2012-05-29

Fish Detection in Underwater Video of Benthic Habitats in Virgin Islands

Nan Wu
University of Miami, n.wu@umiami.edu

UNIVERSITY OF MIAMI

FISH DETECTION IN UNDERWATER VIDEO OF BENTHIC HABITATS IN VIRGIN ISLANDS

By

NAN WU

A THESIS

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Master of Science

Coral Gables, Florida

May 2012

UNIVERSITY OF MIAMI

A thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science

FISH DETECTION IN UNDERWATER VIDEO OF BENTHIC HABITATS IN VIRGIN ISLANDS

NAN WU

Approved:

_____	_____
Shahriar Negahdaripour, Ph.D. Professor of Electrical and Computer Engineering	Terri A. Scandura, Ph.D. Dean of the Graduate School
_____	_____
Mitsunori Ogihara, Ph.D. Professor of Computer Science	Victor Milenkovic, Ph.D. Professor of Computer Science

WU, NAN (M.S., Computer Science)

Fish Detection in Underwater Video of Benthic Habitats in Virgin Islands (May 2012)

Abstract of a thesis at the University of Miami.

Thesis supervised by Professor Shahriar Negahdaripour.
No. of pages in text. (60)

The research of underwater creatures is a fascinating work to marine biologists and environmental experts. However, finding and segmenting underwater creatures in underwater images and videos is time consuming. Automated object detection and segmentation can be applied to such problems to accelerate the tedious process. This work describes an approach to detect fish from images or videos of benthic habitats, recorded in Virgin Islands, during the yellow-tail grouper spawning activities.

The method involves first locating the horizon, separating the seawater and the seafloor, and then using different visual cues to detect fish both within the water column and over the seafloor. The detector can be applied to video data, or a single image (forgoing the visual motion cues). The precision and recall rates of 86.4% and 94.5% have been achieved.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The study of underwater species is a fascinating topic to marine biologists and environmental experts. Video survey is a popular approach to study marine life. To determine the size and distribution, or to study the behavior of species, the researchers need to locate them in images and (or) videos, but this can be very time consuming when processing a large volume of data. Automated image and video segmentation helps to speedup this tedious work.

The purpose of this thesis is to identify fish in the video or images of benthic habitats, by analyzing and exploring various visual information. Some automated fish detection and classification method has already been developed. For example, Benson et al. [5] proposed a fish detection method using Haar Classifiers, and Spampinato [8] suggests combining Adaptive Gaussian Mixture Model [9] and Moving Average Detection Algorithm, achieving 85% accuracy. However, due to the complexity of underwater environments, and the variability of visual cues in images and videos, not all the data can be processed in the same way and (or) with the same level of success.

The data used in this work is a clip of video obtained in the Virgin Islands on March 25, 2011. The video was shot at around 6:30 P.M. by a camera mounted on a remotely operated vehicle (ROV) at depths of 200-250 feet. The video resolution is 480 by 720 pixels. The focal length of the camera is unknown. The scene consists of fish swimming within the water column over the seafloor. The fish in the video are mainly groupers, circling over the reef as they prepare for spawning activity.

Detecting or segmenting individual fish from this video is a challenging problem because of the poor lighting condition, the low resolution and contrast of the video, the size of objects and the blending of the fish with the seafloor.

A frame of the video, demonstrating the general characteristics of the data, is shown in Figure 1.1. Although both fish and the water column have weak texture, the fish is darker than the water column. The seafloor generally has stronger texture than fish (regions labeled as seafloor 1 and fish 2), however, some reef objects on the seafloor (e.g., regions labeled as seafloor 2 and 3) have similar texture and brightness level as the fish, and may not be readily distinguishable (even by human).

However, given that most yellow-tail groupers in our Virgin Island data set are roughly of the same size, the image size of the fish once correcting for the distance from the camera, can be exploited as an additional visual cue to reduce the classification error. The distance cue is established as a byproduct of an image dehazing method that is applied to enhance the underwater image contrast.

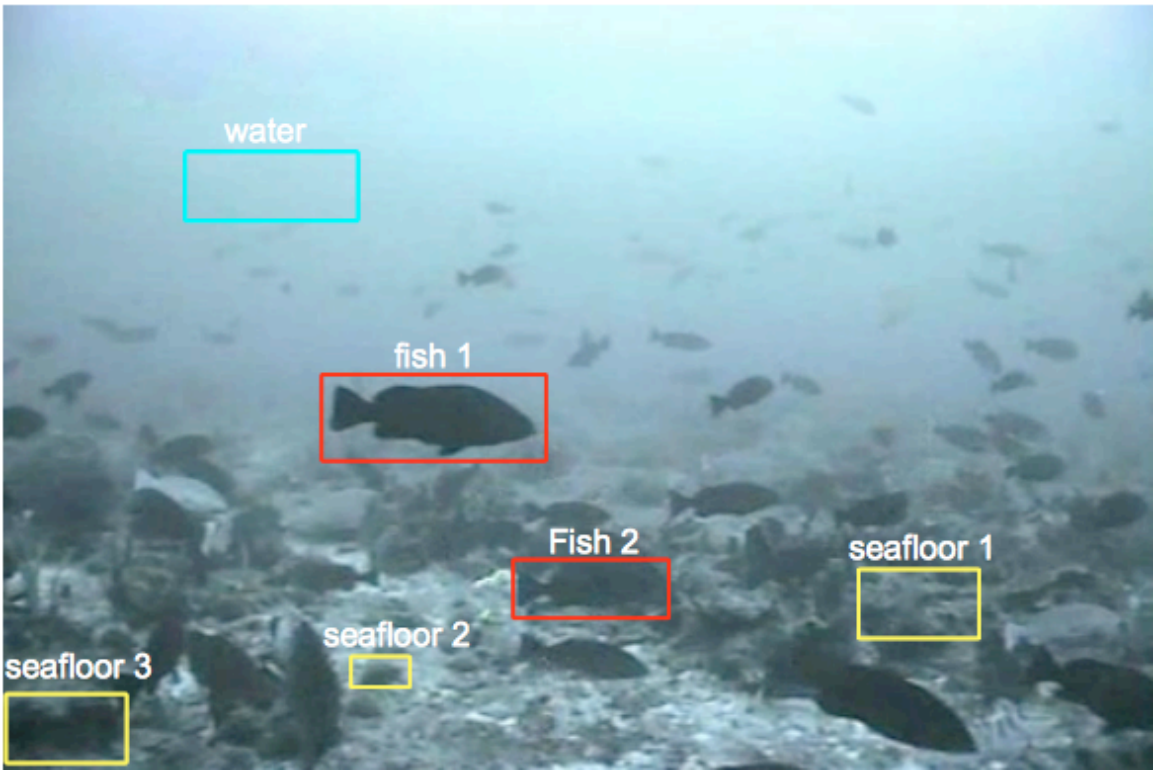

Figure 1.1: Underwater scene from the Virgin Islands video.

Still, some reefs (e.g., region labeled as seafloor 3) have similar texture and brightness (appearance) and size as fish. Here, where video (sequence of frames) is available, motion cue can play a key role in discriminating

between the two types; the swimming fish generate a different optical flow that the non-zero image motion of the stationary background scene, induced by camera motion). Therefore, intensity, texture, and object size and movement are potential cues to detect the fish over the sea floor.

Earlier fish detection methods, such as [5], involve the training of a classifier based on features extracted from fish and samples of other object types. In this work, extracting reliable features has been hard because 1) the video resolution is low; 2) the sizes of fish in the video are small (between 15 to 120 pixels), particularly within the water column; 3) features rely on the texture of objects, and the fish have weak texture due to the poor lighting in the video. Also, the method in [8] is not suitable for our application because of the strong blending of fish with the seafloor.

Two segmenting methods were tested to segment fish from this video or images. The method proposed in [3] provides an image hierarchical segmentation technique according to the local and global brightness, color and texture cues. The video hierarchical segmentation method proposed in [2] utilized the color and optical flow, and the dissimilarity within the spatio-temporal neighborhoods.

The primary difficulty in applying either segmentation method is when the fish swims very near the seafloor, where the intensity and texture characteristics of the seafloor and fish overlap: if the segmentation is too coarse, most small fish are undetectable; if the segmentation is too fine, we cannot differentiate the fish from the background.

Due to the difference between characteristics of the water column and the seafloor, we proposed a detection method that utilizes different cues to detect fish within water column and above seafloor.

1.1 Contribution

This research describes a detection method to locate fish from images or videos of benthic habitats, recorded in Virgin Islands. To improve success rate, domain-dependent cues within benthic scenes are exploited; thus, the method may not work as well in arbitrary underwater environments.

This detection system will do the following:

- Divide an image into two regions along the horizon, which separates the water column and the seafloor;
- Identify fish from the water column using intensity cues;

- Identify fish from the seafloor using texture and the motion cue where video/consecutive frames are available;
- The image is degraded due to the light absorption in underwater environment. To improve the image contrast image, we adopted a single image dehazing method proposed in [4]. A byproduct of the dehazing method is the scene distance form camera (referred to as object depth), which can be used to estimate the image size of the fish at each depth;
- Utilize estimated fish size to establish the rough fish window size (i.e., window size that could potentially contain a fish) at each depth (distance from camera);
- Optical flow is adopted as a motion cue when detecting in video to improve system performance.

1.2 Organization

This thesis is organized as follows:

In Chapter 2, we describe methods for preprocessing and enhancing the images. In Chapter 3, we introduce the motivation and design of the detection system. In Chapter 4, we elaborate on various steps of our method. In Chapter 5 we present the results and assess the accuracy of the

system. Summary concluding comments and potential future improvements are covered in Chapter 6.

Chapter 2

Preprocessing - Image Tag Removal and Dehazing

In the Virgin Islands video, each frame is time stamped and includes heading information from the ROV compass. These complicate the segmentation task, and thus have to be removed. Besides, underwater images are degraded due to light energy absorption and scattering by the medium. The depth-dependent scattering results in a hazy appearance that degrades the image contrast. The hazy appearance is most apparent within the uniform most-distant water column region. The haze component is to be removed to improve the contrast of image.

2.1 Image Tag Removal

Image tags, including the time stamp and compass information, are readily removed using the method of Resynthesizer in Script-Fu[1]. Made available as a GIMP plug-in by Paul Harrison, the Resynthesizer method for removing noise or filling image holes works based a non-linear image texture predictor model [6]. Figure 2.1 shows an example of

[1] http://registry.gimp.org/node25219

a raw image from the Virgin Islands video (a) and the result of tag removal (b).

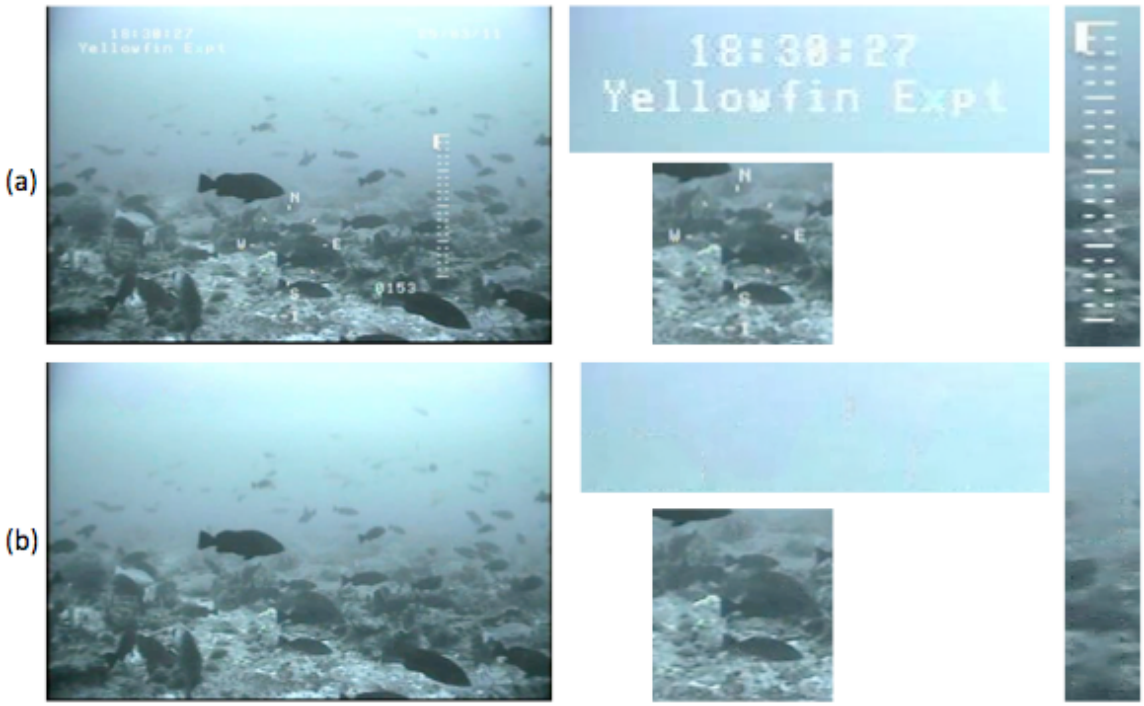

Figure 2.1: Image tag removal.

2.2 Image Dehazing

The image contrast is degraded primarily by medium scattering. A method that addresses the scattering of underwater images, is proposed by Yoav Schechner in [11]. It does not applied because the method is guided by polarization cue. Therefore, we applied a dehazing method [4] which requires no calibration and knowledge of medium properties to eliminate the haze component and recover contrast. The byproduct of this process is the up to scale estimation of the scene depth.

To briefly describe this method, the observed intensity I of a pixel x is modeled as a linear function of the scene radiance and global atmosphere light (so-called air light), according to the model

$$I(x) = t(x)J(x) + (1 - t(x))A \ ,$$

where $J(x)$ is the scene radiance, A is the global atmosphere light and $t(x)$ is the medium transmission.

The idea of the dehazing method is to find a small patch with lowest intensity in all color channels:

$$\min_{\{c=r,g,b\}} I_c(x) = t(x) \min_{\{c=r,g,b\}} J(x) + \big(1 - t(x)\big)A \ .$$

It is hypothesized that the intensity of this patch tends to be zero without haze: $\min_{\{c=r,g,b\}} J_c(x) = \mathbf{0}.$ Therefore, the medium transmission can be estimate by

$$t(x) = 1 - \min_{\{c=r,g,b\}} I_c(x) / A \ .$$

As stated, the scene depth (namely, the distance of each scene feature from the camera) is a byproduct of this dehazing method because the transmission is a function of scene depth $t(x) = e^{-\beta d(x)}$. The scene depth (up to an unknown scale β) can be estimated from the medium transmission by

$$\beta d(x) = -\ln t(x) \ .$$

Figure 2.2 shows a sample image from our data set (a), the dehazed image (b), and the scaled depth map (c).

