Detection and Tracking of Objects in Underwater Video – Paper # 569

Authors omitted for the review process.

Abstract

For oceanographic research, remotely operated underwater vehicles (ROVs) routinely record several hours of video material each 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 and tracks objects that are of potential interest for human video annotators. By pre-selecting salient targets during track initiation using a selective attention algorithm, we reduce the complexity of multi-target tracking and data assignment during tracking. Detection of low-contrast translucent targets is made difficult by variable lighting conditions and the presence of ubiquitous noise from high-contrast "marine snow" particles. We describe the methods we developed to overcome these issues and report our results of processing ROV video data.

1 Introduction

High-resolution video equipment on ocean-going remotely operated vehicles is used by ocean scientists to obtain quantitative data on the distribution and abundance of oceanic animals [1]. High-quality video data supplants the traditional tow net approach of assessing the kinds and numbers of animals in the oceanic water column. Video camera-based quantitative video transects (QVT) are taken through the ocean midwater, from 50 m to 1000 m depths, providing high-resolution data at the scale of the individual animals and their natural aggregation patterns for animals and other objects larger than about 2 cm in length [2, 3, 4]. QVTs are superior to tow nets in assessing the spatial distribution of animals, and in recording delicate gelatinous animals that are destroyed in nets. However, the current manual method of analyzing QVT video is labor intensive and tedious. The approach is limited by the volume of ROV data that needs to be analyzed manually, and by the length and depth increments of QVTs that are practical to analyze as well as the sampling frequency that are practical with ROVs.

The oceanographic research motivation for our work is an order-of-magnitude shift in (1) lateral scale from current 0.5 km to mesoscale levels (5 km); (2) depth increment, from current 100 m to biologically significant 10 m scale; and (3) sampling frequency, from currently monthly to daily, which is the scale of dynamic biological processes. Such an increase in data would enable modeling of the linkage between biological processes and physicochemical hydrography.

2 Algorithms

We have developed an automated system for detecting and tracking animals visible in ROV videos. This task is difficult due to the low contrast of many of the marine animals, their sparseness in space and time, and due to debris ("marine snow") cluttering the scene, which shows up as ubiquitous high contrast clutter in the video.

Our system consists of a number of sub-components whose interactions are outlined in fig. 1. The first step for all video frames is the removal of background. Next, the first frame, and every pth 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.

2.1 Background subtraction

Images captured from the video stream often contain artefacts such as lense glare, parts of the camera housing, parts of the ROV or instrumentation. Also, non-uniform lighting conditions cause luminance gradients that can be confusing. All of these effects have the characteristic that they are constant over short periods of time. Hence we can remove them by background subtraction (x, y and t are assumed to be discrete):

$$I'(x,y,t) = \left[I(x,y,t) - \frac{1}{\Delta t_b} \sum_{t'=(t-\Delta t_b)}^{t-1} I(x,y,t') \right]_+ \quad (1)$$

where I is the image intensity before, and I' after background subtraction. This process is repeated separately for the R, G, and B channels of the color images.

The value of Δt_b should be larger than the typical dwell time of objects in the same position in the camera plane. In our transect videos objects typically move fast. We found

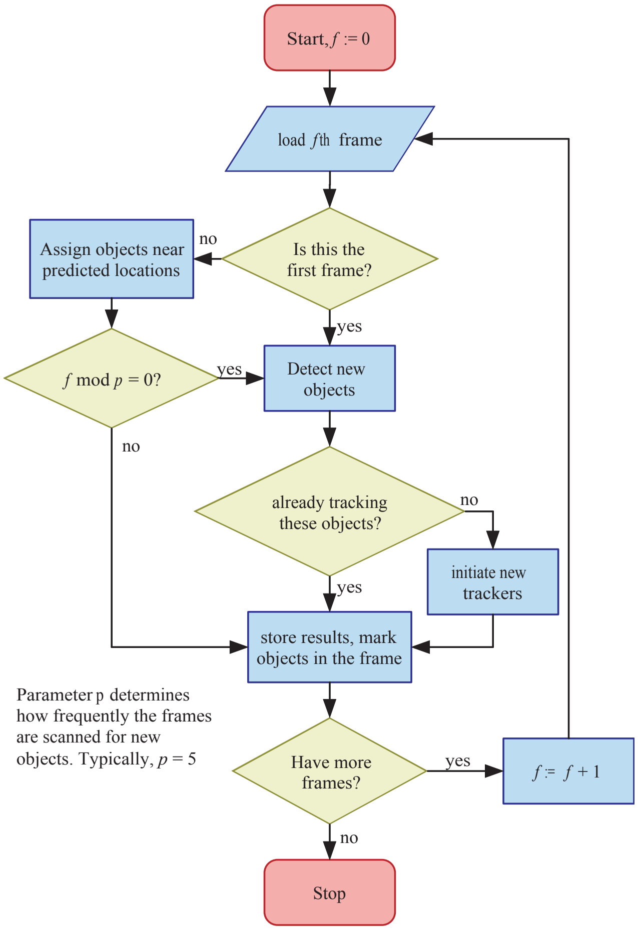

Figure 1: Interactions between the various modules of our system for detecting and tracking marine animals in underwater video. For details of the individual modules please refer to the main text.

