Segmentation methods for visual tracking of deep-ocean jellyfish using a conventional camera

Jason H. Rife, Tufts University & S.M. Rock, Stanford University

>> Accepted Article <<

CITATION:
J. Rife and S.M. Rock (2003). Segmentation methods for visual tracking of deep ocean jellyfish using a conventional camera. IEEE Journal of Oceanic Engineering, 28(4):595-608. DOI: 10.1109/JOE.2003.819315.

CORRESPONDING AUTHOR:
Jason Rife
jason.rife@tufts.edu

COPYRIGHT:
© 2003 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

FINANCIAL SUPPORT:
Monterey Bay Aquarium Research Institute (MBARI), Moss Landing, CA, and Packard Foundation Grants 98–3816 and 98–6228

Segmentation Methods for Visual Tracking of Deep Ocean Jellyfish Using a Conventional Camera

Jason Rife, *Student Member, IEEE,* and Stephen M. Rock, *Member, IEEE*

{jrife, rock}@arl.stanford.edu
MC 4035, Durand 028b
Stanford, CA 94305
(650) 723-3343

***Abstract*—This paper presents a vision algorithm that enables automated jellyfish tracking using ROVs or AUVs. Discussion focuses on algorithm design. Introduction of a novel performance assessment tool, called segmentation efficiency, aids in matching potential vision algorithms to the jelly-tracking task. This general-purpose tool evaluates the inherent applicability of various algorithms to particular tracking applications. This tool is applied to the problem of tracking transparent jellyfish under uneven, time-varying illumination in particle-filled scenes. The result is the selection of a fixed-gradient, threshold-based vision algorithm. This approach, implemented as part of a pilot aid for MBARI's ROV *Ventana*, has demonstrated automated jelly tracking for as long as 89 minutes.**

***Index Terms*—AUV, ROV, jellyfish, segmentation, visual tracking in natural scenes, performance evaluation**

I. Introduction

A. Jelly Tracking: A Visual Servoing Application

The visual jelly-tracking application falls within the broad research area known as position-based visual servoing. The term *visual servoing* implies the use of video as a sensor for automatic control. In many cases, including the jelly-tracking application, visual servoing requires the use of a *visual tracking* algorithm. Visual tracking algorithms are designed to follow projected objects through a 2D video sequence, without any implication of closed-loop motion control of the imaging platform. Visual tracking algorithms implicitly or explicitly address two related imaging problems: segmentation and recognition. The segmentation process clusters pixels into regions that may correspond to the tracked object, while the recognition process distinguishes among these regions to identify the best match to a target profile. By segmenting an image and recognizing the target region, a visual tracking algorithm measures target location. Based on this measurement, a visual servoing algorithm issues a feedback control signal to the imaging platform, as illustrated by Figure 1.

The field of visual servoing has spawned numerous applications. Recent publications that capture the breadth and history of the visual servoing field are [1,2]. Although these reviews of visual servoing make little mention of underwater applications, the ocean community has made substantial progress in the visual navigation of submersible vehicles relative to the ocean floor [3-14].

Each instance of visual servoing uses a different visual tracking strategy suited to the nature of the application. For example, Leahy et al. enabled visual servoing for aircraft refueling by placing easily identified white markers near the fuel port [15].

Amidi et al. report helicopter experiments that identified a ground target using color segmentation or template-based detection [16]. Batavia et al. detected overtaking vehicles in a car's blind spot by propagating an edge map of the background scene and comparing this prediction to the current measurement [17]. Minami et al. used a triangular template strategy to track a fish in a tank using a robotic manipulator [18].

The rich variety of visual tracking methods employed by each of these cases suggests that the selection of a reliable visual tracking algorithm for a new application is nontrivial. In fact, for most visual servoing applications, several tracking algorithms produce viable solutions (of varying quality). This freedom in algorithm choice introduces an important design question, central to this paper. The designer of a visual servoing system must somehow assess the match between tracking algorithms and the visual environment characteristic to an application.

This paper discusses a method for synthesizing a robust and efficient vision strategy for endurance tracking of a single gelatinous animal. To date, no attempt has been made to implement such an experimental visual servoing system, despite the opportunity such a platform offers to extend the science of marine ecology. The lack of published data regarding visual tracking of gelatinous animals, along with the differences between the imaging environments for this application and for other terrestrial, aerial, and marine visual servoing applications, motivates a thorough characterization of the deep ocean imaging environment. Midwater images depict natural, unprepared scenes. Such scenes do not contain manmade features, like corners or straight lines, nor can an observer artificially augment the scene without disturbing the animal behaviors under study. In the absence of such features, the jelly-tracking system must detect a range of

animal specimens with flexible bodies. Transparency, evolved as a defensive adaptation against predation [19], further enhances the difficulty of localizing the jelly target.

These issues somewhat resemble problems encountered by Tang, Fan, Kocak and others in their pursuit of automated systems for the visual detection and classification of marine plankton [20-26]. Nonetheless, the jelly-tracking problem possesses additional characteristics that distinguish it. Light source geometry for remotely operated vehicles (ROVs) changes dramatically from dive to dive. On any given dive, spatial lighting gradients are visible, in addition to temporal derivatives resulting from pan/tilt motion, light source oscillations, and variations in concentration of suspended organic matter, known as marine snow. The automated jelly-tracking system must function despite these noise sources.

Operational constraints further shape the choice of vision algorithm for the jelly tracking application. Both as a pilot aid for ROVs and as a functional component for fully autonomous underwater vehicles (AUVs), the jelly tracker's first priority is reliability. Mean-time-to-failure for the vision system must match the application duration, measured in hours for ROV deployments and in days for AUV deployments. In both cases, it is desirable that the system display low sensitivity to prior selection of algorithm parameters. Also, for endurance AUV operations, limited onboard energy storage restricts sensor power. These energy budget limitations necessitate a strobed vision system [27] with low computational complexity.

This paper develops a quantitative tool that enables an efficient design process for synthesizing a visual servoing algorithm. This approach, called segmentation efficiency, is then applied to the jelly-tracking application. The selected method, an efficient

threshold-based tracking algorithm applied to a gradient prefiltered video stream, was tested successfully in the open ocean using MBARI ROV *Ventana*.

B. Segmentation Efficiency: A Performance Predictive Method

The critical step in synthesizing the jelly tracking sensor involves the assessment of the deep ocean imaging environment in the context of available tracking algorithms. Tools for performance evaluation of computer vision algorithms have arisen in recent research [28-30]. Performance evaluation tools address a wide range of vision issues, including the tracking problem and its subcomponents, the segmentation and recognition problems. Segmentation performance, involving the identification of pixel sets associated with potential targets, dominates the jelly-tracking design problem. Recognition performance, involving correspondence of segments through time, bears a close relationship to segmentation performance, since information that amplifies signal-to-noise for segmentation tends also to amplify pattern-based recognition. A large number of tracking algorithms, furthermore, achieve correspondence based solely on segment position, referenced to a position estimate propagated from prior measurements. Because effective recognition relies so heavily on effective segmentation, performance evaluation for the segmentation component effectively predicts, to a large degree, the performance of the complete tracking algorithm.

This paper introduces a predictive assessment method that differs from existing assessment methods for segmentation and tracking [31-33]. The new method shares in common with other methods a requirement for a segmentation ground truth, as

determined by a human operator or a reference algorithm. Whereas existing assessment tools compare the output of vision algorithms to the ground truth, the new method uses the ground truth to identify the information content in the original image upon which vision algorithms act. This new method addresses feasibility issues associated with the implementation of existing assessment techniques.

From a designer's point of view, implementing existing assessment techniques requires substantial effort. First, to evaluate a number of tracking algorithms, the designer must implement all of the possibilities, often a time-consuming procedure. Second, the engineer must consider a wide range of possible image filters that may enhance signal-to-noise ratio for the application specific visual environment. Third, the engineer must ground truth image sequences in order to apply the metric. The resulting design procedure is combinatorially large. For the case of P prefilters, Q vision algorithms, and M image sequences, each sequence containing N frames, the procedure requires that the designer implement a total of Q algorithms and ground truth a total of $M \bullet N$ frames. The assessment procedure must then analyze $M \bullet N \bullet P \bullet Q$ image frames. The resulting combinatorial explosion is depicted by Figure 2a.

This paper introduces an alternative approach for algorithm performance prediction. The approach contrasts with other performance evaluation methods in that it focuses on the input to vision algorithms rather than the output. This predictive approach studies images specific to an application and identifies image information that enhances target detection. The predictive approach in turn suggests which class of vision algorithm best exploits this information. Figure 2 contrasts the predictive input-centered approach (2b) with the output-focused approach (2a). The predictive assessment procedure does

not require implementation of specific tracking algorithms. Because the procedure focuses on the segmentation and not the recognition component of tracking, the approach requires that only a single image from each video sequence be ground truthed, rather than the entire sequence. Thus, where output-focused approaches require implementation of Q tracking algorithms and ground truthing of $M \bullet N$ frames, the input-focused approach requires implementation of no algorithms and ground truthing of M frames. The number of frames analyzed by the assessment method likewise drops from $M \bullet N \bullet P \bullet Q$ for the output focused approach to $M \bullet P$ for the input focused approach.

II. Linking Image Information to Segmentation Algorithms

A new performance assessment metric aids design of tracking algorithms. The assessment technique is applied to jelly tracking to identify the group of segmentation algorithms that best extract information given the midwater imaging environment. For jelly tracking, this group consists of region-based segmentation methods applied to images prefiltered to extract gradient or background-difference information.