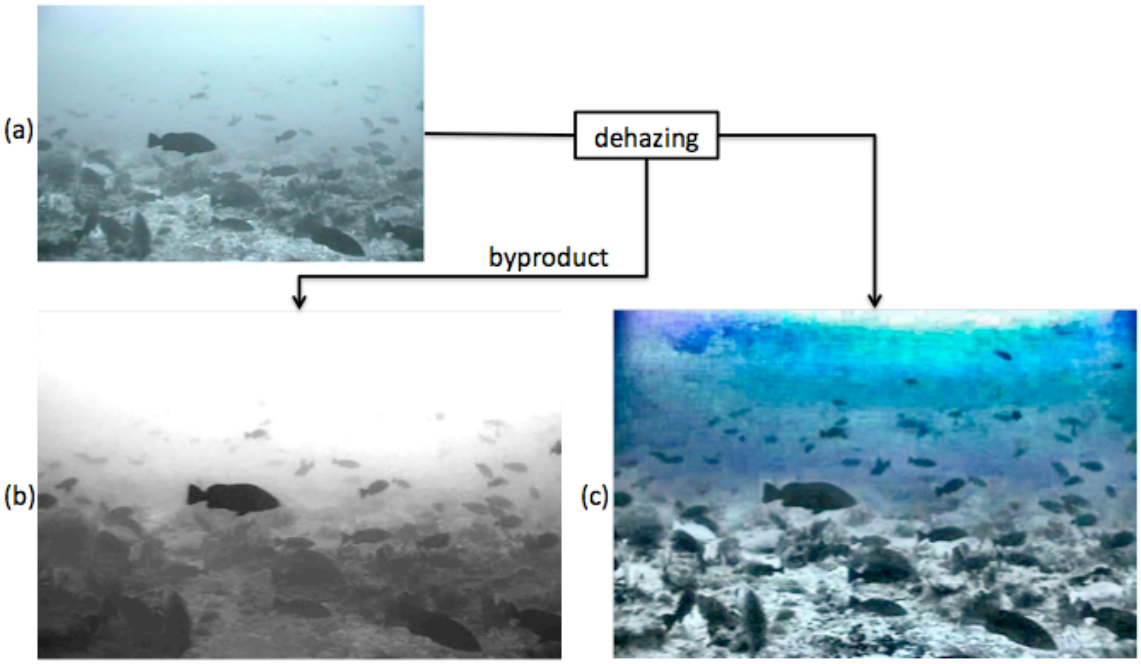

Figure 2.2: Example of an image (a), enhanced through dehazing (c) and the depth map as byproduct (b).

From figure 2.2, one notes that while the contrast has been improved significantly over the seafloor, the noise has also been enhanced over the water column. Much of this noise comes from the red channel because the red wavelength is significantly attenuated with distance due to medium absorption. In additional to improving contrast, enhancement by dehazing also emphasizes the noise in the red channel, leaving very small signal-to-noise ratio (SNR). This is demonstrated in Figure 2.3, which gives the decomposition of the enhanced image into red (a), green (b)

and blue (c) channel components. Compared to the green and blue channels, a large portion of the water column is dark, implying low red component. However, the water column tends to have a relatively uniform intensity distribution, as it should, when averaging only the green and blue channels (d).

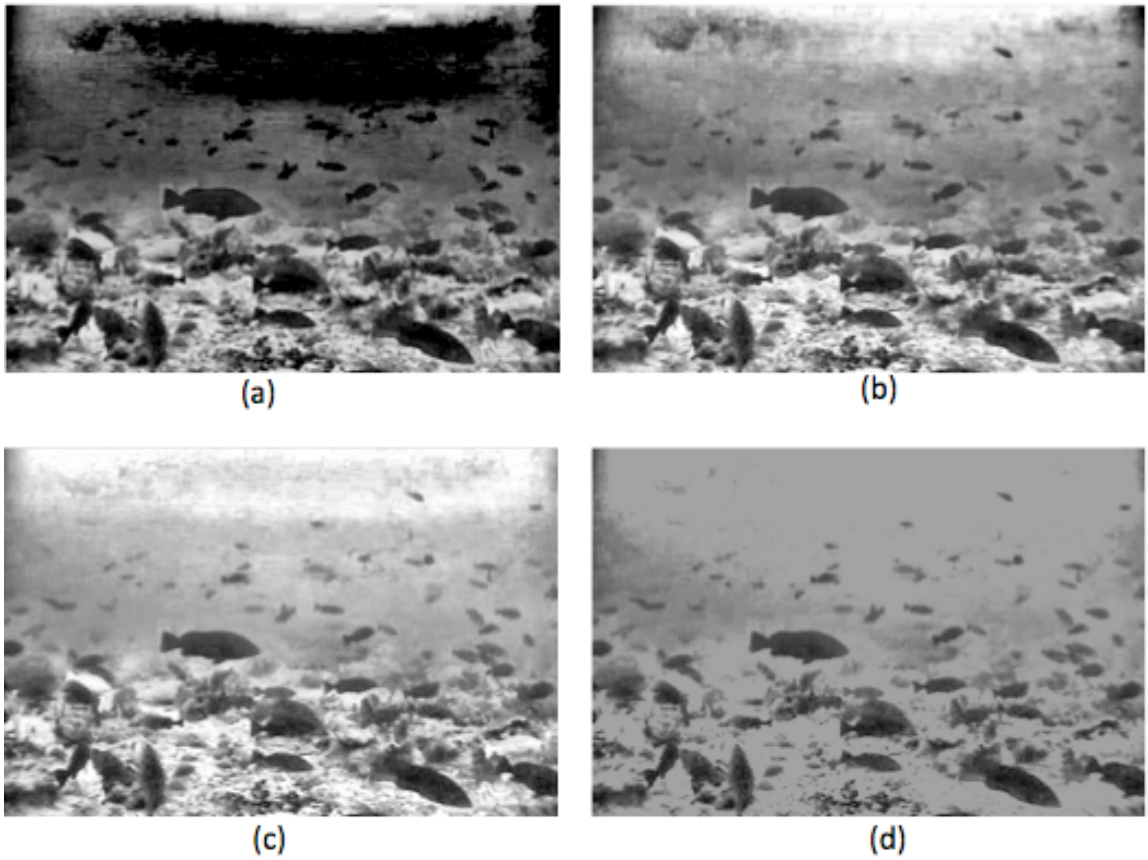

Figure 2.3: (a) Red, (b) Blue, and (c) Green channel decomposition and (d) average intensity of blue and green channels.

We have applied these methods to each of the images in our dataset to remove the image tags, as well as to enhance the image contrast.

Chapter 3

Motivation and Design

In a benthic environment, the background is either the water column or the seafloor, which have very different characteristics. Thus, different complexities exist in detecting fish that swim within the water column and over the seafloor. This chapter describes some of the relevant issues and different visual cues that directly relate to the design of a hybrid fish detector.

3.1 Water Column

In the Virgin Islands video, it is easy for human to identify fish within water column because these have different (lower) intensity level than the brighter water column. This can be verified in Figure 3.1, where water column samples within windows (a, b) have relatively uniform bright intensity distributions, while fish samples (c, d) show bimodal histograms; the first peak due to darker pixels of the fish, and the second comprising bright pixels of the water column. The intensity, thus, can be used to distinguish fish and the water column; the

existence of dark pixels within a window suggests presence of fish.

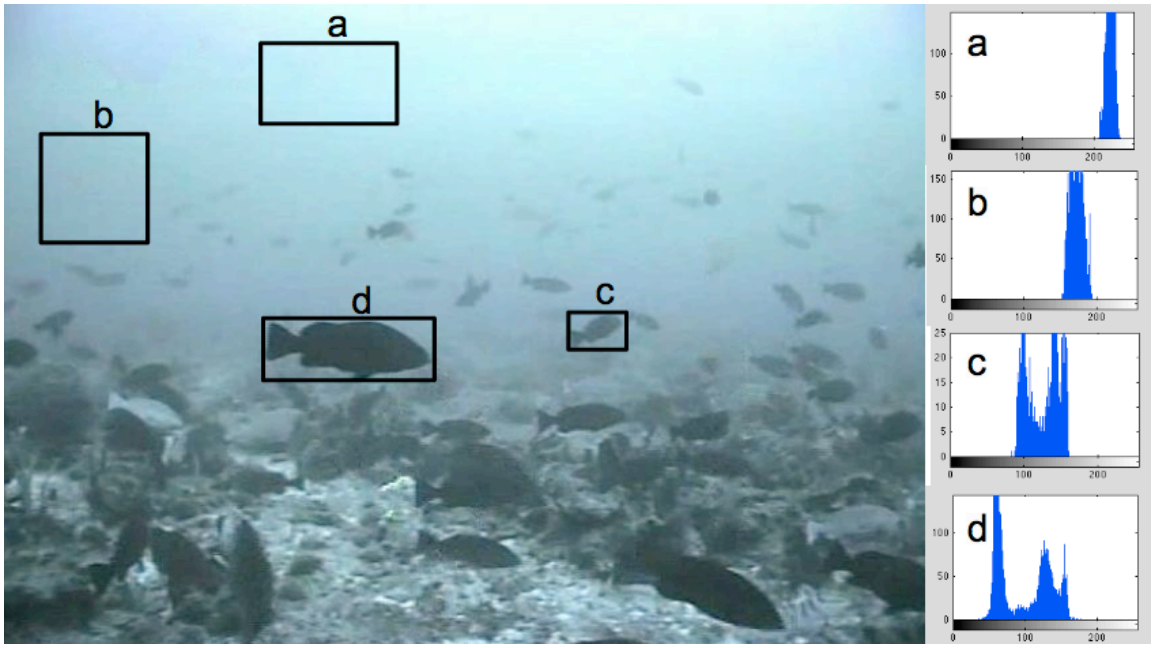

Figure 3.1: Various windows representing samples of the fish and the background water column, and their corresponding histograms.

3.2 Seafloor

It is sometimes hard to locate the fish above the seafloor because they blend in with the background. Different visual cues such as intensity, gradient and dense SIFT, and motion cues are to be considered for robust and accurate detection, as we describe next.

3.2.1 Intensity

The intensity distributions of fish over the seafloor are not typically bimodal, as shown in Figure 3.2. Here, one distribution corresponds to a window that includes a fish,

while the others do not. More precisely, let us take the peak of the dark pixels in (a) to be due to the fish and the brighter pixels correspond to the seafloor. Applying the same rule to (c) could potentially lead to misclassification. Therefore, intensity cue cannot always classify fish and the seafloor correctly.

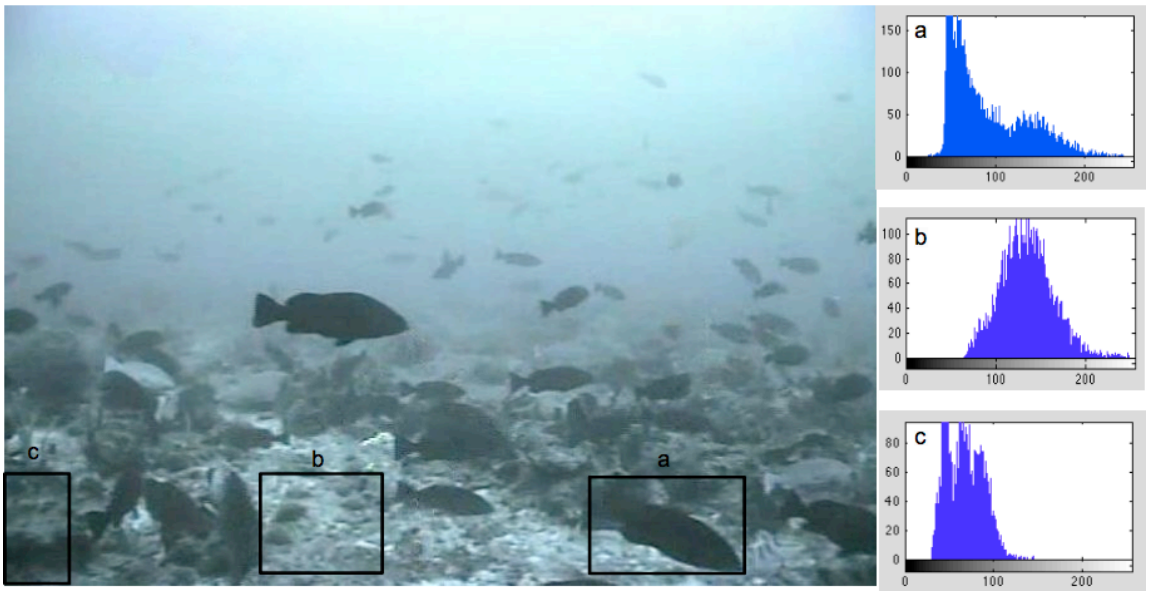

Figure 3.2: Various windows representing samples of the fish and background seafloor, and their corresponding histograms.

The same behavior is observed with the dehazed images. Although the enhanced images have better contrast, one still cannot always differentiate between the two classes according to intensity. As shown in Figure 3.3, the distribution of intensity is more even after dehazing, but dark parts of seafloor in (c) are still similar to the fish in (a). Considering the original and dehazed images, the

intensity characteristic is not enough to correctly label these two classes most of the time.

