Autonomous fish tracking by ROV using Monocular Camera

Jun Zhou Christopher M. Clark

Lab for Autonomous and Intelligent Robotics
Department of Mechanical Engineering
University of Waterloo
200 University Ave. West, Waterloo, ON, N2L 3G1, Canada
jzhou@engmail.uwaterloo.ca, cclark@uwaterloo.ca

Abstract - This paper concerns the autonomous tracking of fish using a Remotely Operated Vehicle (ROV) equipped with a single camera. An efficient image processing algorithm is presented that enables pose estimation of a particular species of fish - a Large Mouth Bass. The algorithm uses a series of filters including the Gabor filter for texture, projection segmentation, and geometrical shape feature extraction to find the fishes distinctive dark lines that mark the body and tail. Feature based scaling then produces the position and orientation of the fish relative to the ROV. By implementing this algorithm on each frame of a series of video frames, successive relative state estimates can be obtained which are fused across time via a Kalman Filter. Video taken from a VideoRay MicroROV operating within Paradise Lake, Ontario, Canada was used to demonstrate off-line fish state estimation. In the future, this approach will be integrated within a closed-loop controller that allows the robot to autonomously follow the fish and monitor its behavior.

Keywords: tracking, monocular vision, underwater, image processing, feature extraction, ROV.

1 Introduction

As the largest unexplored area on earth, the underwater world has unlimited attraction to marine scientists. Due to the complexity of the underwater environment and the limitations of human divers, underwater exploration has been facilitated by the use of submarines, Remotely Operated Vehicles (ROVs) and Autonomous Underwater Vehicles (AUVs) [1] [3]. In many applications, target tracking is of particular interest, especially for enabling short-range applications such as fish tracking, cable following, and docking [8].

Figure 1: The VideoRay Pro III MicroROV

In this research, target tracking of fish via autonomous robots is studied with the purpose of assisting marine biologists in gathering detailed information about the behaviors, habits, mobility, and local and global distributions of particular fish species. To realize this goal, a VideoRay ProIII MicroROV (Fig. 1) is being equipped with a vision servo control system to enable fully autonomous real-time fish tracking.

This paper describes a technique to extract the relative position of the fish using a monocular camera system. Through video image processing, fish features are obtained. Within the image processing, a new technique called *ProjectionCurveSegmentation* has been developed that extracts particular features of the fish. These features are used to calculate the relative position of the fish through feature based scaling and perspective geometric projection methods. In the future, the relative range and bearing obtained with these methods will be used to control the vehicle such that the target will be centered in the image, and automated real-time tracking can be realized.

The paper is organized as follow: Section 2 provides an overview of related research. Section 3 introduces the structure of the proposed tracking sys-

tem. Section 4 presents image processing methods to identify the target fish and extract its features. Section 5 provides calculations to obtain the range and bearing of the target fish. In Section 6, quantitative results and the qualitative analysis are given. Finally, the conclusions and ideas for future work are presented in Section 7 and section 8.

2 Background

Autonomous target tracking is commonly achieved (or partly achieved) by holding some station close to the object over time. This requires knowledge of the relative position of the object with respect to the position of an ROV. Current methods to tracking with ROVs include using optical [9] , acoustic [6] , and laser sensors [1]. Laser techniques unfortunately require high power and large space. In the case of acoustic methods, it is difficult to avoid problems due to multiple path effect and acoustic shading especially in target tracking. Optical methods consume low power and consist of rich environmental information such as color, texture, shape, dynamic properties and geometric properties etc. Despite these advantages, they still have several issues to be addressed. Light attenuates exponentially with distance in water, which makes the quality of underwater images very poor. Feature extraction is complicated and can limit the possibilities for real-time implementation. Also, the vast array of unknown objects in the environment can be misinterpreted for the interested object.

In tracking fish specifically, several additional problems arise. The fish do not appear as exclusive bright against dark backgrounds. Illumination backscatters to the camera, producing a relatively bright and non-uniform background image. Suspended organic particles, known as marine snow, introduce continual small fluctuations to this background image.