The new quantitative assessment tool, called segmentation efficiency, predicts the performance of segmentation algorithms by analyzing the information content of application-oriented video clips. Information content is assessed by applying pixel-level filters to an image. The connection between filtered information and segmentation algorithms lies in the observation that different segmentation algorithms exploit different geometric structures relating image pixels. The segmentation efficiency procedure

quantifies this relationship by computing filtered image statistics over geometric image regions specific to particular segmentation algorithms.

The formal definition of segmentation efficiency (Section II.C) follows the definition of image filters (Section II.A) and of geometric image regions derived from ground truth (Section II.B). Subsequently, Section II.D describes an ensemble averaging process that enables application of the segmentation efficiency tool to a database of video clips. Section II.E applies this tool to a database of gelatinous animals filmed *in situ.*

A. Filters as Image Information Descriptors

Image filters extract or emphasize particular information components in the video stream. In this sense, filtered images can be considered as descriptors of image information content. This section describes fifteen filters chosen to form a partial basis for the space of pixel-level information. Table 1 lists the fifteen filters and their formulae. Since these filters are commonly described in introductory imaging texts such as [34], this section offers only a cursory outline of the notation used in Table 1.

The expression $f^k(x,y)$ denotes the value of an image at pixel location (x, y) for the k^{th} frame of a video sequence. The term f is used here to refer generically to one of the filtered images, all defined on the domain D_f of 320 x 240 images. The pixel argument and k-superscript are dropped when identical for filter input and output. Base images are 24-bit color frames consisting of three color components, c_r, c_g, and c_b.

The notation of Table 1 includes the $\Delta f / \Delta x$ and $\Delta f / \Delta y$ operators, which denote the spatial central difference for approximation of the first derivative. The ** operator

indicates a two-dimensional image convolution. For smoothing operations, the convolution kernel, h, was chosen to be the 3x3 uniform kernel. The morphological operators for erosion ($\ominus$) and dilation ($\oplus$) are defined as follows:

$$(f \oplus q)(x,y) = \max\left\{ f(x+m, y+n) \middle| (x+m, y+n) \in D_f \text{ and } (m,n) \in D_q \right\}$$
$$(f \ominus q)(x,y) = \min\left\{ f(x+m, y+n) \middle| (x+m, y+n) \in D_f \text{ and } (m,n) \in D_q \right\} \quad (1)$$

Effectively the dilation operator enlarges bright regions of a grayscale image, while the erosion operator shrinks them. These morphological filters require a structuring element, q, with domain D_q. For this work, structuring elements were chosen with a domain of 3 x 3 pixels. Erosion and dilation operations are used to compute the morphological gradient and to create the snowless luminance image. The term snowless refers to the result of applying the well known morphological opening-closing filter, which removes small speckles, like marine snow, from an image. The opening operation ($\circ$), which removes small islands and peninsulas of image brightness, and the closing operation ($\bullet$), which fills in small dark holes surrounded by bright pixels, are defined as:

$$(f \circ q) = ((f \ominus q) \oplus q)$$
$$(f \bullet q) = ((f \oplus q) \ominus q) \quad (2)$$

Given these definitions, the snowless image, l_{oc}, is described by the following equation:

(3)

$$l_{OC} = (l \circ q) \bullet q$$

Optical flow velocities, *u* and *v*, were computed using the conservation equation for local luminance and the least squares constraint. The technique solved the following equation in a least squares sense over a 5x5 support region.

$$\left[\left(\frac{\Delta f}{\Delta x} **h\right) \quad \left(\frac{\Delta f}{\Delta y} **h\right)\right]\begin{bmatrix} u \\ v \end{bmatrix} = -d_t \tag{4}$$

The **arctan2** operation, which extracts the direction of the optical flow vector, is the arctangent operator defined with the full four quadrant range, -π to π.

The background difference filter, d_b, requires that an external agent first describe a bounding box around the target. Pixels in the box-shaped hole are interpolated to form an estimated background luminance image, $\hat{l}_b$. Although cubic spline interpolation is often implemented for hole filling, this work used the orthogonal basis set resulting from the solution of the heat equation [35] given the boundary condition of the image pixel values surrounding the hole. Interpolation results comparable with those of a cubic spline fit were obtained over the rectangular hole using a separation of variables approximate solution to the heat equation. The approximation employs the first two series expansion terms superposed with a bilinear function. This yields a solution that requires only 12 coefficients compared with 16 for the cubic spline and, because of the orthogonality of its terms, avoids computation of a pseudoinverse.

B. Geometric Structures Assumed by Segmentation Algorithms

The segmentation efficiency approach attempts to relate image prefilters to classes of segmentation algorithm that best exploit image information content. The approach recognizes that the geometric pattern of pixels amplified by a prefilter varies with filter choice. Similarly, the geometric pattern of image information expected by a segmentation algorithm varies with algorithm choice. Segmentation efficiency uses these geometric patterns to link particular prefilters to particular segmentation algorithms.

To be more specific, segmentation algorithms for object-tracking applications make two fundamental assumptions. Each segmentation algorithm employs: (1) a particular spatial structure for relating target pixels and (2) a decision criterion for classifying pixels as members of such a spatial structure. Prefilters enhance the performance of a given segmentation algorithm if they augment the classification of pixels belonging to the correct geometric structure.

Table 2 groups selected segmentation techniques into three classes based on their assumptions regarding spatial structure of image information. These classes distinguish among region-based, edge-based and hybrid methods. Region-based strategies assume pixel values are related within a segment but distinct between neighboring segments. Edge-based segmentation strategies identify the boundaries around a target segment. A third, hybrid set of strategies identifies segments using both pixel similarity over two-dimensional interior regions and pixel differences along one-dimensional boundaries.

Region-based methods include well known segmentation techniques like (1) the expectation maximization technique, that clusters pixels under an arbitrary number of

parameterized ellipses; (2) the template masking technique, that scales and aligns a template to maximize pixel contrast interior and exterior to the template; (3) threshold techniques, that cluster neighboring pixels above or below a selected threshold; and (4) the correlation technique, that assesses correspondence between a reference image and the current video frame. Correlation algorithms are a special case in that they perform well not only when the region-based signal is strong, but also when the target region exhibits strong local gradients or complexity.

Table 2 lists three edge-based segmentation methods including (1) active contour techniques, that solve a dynamic equation for the target boundary based on a forcing function derived from an edge image; (2) convex edge merging methods, that group edges based on the assumption of a convex target; and (3) Hough transform methods, that extract boundaries from an image given a parameterized contour shape. Finally, Table 2 lists two hybrid segmentation methods, which combine the decision criteria for both edge-based and region-based algorithms. These hybrid techniques include (1) region-merging methods, that join neighboring pixels of similar value, and (2) watershed methods, that filter out internal edges by joining neighboring pixels sharing a common level-set boundary.

By exploring the spatial relationship of pixels extracted by a given filter, the segmentation efficiency approach can match filters to the class of segmentation algorithm that best exploits the information embedded in the filtered image. To assess spatial organization of image information, the segmentation efficiency approach requires that an external agent supply segmentation ground truth. The external agent, which may be a reference segmentation algorithm or a human operator, distinguishes between sets of

pixels that belong to a target, g_t, to the background, g_b, or to a region excluded for the purposes of analysis, g_x. From this ground truth, the target boundary, ∂g_t, can be defined as the pixels interior to g_t that intersect with the dilation of g_b: $g_t \cap (g_b \oplus q)$. Applying a set difference between the target region and the target boundary defines the target interior, $g_t^\circ = g_t \setminus \partial g_t$. The background boundary, ∂g_b, and the background interior, g_b°, are defined similarly to their counterparts for the target region.

Segmentation efficiency relates statistics computed over these geometric regions to the three geometrically defined classes of segmentation algorithm. Comparison of the target and background regions, g_t and g_b, for instance, enables performance assessment for region-based segmentation strategies. Comparison of the interior and exterior edge regions, ∂g_t and ∂g_b, enables performance assessment for edge-based segmentation methods. As some filters cause migration of edge information, comparisons of interior regions to boundary regions, g_t° to ∂g_t or g_b° to ∂g_b, also predict the effectiveness of edge-based and hybrid methods.

C. Segmentation Efficiency

Segmentation efficiency can now be defined, based on the definitions of image filters, f, and image regions, g. The segmentation efficiency approach uses cumulative distribution functions (CDFs) to compute the effectiveness of filters in distinguishing between pairs of image regions. The resulting distribution assigns a value, between zero and one in magnitude, to each possible classifier, δ, in the range space of the filter, f.

In effect, segmentation efficiency plays a similar role to the image histogram, one of the primary tools used for segmentation analysis. Classically, researchers have used bimodal histograms to establish thresholds between cleanly separated peaks associated with a pair of image regions. As early as 1972, Chow and Kaneko derived the optimal segmentation threshold given a bimodal histogram and the assumption of equal weight for misclassification of pixels from either region [36]. The importance of pixels in the two segments is not always equal. A two-objective optimization surface called the receiver operating characteristic (ROC) curve is often employed to express the trade-offs involved with differentially weighted misclassifications [29,30,37].

In segmentation analysis, the relative size of the two segments strongly influences the selection of differential misclassification weights. On their own, image histograms do not account for region size. Regions containing few pixels appear as small bumps on a global histogram, indistinguishable from local maxima associated with histogram noise. This area normalization problem commonly arises when the number of pixels in the background segment far exceeds the number of pixels in a target segment. For this reason, segmentation efficiency makes the assumption that the relative misclassification weights should be assigned such that the number of pixels misclassified in each segment is normalized by the area of that segment.

The segmentation efficiency approach also assumes that image histograms appear "noisy." In many image histograms, correlations between neighboring pixel values create local maxima and minima superposed on the major histogram peak for each image region. The histogram approach often breaks down because of these local extrema.