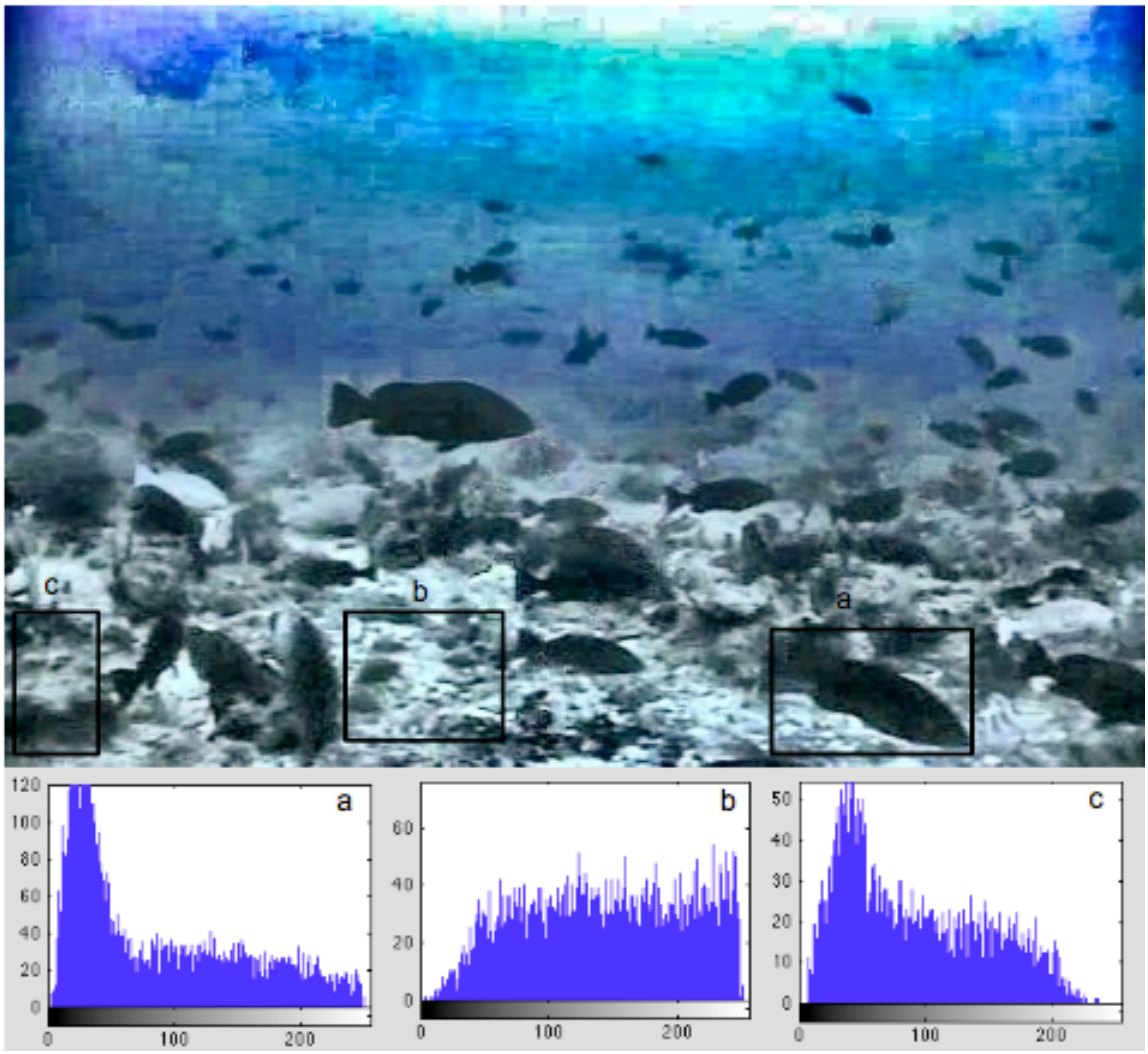

Figure 3.3: Various windows representing fish and seafloor and their corresponding histograms from the enhanced image.

3.2.2 Image Gradient

Although the intensity levels of fish and seafloor overlap significantly, the gradient is a useful cue to differentiate between the fish and seafloor: the fish has a

relatively uniform intensity; on the other side, the intensity of seafloor fluctuates due to rich texture.

Figure 3.4 shows the intensity gradient magnitude and gradient magnitude histogram of the fish and seafloor samples.

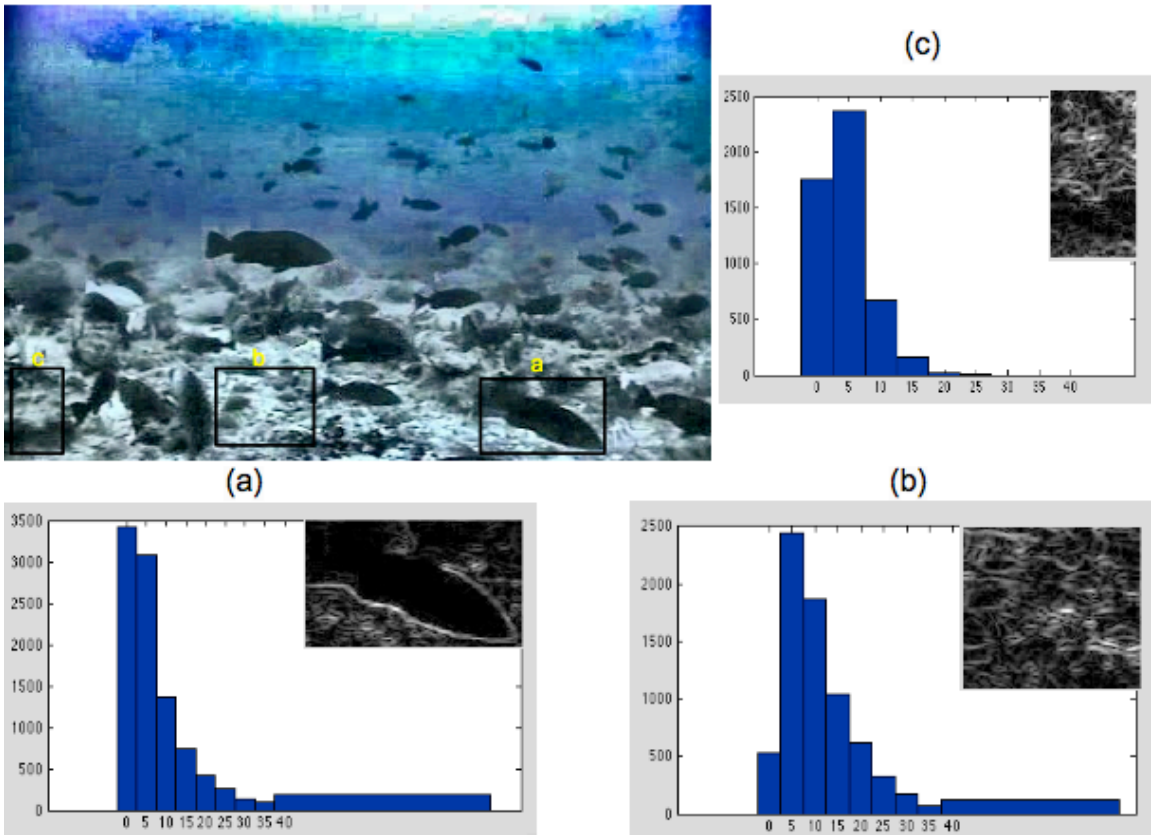

Figure 3.4: Gradient magnitudes and histograms of various fish and seafloor windows.

In the intensity gradient histogram, the peak due to large fish region appears near zero, confirming the uniform intensity of the fish. For the seafloor, the peak appears at around 5 gray levels. Having established the gradient as a potential visual cue, we instead consider a more

descriptive measure that accounts for the gradient within a local region.

3.2.3 Dense SIFT

The SIFT descriptor is a 128-dimensioned vector of gradient distribution over a 16 x 16 pixels area, surrounding each distinct image feature. To be precise, the SIFT descriptor is computed on a 4 x 4 grid, where each grid contains 4 x 4 pixels. At each pixel, the direction and magnitude of the gradient are calculated. An 8-bin[2] gradient histogram of all grids contributes to the SIFT descriptor. Dense SIFT applies the SIFT detector to all image pixels, providing a descriptor at each pixel [7].

The dense SIFT of an image, comprising a vector of length 128 at each pixel, cannot be visualized. The principal component analysis (PCA) can be applied to compress each vector into a much smaller set of values. For example, 3 principle components allow rendering as a color image to visualize. Figure 3.5 shows the intensity gradient magnitude and three-component dense SIFT of an image. Since the dense SIFT descriptor is calculated over a 16 x 16 pixels area, it reflects the variance over a local area.

[2] Eight directions described by corresponding gradient angles: horizontal $(0, \pi)$, vertical $(\frac{\pi}{2}, \frac{3\pi}{2})$ and diagonal $(\frac{\pi}{4}, \frac{3\pi}{4}, \frac{5\pi}{4}, \frac{7\pi}{4})$

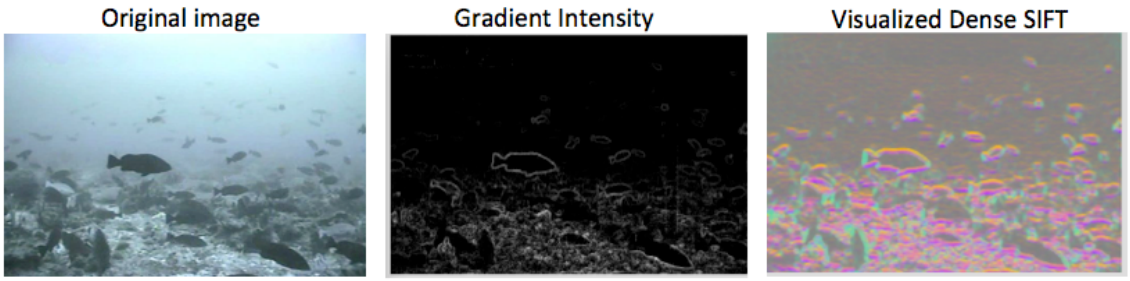

Figure 3.5: An image with intensity gradient and 3 major components of the dense SIFT (from PCA) as a color image.

The histograms (over three regions) of the dense SIFT magnitude, depicted in Figure 3.6, suggest that the fish and the seafloor have different local gradient attributes (as captured by the SIFT descriptor). There are two peaks in the histogram of the fish sample: the first one corresponding to the fish and the second one is for the surrounding seafloor. There is only one peak in the histograms of the seafloor samples. This verifies that the seafloor has stronger texture.

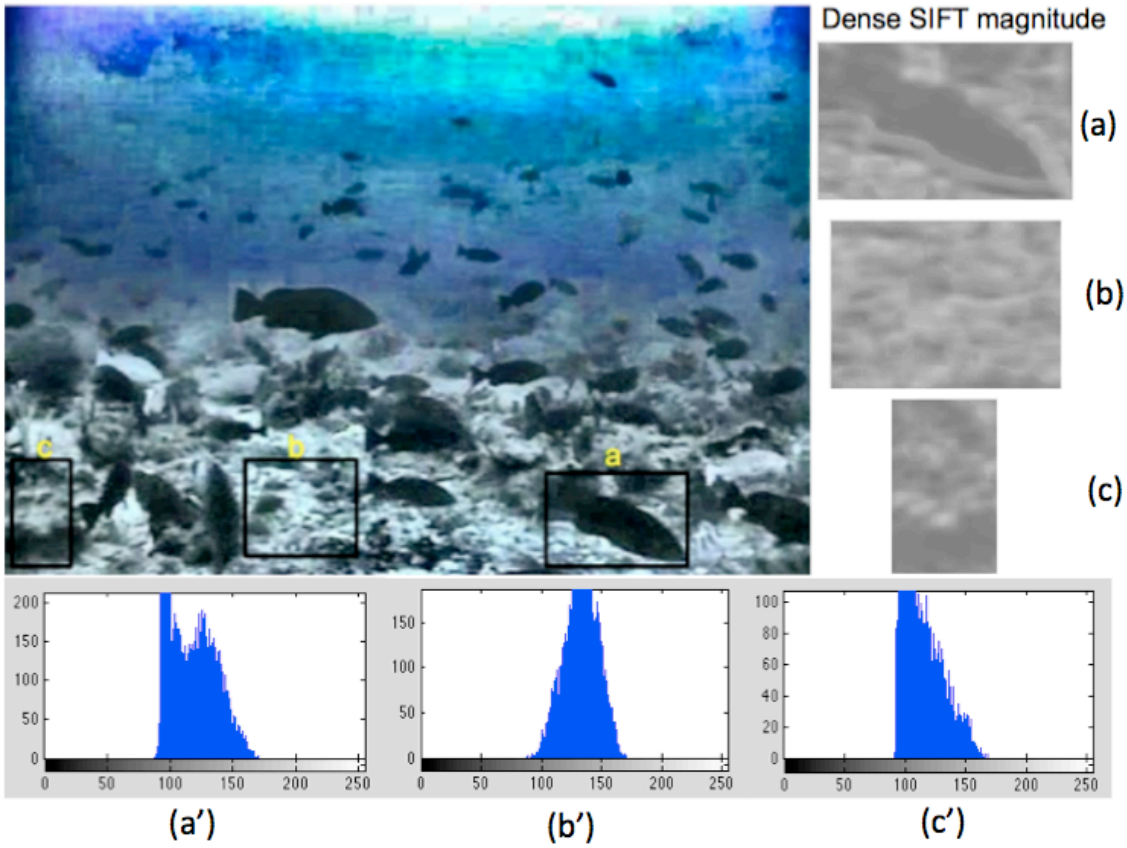

Figure 3.6: Dense SIFT magnitude and histograms of various fish and seafloor windows.

Finally, another example given in Figure 3.7 confirms that the gradient information encoded in the SIFT descriptor is a stronger cue than the intensity to differentiate between fish and the seafloor. The images are segmented using Otsu's method [8]. Otsu's method gives an optimal threshold for a bimodal histogram, by minimizing the intra-class variance, to segment an image region into two parts. Although the distribution is bimodal, the darker coral objects are grouped with the fish into one object (it

is bimodal, however, some of the coral objects are mixed with the fish). Thus, the fish cannot be correctly segmented from the background. However, the optimal threshold for the bimodal histogram of the dense SIFT magnitude, given in (c), correctly segments the fish; see (d). The coral objects with the same texture characteristics form other small independent blobs.

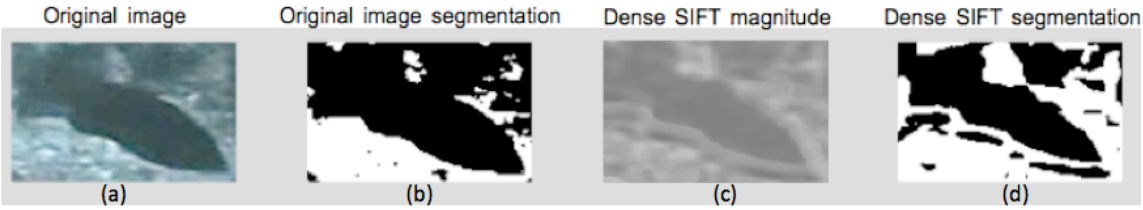

Figure 3.7: Segmentation using Otsu method, based on thresholding of intensity (a,b) and dense sift magnitude (c,d).