Finding gradients is also difficult with fish. Due to the difference of the light reflection ratio of fish scales, the intensity is uneven and the gradient distributions are scattered on the entire body, with some areas of strong intensity and others of weak intensity. Moreover, hotspots on the camera enclosure produce a strong gradient response. Lighting geometries that can result from these bright reflections are difficult to predict in advance.

Color segmentation has success in extracting the fish from the water background, but encounters difficulty in separating the fish from seaweed and the floor.

Background Subtraction methods [9] based on a largely stable background image differences cause moving objects to stand out saliently in sequential images. In this case, this approach works poorly because the background typically changes over time when the ROV is moving, when the fish remains moderately still with respect to the ROV, or in the presence of currents.

The active contour method, such as snake method [11] fails in the various seaweeds and the very uneven intensity on the fish body. Intensity threshold routines, even adaptive ones, proved unreliable. Gradients in the background image create overlap between target and background intensity values. In these cases, no unique threshold level exists.

Region-merging methods also encounter difficulties that result from the similar seaweed and fish body. Expansive regions belonging to the background were often misclassified as target regions, and vice versa.

Nor did watershed methods give reliable results. When applied to the gradient image using bright intensity patches to form initial markers [4], different intensity gradients on the surface of fish body created multiple watersheds for the same target. Attempts to merge these watersheds encountered difficulties similar to those observed for other region-merging methods.

Several papers have touched on the topic of automated animal tracking in natural underwater environments. Jason Rife et al. tackled a robotic tracking of Gelatinous animals in the deep ocean [7]. Other workers have automated visual extraction of marine animals from a video sequence, without closing servo loops. Kocak et al. discuss vision techniques for off-line analysis of bioluminescent zooplankton data [2]. Fan and Balasuriya tested a 20 Hz fish tracking technique off-line, using video collected in the open ocean [5][7]. Other investigators have focused on pattern recognition methods useful for detecting underwater targets [10] [12].

In this paper, an image processing algorithm is presented that uses a Gabor filter followed by a new technique called $ProjectionCurveSegmentation$ to obtain the target fish's obvious features, i.e. the tail and body features. These features are extracted to estimate the relative position of the fish.

3 Control System Architecture

3.1 System structure

This paper presents an image processing based algorithm for estimating the relative position of a fish using monocular vision. The goal is to implement this algorithm into a fish tracking system controlled

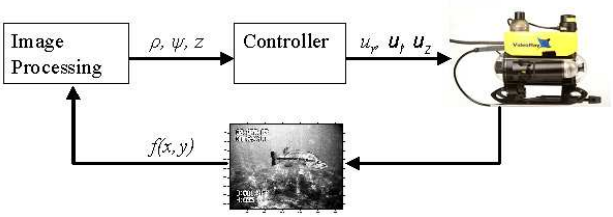

Figure 2: The visual serving control system

using Visual Servoing - the use of visual imagery to control the pose of a robot relative to a target.

To carry out this tracking, a VideoRay Pro III microROV is proposed. The control box used for tele-operation is replaced by a standard PC that interfaces with the ROV sensors and actuators to allow autonomous control. The ROV has three thrusters to actuate the vehicle, two thrusters for differential drive propulsion, and another thruster for depth control. A passive buoyancy moment stabilizes the vehicle around the pitch and roll axes. Also mounted on the ROV are a WDCC-6300 CCD color video camera, depth gauge, compass and two forward looking halogen lights. It is depth rated to 500ft and has 76 m (250 ft) of tether to provide power and control.

In the proposed method, intensity images f(x,y) are processed to extract the target's relative range ρ, bearing ψ, and height z in polar coordinates. The visual servo controller then computes control inputs u_r, u_l, u_z for the right, left and top ROV thrusters respectively, (Fig. 2). In general, this will drive the ROV to hold the target in the center of camera image and at some desired distance.