To enable area normalization and noise rejection, the segmentation efficiency method relies on the calculation of cumulative distribution functions (CDFs) over two segments of interest. Starting from a histogram for each ground truthed segment, this procedure first normalizes the histogram to a probability density function (PDF), addressing, in the process, the area-weighted normalization issue. Second, the procedure integrates the PDF to form a CDF, thereby automatically smoothing local maxima and minima. The following equation describes the generation of a CDF, χ, based on an underlying histogram, H, calculated for the filtered image, f, over the region, g. Here N_g refers to the number of pixels in region g, and $D_{H_{f,g}}$ refers to the set of histogram bins.

$$\chi(m; f, g) = \left\{ \sum_{n \le m} \left(\frac{H(n; f, g)}{N_g} \right) \;\middle|\; m, n \in D_{H_{f,g}} \right\} \tag{5}$$

Computed over two ground truthed segments, g_A and g_B, the CDF describes the fraction of pixels in each segment correctly identified by a point classifier, $\delta = m$. Applying the classifier gives estimates, $\hat{g}_A$ and $\hat{g}_B$, of these pre-surveyed segments:

$$\begin{aligned} &\left\{ \hat{g}_A = (x, y) \;\middle|\; f(x, y) \le \delta, (x, y) \in D_f \right\} \\ &\left\{ \hat{g}_B = (x, y) \;\middle|\; f(x, y) > \delta, (x, y) \in D_f \right\} \end{aligned} \tag{6}$$

Given a particular classifier, the fraction of pixels correctly identified in each region is

$$\Theta\left(\delta; g_j\right) = \frac{\mathbf{area}\left(\hat{g}_j \cap g_j\right)}{\mathbf{area}\left(g_j\right)} \tag{7}$$

When a classifier is applied globally, the correctly identified pixel fractions, Θ, over each region are directly related to the CDF functions calculated over the regions.

$$\begin{aligned}\Theta\left(\delta; f, g_A\right) &= 1 - \chi\left(\delta; f, g_A\right) \\ \Theta\left(\delta; f, g_B\right) &= \chi\left(\delta; f, g_B\right)\end{aligned} \tag{8}$$

These scalars, each with magnitude between zero and one, can be combined into a single scalar function: segmentation efficiency, $\eta\left(\delta; f, g_A, g_B\right)$.

$$\begin{aligned}\eta\left(\delta; f, g_A, g_B\right) &= \Theta\left(\delta; f, g_B\right) + \Theta\left(\delta; f, g_A\right) - 1 \\ &= \chi\left(\delta; f, g_B\right) - \chi\left(\delta; f, g_A\right)\end{aligned} \tag{9}$$

The peak value of $\left|\eta\left(\delta\right)\right|$ identifies the maximum possible fraction of pixels, weighted by region size, correctly identified for some choice of classifier, δ_{max}, given a particular prefilter and region pairing. A classifier that achieves unity segmentation efficiency perfectly distinguishes the two regions. Zero efficiency means that a classifier makes no distinction between two segments. The sign of $\eta\left(\delta\right)$ distinguishes the region for which the classifier is an upper bound and is otherwise arbitrary; that is:

$$\eta(\delta; f, g_B, g_A) = -\eta(\delta; f, g_A, g_B) \tag{10}$$

Figure 3 plots a sample segmentation efficiency distribution and associated CDFs, as a function of image luminance, *l*, for a gelatinous ctenophore.

D. Ensemble Averaged Segmentation Efficiency

Given a large number of video clips imaged under specific environmental conditions, segmentation efficiency extracts the most useful information from each image and maps this information to appropriate vision algorithms. Thus, the ensemble averaged efficiency serves as a convenient tool for analysis of the image database:

$$\left\langle \eta(\delta; f, g_A, g_B) \right\rangle = \frac{1}{M_i} \sum_{i=1}^{M_i} \eta(\delta; f, g_A, g_B) \tag{11}$$

The ensemble averaging process compares segmentation efficiency across the database of M_i samples for each classification level, δ, given a particular choice of filter, *f*, and the spatial comparison embodied by the choice of image regions, g_A and g_B.

The argument that maximizes the magnitude of $\left\langle \eta(\delta) \right\rangle$ is the classifier, δ_{max}, that optimizes the area-weighted fraction of correctly identified pixels for the two image regions across the entire video database, given a particular choice of prefilter. Thus the peak magnitude of $\left\langle \eta(\delta) \right\rangle$ serves as a useful metric for comparing the quality of various

filters computed over specific spatial geometries. Confidence limits around $\langle \eta(\delta) \rangle$ establish the consistency of each classifier. Both quality and consistency of a classifier are important considerations in the synthesis of a visual servoing system.

E. Application of Segmentation Efficiency to Jelly Tracking

Computation of segmentation efficiency across a marine life database aids design of an automated jelly-tracking system. The marine life database was populated with short (half-second) clips of animals filmed *in situ* by ROV *Ventana*. Video clips included 182 samples of gelatinous animals filmed under a variety of lighting conditions. Variations include lamp configuration, camera zoom, camera gain, and marine snow backscatter conditions. For each clip in the database, a human operator provided ground truth by defining the target region, g_t, with a spline fit to the target interior edge. The background region, g_b, was defined as the image complement to the target animal region(s) and to the excluded region, g_x, consisting of small snow particles:

$$g_b = \left(g_x \cup g_t \right)^C \tag{12}$$

Here snow pixels were identified automatically, without human input, according to the relation:

$$g_x = \left\{ (x,y) \;\middle|\; \Phi(x,y) \geq \delta_{snow} \text{ and } (x,y) \in D_f \right\} \tag{13}$$

$$\Phi(x,y) = \frac{\sum_{m,n \in D_q} \left(l(x-m, y-n) > l(x,y) \right)}{\sum_{m,n \in D_q} 1}$$

In effect, this operator recognizes marine snow by identifying small pixel regions with high contrast to their neighbors. For this work q was chosen on a 5x5 square grid, with the domain, D_q, comprised of the grid's 16 border elements. δ_{snow} was set equal 0.75.

Table 3 shows the peak magnitude of ensemble averaged segmentation efficiency, $\left| \left\langle \eta(\delta_{\max}) \right\rangle \right|$. Peak height is listed for the fifteen filters described in Section II.A and four geometric structures described in Section II.B.

The first column of Table 3 uses $\left\langle \eta(\delta_{\max}; f, g_b, g_t) \right\rangle$ to describe regional contrast between the entire target and background segments. The background difference filter, the gradient filters, and the monochrome luminance filter have the highest peak values for distinguishing target and background pixels. For most vision applications, high gradient is associated primarily with target edges. Segmentation efficiency analysis highlights the unexpected result that, for scenes containing gelatinous animals, gradient filters detect the entire target interior region. This phenomenon occurred broadly for a variety of gelatinous targets at ranges of 0.2 - 3 m and camera viewing cone angles between 10° and 60° for a constant video resolution of 320x240 pixels.

The second, third, and fourth columns of Table 3 assess information along the target boundary. Of these boundary data, the strongest responses were observed for $\left\langle \eta(\delta_{\max}; d_b, \partial g_b, \partial g_t) \right\rangle$, the strict-edge comparison using the background difference filter,

and for $\left\langle \eta\left(\delta_{\max}; \left\|\nabla l_{OC}\right\|_M, g_b^{\circ}, \partial g_b\right)\right\rangle$, the background edge-to-interior comparison using the snowless gradient filter.

Of all the entries in Table 3, the highest peak value of segmentation efficiency corresponds to the background difference filter applied regionally, $\left\langle \eta\left(\delta_{\max}; d_b, g_b, g_t\right)\right\rangle$. This peak value indicates the high quality of the background difference signal for region-based segmentation. The restriction that an external agent initialize the filter (see Section II.A), however, limits its use in real-time tracking applications.

In contrast, the gradient filters achieve high regional segmentation efficiency peaks without any need for special initialization. Moreover, segmentation efficiency analysis indicates that a fixed-level gradient classifier consistently describes the target region associated with gelatinous animals, despite the wide range of lighting conditions included in the marine life video database. Although the gradient filter performs almost as well for boundary comparisons as for regional ones, segmentation efficiency is slightly higher for regional comparisons. Consequently, region-based segmentation strategies applied to gradient prefiltered images arise from segmentation efficiency analysis as primary candidates for use in a visual jelly-tracking algorithm.

Tight confidence limits on efficiency further buoy this recommendation. Confidence limits indicate excellent consistency for the gradient filter, especially in comparison with other competing filters such as the luminance filter. Figure 4 depicts the mean distribution and 90% confidence interval for segmentation efficiency as a function of classifier, δ, for both the luminance and morphological gradient filters. Confidence limits are significantly tighter for morphological gradient than those for the luminance distribution. Significantly, the peak value of segmentation efficiency always occurs in

approximately the same location, δ_{max}, for morphological gradient distributions across the database. At 320x240 resolution, the gradient classifier, δ_{max}, (equal 10 gray levels per pixel) yields an efficiency of at least 0.30 for 95% of transparent animal samples drawn from the marine life database. By comparison, the lower confidence bound for the luminance distribution is nearly flat at zero efficiency.

Inconsistent peak height for luminance information results from scene-to-scene luminance variations and from mild image gradients (smaller than one gray level per pixel) across typical ROV-imaged scenes. A single luminance classifier that distinguishes target pixels from background pixels does not always exist given uneven lighting. Figure 5 illustrates this phenomenon by displaying a luminance contour that envelops both a target ctenophore and a section of the background image. By contrast, gradient and background difference filters cleanly distinguish the target in this case.