The seafloor has stronger texture than the fish over larger areas, however, may not be the case over smaller regions. As an example, the two windows (a) and (b) in figure 3.8 are alike in terms of the histogram of dense SIFT magnitude. However, if we assume that all fish are about the same size, window (a) can be ruled out as a fish because it is too small in comparison to (b). Therefore, estimating the fish size could avoid some potential misclassification. Finally, window (c) not only has similar appearance as fish, but also comparable size. The

discrimination based on appearance and size can be difficult (even for human!).

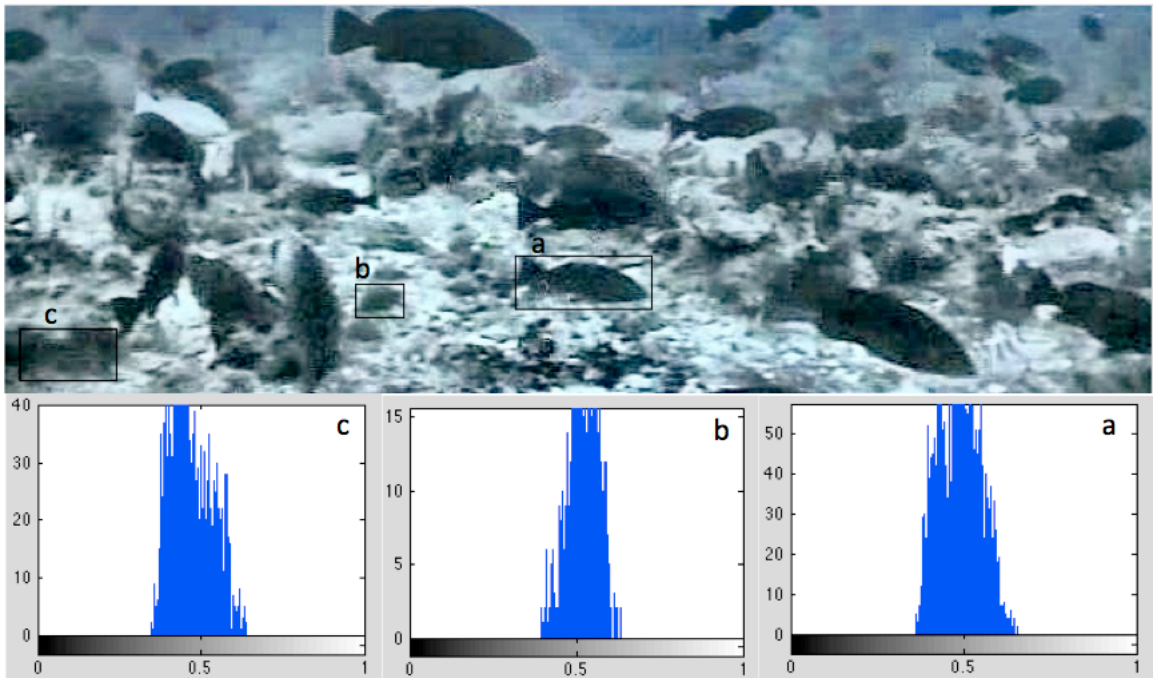

Figure 3.8: (b,c) Examples of reef objects with histograms that are similar to that of the fish (a).

3.2.4 Optical Flow

Motion is a very strong visual cue for object and background segmentation [3]. For example, many species utilizing camouflage for protection against predators (e.g., leaf-mimic butterfly) can only be detected once they move. In a video clip, we can make use of the difference between the fish movement and the background coral objects with similar appearance.

Given a pair of consecutive images in a motion sequence, the optical flow provides an estimate of object movement.

When the lighting is unchanging, the intensity I of point (x,y) at time t is expected to remain unchanged when it moves to a new position $(x+dx,y+dy)$ at time $t+dt$:

$$I(x,y,t) = I(x+dx,y+dy,t+dt),$$

where (dx,dy) is the image motion of the point (x,y) during dt .

Optical flow is a good estimate of an object's image motion when object texture is strong and the lighting variation is insignificant. In this work, the optical flow is more reliable over the seafloor due to the strong texutre, but not often the case within the water column. This is not a drawback since the detection complexity persists mainly when the fish swim near the seafloor.

Figure 3.9 shows the optical flow field superimposed on the first frame of a motion sequence, where the arrows show motion directions, and the length of arrows are proportional to the speed (magnitude of displacements).

As one notes, there is an optical flow over the stationary background, which is induced by the camera motion. It changes smoothly over most of the image. In addition, the movement of each fish generates an independent image motion, the direction and magnitude of optical flow changed wherever moving fish appears.

Therefore, the discontinuity of optical flow is a useful cue when video is available.

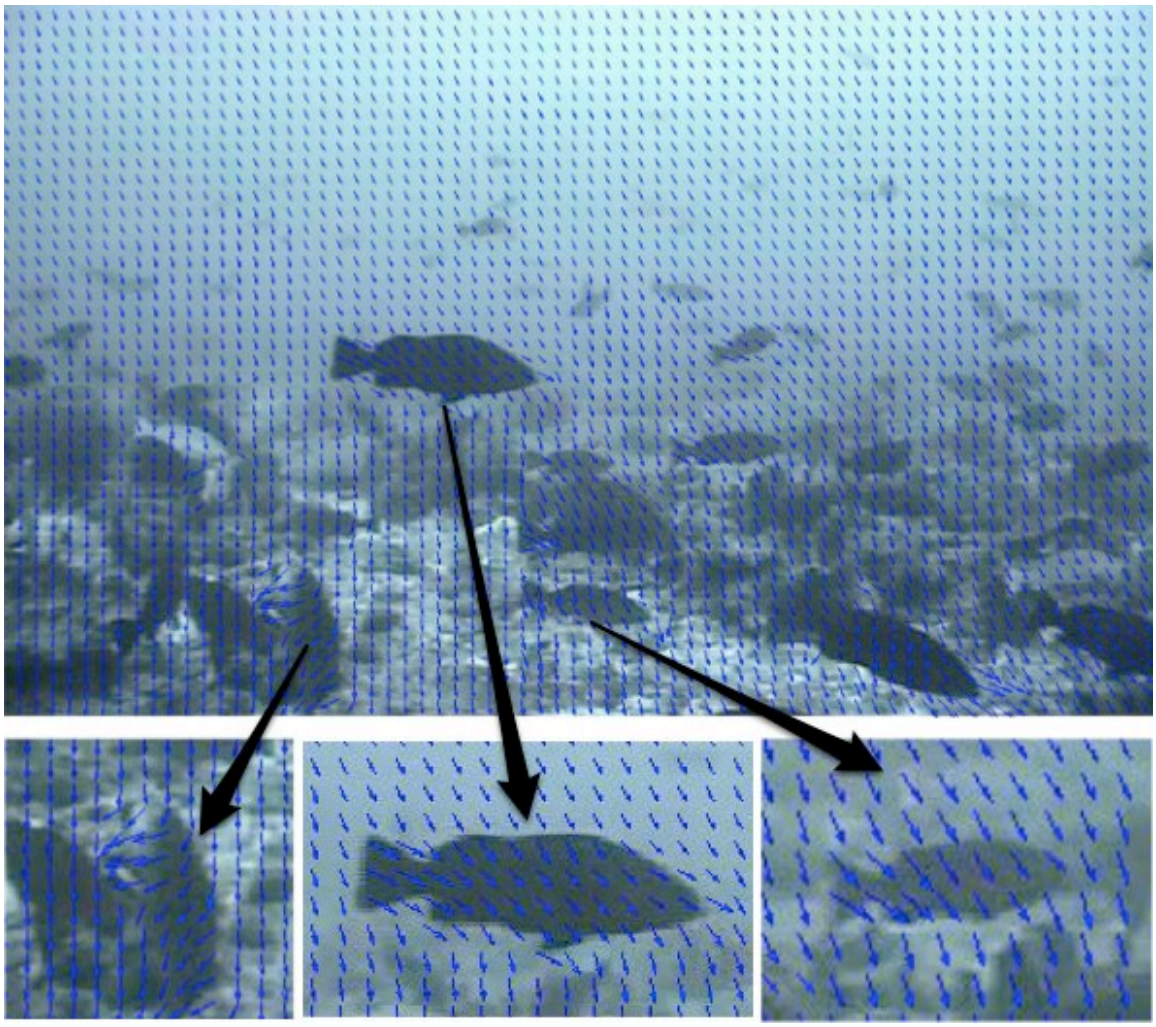

Figure 3.9: Induced optical flow (shown by blue arrows) by camera motion over the background is different over the fish, both in magnitude and direction, due to independent fish motion.

In conclusion, although the brightness distribution is not sufficient to distinguish the fish over the sea floor, the local image gradient information encoded in the dense

SIFT magnitudes and the optical flow are rich visual cues that can be utilized to locate fish.

3.3 Hybrid Detection Method

As stated above, the fish and water column have distinct intensity characteristics, but this is not so for the fish over the seafloor. Due to the difference between water column and the seafloor, this thesis proposes a hybrid detection method, as depicted in Figure 3.10, involving the following computational steps:

- Detection of the horizon, to divide the background regions of the image into two parts, either the water column and seafloor;
- Detection of the fish within the water column using mainly the intensity cue;
- Detection of the fish over the seafloor by utilizing intensity gradient and optical flow cues;
- Utilizing the depth cue to establish the applicable window size containing fish at each depth.

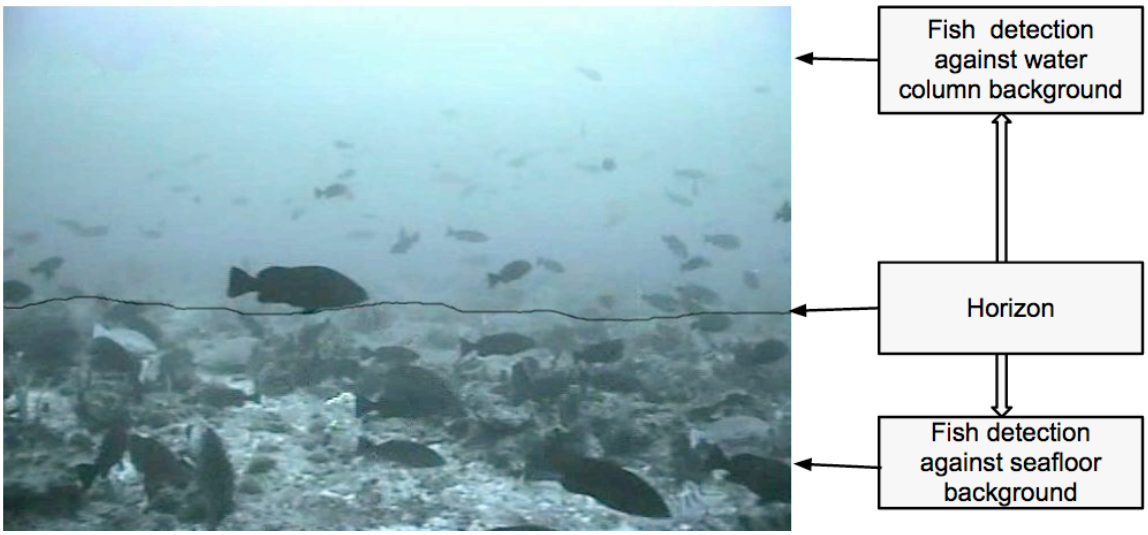

Figure 3.10: Scene segmentation by the detection of the horizon.

Chapter 4

Hybrid Detection Method

As discussed in previous chapter, our approach starts with locating the horizon (line dividing the image into the water column and the seafloor regions), in order to detect the fish within each region based on different visual cues. This chapter elaborates on the method(s) for detecting the horizon, and for utilizing different visual cues to locate the fish.

4.1 Horizon Detection

Detection of the horizon, where the seafloor and the water column meet, as depicted in Figure 4.1, is the first step in our method. We can utilize different cues to locate the horizon, as we describe next.

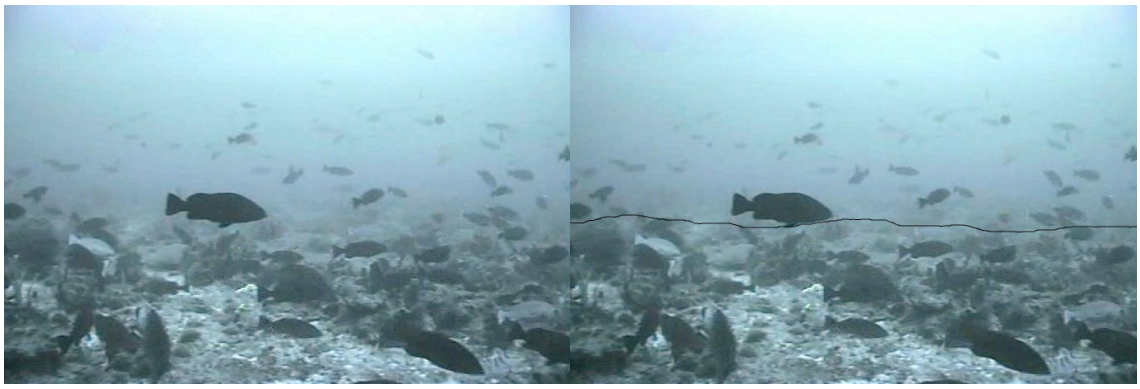

Figure 4.1: Horizon line drawn manually

4.1.1 Gradient-Based Horizon Detection

The water column and the seafloor have different appearance characteristics. The water column has low intensity variations due to the uniform color, while the seafloor has high variations due to the strong texture. Therefore, intensity variations can be a cue of horizon detection.

The image gradient encodes information about the local intensity variations. The horizon can be declared as the line separating the weak from the high gradient magnitude regions. Figure 4.2 shows the horizon drawn by human and the average intensity gradient magnitude of each row of the same image. The horizon drawn by human is at about row 300, where the average gradient magnitude is relatively high.

