Detecting, Tracking and Classifying Animals in Underwater Video

Duane R. Edgington, Danelle E. Cline, Daniel Davis,
Ishbel Kerkez, Jérôme Mariette
Monterey Bay Aquarium Research Institute
7700 Sandholdt Road
Moss Landing, CA 95039, USA

Abstract **– For oceanographic research, remotely operated underwater vehicles (ROVs) and underwater observatories routinely record several hours of video material every day. Manual processing of such large amounts of video has become a major bottleneck for scientific research based on this data. We have developed an automated system that detects, tracks, and classifies objects that are of potential interest for human video annotators. By pre-selecting salient targets for track initiation using a selective attention algorithm, we reduce the complexity of multi-target tracking. Then, if an object is tracked for several frames, a visual event is created and passed to a Bayesian classifier utilizing a Gaussian mixture model to determine the object class of the detected event.**

I. INTRODUCTION

Remotely operated vehicles (ROVs) have revolutionized oceanographic research, supplanting traditional technologies of towed nets and acoustics as tools for the assessment of animal diversity, distribution and abundance [1]. Video equipment deployed on ROVs and underwater observatories enable quantitative video transects (QVTs) to be obtained in midwater and benthic habitats, providing high-resolution data at the scale of the individual organisms and their natural aggregation patterns [1, 2]. Compared to previous conventional methodologies, these underwater platforms are non-invasive and *in situ*, enabling a more accurate assessment of fragile gelatinous animals that are easily destroyed by sampling equipment and thus were until recently grossly under-sampled.

The great success of ROVs in collecting high-resolution video during dives and the promise of using underwater observatories encourages a search for ways that will optimize both the amount of information and knowledge gathered from underwater cameras. Currently, the application of this methodology to marine ecological research is hampered by our ability to process the tapes. Long-term time series are rare and invaluable [4, 5] as is the timely analysis of these data. The continued use of ROVs and growing use of fixed underwater observatories to collect long-term continuous observations offer potential for even more data, but that potential is constrained by the time and effort currently necessary to view and quantify data. Hence, there is tremendous potential benefit in automating portions of the analysis.

To overcome the bottleneck in analyzing ROV dive videos and to anticipate the emerging field represented by fixed underwater observatories, an automated system for detecting animals (events) visible in the videos is being developed. The videos are processed with an attentional selection algorithm [6] that has been shown to work robustly for target detection in a variety of natural scenes [7].

The candidate locations identified by the attentional selection module are tracked across video frames using linear Kalman filters [8]. If objects can be tracked successfully over several frames, they are stored as potentially 'interesting' events. Based on low-level properties, interesting events are identified and marked in the video frames. Interesting events are then processed by a classification module trained to classify specific animal categories.

The demonstrated attentional selection, tracking and classification modules are our first steps towards an integrated automated video annotation system. Our work enables follow-on development of automated ocean observatory cameras and automated processing of video from cameras on autonomous underwater vehicles.

II. METHODS

The automated system for detecting and tracking animals in underwater video consists of a number of steps which are outlined in fig.1. First, the foreground must be separated from the background for all video frames. Secondly, the first frame and every *p*th frame after that (typically, p=5) are processed with an attentional selection algorithm to detect salient objects. Detected objects that do not coincide with already tracked objects are used to initiate new tracks. Objects are tracked over subsequent frames, and their occurrence is verified in the proximity of the predicted location. Finally, detected objects are marked in the video frames.

A. Hardware

At MBARI, two ROVs are in use for deep sea exploration, the *ROV Ventana* and the *ROV Tiburon*, data analyzed come from dives of both of these ROVs. *ROV Ventana*, launched from *R/V Point Lobos*, used a Sony HDC-750 HDTV (1035i30, 1920x1035 pixels) camera for video data acquisition, and data were recorded on a DVW-A500 Digital BetaCam video tape recorder (VTR) onboard the *Point Lobos*. *ROV Tiburon* operated from *R/V Western Flyer* using a Panasonic WVE550 3-chip CCD (625i50, 752x582 pixels) camera, and video also recorded on a DVW-A500 Digital BetaCam VTR.