4 Image Processing Algorithm

This section describes the vision processing algorithm used to track a Large Mouth Bass in natural environments. The algorithm combines a series of existing filters commonly found in the vision literature, with a new segmentation filter called Projection Curve Segmentation, (see Fig. 3).

4.1 Image Scaling

To reduce computation, the input original color images are converted to greyscale and the pixel values are limited in the interval [0,1].

Due to underwater light limitations, images are underexposed and blurry. The poor contrast forces grey values to concentrate into a small range. To remedy this, intensities are adjusted linearly to maximize the range, and histogram equalization method

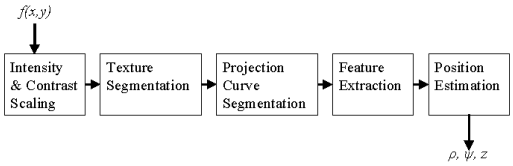

Figure 3: Image Processing Overview

is used to stretch contrast so that all grey-levels have similar likelihoods [], (see Fig. 4).

4.2 Texture segmentation by Gabor filter

Texture segmentation is the problem of breaking an image into components within which the texture is constant. In this case, the target fish's tail and body consist of obvious and regular orientation stripes. To extract these features, a single oriented Gabor filter of spatial-frequency is proposed. The method is not only effective in extracting the patterns, but is efficient since only a single texture extraction filter is required.

The Gabor filter is orientation selective. Its kernels are Fourier basis elements that are multiplied by Gaussians, meaning they respond strongly at image points where there are components that locally have a particular spatial frequency and orientation.

If $s(x, y)$ is a complex sinusoidal known as the carrier, and $w_r(x, y)$ is a 2-D Gaussian-shaped function known as the envelope, the Gabor filter is a complex function $g(x, y)$:

$$g(x, y) = s(x, y) \times \omega_r(x, y) \tag{1}$$

The sinusoidal is defined in terms of the spatial frequencies (u_0, v_0) and the carrier phase P as follows:

$$s(x, y) = \exp(j2\pi(u_0 x + u_0 y) + P) \tag{2}$$

The Gaussian envelope is defined in Eq. 3, where K scales the envelope magnitude, (a, b) scale the envelope axis, θ defines the envelope rotation angle, and (x_0, y_0) defines the peak location of the envelope.

$$\omega_r(x, y) = K\exp(pi(a^2(x - x_0)_r^2 + b^2(y - y_0)_r^2)) \tag{3}$$

Note that the subscript r represents a rotation operation such that:

$$\begin{aligned}(x - x_0) &= +(x - x_0)\cos\theta + (y - y_0)\sin\theta \\ (y - y_0) &= -(x - x_0)\sin\theta + (y - y_0)\cos\theta\end{aligned} \tag{4}$$

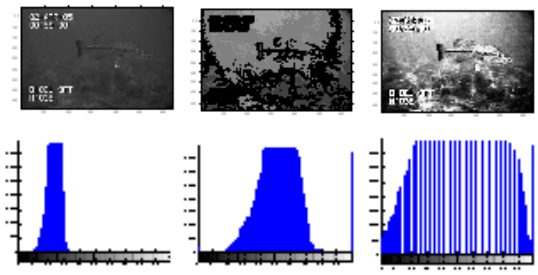

Figure 4: Images and related histograms (a) original image, (b) intensity scaled image, and (c) the contrast scaled image.

Each complex Gabor consists of two functions in quadrature (out of phase by 90 degrees), conveniently located in the real and imaginary parts of a complex function.