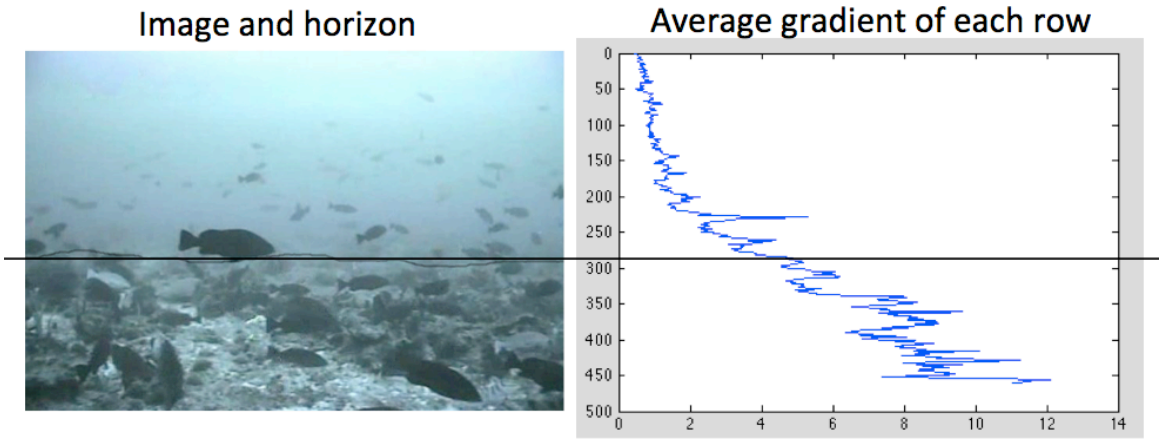

Figure 4.2: Relationship between horizon drawn by human and the gradient of average intensity (along each scan line).

Finding the peaks of gradient intensity enables locating the horizon. Figure 4.3 (a) shows the peaks with gradient intensity greater than 5, the line on the right shows the position of the first peak. Figure 4.3 (b) shows the peaks with gradient intensity greater than 10, the line on the right shows the position of the first peak.

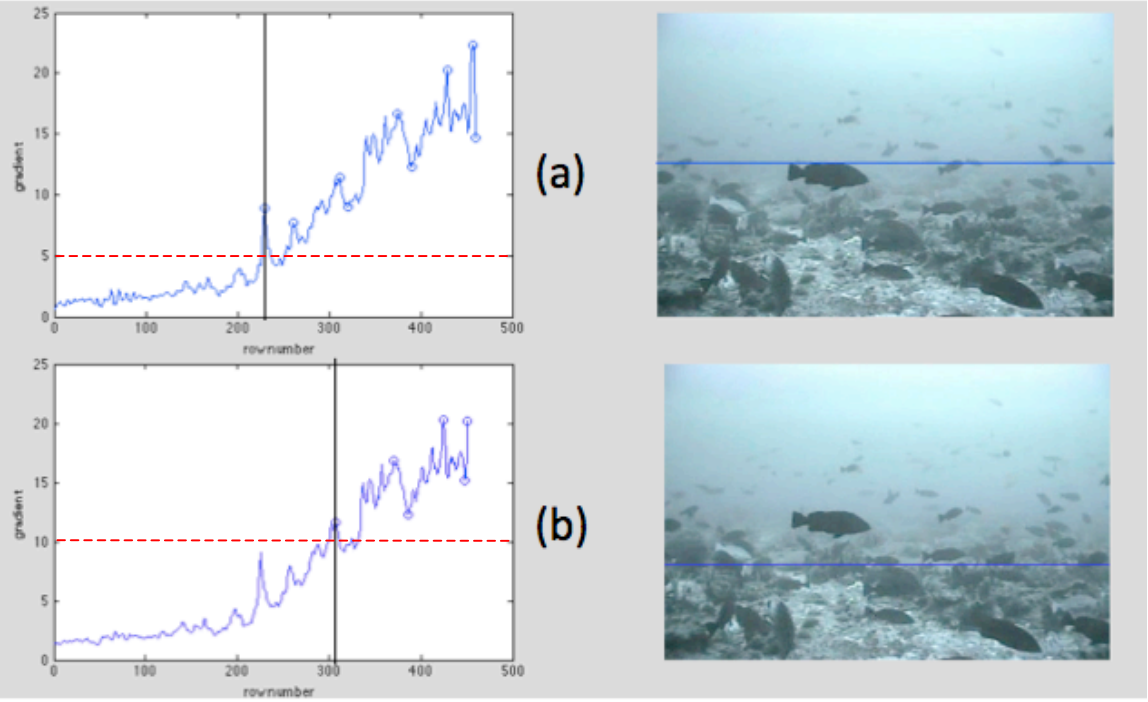

Figure 4.3: Gradient based horizon detection by fist peak of gradient above a certain threshold (a) T=5; (b) T=10.

4.1.2 Depth-Based Horizon Detection

Similar to the skyline, the benthic horizon is where the seafloor appears to meet the water column. It can be detected as the line where there is a depth discontinuity. As shown in Figure 4.4, moving from the sea floor towards

the water column, the depth drops gradually until we reach a discontinuity, corresponding to the sought after horizon.

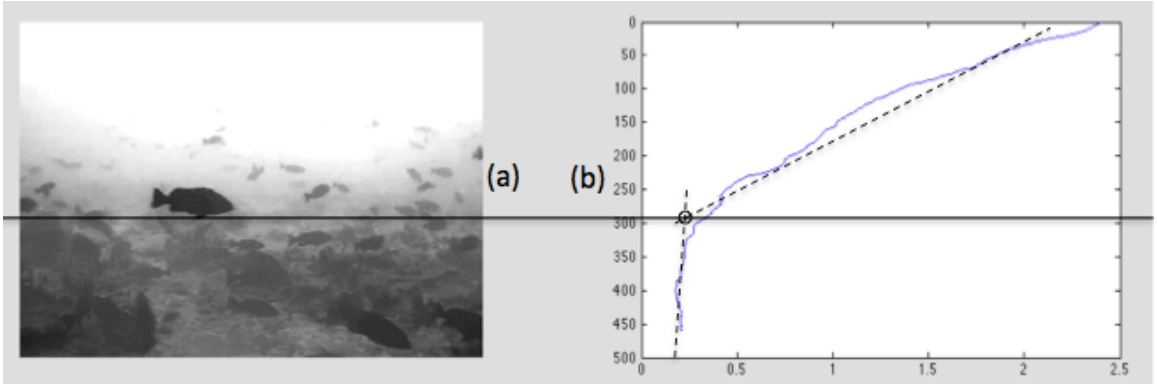

Figure 4.4: Depth map (a) and average depth of each row (b). The line shows location of sharp change is depth variation.

As shown in Figure 4.5 (a), the depth histogram has two modes with sharp and mild peaks. The scene can be segmented into two parts according to the depth values using the Otsu's method.

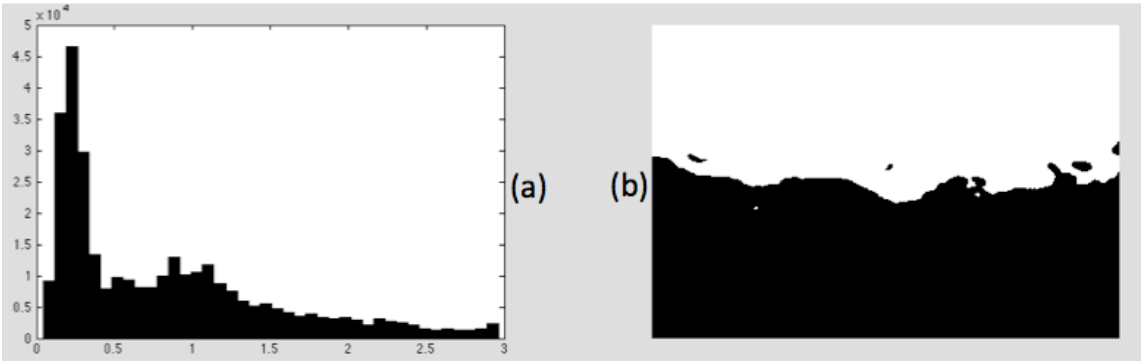

Figure 4.5: Depth histogram (a) and locating horizon for scene segmentation using Otsu method (b).

However, Otsu's method does not guarantee consistent two-region segmentation; isolated fragments can exist. Since the depth variations are more or less monotonic, as shown

in Figure 4.4 (b), the isolated small fragments can be removed. To obtain a continuous horizon, we can sweep each vertical line from the bottom, searching for the transition from black (sea floor) to white (water column). We then set all the points on the column above the transition point to white. The estimation horizon is shown in Figure 4.6 (b).

This approximation works when the depth is approximately consistent, so fragments only appear around the horizon.

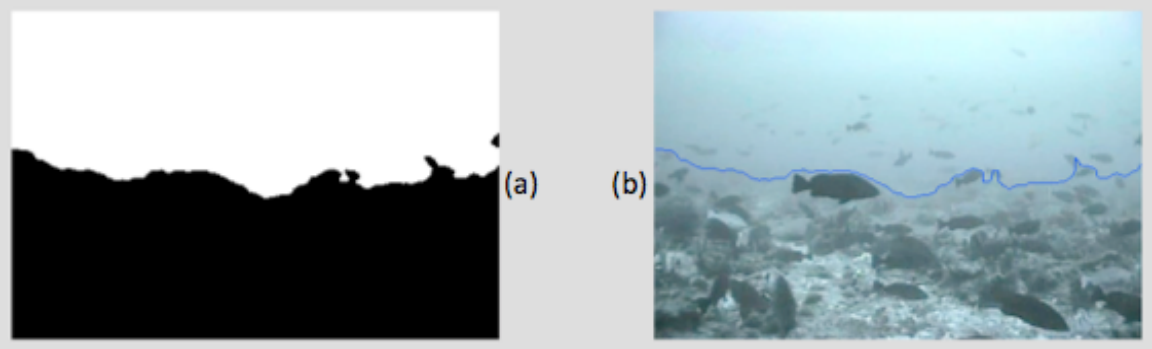

Figure 4.6: Horizon based on discontinuity (sharp change) in depth variation.

Figure 4.7 compares the horizon line drawn by human (a), and those found by using a gradient magnitude threshold of 5 (b) and depth discontinuity (c). The horizon found using the depth discontinuity is more natural to use.

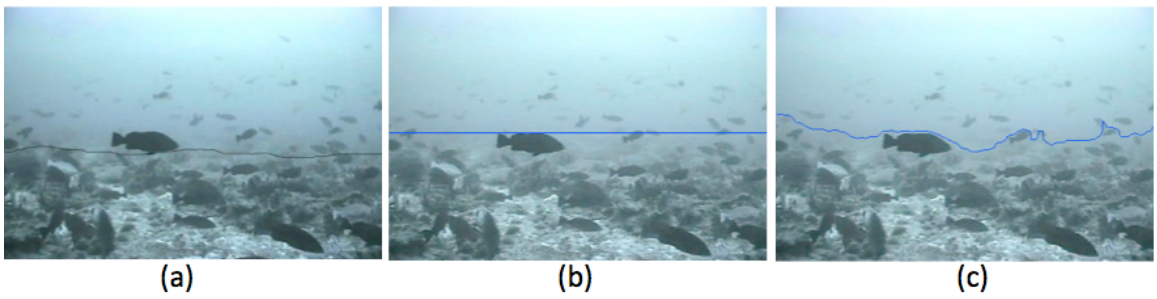

Figure 4.7: Comparing horizons drawn manually (a), and detected according to image gradient (b) and depth gradient (c).

4.2 Adjacent Neighborhood-Based Segmentation

To produce the initial regions for fish detection, we have adopted the graph-based segmentation method of Felzenszwalb and Huttenlocher's [1], where the image is segmented according to the color dissimilarity. This method can be applied to segment the image according to the intensity, intensity gradient magnitude, and optical flow magnitude and (or) direction.

In this method, each pixel is initially a unique region that is connected to its 8 neighboring pixels by edges with weight related to the color differences. The internal variance of a region is defined as the largest weight of the minimum spanning tree of the region:

$$IntR = \max_{e \in MST(R)} \omega(e).$$

The color difference of two regions is the minimum weight of the edge connecting the two regions,

$$Dif(R_1, R_2) = \min_{v_1 \in R_1, v_2 \in R_2, (v_1, v_2) \in E} \omega(v_1, v_2).$$

Two regions will be merge if the color difference is smaller than the minimum internal variance of the two regions,

$Dif(R_1, R_2) < MinInt(R_1, R_2)$.

The minimum internal variance is the minimum value of the sum of the internal variance and a threshold of two regions,

$$MinInt(R_1, R_2) = \min\left(IntR(R_1) + \frac{k}{|R_1|}, IntR(R_2) + \frac{k}{|R_2|}\right),$$

where $|R_i|$ denotes the size of region and k is a constant parameter that allows the regions with very small number of pixels to merge. Larger k produces larger regions, while smaller k leads to smaller regions.

As stated, this method can be applied to other properties under consideration. In this case, $\omega(v_1, v_2)$ represents the dissimilarity of the intensity, gradient magnitude, or optical flow for two neighboring pixels v_1, v_2.

4.3 Fish Size Estimation

After initial regions are formed, those initial regions with inconsistent size are excluded. To determine this, let us assume that all fish are of the same actual size. The image size of the fish varies inversely with the distance from the camera. This can be modeled by $s = \frac{f}{d}$ where f is a constant. The height of fish h is proportional to the

length of fish $h = as$, where a representing the height to length ration, is a constant.

Figure 4.8(a), showing the plot of fish size –vs- depth (distance from camera) of 80 fish samples, verifies the assumed relationship. The estimate coefficient is $f = 10.9116$, error $\sigma = 17.6413$ pixels. Aside from possible size variations, we note a biased error at larger distances. This can be readily explained by noting that these data points represent fish within the water column. Affected by image blur due to scattering, the fish sizes are over-estimated. In (b), the linear relationship between the height and width of the fish is confirmed. Applying $h = as$, where the coefficient $a = 0.3274$, error $\sigma = 3.1739$ pixels.

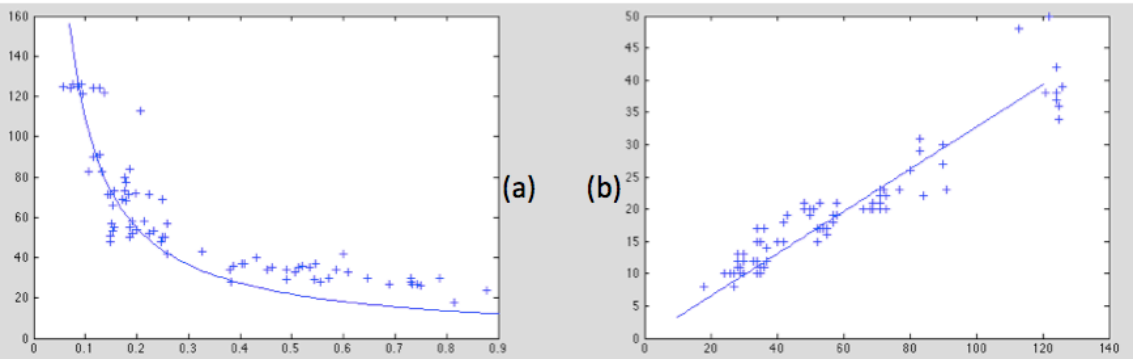

Figure 4.8: Fish length –vs- depth (a), and fish length –vs- height (b).