1-4244-0115-1/06/$20.00 ©2006 IEEE

On shore a Matrox RT.X10 video editing card in a Pentium P4 1.7 GHz personal computer (PC) running the Windows 2000 operating system was used to capture the video as AVI movie files at a resolution of 720 x 480 pixels and 30 frames per second. The video processing described below was performed on several state-of-the-art PCs running Red Hat Linux. All software development is done in C++ and Matlab under Linux. To be able to cope with the large amount of video data that needs processing at MBARI in a reasonable amount of time, we have deployed a computer cluster with 8 Rack Saver rs1100 dual Xeon 2.4 GHz servers, configured as a 16 CPU Gigabit Ethernet Beowulf cluster. The captured AVI movie clips are decomposed into single frames that can be processed by the special-purpose video analysis software.

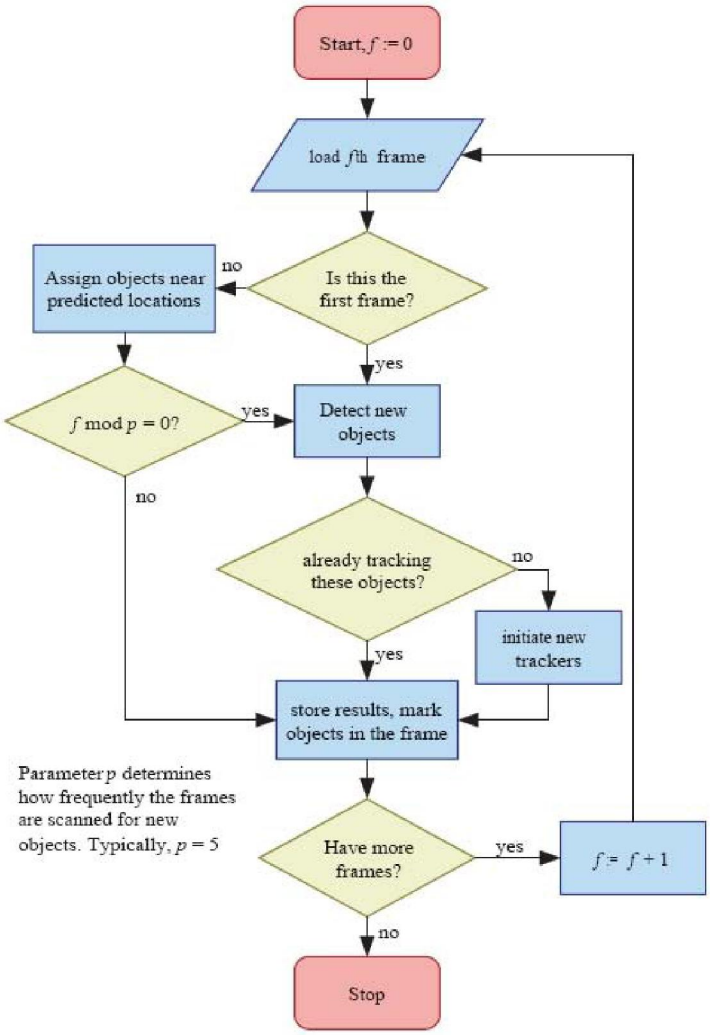

Figure1: Interactions between the various modules of our system for detecting and tracking marine animals in underwater video.

B. Extracting Foreground from Background: Determining the location of objects.

Like many computer vision applications, the foreground must be segmented from the background. Estimating the background is done differently for a fixed camera versus a moving camera.

Fixed camera

With fixed cameras that are used in underwater observatories a static background is obtained by calculating an average of the first few consecutive frames. This static background is then used to segment moving objects by comparing the background from the current frame using the graph cut algorithm [9]. This algorithm begins by building a graph based on the image. Each pixel in the image is linked by a weight to both the foreground and the background. This weight is directly derived from the difference between the current frame and the background at the corresponding pixel. Moreover, each pixel is linked to its four-connected neighboring pixels making the segmentation depend on values of a group of pixels, rather than on one individual value, leading to an output with larger clusters of homogeneity.