A method exploiting luminance information would thus need to adapt spatially and temporally to compensate for poor consistency of luminance classifiers; adaptation introduces concerns of convergence and processing requirements for real-time, power-constrained applications. By comparison, a gradient algorithm can use a fixed-level classifier to identify gelatinous targets consistently. This characteristic of gradient information enables the implementation of a non-iterative, bounded-time segmentation component in a visual tracking algorithm.

This focus on gradient and background difference information contrasts with other biology-inspired studies conducted using marine video. These studies have successfully employed, for example, region-based optical flow methods for fish tracking [20], edge-based luminance gradient methods for study of bioluminescence [21], and

region-based luminance methods for classification of plankton [18,22,24]. Differences among applications motivate the choice of these particular filtering and tracking strategies. Such differences emphasize the importance of a design tool, such as segmentation efficiency, for matching application-oriented image information to specific tracking algorithms available in the vision literature.

III. Synthesis of a Segmentation Technique for Jelly Tracking

This section describes two segmentation strategies that rely on gradient image thresholds. Both methods were synthesized by placing segmentation efficiency results in the context of operational jelly tracking. The first method relies solely on gradient information. The second method refines the first using background difference information. Though more computationally complex, the second technique improves segmentation quality to enhance reliability for low sample rate applications.

A. Constraining Algorithm Choice Based on System Requirements

Segmentation efficiency, alone, determines only the general class of segmentation algorithm suited to a particular application. An understanding of application constraints completes the design process, enabling the selection of a specific vision algorithm from the general class indicated by segmentation efficiency analysis.

For jelly tracking, the first priority of a vision system is robustness to variable lighting conditions and target identity. Considering prior discussion, gradient-based

regional segmentation methods have a strong advantage in terms of signal quality, as predicted by peak segmentation efficiency height, and in terms of consistency, as predicted by tight efficiency confidence limits. Likewise, the high efficiency peak for the background difference filter suggests that this information could enable accurate jelly segmentation, given automation of the filter's external agent requirement. These results narrow the search for a jelly-tracking algorithm to the class of region-based segmentation methods applied to images filtered with a gradient operator, or possibly with the background difference operator.

Within this region-based class, segmentation methods may be distinguished primarily by two characteristics: (1) incorporation of shape knowledge and (2) parameter adaptation. As summarized in Table 2, certain segmentation methods incorporate a definite notion of target shape (the template masking technique) or target pixel pattern (the correlation technique). Other region-based methods possess some ability to handle varying image conditions through adaptation (the adaptive threshold and expectation maximization techniques). For the jelly-tracking application, the target's flexible, three-dimensional structure makes explicit incorporation of shape knowledge a difficult problem. Furthermore, the segmentation efficiency analysis indicates that, given the appropriate choice of image prefilter, a fixed-parameter, non-adaptive technique can robustly segment targets over a wide range of lighting conditions. Both adaptation and shape-constraints add complexity to a segmentation algorithm. As neither characteristic clearly benefits the jelly-tracking application, an appropriate selection criterion among region-based segmentation methods is simplicity. Of the region-based methods, fixed-

parameter global thresholding is the least complex, with no capability for adaptation and with the assumption of a fully arbitrary target shape.

This segmentation method also fits the requirements for both ROV and AUV operations. The application of a global threshold to an image, along with assignment of neighboring pixels to amorphous segments, results in an easily implemented, robust segmentation strategy with excellent computational efficiency. Because the method does not require parameter tuning, it behaves reliably and repeatably upon activation. Furthermore, global threshold segmentation carries no information from sample step to subsequent sample step. This characteristics makes the method well suited for the range of sample rates expected for field operations, as high as 30 Hz for an ROV pilot assist and as low as a fraction of a Hertz for a strobed AUV application.

B. Segmentation with a Gradient Based Threshold

Based on a segmentation efficiency analysis in the context of application constraints, a gradient-based global threshold method was implemented, along with a pattern-vector recognition routine, as the visual tracking algorithm for field operations. The global threshold method relies on smoothed morphological gradient information, extracted by the $\|\nabla I\|_{M,S}$ filter, since this information has the highest peak of the gradient filters in Table 3. For this filter, the choice of a fixed gradient threshold matches the value of δ_{max}, at 10 gray levels per pixel (for 320x240 resolution).

The complete segmentation algorithm is summarized as follows:

1. Apply a 3x3 uniform filter to the monochrome luminance image
2. Calculate morphological gradient for the smoothed luminance image
3. Apply a global threshold to identify potential target regions
4. Calculate size of connected regions and filter out small segments (snow)

The algorithm has low computational complexity. For an image containing P pixels, the uniform smoothing filter, which is separable, requires $4P$ addition operations. Morphological erosion and dilation operators, used to calculate morphological gradient, are also separable when a square, in this case 3 x 3, structuring element is applied. It follows that the operations count to compute morphological gradient involves $8P$ comparisons and P subtractions. Application of a global threshold requires P pixel comparisons. The total algebraic operations count for the method is $5P$ additions and no multiplications. No iteration is required. Because the computational burden associated with the method is quite low, the algorithm is well suited for a real-time, processor-constrained application.

Although gradient filters consistently recognize jelly targets, they also amplify snow particles. Step 4 of the segmentation algorithm, above, addresses this issue and removes snow particles by filtering potential target segments based on pixel count. A size filter of 25-30 total pixels (given 320x240 resolution and typical zoom and marine lighting configurations) removes the majority of marine snow pixels from consideration. Table 4 emphasizes the small size of the majority of snow particles and validates the use of a snow filter based on size information. Removing snow particles prior to recognition reduces computational requirements for the recognition step.

An additional theoretical concern associated with gradient based segmentation addresses extreme camera resolution and lighting settings. At some limit of high zoom or low lighting, gradients over the target must fall below camera sensitivity. In practice however, gradient threshold segmentation works very well at a range of zoom settings, ranges, and target sizes. Table 4 describes these variations in terms of target size in pixels across the set of video clips in the marine life database.

For ROV applications, a simple recognition scheme complements gradient based segmentation to form a complete tracking solution. The recognition component computes a pattern vector for each identified segment and finds the segment best matching a target profile vector. Elements of the pattern vector include the image plane coordinates of the segment centroid, the segment's mean luminance, its pixel area, and its aspect ratio. These statistics sufficiently establish correspondence of the target region through time, given the 10 Hz sample rate used for field demonstrations.

C. Augmenting Gradient Threshold with Background Difference Information

For high sample rate applications the primary information useful for recognition, given multiple segments per frame, is knowledge of the target segment's location in previous frames. Converging animal trajectories or low sample rates, however, place recognition techniques relying on position estimation in jeopardy of failure. These situations require refined, preferably time-invariant recognition statistics. Examples of refined metrics include shape descriptors, granulometries [22], and pixel value histograms. Histogram descriptors are especially relevant to recognition of gelatinous

animals, as histograms offer potential invariance to target orientation. A high quality segmentation algorithm aids in extracting refined recognition statistics.

Background difference information offers potential for higher quality segmentation, since the d_b filter displayed the highest efficiency of the filters compared by Table 3. The background difference filter cannot, however, be applied to an image without external input. A successive filter approach automates generation of this external input. First, a gradient based method akin to the technique described in Section III.B produces a rough segmentation. The snowless gradient filter, which reliably captures target edge information while automatically eliminating small snow particles, works well for this first step. Bounding boxes are calculated around high gradient regions and passed to the background difference filter as external inputs. Within each bounding box, a second step calculates background difference and thresholds to enhance quality of the target segmentation. Figure 6 shows typical segmentation results using gradient information only and using the augmented background difference method.

The refined algorithm involves the following steps:

1. Apply the snowless filter by first opening and then closing the monochrome luminance image
2. Calculate morphological gradient for the snowless image, $\|\nabla l_{OC}\|_M$
3. Apply a global threshold to identify potential target regions
4. Calculate bounding boxes for each segmented region
5. Synthesize a background image for each bounding box
6. Calculate the background difference image, d_b, in each bounding box

7. Apply a background difference threshold within the bounding box

This augmented segmentation algorithm requires more computational effort than the simple gradient threshold algorithm. In exchange the algorithm produces a higher quality segmentation. For an image containing P pixels, the opening and closing operations, based on 3x3 square structuring elements, require $16P$ comparisons. The operations count to compute morphological gradient requires $8P$ comparisons and P subtractions. Application of a global threshold requires P pixel comparisons. The augmentation step considers bounding boxes enclosing Q pixels, $Q < P$. Synthesizing the background image requires $11Q$ multiplications and additions and $8Q$ table lookups. Calculating the background difference image requires Q subtractions. The final threshold step requires an additional Q comparisons. The final algebraic operations count is $12Q+P$ additions and $11Q$ multiplications. No iteration is required. Thus, if Q approaches P, the algorithm's computational cost greatly exceeds the $5P$ additions required for the basic gradient threshold method, described in Section III.B.

Neither the background difference segmentation algorithm, nor a refined recognition method, has yet been implemented for ocean testing.

IV. Ocean Experiments using an ROV Platform

Field experiments demonstrate that the gradient-based vision algorithm, synthesized for the jelly tracking application using the segmentation efficiency technique, performs under ocean conditions. The success of these experiments illustrates the power

of the predictive performance assessment approach. The examination of a jellyfish database in search of strong signals and their relationship to specific classes of segmentation algorithm produced a computationally simple, but nonetheless robust, technique to enable automated jellyfish tracking.