Another source of error is due to variations in the estimation of scene depth. Recall that the scene depth is an estimation obtained as a byproduct of dehazing. Although

the up-to-scale depth model applies within one frame, the scaling of depth is most likely different across frames. To confirm, we have identified a large fish in the near field that is swimming parallel to the camera in a few frames. Ideally, the depth of the fish must be constant. However, the estimated depth values over 9 consecutive frames are 0.0750, 0.0694, 0.0875, 0.1092, 0.0912, 0.0935, 0.1153, 0.1147 and 0.1543, respectively. We may instead assume the model $s = \frac{f'}{d + a}$, where *a* represents the variations in the reference point for the depth values for each frame. We now estimate a, f' of each frame, and then scale and shift the depth for each frame to the same reference point. The adjusted depth values now satisfy the original model $s = \frac{f}{d}$ as depicted in Figure 4.9. The new model coefficient is $f = 10.4217.$ with error standard deviation $\sigma = 13.1550$ pixels. We note that the size bias at distant fish, due to blurring, still exists.

In Figure 4.9, we have analyzed potential variations in fish size, in addition to size uncertainty due to blurring. (where up-to-scale depth exceeds 0.4). The sizes of most fish within the water column are within twice the estimate

by the model (shown in red). For the fish at closer distance, we have computed a variation of about 20% from the size predicted by the model (see dashed green lines). This can be attributed to the actual variations in fish size, where most of the samples fall within this 20% range.

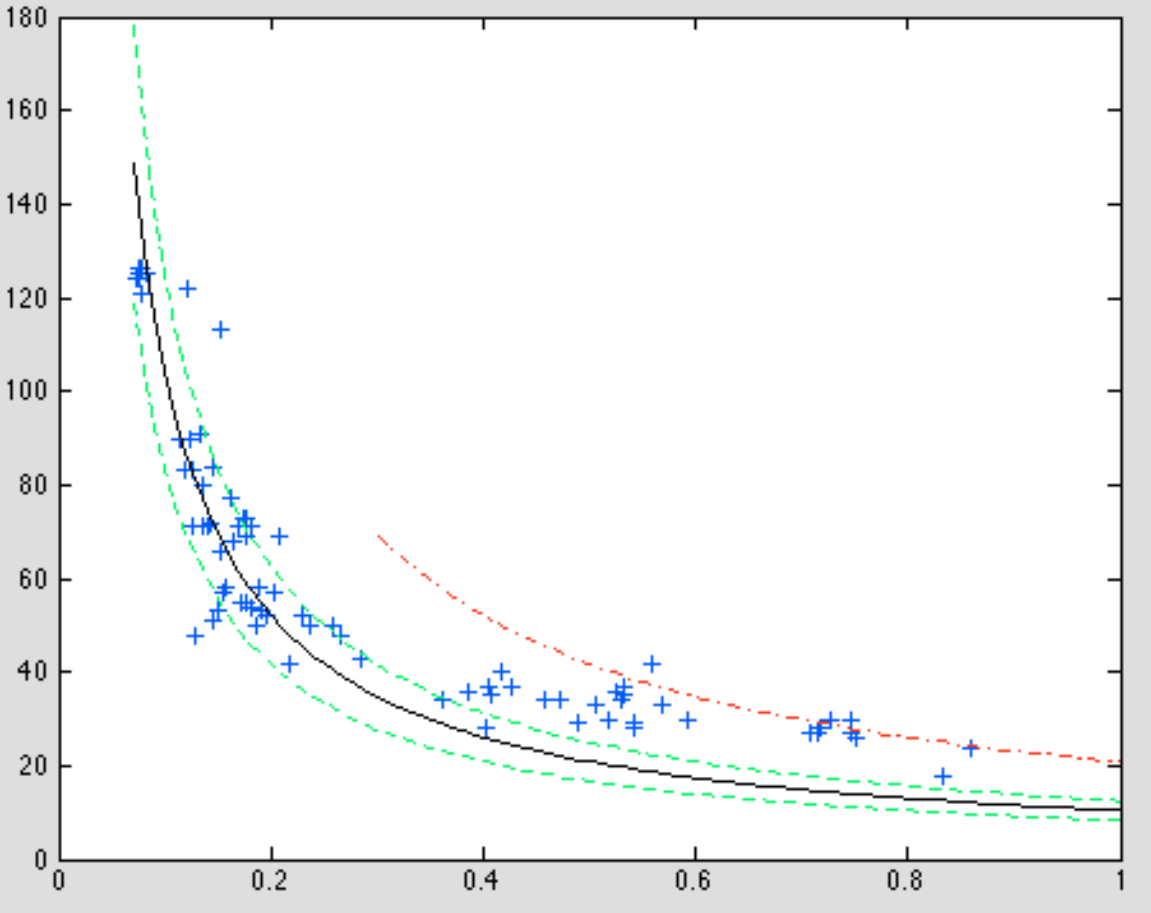

Figure 4.9: Estimate size with 20% variation at close distance (dashed green line) and 100% at far distance (dashed red line).

4.4Detect Fish within Water Column

To detect fish within the background water column, the initial regions are produced by the adjacent neighborhood-

based segmentation. Then, regions (windows) with suitable size at each depth are considered as ones containing fish.

4.4.1 Intensity-Based Detection

As discussed in chapter 3.1, the water column and fish have distinct intensity distributions. Using adjacent neighborhood-based segmentation, water is merged into a large region (shown in orange in Figure 4.10). This large region is considered as background and excluded. The remaining small blobs are potentially fish, as well as some formed due to noise. We can exclude noise regions by utilizing the fish size estimates. As determined in section 4.3, the fish at depth greater than 0.4 (within water column) can vary by up to 100% with respect to the size predicted by the model. If the fish is closer, the size should fall within roughly 0.8 times to 1.2 times of the estimated size. All regions violating the size limit are excluded. The rest regions are shown in windows in Figure 4.10.

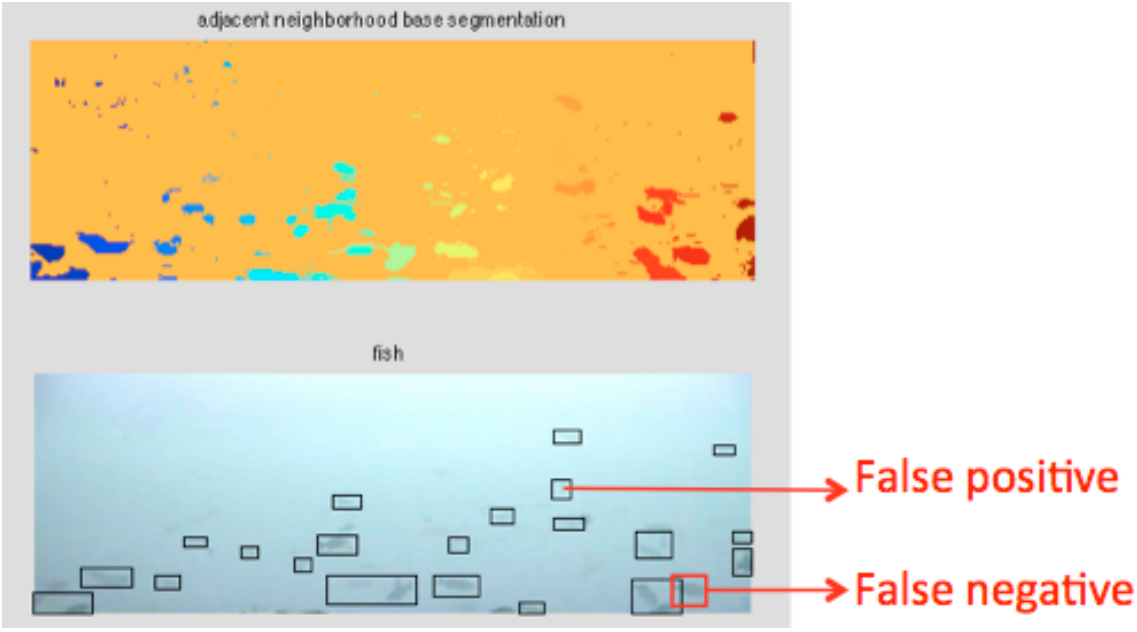

Figure 4.10: Detected fish within the water column using intensity cue.

In this example, there are 25 fish (Targets=25) in the view, the system identifies 24 of them (Hits=24) and with 1 false positive (False Positive=1). The precision is 96% where precision is defined as

$$Precison = \frac{Hits}{Targets},$$

and the recall is 96%, where recall is defined as

$$Recall = \frac{Hits}{Hits + FalsePositive}.$$

4.4.2 Intensity-Based Detection in Dehazed Image

To assess the performance by utilizing the enhanced image, the same method is applied to the dehazed image. As stated, although the dehazing improves the contrast of the image, it leads to a remarkable level of noise in red

channel within the water column. The average image of the green and blue channels is more reliable for segmentation. Figure 4.11 shows the detection results using the average image.

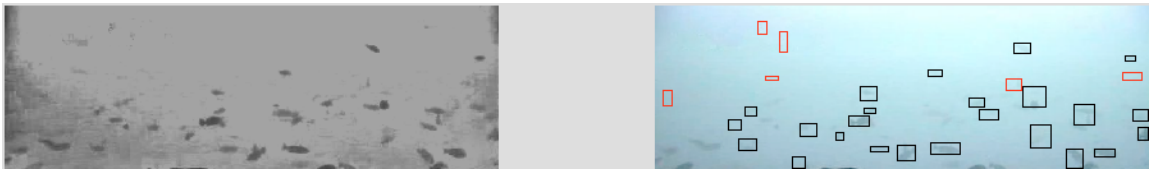

Figure 4.11: Detected fish in the blue and green channels of enhanced image. Red boxes shows false positives.

Out of the 25 fish in this view, we are able to detect 20, but there are also 6 false positives. The precision is 80% and the recall is 76.9%, in this example. Note that most of the false detections are over regions that are most distant from the camera.

For this example, compared the fish detection precision and recall rates by using 1) intensity, 2) average intensity of the blue and green channels of the enhanced image.

Table 4.1: Compare the precision and recall to detect fish in water

	Intensity of original image	Average intensity of blue and green channel
Precision	96%	80%
Recall	96%	76.9%

Based on this, detecting fish using the intensity cue in the original image gives the best precision and recall rates.

4.5 Detect Fish Over Seafloor

As elaborated, the fish over the seafloor is not as distinct as the fish within the water column (based on intensity distribution only). Other visual cues, such as dense SIFT magnitude and the optical flow are employed, and the performance with each and both cues are compared.

4.5.1 SIFT-Based Detection

As explained in section 3.2, the fish and seafloor have similar intensity distributions, while the seafloor has stronger texture than fish. Therefore, variations in the dense SIFT magnitude was proposed as a suitable cue for detecting fish over the seafloor. As described in section 4.4.1, the initial regions are created using adjacent neighborhood-based segmentation and the ones violating the size cue are excluded. An example of the detection based on the similarity of the dense SIFT magnitudes is shown in Figure 4.12. We are able to found 18 fish with 9 false alarms. This gives a precision of 78.3% with a recall rate of 66.7%.

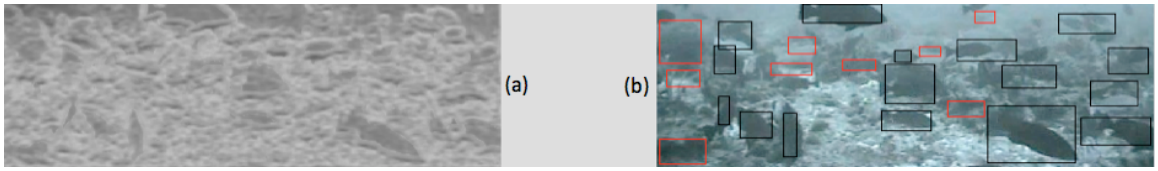

Figure 4.12: Detecting fish using dense SIFT magnitude.

4.5.2 Optical Flow-Based Detection

When video frames are available, optical flow between two images can be computed and utilized. The detection method can utilize the similarity and (or) discrepancy of optical flow. As an example, the regions with different optical flow have been detected in Figure 4.13.

Figure 4.13: Detecting fish using optical flow cue.

Here 10 fish are found entirely, 8 more fish are found in part and there are only 3 false positives. False positives are caused by the inaccuracy of optical flow estimation. There are at least two reasons why only part(s) of a fish are found. First, fish is not a rigid object and sometimes, the tail moves more noticeably than the body. Second, the texture of fish is not as strong as the seafloor, and thus the optical flow is reliable mainly on the edges of the fish.

On the positive side, a fish may be fully blended with the background (marked in red in Figure 4.14), which can go undetected (even by human) and (or) our automated detector based on the dense SIFT magnitude. However, it is detected by utilizing the optical flow because the fish is swimming against the direction of camera movement. However, not all fish in the view can be detected based on optical flow. For example, the fish marked in yellow in Figure 4.14 is not detected because its image motion is not sufficiently different from the background.

Figure 4.14: Moving fish (red) and still one (yellow).

4.5.3 Variation and Optical Flow-Based Detection

As the results in Figure 4.12 show, the dense SIFT magnitude detects the fish boundary precisely, but could produce some false positives for any other region with similar appearance and suitable size, including some darker coral objects. Based on the results in Figure 4.13, the

detection using the optical flow produces fewer false alarms but some fish are identified partially (only part of fish body is detected).

By integrating the optical flow with appearance cues we can overcome these shortcomings; the appearance cue could overcome the broken fish regions and the optical flow could verifies potential regions and reduce the number of false positive. Integrating optical flow detection can also improve precision. To reduce false positive and improve precision, we can devise a very simple method as following.

a) Denote R_a regions obtained from appearance cue, and R_f from optical flow

b) For each region $r \in R_a$, if it does not contain any region in R_f, then remove region r from R_a

c) For each region $r' \in R_f$, if it is not contained by any region in R_a, then add it to R_a

d) R_a is the set of regions with fish

Figure 4.15 (c) shows the result of applying this strategy. There are 23 targets, 19 fish are found with only 3 false alarms (in red boxes). The precision is 82.6% and the recall is 86.4%.