Once the graph is constructed, the task is to find the minimum cost involved in separating the source from the sink. Nodes still connected to the source are then flagged as foreground, while those connected to the sink are flagged as background.

Transect camera

Non-uniform lighting conditions cause luminance gradients that can be confusing to a contrast-based detection algorithm. All of these effects are constant over courses of a dive, and hence they can be removed by background subtraction (x, y and t are assumed to be discrete):

$$I'(x,y,t)=[I(x,y,t)-\frac{1}{(\Delta tb)}\sum_{t'=(t-\Delta tb)}^{t-1} I(x,y,t')]$$

where I(x,y,t) is the intensity at image location (x,y) at time t in the original image, and I' is the intensity after background subtraction. Only the non-negative part of the background subtraction is retained. This process is repeated separately for the red, green, and blue channels of the RGB color images.

C. Finding salient objects

With the localization of foreground objects known from the segmentation, locations of salient objects are given by using the model for saliency-based attention in humans by Itti & Koch [6 14]. In this model, each frame is decomposed into seven channels (intensity contrast, red/green and blue/yellow double color opponencies, and the four canonical, spatial orientations) at six spatial scales, yielding 42 "feature maps". After iterative spatial competition for salience within each map, only a sparse number of locations remain active, and all maps are combined into a unique "saliency map" (fig.2c). The saliency map is scanned by the focus of attention in order of decreasing saliency, through the interaction between a winner-take-all neural network (which selects the most salient location at any given time) and an inhibition-of-return mechanism (transiently suppressing the currently attended location from the saliency map). At the salient locations found, the centroids of the foreground objects are used to initiate tracking (see *section D*).

We found that oriented edges are the most important feature for detecting marine animals. In many cases, animals that are marked by human annotators have low contrast but

are conspicuous due to their clearly elongated edges (fig. 2a). In order to improve the performance of the attention system in detecting such faint yet clearly oriented edges, we use a normalization scheme for the orientation filters that is inspired by the lateral inhibition patterns of orientation-tuned neurons in the visual cortex. We compute oriented filter responses in a pyramid using steerable filters [10, 11] at four orientations. However, high-contrast "marine snow" particles that lack a preferred orientation often elicit a stronger filter response than faint string-like animals with a clear preferred orientation (fig. 2c). To overcome this problem, we normalize the response of each of the oriented filters is with the average of all of them:

$$O'_i(x,y) = [O_i(x,y) - \frac{1}{N}\sum_{j=1}^{N} O_j(x,y)]$$

where $O_i(x,y)$ denotes the response of the ith orientation filter ($1 \le i \le N$) at position (x,y), and $O'_i(x,y)$ is the normalized filter response (here, N=4). This across-orientation normalization significantly improved the detection of faint elongated objects (fig. 2d).

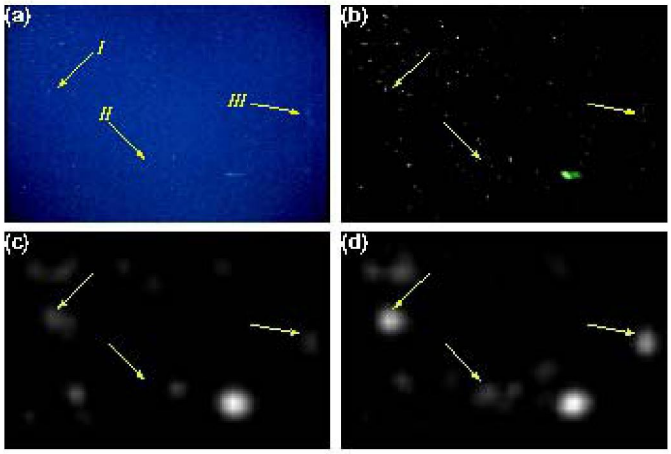

Figure 2: Example for the detection of faint elongated objects using across-orientation normalization. (a) original video frame with three faint, elongated objects marked; (b) the frame after background subtraction according to eq.1 (contrast enhanced for displaying purpose); (c) the orientation conspicuity map (sum of all orientation feature maps) without across-orientation normalization; (d) the same map with across-orientation normalization. Objects I and III have a very weak representation in the map *without* normalization (c), and object II is not represented at all. The activation to the right of the arrow tip belonging to object II in (c) is due to a marine snow particle next to the object, and not due to object II. Compare with (d), where object II as well as the marine snow particle create activation. In the map *with* normalization (d), all three objects have a representation that is sufficient for detection.