A. The Experimental System

A jelly-tracking system was implemented as a pilot aid on MBARI ROV *Ventana* (Figure 7). During midwater experiments, *Ventana*'s cameras film jellies in their natural habitat at depths between 100 and 1000 m. A fiber optic link carries these images from the submerged ROV to the surface support vessel, R/V *Point Lobos*. Video from a second camera, mounted approximately one meter above the main camera, is also transmitted. A 700 MHz Pentium III computer acquires the twin video signals in NTSC format using Matrox Meteor cards. The gradient-based segmentation method identifies potential targets in frames from each of the two video streams. A pattern vector recognition component identifies the segment best matching the target profile for each stream. After checking for possible false positives, the algorithm uses triangulation to generate a three dimensional position vector for the target, relative to the ROV. The computer then applies an axis-decoupled linear control law based on the relative position vector. These control commands are routed through the pilot joystick to allow manual override in a crisis situation. The tether carries the summed pilot and computer commands back to the submersible to close the control loop. The block diagram of

Figure 8 encapsulates this system description; further details regarding the experimental hardware are described in [38].

A single button press by the operator activates the experimental jelly-tracking system. A two-second training period sets the initial value of the pattern vector used for recognition. During the training period, the target is identified with an abbreviated profile that considers target region size and proximity to the center of the image. The trainable pattern vector recognition approach permits application of the system with a wide range of gelatinous specimens, including larvaceans, cnidaria and ctenophores.

The control law acts in cylindrical coordinates to regulate ROV range to target, relative depth, and relative yaw. Quasi-steady tether forces introduce an undesirable dynamic response during system initialization and an offset error at steady state. To counteract these effects, a dynamic estimator infers bias force and enables a disturbance accommodating control term.

The three coordinate control law (range, depth, and yaw) does not act on pitch, roll, or the ROV's circumferential position about the jelly. Passive buoyancy stabilization restricts pitch and roll motion. The remaining free coordinate constitutes a control law null space. The control system can exploit the null space dimension to accomplish goals secondary to the primary goal of holding the ROV camera centered on the jelly. Also, the human pilot can also issue commands in the null space without disturbing the automated jelly-tracking control law. In the field, a human pilot has demonstrated this capability, issuing circumferential commands to select the orientation of a gelatinous specimen relative to the ROV cameras. Additional details of the jelly tracking control system are discussed in [39].

B. Experimental Results

Open ocean experiments performed the first ever demonstration of fully automated robotic tracking of a midwater animal. Tests further demonstrated that gradient based segmentation successfully enables long duration autonomous jelly tracking. Performance of the jelly-tracking system was evaluated over multiple *Ventana* dives during the course of 2001-2002.

Tests explored endurance tracking with the ROV under full computer control. Several long, successful runs confirm the robustness of gradient-based segmentation algorithms under field conditions. Four notable runs included the tracking of a *Solmissus* specimen for 25 continuous minutes, of a *Benthocodon* for 29 minutes, of a sinking larvacean house for 34 minutes, and of a *Ptychogena* for 89 minutes. All runs were conducted with no intervention by a human pilot. During the runs, the control system countered unmodeled tether and buoyancy forces to maintain lock on the jelly target. All four of these specimens were in motion relative to the water column. Figure 9 depicts the *Solmissus* experiment with an overlay image that combines 40 samples of the tracked target at even intervals through the run. The figure demonstrates the success of the servoing system in maintaining small relative error throughout the *Solmissus* run.

In all four cases, a number of ocean animals, including squid and salps, wandered inside the ROV camera's field of view without disturbing the jelly-tracking system. Thus the recognition component performed ably, despite its simplicity. Only one of the four long runs ended as a result of a visual tracking system error. This error terminated the

Solmissus experiment when a small salp passed in front of the jelly target and corrupted the recognition profile. Before the error, the recognition system successfully distinguished between the jelly target and several nearby squid. Additional recognition errors were observed during an experiment inside an unusual, dense swarm of krill. During these experiments, squid initially feeding on krill began to approach the ROV. In cases when multiple squid subsequently overlapped the target segment and when more than three squid were simultaneously visible in the vision cone, the recognition algorithm twice failed after tracking a target for only 15 minutes. These notable recognition breakdowns indicate the limitations of the current pattern vector approach and motivate, as future research, the investigation of improved recognition strategies.

In addition to endurance tracking results, field experiments highlight the practical difficulties involved with deploying a visual robotic system in the ocean. Challenges arise from the ROV platform and from harsh ocean conditions. These challenges include:

- Nonuniform lighting conditions and marine snow
- Recognition of the target in scenes containing multiple animals
- Control of a vehicle platform given a poor plant model
- Control of a vehicle with unknown buoyancy force and tether dynamics
- Transition between pilot and automated control modes
- Frequent hardware failure caused by harsh operational environment (serial port communication, video transmission, camera angle sensors)
- Limited ability to make system modifications at sea

The requirement to handle these challenges implies that hardware and software for a jelly-tracking system must be not only reliable and robust, but also easy to operate under harsh conditions.

V. Summary

A method for extracting useful information from a database of natural images was introduced. The technique, called segmentation efficiency, detects information inherent to an application and maps the information to classes of segmentation algorithm available in the vision literature. Morphological gradient and background difference filters were found to be highly effective at extracting image information for the jelly-tracking application. Use of these prefilters enabled the synthesis of a robust segmentation system based on global thresholding rather than other, more complex vision algorithms. The low operations count and high consistency of gradient-based global thresholding match the operational requirements for reliability and efficiency demanded for the ROV pilot aid application and for future AUV implementation. An ROV pilot aid system was fielded to demonstrate, for the first time, the use of an automated marine vehicle to track an animal in the deep ocean. Subsequent experiments tracked jellies for as long as 89 minutes under full computer control.

Acknowledgement

The authors would like to thank the Monterey Bay Aquarium Research Institute and Packard Foundation grants 98-3816 and 98-6228 for supporting this work. In particular, we wish to recognize Bruce Robison and Craig Dawe of MBARI for their insightful input to this project.

References

[1] P.I. Corke and S.A. Hutchinson. "Real-time vision, tracking and control," *Proc. IEEE ICRA 2000*, pp. 622-623, 2000.

[2] M. Vincze and G.D. Hager, Ed. *Robust Vision for Vision-Based Control of Motion*, SPIE Optical Engineering/IEEE Press, 2000.

[3] S.D. Fleischer, S.M Rock, and R. Burton, "Global position determination and vehicle path estimation from a vision sensor for real-time video mosaicking and navigation," *IEEE/MTS OCEANS 1997*, vol. 1, pp. 641-647, 1997.

[4] R. Garcia, J. Batlle, and X. Cufi, "A system to evaluate the accuracy of a visual mosaicking methodology," *IEEE/MTS OCEANS 2001*, vol. 4, pp. 2570-2576, 2001.

[5] N. Gracias and J. Santos-Victor, "Underwater mosaicking and trajectory reconstruction using global alignment," *IEEE/MTS OCEANS 200*1, vol. 4, pp. 2557-2563, 2001.

[6] A. Huster, S.D. Fleischer, and S.M. Rock, "Demonstration of a vision-based dead-reckoning system for navigation of an underwater vehicle," *Proc. 1998 World Conference on Autonomous Underwater Vehicles*, pp. 185-189, 1998.

[7] J.-F. Lots, D.M. Lane, E. Trucco, and F. Chaumette, "A 2D visual servoing for underwater vehicle station keeping," *Proc. IEEE ICRA 2001*, vol. 3, 2001.

[8] J.-F. Lots, D.M. Lane, and E. Trucco, "Application of a 2 ½ D visual servoing to underwater vehicle station-keeping," *IEEE/MTS OCEANS 2000*, vol. 2, pp. 1257-1264, 2000.

[9] R.L. Marks, H.H. Wang, M.J. Lee, and S.M Rock, "Automatic visual station keeping of an underwater robot," *IEEE OCEANS 1994*, vol. 2, pp. 137-142, 1994.

[10] S. Negahdaripour and Xun Xu. "Mosaic-based positioning and improved motion-estimation methods for automatic navigation of submersible vehicles," IEEE Journal of Oceanic Engineering, vol. 27, no. 1, 2002, pp. 79-99.

[11] S. Negahdaripour and P. Firoozfam. "Positioning and photo-mosaicking with long image sequences; comparison of selected methods," *IEEE/MTS OCEANS 2001*, vol. 4, pp. 2584-2592, 2001.

[12] C. Roman and H. Singh, "Estimation of error in large area underwater photomosaics using vehicle navigation data," *IEEE/MTS OCEANS 2001*, vol. 3, pp. 1849-1853, 2001.

[13] H. Singh, C. Roman, L. Whitcomb, and D. Yoerger, "Advances in fusion of high resolution underwater optical and acoustic data*," Proc. 2000 International Symposium on Underwater Technology*, pp. 206-211, 2000.

[14] S. van der Zwaan and J. Santos-Victor, "Real-time vision-based station keeping for underwater robots," *IEEE/MTS OCEANS 2001*, vol. 2, pp. 1058-1065, 2001.

[15] M.B. Leahy, V.W. Milholen, and R. Shipman. "Robotic aircraft refueling: a concept demonstration," *Proc. of the Aerospace and Electronics Conference, 1990*, vol. 3, pp. 1145-1150, 1990.

[16] O. Amidi, T. Kanade, and R. Miller. "Vision-based autonomous helicopter research at Carnegie Mellon Robotics Institute (1991-1998)," in *Robust Vision for Vision-Based Control of Motion*, M. Vincze and G.D. Hager, Ed., SPIE Optical Engineering/IEEE Press, pp. 221-232, 2000.

[17] P.H. Batavia, D.A. Pomerleau, and C.E. Thorpe. "Overtaking Vehicle Detection Using Implicit Optical Flow," *Proc. IEEE Transportation Systems Conference*, pp. 729-734, 1997.