Now we have the complex Gabor function in space domain.

$$\begin{aligned} g(x,y) =& K\exp(-\pi(a^2(x-x_0)_r^2 + b^2(y-y_0)_r^2)) \\ & \exp(\jmath(2\pi(u_0x + v_0y) + P)). \end{aligned} \tag{5}$$

The 2-D Fourier transform of this Gabor is as follows

$$\begin{aligned} \hat{g}(u,v) =& \frac{k}{ab}\exp(\jmath(-2\pi(x_0(u-u_0) + y(v-v_0) + p)) \\ & \exp(-\pi(\frac{(u-u_0)_r^2}{a^2} + \frac{(v-v_0)_r^2}{b^2})) \end{aligned} \tag{6}$$

The Gabor filter is used as a kernel to convolve with the input image $I(x,y)$, input image to produce:

$$\begin{aligned} imagabout(x,y) &= I(x,y) \bigotimes imag(g(x,y)) \\ regabout(x,y) &= I(x,y) \bigotimes real(g(x,y)) \end{aligned} \tag{7}$$

By applying the Gabor filter, the majority of the fish and its local background are removed except for the tail and body features. This establishes a good basis for the following feature projection segmentation.

4.2.1 Projection Curve Segmentation

In this step of the vision processing, the body and tail features are extracted from the remaining background.

After the image is processed by the Gabor Filter, a threshold is applied to force pixels to take on values of 0 or 1. In observing the resulting image (see the Fig. 5 a), only the fish tail pattern, body center pattern, and some background patterns (i.e. underwater grass) remain. The fish patterns have limited overlap with the background.

Projecting the threshold image into a vertical histogram $H_v(y)$, i.e. summing the number of black pixels in each row of the image, results in two separate shapes. The first is the background curve with no defining shape. The second is a sharp and narrow spike protruding from a smooth and low curve. This second shape is a projection of the tail and body features, (Fig. 5 b).

With this histogram, a search for the tail and body patterns is conducted to produce an interval of rows in which the fish is located. If A is a predetermined threshold that characterizes the tail width, the tail interval is defined as rows belonging to $[y_{tailstart}, y_{tailstop}]$ such that a scan from the top of the image produces:

$$\begin{aligned} y_{tailstart} &= max(y|H_v(y) > A) \\ y_{tailstop} &= max(y|H_v(y) < A, y < y_{tailstart}) \end{aligned} \tag{8}$$

The peak within this interval is determined by:

$$y_{max} = max(y|y \in [y_{tailstart}, y_{tailstop}]) \tag{9}$$

If the slope of the histogram within intervals $[y_{max} - \delta, y_{max}]$ and $[y_{max}, y_{max} + \delta]$ have magnitudes less than m_{min}, it is determined that the fish tail feature is found.

If the slope conditions are satisfied, rows outside the interval $[y_{tailstart}, y_{tailstop}]$ are subtracted from the image, effectively eliminating background in the top and bottom portions of the image, (see Fig. 5 c).

In a similar fashion, the image is projected into a horizontal histogram $H_h(x)$, i.e. summing the number of black pixels in each column of the image. The tail pattern dominates the histogram with an obvious spike. The body pattern is also evident as a region of constant amplitude adjacent to the tail spike. In this case, a search for these two features is conducted to define an interval of columns in which the fish resides. Columns outside this interval are subtracted to remove background on the two sides of the fish, (see Fig. 5d). What remains is an image with only the tail and body features.

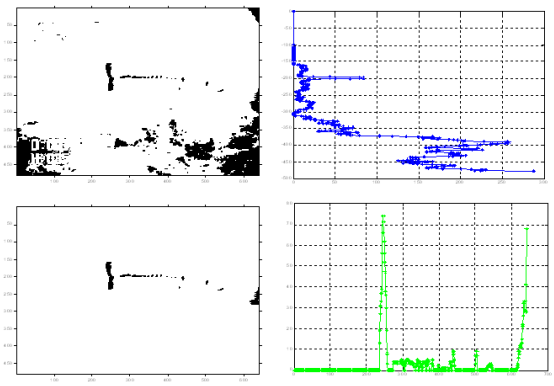

Figure 5: Curve Segmentation: (a) Image after Gabor filter and threshold, (b) the vertical projection curve, (c) the image after subtracting top and bottom background, and (d) the horizontal projection curve.