D. Tracking and discriminating visual events: Are these objects visual events ?

Having identified "interesting" objects in single frames, we attempt to track objects ("visual events") across frames. Assume that we have already tracked a number of visual events over the last few frames. We then face the problem of assigning the salient objects in the current frame to those tracked events. This is done by comparing each object's position with the expected position for each event, extrapolated from its positions in the past. Each object is assigned to the event that it matches best. If no match is found for an object, a new visual event is created with this object as its first data point.

Sometimes the algorithm fails to detect an object in a particular frame, but it locates it again in the next frame. In this case we interpolate the missing data point using the one before and the one after the respective frame. If events have not been assigned objects in two consecutive frames, we declare these events "closed", i.e. they are not considered for data assignment anymore. Closed events that could not be tracked for more than seven frames are discarded as noise.

Some of the tracked visual events are "boring" (particles of marine snow); others are "interesting" (animals).

Once we have established the visual events in the scene and deem which ones are "interesting", they can be marked in the video by drawing a boundary box for the objects into the frames. Occasionally, video has several seconds or even minutes in which no salient object appears. There is now the option of omitting those long stretches of "boring" video from being passed on to the annotators.

E. Towards Classification

Ultimately, it is our goal to not only detect and track salient objects, but also to classify them into biological taxonomies. This may seem very difficult, given that currently on the order of 400 species and families are being annotated routinely. However, in a recent study [12] it was determined with professional annotations between 1997 and 2002, the ten most common midwater animals correspond to 60%, and the 25 most common midwater animals to 80% of all annotations. Thus, if the 25 most common midwater animals were to be reliably recognized, many scientific missions concentrating on those most abundant animals could be automated.

This step for our system is still in its early development stage but is showing promising results. For each tracked object, a binary mask is obtained with segmentation algorithms described in II.b, allowing features to be extracted. The detected object is then classified using a Gaussian mixture model [13] of feature vectors based on Schmid invariants using a training set obtained with the help of professional annotators at MBARI.

Automated classification of benthic animals would be a powerful tool for determining their abundance and distribution. As a proof of concept test, a benthic image training data base was created with a total of about 6000 frames and 200 events. Grayscale square subimages containing an example of object classes were processed, local jets were used to extract features [13] and the training data was modeled with a mixture of Gaussians [13]. Classification tests were carried out using Matlab using three training classes: *Rathbunaster californicus, Parasticopus leukothete,* and "other" (containing sub-images representing others events *e.g.* flatfish, rockfish, particles, etc.). The test events were randomly sampled from objects detected by the

program. Events were taken from dive video that had not been used in the training and the performance of the classification module was examined. Each frame of an event was independently classified and the overall plurality for each event taken. Final classification was deemed the overall plurality for each event ("majority rule").

III. RESULTS

The attentional selection algorithm shows very promising results for single image NTSC frame-grabs obtained from midwater dives of ROVs (ROV *Tiburon* and *Ventana*). In single frames in which human observers could identify one or more animals, the most salient (first attended) location found by the attention algorithm coincides with animals in about 90% of cases. The processing of video clips shows similarly promising results [14].

A recent evaluation of system performance with clips obtained from ROVs on benthic dives show a strong positive correlation between professional annotations and successful detections made by the system. In a processed 45 minute segment of video, the program detected 489 out of a total 520 professionally annotated objects with a successful detection rate of 94%.