[18] M. Minami, J. Agbanhan, and T. Asakura. "Manipulator visual servoing and tracking of fish using a genetic algorithm," *Industrial Robot*, vol. 26, no. 4, pp. 278-289, 1999.

[19] S. Johnsen and E. Widder. "The physical basis of transparency in biological tissue: ultrastructure and minimization of light scattering," *Theoretical Biology*, vol. 199, pp. 181-198, 1999.

[20] Y. Fan and A. Balasuriya. "Autonomous target tracking by AUVs using dynamic vision", *Proc. of the 2000 International Symposium on Underwater Technology.* pp. 187-192, 2000.

[21] D. Kocak, N. da Vitoria Lobo, and E. Widder. "Computer vision techniques for quantifying, tracking, and identifying bioluminescent plankton", *IEEE Journal of Oceanic Engineering*, 24, vol. 1, pp. 81-95, 1999.

[22] X. Tang, W.K. Stewart, L. Vincent, H. Huang, M. Marra, S.M. Gallager and C.S. Davis. "Automatic plankton image recognition," Artificial Intelligence Review, vol. 12, pp. 177-199, 1998.

[23] X. Tang and W. K. Stewart. "Plankton image classification using novel parallel-training learning vector quantization network", *Proc. IEEE/MTS OCEANS '96.* v.*3,* pp. 1227-1236. 1996.

[24] R.A. Tidd and J. Wilder, "Fish detection and classification system," Journal of Electronic Imaging, vol. 10, no. 6, pp. 283-288, 2001.

[25] S. Samson, T. Hopkins, A. Remsen, L. Langebrake, T. Sutton, and J. Patten, "A system for high-resolution zooplankton imaging," IEEE Journal of Oceanic Engineering, vol. 26, no. 4, 2001, pp. 671-676.

[26] L.B. Wolff, "Applications of polarization camera technology," *IEEE Expert*, 1994.

[27] J. Rife and S. Rock. "A low energy sensor for AUV-based jellyfish tracking," *Proc. of the 12th International Symposium on Unmanned Untethered Submersible Technology*, August, 2001.

[28] K.W. Bowyer and J.P. Phillips. "Overview of work in empirical evaluation of computer vision algorithms," in *Empirical Evaluation Techniques in Computer Vision*, K.W. Bowyer and J.P. Phillips, Ed., IEEE Computer Press, 1998.

[29] K.W. Bowyer. "Experiences with empirical evaluation of computer vision algorithms," in *Performance Characterization in Computer Vision*, R. Klette, H.S. Stiehl, M.A. Viergever, and K.L. Vincken, Ed., Kluwer Academic Publishers, 2000.

[30] P. Courtney and N.A. Thacker. "Performance characterization in computer vision: the role of statistics in testing and design," in *Imaging and Vision Systems: Theory,*

Assessment and Applications, J. Blanc-Talon and D.C. Popescu, Ed., Nova Science Publishers, Inc., pp. 109-128, 2001.

[31] Ç.E. Erdem and B. Sankur, "Performance evaluation metrics for object-based video segmentation," *Proc. X European Signal Processing Conference*, vol. 2, pp. 917-920, 2000.

[32] P. Villegas, X. Marichal and A. Salcedo, "Objective evaluation of segmentation masks in video sequences," *WIAMIS'99*, pp. 85-88, 1999.

[33] Y.J Zhang, "A survey on evaluation methods for image segmentation," Pattern Recognition, vol. 29, no. 8, pp. 1335-1346, 1996.

[34] R.C. Gonzalez and R.E. Woods, *Digital Image Processing.* Addison-Wesley Publishing Company, Inc., 1993.

[35] W.E. Boyce and R.C. DiPrima, *Elementary Differential Equations and Boundary Value Problems*, Wiley, 2001.

[36] C.K. Chow and T. Kaneko, "Automatic boundary detection of the left ventricle from cineangiograms," *Computers and Biomedical Research*, vol. 5, pp. 388-410, 1972.

[37] S. Dougherty and K.W. Bowyer, "Objective evaluation of edge detectors using a formally defined framework," in *Empirical Evaluation Techniques in Computer Vision*, K.W. Bowyer and J.P. Phillips, Ed., IEEE Computer Press, 1998.

[38] J. Rife and S. Rock. "A pilot-aid for ROV based tracking of gelatinous animals in the midwater," *Proc. IEEE/MTS OCEANS 2001,* vol. 2*,* pp. 1137-1144, 2001.

[39] J. Rife and S. Rock. "Field experiments in the control of a jellyfish tracking ROV," *Proc. IEEE/MTS OCEANS 2002,* vol. 4*,* pp. 2031-2038, 2002.

Biographies:

Jason Rife received his B.S. degree in mechanical and aerospace engineering from Cornell University, Ithaca, NY, in 1996, and his M.S. degree in mechanical engineering from Stanford University, Stanford, CA, in 1999. Before commencing the M.S. degree program, Mr. Rife spent one year working in the turbine aerodynamics group of the commercial engine division of Pratt & Whitney in East Hartford, CT. He is currently a Ph.D. candidate at Stanford University, investigating sensing and control technologies required to enable a jellyfish-tracking underwater vehicle. This work is part of a joint collaboration between the Stanford Aerospace Robotics Laboratory and the Monterey Bay Aquarium Research Institute to study advanced underwater robot technologies.

Stephen M. Rock received the B.S. and M.S. degrees in mechanical engineering from the Massachusetts Institute of Technology, Cambridge, MA, in 1972, and the Ph.D. degree in applied mechanics from Stanford University, Stanford, CA in 1978. Dr. Rock joined the Stanford faculty in 1988, and is now an Associate Professor in the Department of Aeronautics and Astronautics. He is also an adjunct engineer at the Monterey Bay Aquarium Research Institute. Prior to joining the Stanford faculty, he led the Controls and Instrumentation Department of Systems Control Technology, Inc. In his 11 years at SCT, he performed and led research in integrated control; fault detection, isolation, and accommodation; turbine engine modeling and control; and parameter identification. His

current research interests include the development and experimental validation of control approaches for robotic systems and for vehicle applications. A major focus is both the high-level and low-level control of underwater robotic vehicles.