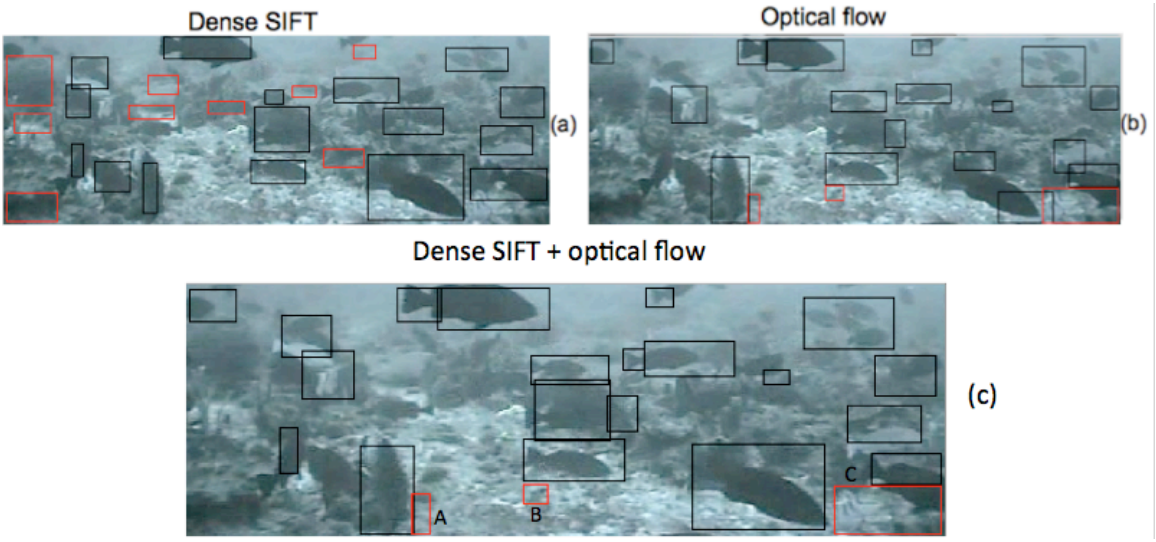

Figure 4.15: Detecting fish using dense SIFT magnitude and optical flow.

In these results, the number of false positives decreases, as compared to merely using the dense SIFT magnitudes. The remaining false positives are due to the inaccuracy of computed optical flow. These are shown with windows (A),(B) and (C) in Figure 4.15 (c). They are added to the result because of noisy optical flow, which is different from the image motions of the neighboring pixels. In summary, incorporating optical flow can reduce the false positives from the dense SIFT magnitude quite effectively, although false positives may be introduced due to the inaccuracy in estimating the optical flow.

4.5.4 Region Growing

Although integrating dense SIFT and optical flow alleviated the broken fish, there are some left because of

step (c) in the procedure described in section 4.5.3. To further resolve this problem, we post-process the results to determine if a region can be grown to completely bound from "a broken fish."

The region growing method examines the neighboring pixels of the initial region, adding pixels with similar color/intensity.

In step (c), for a region $r' \in R_f$, that it is in any region in R_a, we feed it to the region growing method, instead of adding it to R_a directly. If the grown region r'' does not violate the estimated size, we add it to R_a.

Region growing also helps reducing false positives. The seafloor has strong texture, and regions grown according to intensity or color are usually small; the fish has dark uniform intensity, and grown regions can be relatively large. Figure 4.16 provides two examples with seed regions from the sea floor and a fish. The region grown from the seafloor (a) is quite smaller than the fish (b). The region on the right conforms to the estimated fish size, thus is added to R_a.

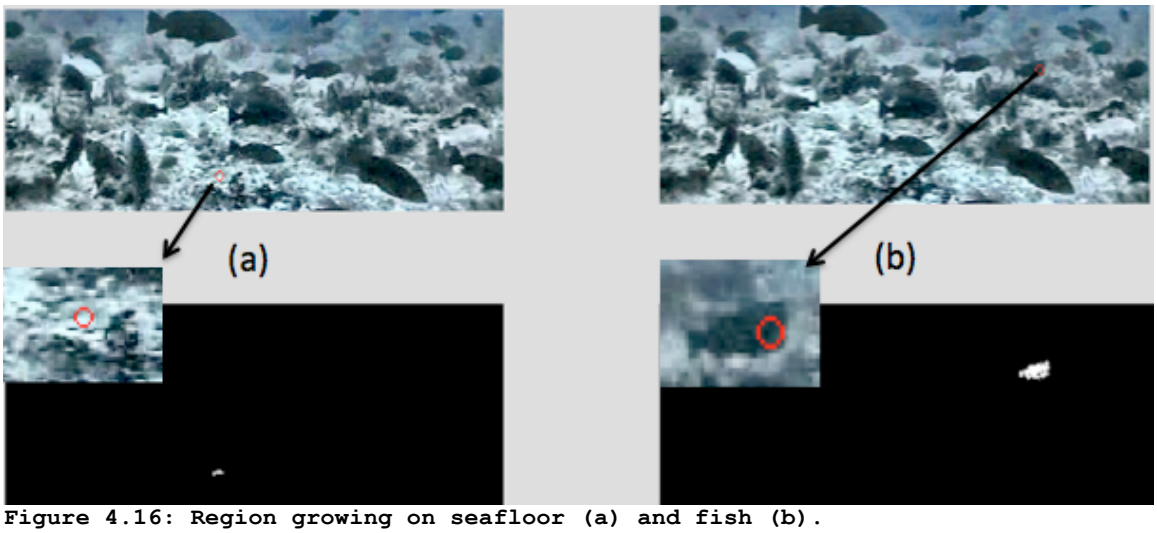

Figure 4.16: Region growing on seafloor (a) and fish (b).

The result of integrating region growing is shown in Figure 4.17. We see three kinds of change with region growing: 1) Window (a, b) over the seafloor is removed in the right image by applying the size cue after region growing, thus resolving this one false positive. 2) A part of fish in window (d) in the left image is grown to a whole fish within window (d') in the right image. 3) We only try to make regions smaller than fish to grow. A false positive (c), which has size similar to fish, is not excluded.

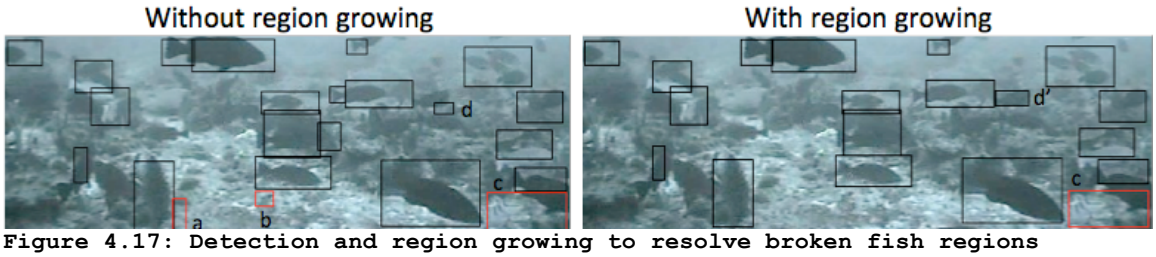

Figure 4.17: Detection and region growing to resolve broken fish regions

For The region growing method is only applied to small regions in R_f (from optical flow-based detection) in this example. It may also be applied to other steps (e.g. dense SIFT-based detection) to obtain a more precise boundary.

4.5.5 Fish tracking

Optical flow provides an estimate of how object moves. We can keep tracking of fish using optical flow. A fish at position (x,y) in current frame should be at $(x+dx, y+dy)$ in next frame, and $(x-dx', y-dy')$ in previous frame, where dx, dy is the optical flow between current frame and next frame and dx', dy' is the optical flow between previous frame and current frame. . Figure 4.18 shows the estimated positions of a fish in three consecutive frames.

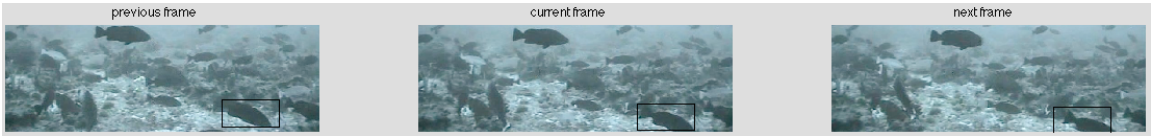

Figure 4.18: The estimated positions of the fish in 3 consecutive (previous, current and next) frames.

Keeping track of detected windows can improve the detection performance. For example, in Figure 4.19, assume a fish is detected in frame 1 in some window (a) and in frame 3 in window (b). The estimated positions of windows (a) and (b) in the current frame are (a') and (b'). For the

same fish, we expect (a') and (b') to be roughly in the same location. This suggests that we would expect a fish at or near this predicted position when it is detected in both the previous and next frames. Conversely, if a fish is detected within some window (c) in the current frame, but not in the expected positions (c') and (c'') according to optical flow in the previous and next frames, we remove (c) as a false positive.

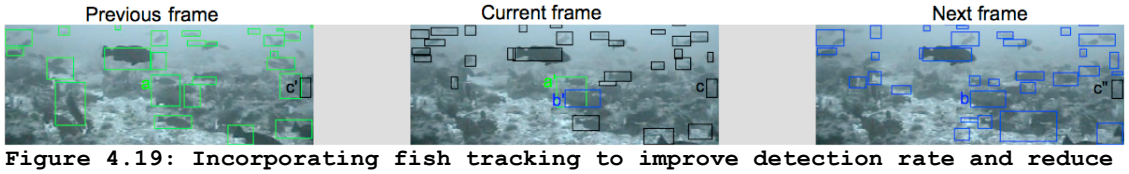

Figure 4.19: Incorporating fish tracking to improve detection rate and reduce false detection. Blue and green boxes show window from previous frame and next frame.

The tracking of detected fish windows can reduce the false positives and increase the hits. In Figure 4.20, there are 13 fish detected with 2 false positives before incorporating the tracking scheme. With tracking, there are 15 detected fish and 1 false positive. The number of hits increases by 2, and the number of false positives decreases by 1.

Figure 4.20: Detection results with and without tracking. Red box shows false positives

To summarize, we have described our method of detection fish in benthic habitats by dividing images into two regions, namely the water column and the seafloor. We have implemented fish detection within the water column based on intensity cue and based on dense SIFT and optical flow over the seafloor. We have demonstrated through a sample image and a motion sequence, why various components of our detection method such as size estimation, region growing and fish tracking are necessary in order to improve the hit rate, while simultaneous reduce the number of false positives. In the next chapter, we present the results of applying our method to a larger data set.

Chapter 5

Experimental Results and Assessment

As we have described in detail, our fish detection method divides the image into two parts, namely, the water column and the seafloor. Within the water column, we detect the fish using the intensity cue. Over the seafloor, we detect fish using the dense SIFT magnitude and optical flow (where video or consecutive frames are available). We utilize the expected fish size at different depths and tracking of detected windows to improve the results. In this chapter, we test the method on 6 new images.

In our method, initial regions are created based on intensity (within water column) or dense SIFT (above seafloor) are not necessary fish regions. Noise can be detected as fish when detecting the intensity dissimilarity, e.g. window (A) in Figure 5.1(a), low texture seawater and small part of reefs can be detected as fish in dense SIFT base detection. Estimated fish size can reduce those confusions. Small regions caused by noise (A), large regions such as the seawater (B) and small reefs (C) can be rejected by estimated size at current depth.

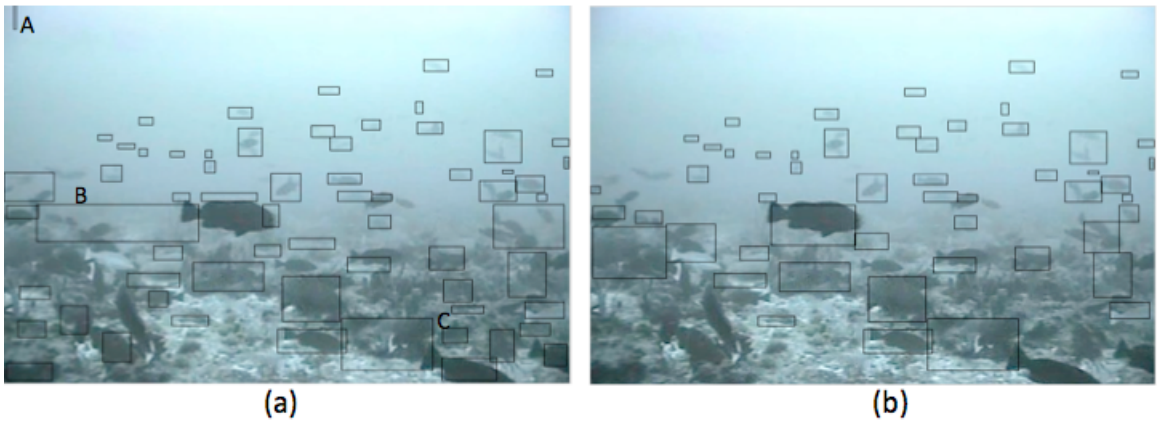

Figure 5.1: Comparing detection based on dense SIFT (a) and post-processing based on expected fish size (b). (same here, don't use no roman numerals)

Regions with suitable size, e.g., in Figure 5.2(b), are not always fish either. Therefore, the motion cue is utilized to reduce false positives and increases hits. The image emotion of reef objects, induced by camera motion, varies smoothly over the image. However, the optical flow of fish shows sharp variations from the background, due to the independent fish motion. Figure 5.2 (b) illustrates the detection result after integrating the optical flow cue. The non-fish regions such as (A, B) are excluded due to the lack of independent movement. A whole fish (D) and two partial fish parts (C, E) are detected. To obtain a complete fish part, the region growing method is applied to regions (C, E). However, because region (E) is blended with the reef, it does not grow correctly to complete the fish, as for region (C') in Figure 5.2 (c).