4.3 Feature Extraction

In this stage of the image processing, the remaining black pixels of the image are modelled with two lines, one representing the tail feature and one representing the body feature. These two lines are later used to describe the position and orientation of the fish.

The leftmost and the right-most pixels $(U_1(i), V_1(i))$, $(U_2(i), V_2(i))$ are determined for each row in the tail interval. The central points $(U_0(i), V_0(i))$ of the tail are defined as follows:

$$\begin{aligned} U_0(i) &= (U_1(i) + U_2(i))/2 \\ V_0(i) &= (V_1(i) + V_2(i))/2 \end{aligned} \tag{10}$$

A least squares linear regression is then used to fit a straight line to the tail. A similar process is used to find the body's central line. These two lines are used to extract the position of the fish as discussed in the next section.

5 Position Estimation From Monocular Camera

Given the position of the fish features within a video image, the position of the fish relative to the ROV can be obtained. With the relative coordinate system shown in Fig. 6, it is assumed that the three axis of the camera coordinate frame coincide with the ROV body-fixed frame. After transforming this to polar coordinates, feature based scaling is used to produce a relative range measurement based on some predetermined length scale. Specifically, statistical data of the target fish size is used to relate

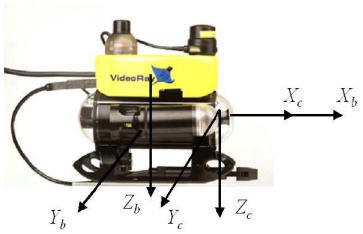

Figure 6: ROV body and camera coordinate frames.

a relative length l of the fish tail (or body) line in pixels, to predetermined length L in meters.

If the camera has been calibrated, then the depth D of a line (e.g. the fish tail) relative to the camera is calculated with:

$$D = \frac{kf}{l} \times L \tag{11}$$

In Eq. 11, the real focus is f (mm), k is the scaling factor that transforms f into the image plane, and D is the distance from the fish plane to the camera.

The accuracy of both the tail length and body features will suffer from varying light intensity, the tail swaying, and the body deforming. To help remedy such disturbances, the depth estimation information provided by the two features are combined by a simple Kalman filter that weights the fusion based on variance. If D_t is the depth calculated by the length of tail with variance σ_t^2, and D_b is the depth calculated by the width of body line of fish with variance σ_b^2 , then D_k is the optimal depth calculated by the Kalman filter equations.

$$D_k = D_t + K(D_b - D_t) \tag{12}$$

$$K = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_b^2}; \tag{13}$$

The position estimation is calculated by perspective geometry projection relation. In Fig. 7, a point on the target is described by coordinate (P_x, P_y, P_z). The position of this point's light ray on the camera's image is defined in camera coordinates as (p_u, p_v). The range to camera is ρ, the yaw bearing is ψ, and the relative depth to camera Z can then be calculated with:

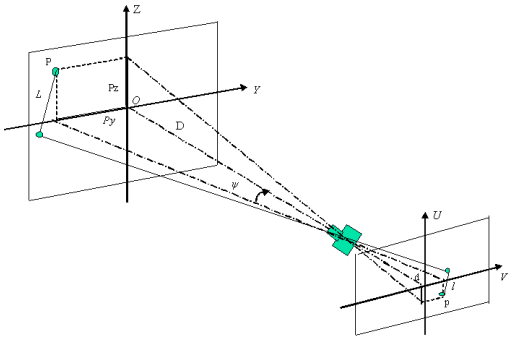

Figure 7: Camera geometry projection

$$
\begin{aligned}
P_x &= D_k \\
P_y &= \frac{p_v}{kf} \times D_k \\
\rho &= \sqrt{P_x^2 + P_y^2} \\
\psi &= \arctan(\frac{p_u}{kf}) \\
Z = P_z &= \frac{p_u}{kf} \times D_k
\end{aligned}
\tag{14}
$$

Errors in the feature-based range estimate result from difficulties in precisely determining target lengths within images, and from the uncertain size of the target. Assuming the covariances of the image plane measurements are σ_u^2, σ_v^2 , σ_l^2 , σ_s^2 respectively, and the covariance of the size of fish is σ_L^2, then we can compute their relative variance $\Sigma\psi$, ΣD, ΣZ, and $\Sigma\rho$ according to the error propagation law:

$$\Sigma f = \nabla f C_f \nabla f^T \tag{15}$$

6 Results And Analysis

Video data images of a Large Mouth Bass were acquired using the WDCC-6300 CCD camera installed on human driven VideoRay ROV. Images were of dimensions 480x640, and were grabbed at a frame rate 20Hz.