Figure 1

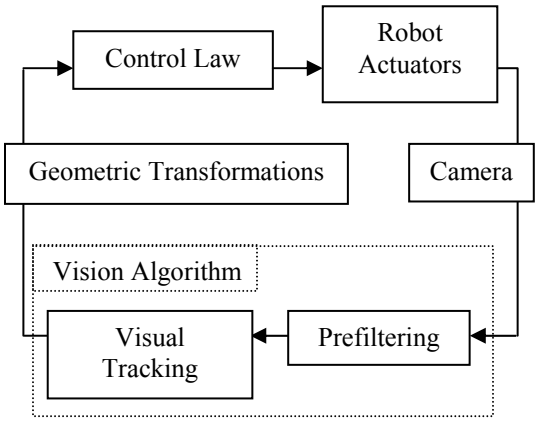


Figure 2

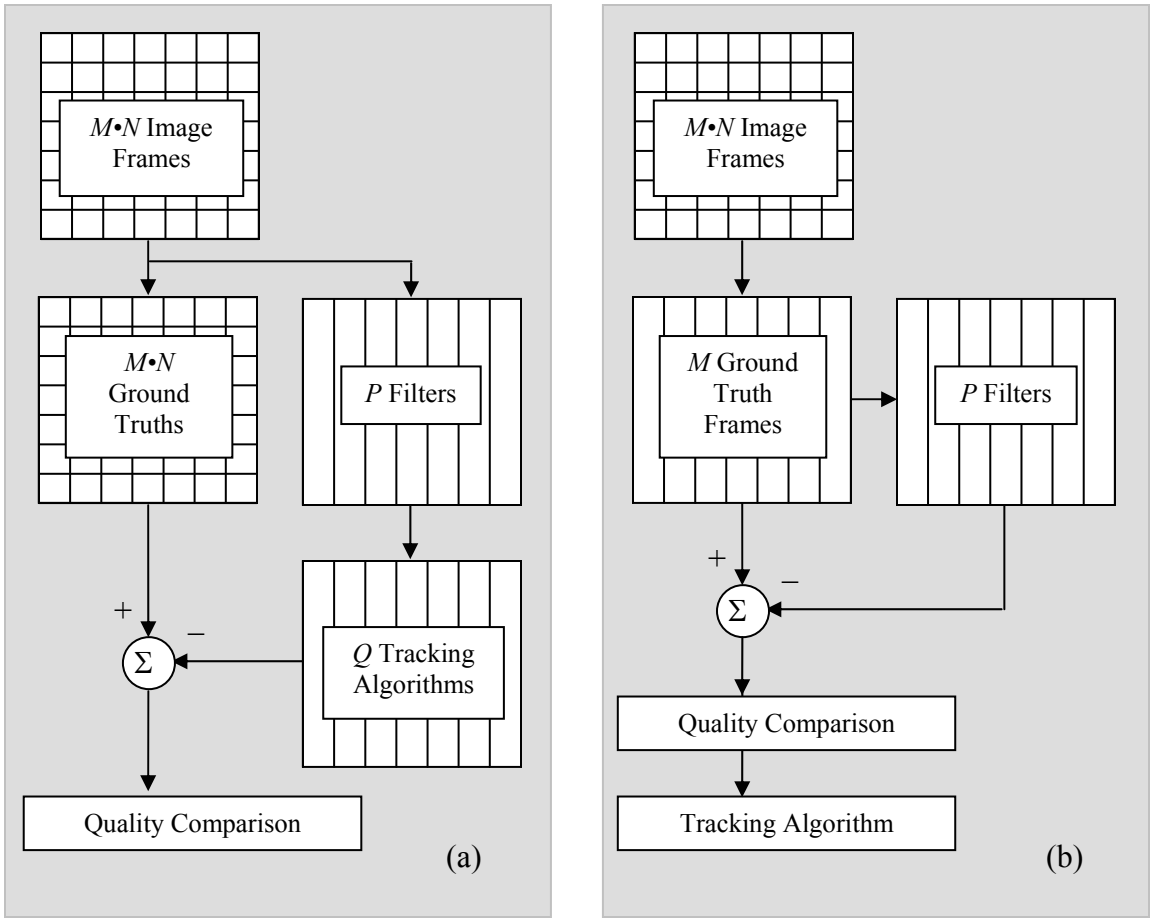


Figure 3

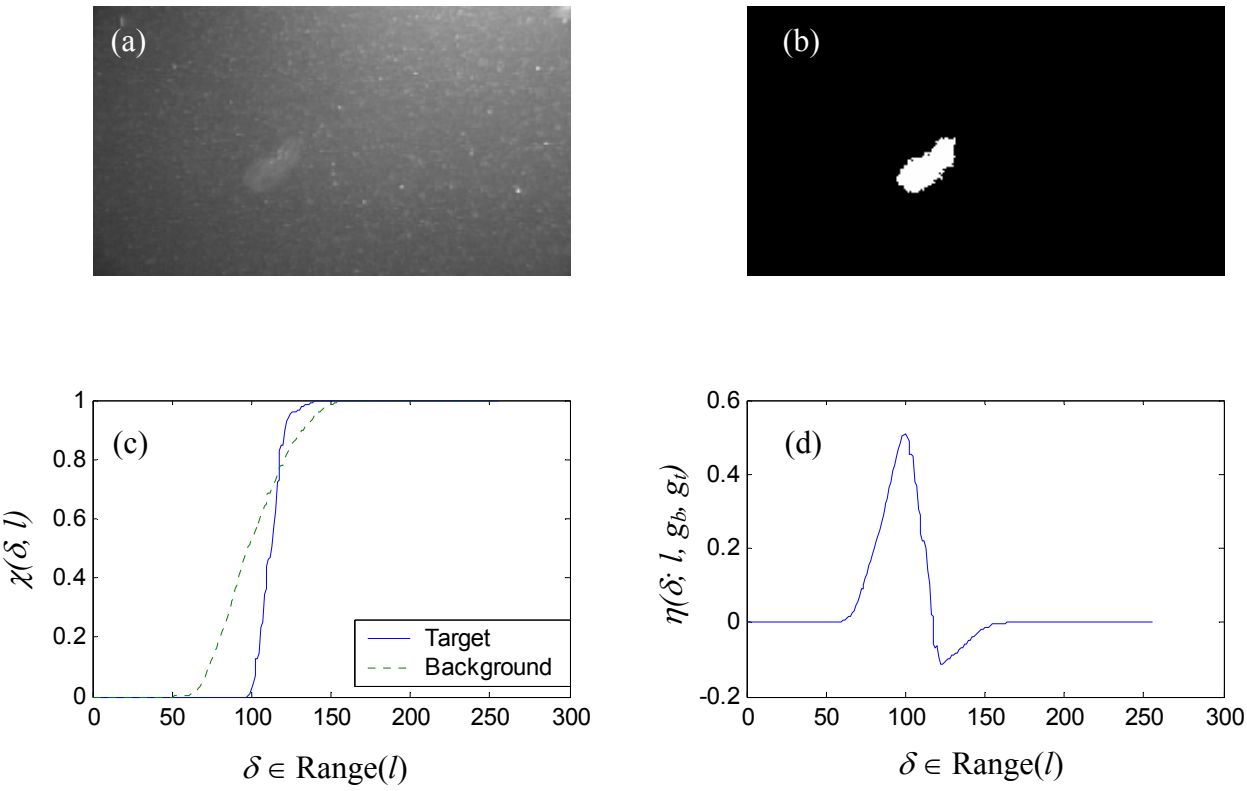


Figure 4

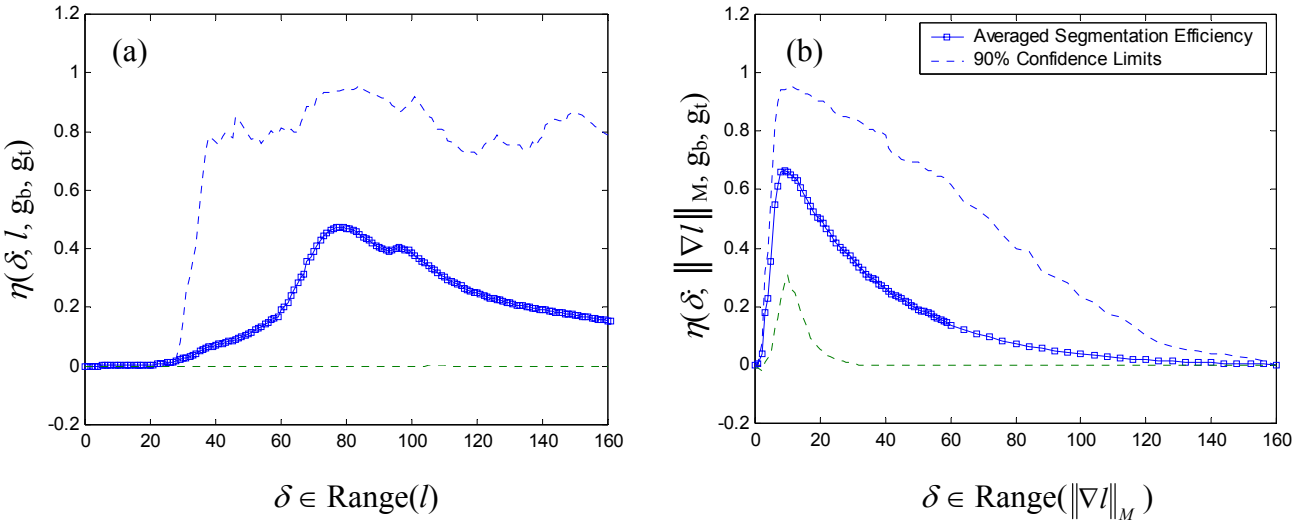
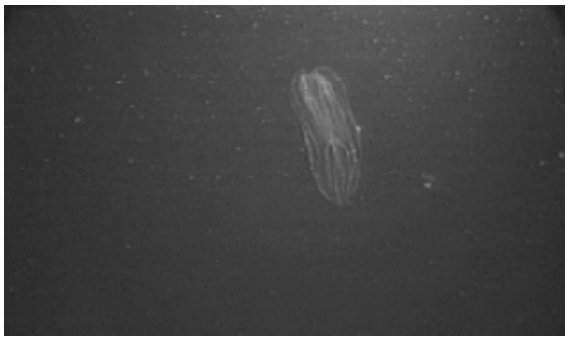
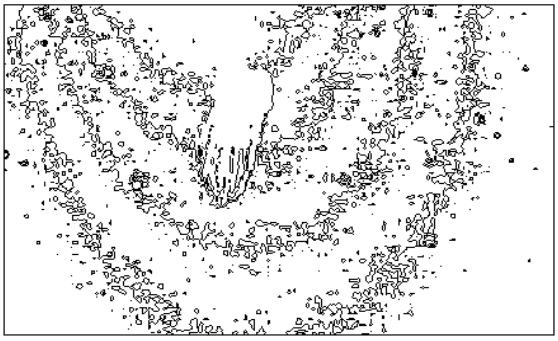


Figure 5

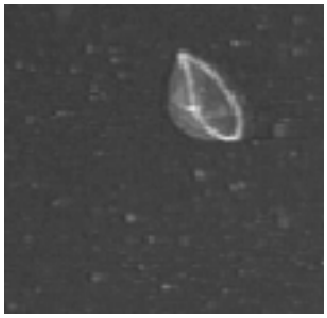


(a)

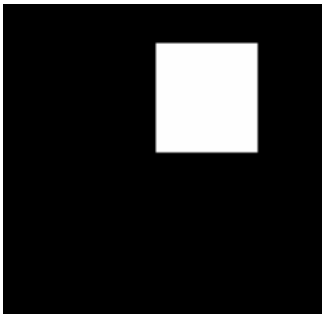


(b)

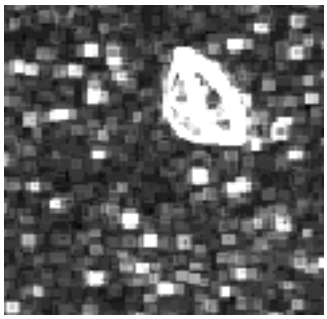
Figure 6



(a)

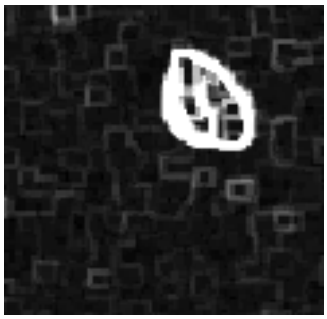


(b)

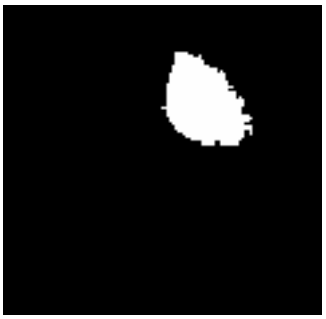


(c)

(d)



(e)



(f)