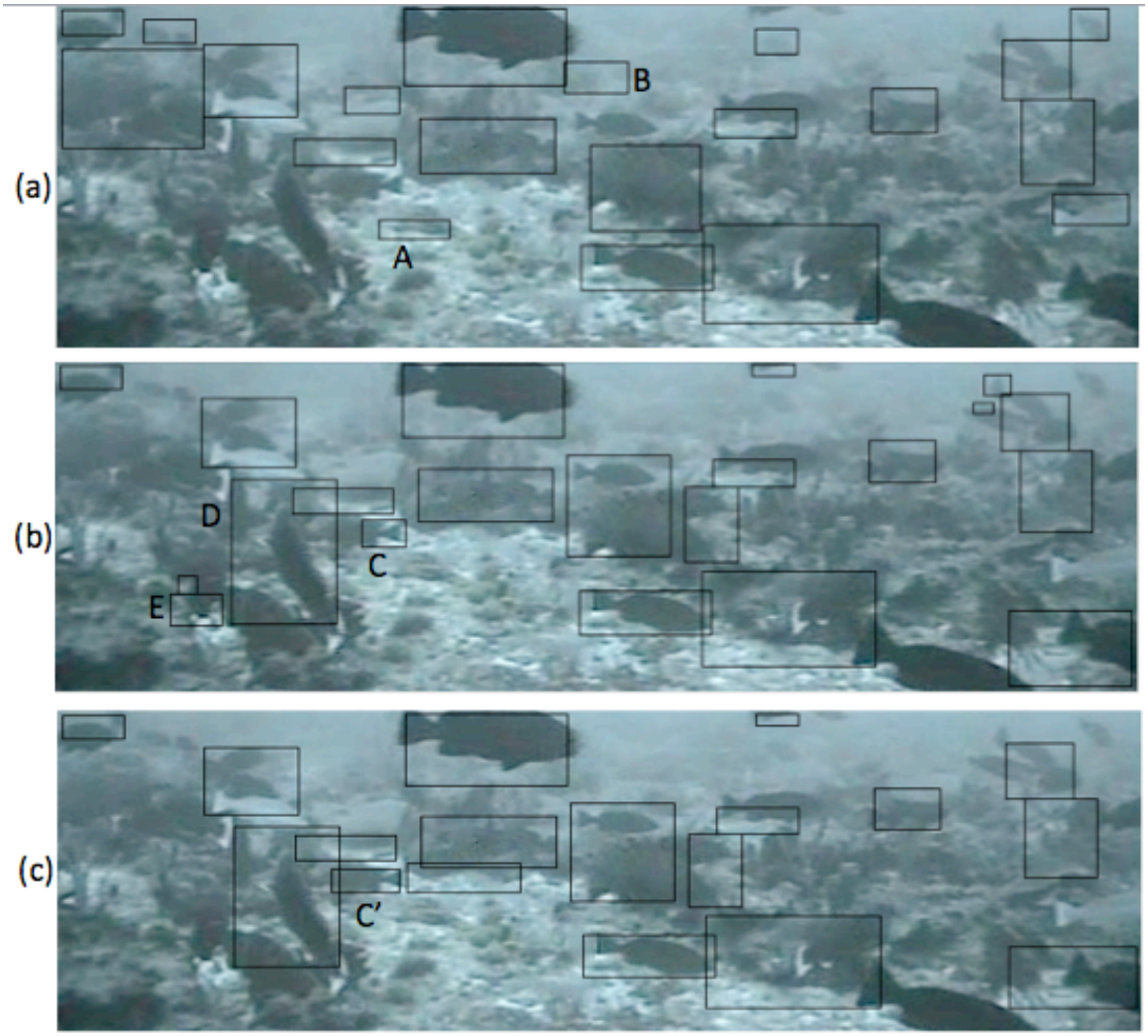

Figure 5.2: Detecting results using dense SIFT cue only (a), dense sift and optical flow cues (b), and dense SIFT and optical flow cues with region growing (c).

The tracking of fish can reduce the false positives and increase the precision. In two fish tracking examples shown in Figure 5.3, two more regions (A, B) were added and a false positive in (C) is resolved by tracking.

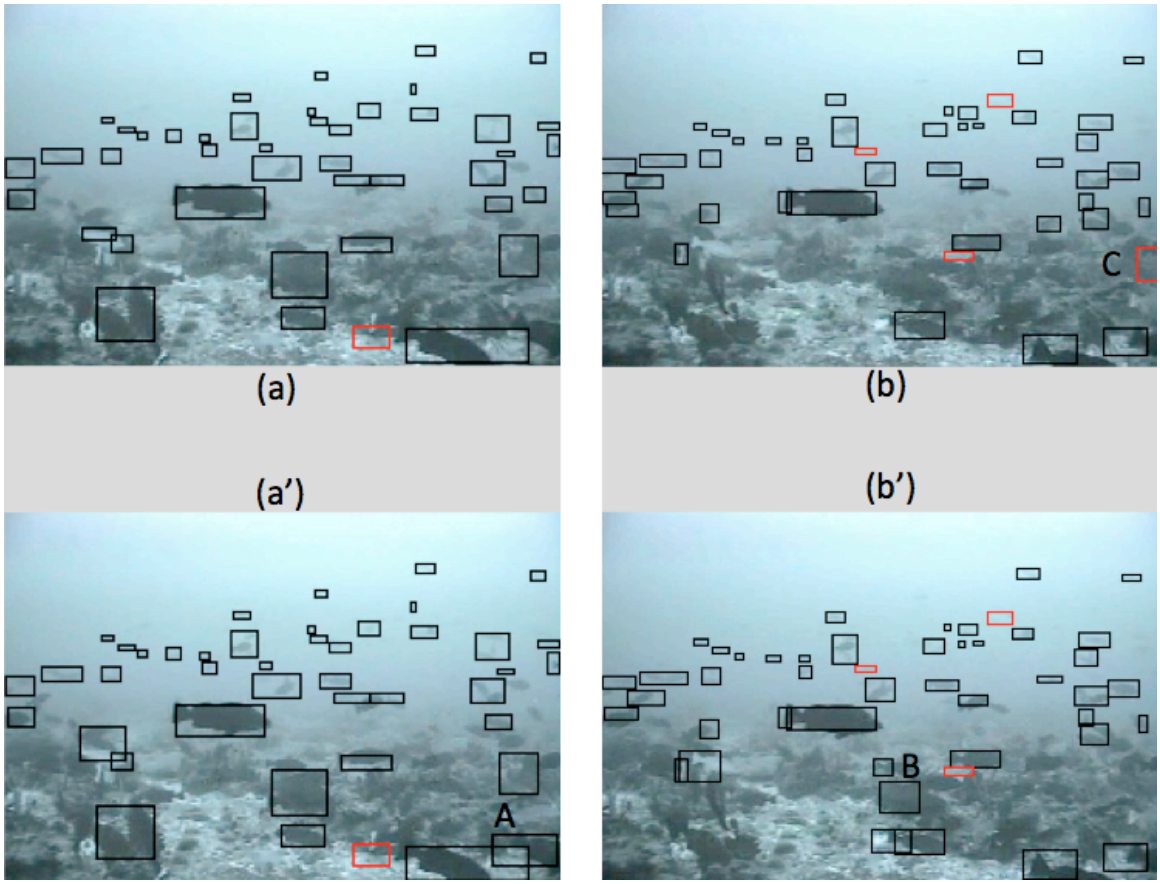

Figure 5.3: Comparing the detecting result of without tracking (a, b) and the result with tracking (a, b'). Red boxes show false positives.

Figure 5.4 depicts the segmentation results for 6 images. The number of targets, the hits and false alarms of detectors of each image is shown in Table 5.1. The precision and recall of each image is shown in Figure 5.5. The average precision of the detector is 86.4% and the average recall is 94.5%. If a minimum detectable size is set, the number of false positives within water column will reduce, so the recall rate of frame 3, 4 will improve.

Figure 5.4: Detect fish on 6 images. Red boxes show false positives.

Table 5.1: Targets, hits and false alarms

	1	2	3	4	5	6
Targets	58	57	49	52	54	50
Hits	47	49	44	47	47	42
False alarms	1	1	3	5	3	3

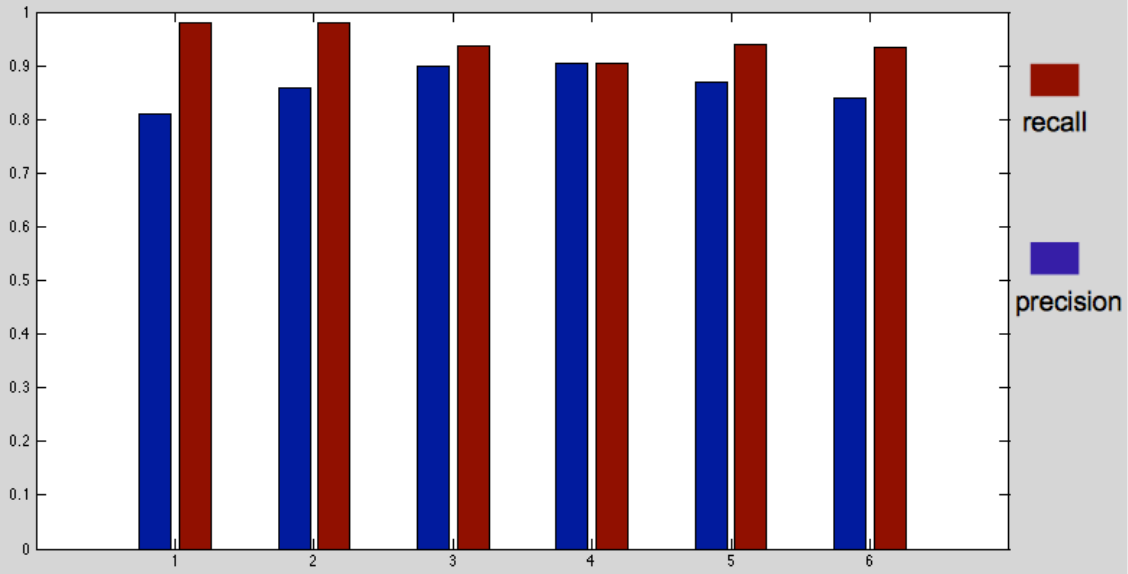

Figure 5.5: Precision and recall of each image

Based on our results and while the achieved performance cannot be assessed in absolute terms, the ability to accurately detect a large percentage of fish in each image (over 85% on the average), with very small number of false alarms, provides the marine scientists with an effective tool to process a large volume of images and video data automatically.

Chapter 6

Conclusion

Automated detection of underwater species from a large volume of underwater images or video is useful in marine studies, e.g., marine biology, fisheries, and coral reef research. However, the task can be difficult because of the complexity and variability within the benthic environments. Some key factors include the water depth and visibility, lighting condition, time of day, and certainly the image resolution.

This thesis described a fish detection system that has been developed for the automated processing of a data set, collected in the Virgin Islands in Spring 2012, during the spawning activity of the yellow-tail grouper fish. A digital camera was mounted on a remotely operated vehicle, operating at depths of 200-250 ft. The recording time is around 6:30 pm, shortly before the sunset. Thus, limited lighting has been a serious issue in adversely impacting data quality.

The goal is to explore and analyze certain visual information and their effectiveness in detecting the fish. The system was implemented using Matlab.

Within a benthic environment, some fish swimming at high enough altitudes have to be discriminated against the water column background. In contrast, other fish near the sea bottom need to be distinguished from the reef. Thus, the task is to separate the fish detection within the water column and over the seafloor because of variation in visual cues over these regions. Detecting fish within water column according to intensity is promising because their distributions are distinct, but is not so for detecting fish near the seafloor. By dividing the image, the detector utilized different visual characteristics of the water column and the seafloor to locate fish.

In our method, we start by identifying the horizon as the line between the high texture area of the seafloor and the low texture region of the water column. To implement, we identified the first peak on each row (averaged over columns) of the gradient magnitude to locate the horizon. The scene depth information, which is the byproduct of the dehazing process, is another approach to locate horizon.

The intensity characteristics of fish and the water column are so different; that the water column is typically brighter than fish, so it is used to detect fish within the

water column. The precision and recall rates for detection were 96% and 96% respectively.

The dense SIFT was used to detect fish above the seafloor since the seafloor and fish have distinct variance features: the seafloor has strong texture, while fish have weaker ones. Rather than a still image, video offers optical flow information, which assisted the detector to reduce false alarms because fish typically have different movements than the seafloor.

Fish size is estimated according to depth, it provides a criterion in selecting regions. Region growing is adopted to solve the broken fish and reduce the false positives caused by optical flow inaccuracy. We also tracks windows between frames to assure the precision of detection and reduce false positives.

In the results for 6 images, 86.4% of fish were detected, with an average recall rate of 94.5%. These suggest that the proposed system can offer the marine scientists with an effective tool to process a large volume of images and video data automatically.

Bibliography

[1] Efficient Graph-Based Image Segmentation. Pedro F. Felzenszwalb and Daniel P. Huttenlocher, International Journal of Computer Vision, Volume 59, Number 2, September 2004.

[2] Contour Detection and Hierarchical Image Segmentation. P. Arbelaez, M. Maire, C. Fowlkes and J. Malik, IEEE TPAMI, Vol. 33, No. 5, pp. 898-916, May 2011.

[3] Efficient Hierarchical Graph-Based Video Segmentation. Computer Vision and Pattern Recognition, IEEE Computer Society Conference on, Vol. 0, pp. 2141-2148, 2010.

[4] Single Image Haze Removal Using Dark Channel Prior. Kaiming He; Jian Sun; Xiaoou Tang. Computer Vision and Pattern Recognition, 2009. CVPR 2009.

[5] Field Programmable Gate Array Based Fish Detection Using Haar Classifiers. B. Benson, J. Cho, D. Goshorn, and R. Kastne, In American Academy of Underwater Science, 2009.

[6] Image Texture Tools. P. Harrison, PhD thesis, Monash University, 2005.

[7] Distinctive Image Features from Scale Invariant Features, Lowe, David G., International Journal of Computer Vision, Vol. 60, No. 2, pp. 91-110, 2004.

[8] A Threshold Selection Method from Gray-Level Histograms, Otsu, N., IEEE Transactions on Systems, Man, and Cybernetics, Vol. 9, No. 1, pp. 62-66, 1979.

[9] Improved adaptive Gaussian Mixture Model for Background Subtraction. Zivkovic Z, Proceedings of ICPR 2004, August 23-26, 2004.

[10] Detecting, Tracking and Counting Fish in Low Quality Unconstrained Underwater Videos. C. Spampinato, Y. H. Chen Burger, G. Nadarajan, and R. B. Fisher. VISAPP, pages 514-519, 2008.

[11] Recovery of Underwater Visibility and Structure by Polarization Analysis. Yoav Y. Schechner and Nir Karpel, IEEE Journal of Oceanic Engineering , Vol. 30 , No. 3 , pp. 570-587, 2005.