6.1 Image Processing

The image processing algorithm was applied to each frame of each sequence. The series of filters including texture, projection curve segmentation, geometrical shape feature extraction proved simple, efficient and effective if several conditions were met. These conditions included that the body side face toward the camera, the tail is clearly visible, and the fish swims some minimal distance above the underwater grass.

An example of a typical image being processed is shown in Fig. 8. In Fig. 9, the fish motion represented by the geometrical feature in the 9 sequential images taken from the ROV. The results indicate that the image segmentation and feature extraction method provide sufficient relative pose estimates for fish tracking.

6.2 Position estimation Results and Analysis

The relative position between the target fish and the ROV is calculated for ten successive images taken across a time span of 2 seconds. Results are displayed in Fig. 10. At present, there is no truth data for comparison. Error results are based on the theoretical calculations of propagation error. Seen from Fig. 10 (c1), the trend of yaw in the ten images match that of the fish shown in Fig. 9. Because the distance of camera lens to image plan can be gained from camera system calibration, its error in the system can be eliminated. Hence the accuracy in yaw is only affected by the error in measuring the distance between the target point to the origin in the image coordinates. Since the propagation error is small. it is expected that the yaw will have higher accuracy.

Fig. 10 (b1) shows the calculated depth or relative position of fish along the Z_b axis. Compared with Fig. 9, the trends coincide. However, because the size of this species of adult fish is undetermined and can only be obtained from statistical data, higher error in the depth estimation occurs. This will affect the accuracy in estimating relative vertical and range positions.

For example, assume that the length of the fish's tail is 9cm with σ^2=0.5cm^2 , and the width of fish body central pattern is 0.7cm with σ^2=0.05cm^2. When the range position calculated from the image has maximal value 0.95m, the propagation error is 0.2m. When the relative depth calculated has maximal value 0.1m, the propagation error is 0.02m. When implementing this within the proposed fish tracking system, these errors should be acceptable.

Fig. 11 shows the relative depth estimation results and the corresponding propagation errors from using 1) feature scaling the fish tail, 2) feature scaling of the body line, and 3) fusion of the two previous results via the Kalman filter. While the monocular vision system presented does have inherent difficulties in predicting depth, the fusion of depth measurements obtained from both features aids in decreasing errors.