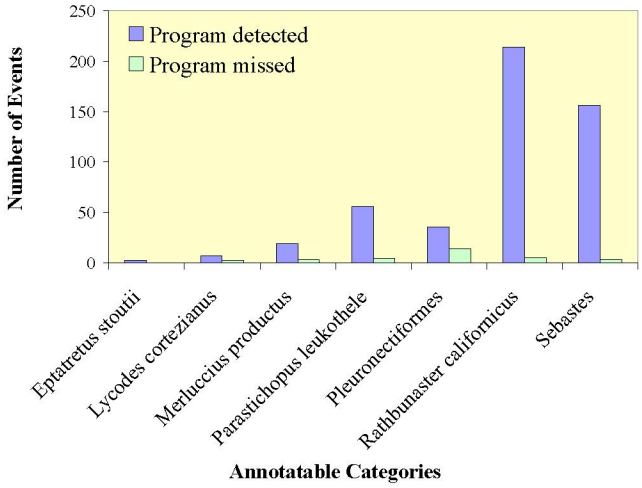

Figure 3: A comparison of automated event detections with professional human annotations for 45 minutes of processed benthic video.

The animal most commonly observed in the benthic transects was the echinoderm *Rathbunaster californicus* with a total of 219 recorded observations (fig. 3). The program was successful in detecting these animals 97.7% of the time.

Preliminary results of processing clips recorded by underwater observatories encourage further development of the system for this application (fig. 4). Video was provided by researchers responsible for a fixed underwater observatory off the Georgia coast [15] and from the *Eye-In-The-Sea* project [16]. While no statistics have been calculated, a high rate of detection and a low rate of false detection are evident in all videos processed so far.

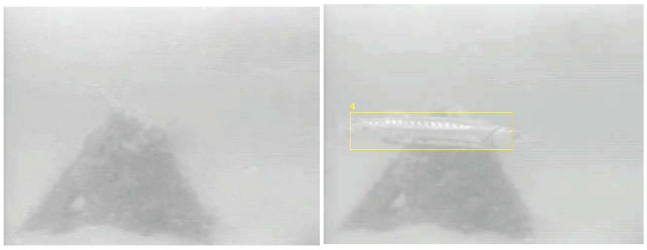

Figure 4: (a) empty frame recorded by an underwater observatory [15] (b) event (barracuda) flagged by the system.

As a first test of the classification module, we analyzed 7.5 minutes of benthic transect data from one ROV *Ventana* dive. We trained the classifier with grayscale square sub-images of segmented frames, where each sub-image contained an example object of some class (see METHODS IIE). For testing, we extracted 210 events detected by our system (about 7250 images).

The recognition module successfully classified 38 of 42 (90%) *Rathbunaster californicus* tagged by professional annotators (90% recall: the rate the system classified the object as a *Rathbunaster* and it was a *Rathbunaster*). There were no instances in which any other events were falsely classified as *Rathbunaster californicus* (100% precision: the rate the event was a *Rathbunaster* and the system classified it as such).

IV. DISCUSSION AND CONCLUSION

We report here our progress in developing a new method for processing video streams from underwater videos automatically. This technology has potentially significant impact on the work of video annotators by aiding the annotators in looking for noteworthy events in the videos. Eventually, we hope that the software will be able to perform a number of routine tasks fully automatically, such as "outlining" video, analyzing QVT videos for the abundance of certain easily identifiable animals, and marking especially interesting episodes in the videos that require the attention of the expert annotators. Most of this work can be done in an unsupervised, fully automated fashion. We are constantly improving our software towards this goal, and we are setting up the necessary hardware.

Additionally, it is recognized that if this method were to be applied to autonomous underwater vehicles (AUVs), the collection and analysis of quantitative video transects could be obtained much more frequently, at an ecologically significant finer spatial resolution and at a greater spatial range than is currently practical and economical for ROVs.

There is great benefit in automating portions of the analysis of video from fixed observatory cameras, where autonomous response to potential events (e.g. pan and zoom to events), and automated processing for science users of potentially largely "boring" video streams from 10s or even 1000s of network cameras could be key to those cameras being useful practical scientific instruments.