that $\Delta t_b = 0.33\ s$ (10 frames) works quite well, giving us enough flexibility to adjust to changes in the artefacts quickly (fig. 2b).

2.2 Detection

For the detection of new objects we use the saliency-based bottom-up attention system by Itti & Koch [5] that has been shown to work well for a variety of applications [6, 7].

Following background subtraction, input frames are 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 this way objects are segmented from the image, and their centroids are used to initiate tracking (see section 2.3).

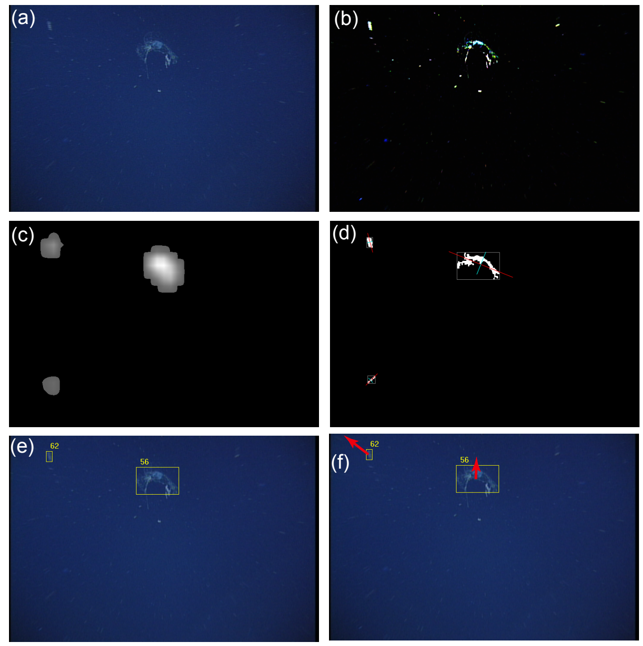

Figure 2: Processing steps for detecting objects in video frames. (a) original frame (720×480 pixels, 24 bits color depth); (b) after background subtraction according to eq. 1 (contrast enhanced for displaying purpose; (c) saliency map for the preprocessed frame (b); (d) detected objects with bounding box and major and minor axes marked; (e) the detected objects marked in the original frame and assigned to tracks; (f) direction of motion of the object obtained from eq. 11.

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 clear elongated edges (fig. 3a). In order to improve the performance of the attention system in detecting such faint yet clearly oriented edges, we used a normalization scheme for the orientation filters.

We compute oriented filter responses in a pyramid using steerable filters [8] 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. 3c). To overcome this problem, we normalize the response of each

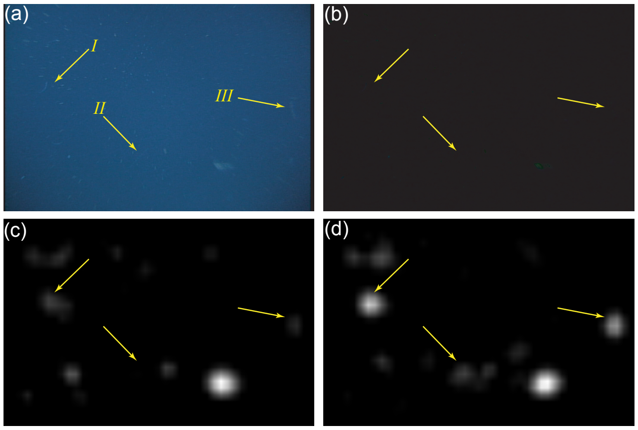

Figure 3: Example for the detection of faint elongated objects using across-orientation normalization. (a) original video frame with three faint, elongated objects marked; (b) the frame after background subtraction (eq. 1); (c) the orientation conspicuity map (sum of all orientation feature maps) without 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.

of the oriented filters with the average of all of them:

$$O'_i(x,y) = \left[O_i(x,y) - \frac{1}{N} \sum_{j=1}^{N} O_j(x,y) \right]_+ \tag{2}$$

where $O_i(x,y)$ denotes the response of the ith orientation filter ($1 \leq i \leq N$) at position (x,y), and $O'_i(x,y)$ is the normalized filter response (here, $N = 4$). This across-orientation normalization leads to a clear improvement in detecting faint elongated objects (fig. 3d).