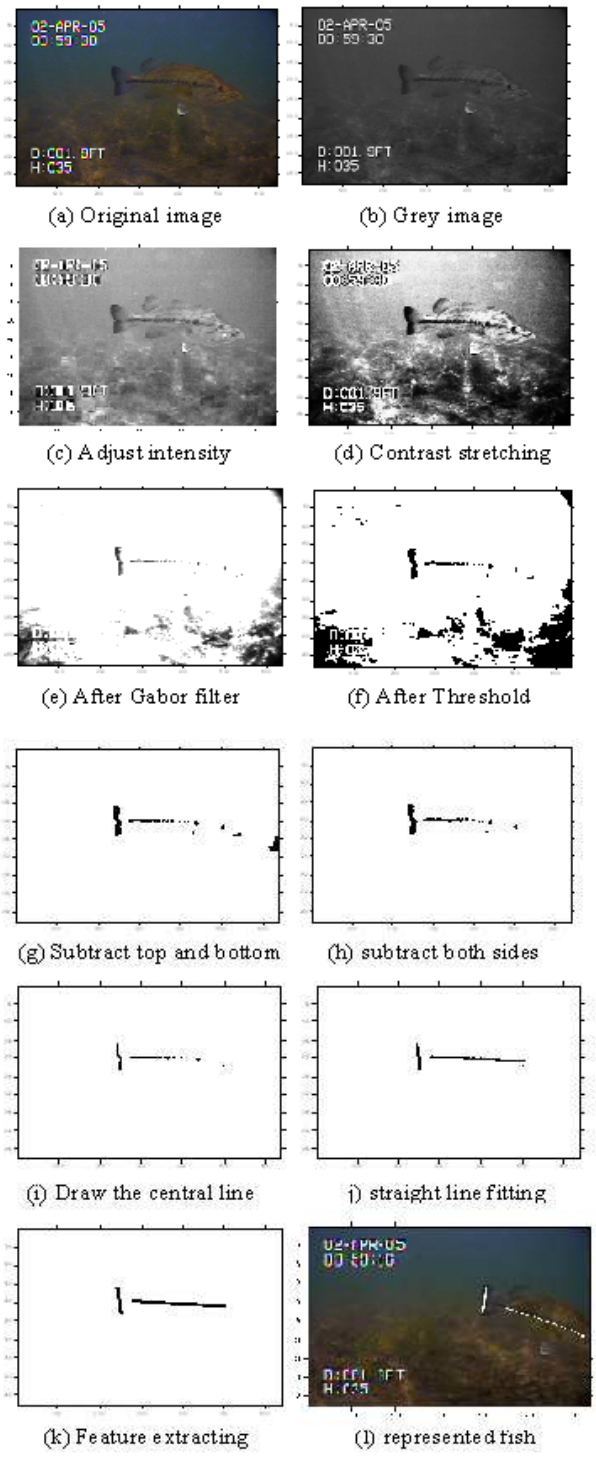

Figure 8: Example results of the image processing algorithm

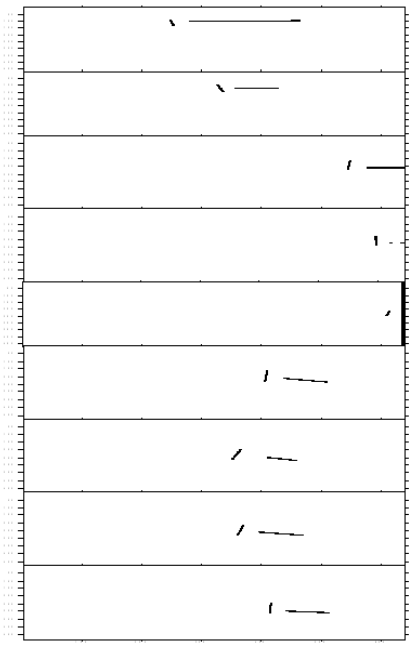

Figure 9: The feature extracted shows the motion of the fish in ten successive images of the fish taken by ROV in the lake in Waterloo.

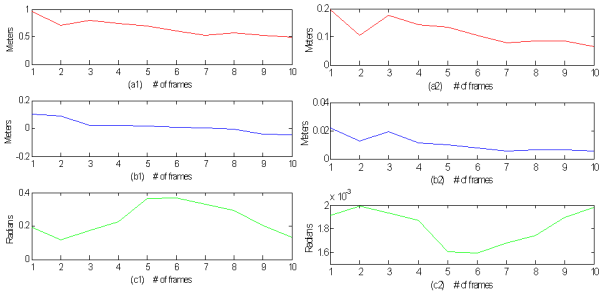

Figure 10: Relative position of fish across ten successive frames. (a1) Range position(ρ) , (b1) Vertical position (Z_b), (c1) Yaw position (ψ). The right diagrams (a2) (b2) (c2) depict the propagation error of three axes respectively.