ACKNOWLEDGMENT

We thank the David and Lucille Packard Foundation for their generous support. This project originated at the 2002 Workshop for Neuromorphic Engineering in Telluride, CO, USA in collaboration with Dirk Walther, California Institute of Technology (Caltech), Pasadena, CA, USA. At MBARI, Karen A. Salamy and Dorothy Oliver provided technical assistance and Michael Risi provided engineering assistance; we thank Bruce Robison, Linda Kuhnz, Rob Sherlock, Nancy Jacobsen Stout, and Joshua Kroll for helpful discussion. We thank Marc'Aurelio Ranzato (New York University) and Pietro Perona (Caltech) for providing the code for the classifier, Laurent Itti (University of Southern California) for providing the *iLab Neuromorphic Vision C++ Toolkit*, Dirk Walther for engineering the detection and tracking modules, and Nicholas Howe (Smith College) for Foreground / Background segmentation code.

REFERENCES

[1] Robinson, B.H. 2000. *The coevolution of undersea vehicles and deep-sea research*. Marine echnology Society Journal, **33**:p. 69-73.

[2] Robinson, B.H., K.R. Reisenbichler, R.E. Sherlock, J.M.B. Silguero, and F.P. Chavez, 1998. *Seasonal abundance of the siphonophore, Nanomia bijuga, in Monterey Bay*. Deep-Sea Research II, **45**:p. 1741-1752.

[3] Siguero, J.M.B. and B.H. Robinson. 2000. *Seasonal abundance and vertical distribution of mesopelagic Calycophoran siphonophores in Monterey Bay, CA*. Journal of Plankton Research, **22**:p. 1139-1153.

[4] Estes, J.A. and C.H. Peterson. 2000. *Marine ecological research in seashores and seafloor systems: Accomplishments and future directions*. Marine Ecology Progress Series, **195**: p. 281-189.

[5] Southward, A.J. 1992. *The importance of long time-series in understanding the variability of natural systems. In Internatioanal Helgoland Symposium "The Challenge to Marine Biology in a Changing World"*. Helgoland, Hamburg, Germany.

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

[7] Itti, L. and C. Koch. 1999. *Target Detection using Saliency-Based Attention. In Proc. RTO/SCI-12 Workshop on Search and Target Acquisition (NATO Unclassified)*. Utrecht, The Netherlands.

[8] Kalman, R.E. And R.S. Bucy, 1961. *New Results in Linear Filtering and Predicting Theory*. Journal of Basic Engineering, **83**(3): p. 95-108.

[9] Howe, N.R. And Deschamp A, *Better Foreground Segmentation Through Graph Cuts*. Smith College. Northampton, MA, unpublished. http://maven.smith.edu/~nhowe/research/pubs/bgsub.pdf

[10] E.P. Simoncelli and W.T. Freeman. *The steerable pyramid: a flexible architecture for multi-scale derivative computation*. In IcIP, 1995.

[11] R. Manduchi, P. Perona, and D. Shy. *Efficient implementation of deformable filter banks. IEEE Transactions on Signal Processing*, **46**(4): 1168-1173, 1998.

[12] A. Wilson and D.R. Edgington. 2003. First steps towards autonomous recognition in Monterey Bay's most common mid-water organisms: Mining the ROV video database on behalf of the Automated Visual Event Detection (AVED) system. Technical report, MBARI, unpublished.

[13] Ranzato, M.A. 2004. Automatic visual recognition of biological particles. California Institute of Technology: Pasadena, CA. unpublished. http://vision.caltech.edu/ranzato/

[14] Walther, D., D.R. Edgington, and C.Koch. 2004.Detection and Tracking of Objects in Underwater Video. In IEEEE International Conference on Computer Vision and Pattern Recognition (CVPR). Washington, D.C.

[15] Barans, C.A., M.D. Arendt, T. Moore, and D. Schmidt. 2005. Remote video revisited: *A visual technique for conducting long-term monitoring of reef fishes on the continental shelf*. Marine Technology Society Journal, **39** (2): 80 - 88.

[16] E.A. Widder, B.H. Robison, K.R. Reisenbichler and S.H.D. Haddock. 2005. *Using red light for in situ observations of deep-sea fishes*. Deep-Sea Research I, **52**: 2077–2085.