2.3 Tracking

Once objects are detected, we extract their outline and track their centroid across the image plane using separate linear Kalman filters to estimate the x and the y coordinates.

During QVTs the ROV is driven through the water column at constant speed. While there are some animals that propel themselves with a speed that is comparable to or faster than the speed of the ROV, most objects either float in the water passively, or they move on a much slower time scale than the ROV. Hence, we can approximate that the camera is moving at a constant speed through a group of stationary objects.

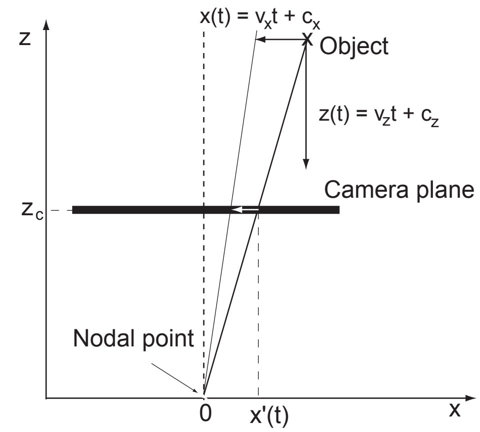

Figure 4: Geometry of the projection problem in the camera reference frame. The nodal point of the camera is at the origin, the camera plane is at z_c. The object appears to be moving at a constant speed into the x and z direction as the camera moves towards the object. Eq. 3 describes how the projection of the object onto the camera plane moves in time.

Fig. 4 illustrates the geometry of the problem in the reference frame of the camera. In this reference frame, the object is moving at a constant speed in the x and z directions. The x coordinate of the projection onto the camera plane is:

$$x'(t) = \frac{x(t) \cdot z_c}{z(t)} = \frac{v_x z_c \cdot t + c_x z_c}{v_z \cdot t + c_z} \tag{3}$$

To a second order approximation, the dynamics of this system can be described by a model that assumes constant acceleration:

$$\frac{d\underline{x}}{dt} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \underline{x} \tag{4}$$

with

$$\underline{x} = \begin{pmatrix} x' \\ v \\ a \end{pmatrix} \tag{5}$$

which results in a fundamental matrix that relates $\underline{x}(t)$ to $\underline{x}(t+\tau)$:

$$\Phi(\tau) = \begin{bmatrix} 1 & \tau & \frac{\tau^2}{2} \\ 0 & 1 & \tau \\ 0 & 0 & 1 \end{bmatrix} \tag{6}$$

$\Phi(\tau)$ is used to define a linear Kalman filter [9, 10]. The deviations of this simplified dynamics from the actual dynamics are interpreted as process noise Q. The resulting iterative equations for the Kalman filter are:

$$M_k = \Phi_k P_{k-1} \Phi_k^T + Q \tag{7}$$
$$K_k = M_k H^T \left(H M_k H^T + R_k \right)^{-1} \tag{8}$$
$$P_k = (I - K_k H) M_k \tag{9}$$

where P_0 is initialized to a diagonal matrix with large values on the diagonal, $R = [\sigma_m^2]$ is the variance of the measurement noise, $H = [1\ 0\ 0]$ is the measurement matrix, relating the measurement x^M to the state vector $\underline{x}$, and Q is the process noise matrix:

$$Q(\tau) = \sigma_p^2 \begin{bmatrix} \frac{1}{20}\tau^5 & \frac{1}{8}\tau^4 & \frac{1}{6}\tau^3 \\ \frac{1}{8}\tau^4 & \frac{1}{3}\tau^3 & \frac{1}{2}\tau^2 \\ \frac{1}{6}\tau^3 & \frac{1}{2}\tau^2 & \tau \end{bmatrix} \tag{10}$$

where σ_p^2 is the variance of the process noise. For our particular tracking problem we found $\sigma_m^2 = 0.1$ and $\sigma_p^2 = 10$ to be convenient values.

Once the Kalman matrix K_k is obtained from eq. 8, an estimation for $\underline{x}_k$ can be computed from the previous state estimate $\hat{\underline{x}}_{k-1}$ and the measurement x_k^M:

$$\hat{\underline{x}}_k = \Phi_k \hat{\underline{x}}_{k-1} + K_k (x_k^M - H \Phi_k \hat{\underline{x}}_{k-1}) \tag{11}$$

When no measurement is available, a prediction for $\underline{x}_k$ can be obtained as:

$$\hat{\underline{x}}_k = \Phi_k \hat{\underline{x}}_{k-1} \tag{12}$$

For the initiation of the tracker we set $\hat{\underline{x}}_0 = [x_0^M\ 0\ 0]^T$, where x_0^M is the coordinate of the object's centroid obtained from the saliency-based detection system described in the previous section.