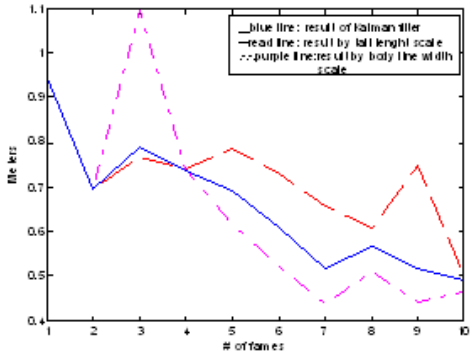

Figure 11: Results of depth estimation using the length and width scaling method, and fused with a Kalman filter;

7 Conclusions

This paper describes a system for automated fish tracking by an ROV using visual servoing control. The core of the paper focuses on an efficient image processing algorithm used to extract the relative position of a fish, i.e. a Large Mouth Bass. The algorithm uses a Gabor filter to extract texture, a new filter called projection curve segmentation to remove background, and a linear regression based feature extraction method.

To validate the algorithm, offline image processing was conducted on video footage obtained by piloting the ROV around Paradise Lake, Ontario, Canada. While the uncertainty in fish size and feature lengths decreased the accuracy of relative range estimation, it is expected that the errors will be small enough to allow fish tracking via visual servoing control.

Despite the success in tracking fish over several images, the algorithm has several limitations. First, it is assumed that only one fish be present in each frame. Second, it is assumed that the fish swim perpendicular to the camera lens. Lastly, the fish cannot be occluded (e.g. by seaweed).

8 Future Work

As for future work, fusing the relative fish position obtained from monocular vision with high-resolution imaging sonar data is already under investigation. Further improvements are also necessary for image processing, including more robust algorithm to permit better recognition and false positive detection, increasing the accuracy of the feature to improve the precise of range, and ensuring fast processing - a requirement for real-time processing in natural underwater environment.

References

[1] R.C.Michelson A. L. Meyroqitz, D.R. Blidberg. Autonomous vehicles. In *Proceedings of the IEEE*, Bol.84,No.8,August 1996.

[2] E. Widder D. Kocak, N. da Vitoria Lobo. Computer vision techniques for quantifying, tracking, and identifying bioluminescent plankton. *IEEE Journal of Oceanic Engineering, 24*, 1999.

[3] S.G.Chappell D.R. Blidberg, R. M. Tumer. Autonomous underwater vehicles: Current activities and research opportunities. *Robotics and Autonomous Systems*.

[4] E. Dougherty (ed.). Mathematical morphology in image processing. marcel dekker. 1992.

[5] Y. Fan and A. Balasuriya. Autonomous target tracking by auvs using dynamic vision. In *Proc. of the 2000 International Symposium on Underwater Technology*, pages 197–192, 2000.

[6] J.A.Catipovic. Performance limitations in underwater acoustic telemetry. *IEEE, Journal of Oceanic Engineering*, July,1990.

[7] T. Asakura M. Minami, J. Agbanhan. Manipulator visual servoing and tracking of fish using a genetic algorithm. *Industrial Robot*, 1999.

[8] X. Cufi R. Garcia and M. Carreras. Estimating the motion of an underwater robot from a monocular image sequence. In *in IEEE/RSJ Int. Conf. on Intelligent Robots and Systems IROS '01*, volume 3, pages 1682–1687, Maui,Hawaii, USA,2001.

[9] J. Rife. Automated robotic tracking of gelatinous animals in the deep ocean. *PhD thesis, Stanford University, Stanford,California,*, December,2003.

[10] X. Tang and W. K. Stewart. Plankton image classification using novel parallel-training learning vector quantization network. In *Proc. IEEE/MTS OCEANS '96*, volume 3, pages 1227–1236, 1996.

[11] E. Trucco and A. Verri. Introductory techniques for 3-d computer vision. In *Prentice Hall*, pages 108–112, 1998.

[12] J. Chen R. Chen P. Liu X. Yuan, Z. Hu. Online learning and object recognition for auv optical vision. In *IEEE Int. Conf. on Systems, Man, and Cybernetics, 1999*, volume 6, pages 857–862, 1999.