Figure 7

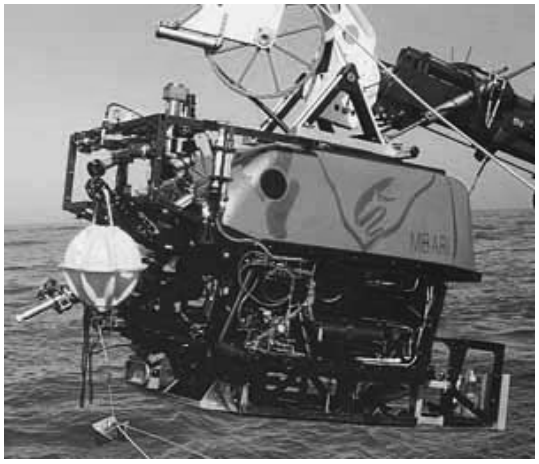


Figure 8

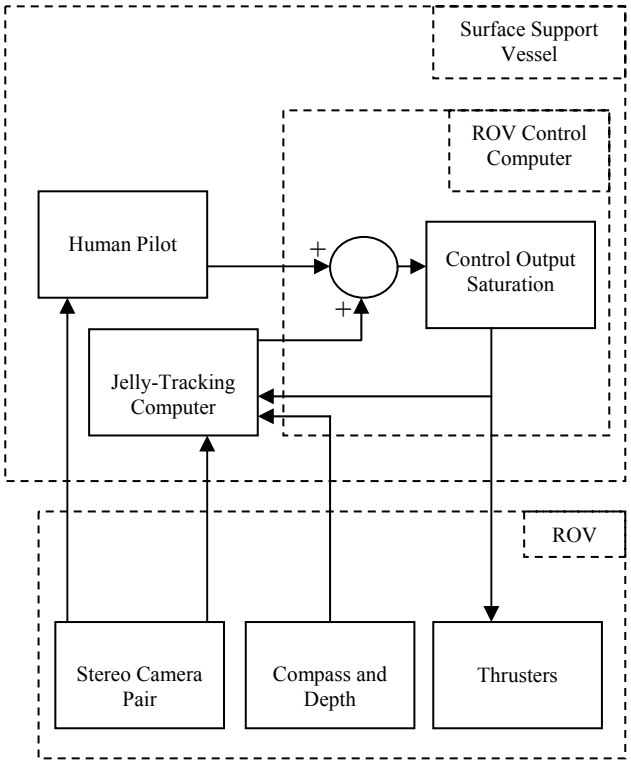
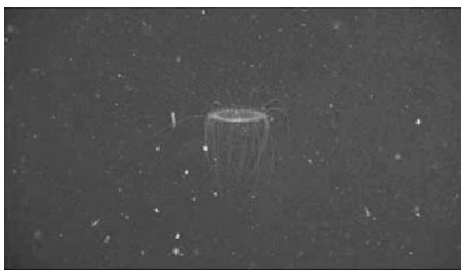
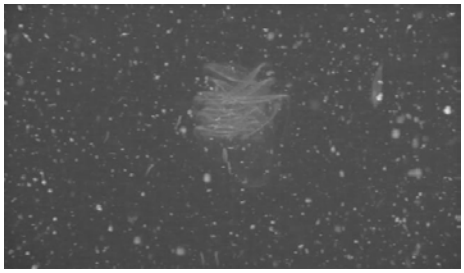


Figure 9





(a) (b)

Table 1 – Local Image Filters

Symbol	Filter Title	Equation
c_r	Red	
c_g	Green	
c_b	Blue	
l	Luminance	$l = [0.30 \quad 0.59 \quad 0.11]\begin{bmatrix} c_r & c_g & c_b \end{bmatrix}^T$
$\|\nabla l\|_2$	Euclidian Gradient	$\|\nabla l\|_2 = \sqrt{\left(\frac{\Delta l}{\Delta x}\right)^2 + \left(\frac{\Delta l}{\Delta y}\right)^2}$
$\|\nabla l\|_{2,S}$	Smoothed Euclidian Gradient	$\|\nabla l\|_{2,S} = \|\nabla l\|_2 ** h$
$\|\nabla l\|_M$	Morphological Gradient	$\|\nabla l\|_M = (l \oplus q) - (l \ominus q)$
$\|\nabla l\|_{M,S}$	Smoothed Morph. Gradient	$\|\nabla l\|_{M,S} = \|\nabla l\|_M ** h$
$\|\nabla l_{OC}\|_M$	Snowless Morph. Gradient	$\|\nabla l_{OC}\|_M = (l_{OC} \oplus q) - (l_{OC} \ominus q)$
$\nabla^2 l$	Laplacian	$\nabla^2 l = l ** \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$
$\nabla^2 l_S$	Smoothed Laplacian	$\nabla^2 l_S = \nabla^2 l ** h$
d_t	Time difference	$d_t = l^k - l^{k-1}$
d_b	Background difference	$d_b = l - \hat{l}_b$
$\|p\|$	Optical Speed	$\|p\| = \sqrt{u^2 + v^2}$
$\angle p$	Optical Flow Direction	$\angle p = \mathbf{arctan2}(u, v)$

Table 2 – Grouping Segmentation Methods

Grouping	Examples	Pixel-level Distinction	Shape Assumptions
Regional	Expectation Maximization (EM) over elliptical regions	Ellipse interior vs. exterior	Union of ellipses describes target
	Template Masking	Mask interior vs. exterior	Target shape known
	Threshold	Blob interior vs. exterior	No shape assumptions
	Adaptive Threshold	Blob interior vs. exterior	No shape assumptions
	Correlation	Under reference image vs. exterior to it	Target shape described by reference image
Edge	Active Contours (Snakes)	Edge vs. non-edge pixels	Target contour connects edges with minimum length and curvature
	Convex Edge Merging	Edge vs. non-edge pixels	Target contour connects convex set of edges
	Hough Transform	Edge vs. non-edge pixels	Target shape known
Hybrid	Region Merging	(1) Initial Seed: pixels interior vs. exterior to amorphous target region (2) Termination Criterion: edge vs. non-edge pixels	Target shape arbitrary, but characterized by well defined edges at regional boundaries
	Watershed	As above	As above

Table 3 – Peaks of the Ensemble Averaged Segmentation Efficiency Distribution

Input Filter	Region Based Comparison	Edge Based Comparisons		
	$\left\langle \eta_{g_b - g_t, f} \right\rangle$	$\left\langle \eta_{\partial g_b - \partial g_t, f} \right\rangle$	$\left\langle \eta_{\partial g_t - g_t^\circ, f} \right\rangle$	$\left\langle \eta_{g_b^\circ - \partial g_b, f} \right\rangle$
c_r	0.2353	0.2139	0.0909	0.0312
c_g	0.0709	0.0763	0.0300	0.0284
c_b	0.1888	0.1720	0.0697	0.0330
l	0.4755	0.3253	0.1390	0.1626
$\|\nabla l\|_2$	0.5398	0.3344	0.1304	0.3052
$\|\nabla l\|_{2,S}$	0.6688	0.2989	0.1427	0.5102
$\|\nabla l\|_M$	0.6668	0.3571	0.1555	0.4512
$\|\nabla l\|_{M,S}$	**0.7212**	0.2819	0.1453	0.5886
$\|\nabla l_{OC}\|_M$	0.5904	0.1291	0.2343	**0.6296**
$\nabla^2 l$	0.228	0.2553	0.1150	0.1848
$\nabla^2 l_S$	0.3647	0.4392	0.1316	0.3871
d_t	0.1876	0.1590	0.0572	0.1043
d_b	**0.8678**	**0.6068**	0.2199	0.2680
$\|p\|$	0.1011	0.0206	0.0328	0.0746
$\angle p$	0.0228	0.0060	0.0164	0.0228

Table 4 – Target size distribution

Target Size Classifications	
10^0-10^1 Pixels	0
10^1-10^2 Pixels	18
10^2-10^3 Pixels	149
10^3-10^4 Pixels	77
10^4-10^5 Pixels	19

Snow Size Classifications	
10^0-10^1 Pixels	156487
10^1-10^2 Pixels	23950

Figure Captions:

Figure 1: Block diagram describing visual servoing

Figure 2: Comparison of processing and preparation requirements for (a) existing assessment techniques and (b) a new input-focused technique. Analysis considers application of P prefilters and Q tracking algorithms to a database consisting of M image sequences, each with N frames.

Figure 3: Region statistics for a sample from the marine life database. (a) Ctenophore imaged in dense snow and nonuniform lighting. (b) Background difference based segmentation of the ctenophore. (c) Cumulative distribution functions (CDFs) for the luminance image over target and background regions. (d) Segmentation efficiency for the luminance image, $\eta(\delta; l, g_b, g_t)$.

Figure 4: Ensemble averaged segmentation efficiency distribution and confidence limits for transparent animal samples from the marine life video database. Distributions were calculated over the target and background regions for (a) luminance images and (b) smoothed morphological gradient images.

Figure 5: No unique luminance threshold separates this ctenophore target from the background. (a) Luminance image of a ctenophore observed under nonuniform illumination. (b) Contours for the luminance image at intervals of 8 gray levels.

Figure 6: Comparison of two segmentation methods. (a) Luminance image. (b) Externally defined bounding box. (c) Smoothed gradient image. (d) Segmentation with gradient threshold. (e) Snowless gradient image. (f) Segmentation augmented with background difference information.

Figure 7: MBARI *ROV Ventana*.

Figure 8: Block diagram for ROV-based jelly-tracking pilot aid

Figure 9: Images filmed during the 25 minute run tracking a *Solmissus* specimen. (a) Sample shot of specimen. (b) Overlay image combining 40 samples imaged at even intervals during the tracking run.