with salmon. In *Americas Conference on Information Systems*, volume 1, pages 1482–1490.