We employ the same mechanism to track the y coordinate of the object in the camera plane. Whenever the x or the y coordinate tracker runs out of the camera frame, we consider the track finished and the corresponding object lost. Re-entry of objects into the camera frame almost never occur in our application. We require that an object can be successfully tracked over at least five frames, otherwise we discard the measurements as noise.

In general, we are tracking multiple objects all at once. Normally, multi-target tracking raises the problem of assigning measurements to the correct tracks [11]. Since the attention algorithm only selects the most salient objects, however, we obtain a sparse number of objects whose predicted locations are usually separated far enough to avoid ambiguities. If ambiguities occur, we resolve them using a measure that takes into account the Euclidean distance of the detected objects from the predictions of the trackers and the size ratio of the detected and the tracked objects.

2.4 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 difficuls, given that currently on the order of 400 species and families are being annotated routinely. However, in a recent study [reference omitted for the review process] we determined that in the human annotations between 1997 and 2002, the ten most common animals correspond to 60%, and the 25 most common animals to 80% of all annotations. Thus, if we succeed in recognizing the 25 most common animals reliably, we can automate many scientific missions that concentrate on those most abundant animals. To date, our experiments using low-level features such as major and minor axes, aspect ratio, total area size, maximum and average luminance over the shape of the object to classify detected objects into the broad categories "interesting" and "not interesting" are inconclusive. The classification into species remains an open issue for us.

2.5 Implementation

We use two ROVs for deep sea exploration, the ROV Ventana and the ROV Tiburon [12, 13]. ROV Ventana, launches from R/V Point Lobos, uses a Sony HDC-750 HDTV (1035i30, 1920x1035 pixels) camera for video data acquisition, and the data are recorded on a DVW-A500 Digital BetaCam video tape recorder (VTR) on board the R/V Point Lobos. ROV Tiburon operates from R/V Western Flyer; it uses a Panasonic WVE550 3-chip CCD (625i50, 752x582 pixels) camera, and video is also recorded on a DVW-A500 Digital BetaCam VTR. On shore, a Matrox RT.X10 and a Pinnacle Targa 3000 Serial Digital Interface video editing card in a Pentium P4 1.7 GHz personal computer (PC) running the Windows 2000 operating system and Adobe Premier are used to capture the video as AVI or QuickTime movie files at a resolution of 720 x 480 pixels and 30 frames per second. All software development is done in C++ under Linux. To be able to cope with the large amount of video data that needs processing 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. After converting frames to Netpbm color image format, we currently process approximately three frames per second on each of the Xeon nodes at a resolution of 720x480 pixels. We are optimistic that additional hardware and software optimization will enable us to process 30 frames per second in real time soon. See the supplementary material for an example video before and after processing with our system.

3 Results

We present two groups of results – an assessment of the attenional selection algoithm for our purpose, and a comparison of a processed 10 min video clip with expert annotations.

3.1 Single frame results

In order to assess the suitability of the saliency-based detection of animals in video frames in the early stage of our project, we did an analysis of a number of single video frames. We captured the images from a typical video stream at random. We did this analysis for two images sets – one with 456 images from video recorded by ROV Tiburon on June 10, 2002, and one with 1004 images from video recorded by ROV Ventana on June 18, 2002. Only some of the images in the image sets contain animals. We used the attentional detection system described in section 2.2 to evaluate its performance on these images. We counted in how many of the images the most salient location, i.e. the location first attended to by the algorithm, coincides with an animal. The results are displayed in table 1.

Table 1: Single frame analysis results.

	Image set 1	Image set 2
Date of the dive	06/10/2002	06/18/2002
ROV used for the dive	Tiburon	Ventana
Number of images obtained	456	1004
Images without animals	205	673
Images with detected animals	224	291
Images with missed animals	27	40

In the images that did not contain animals, the saliency mechanism identified other visual features as being the most salient ones, usually particles of marine snow. Originally, the system had no ability to identify frames that did not contain any objects of interest. We introduced this concept when we implemented tracking of objects (see section 2.3). For the majority of the images that did contain animals, the saliency program identified the animal (or one of the animals, if more than one were present) as the most salient location in the image. In image set 1, the animals were identified as the most salient objects in 89% of all images that contained animals. In image set 2 this was the case for 88% of the images with animals. Ground truth was established by individual inspection of the test images.

3.2 Video processing

As a test of the video processing capabilities of our system, we processed a 10 min video segment that had been annotated by scientists before. The scientists annotated 57 object in this video clip. Our system detected 40 of those. In two cases, our automated detection system detected animals that the annotators had missed. The program also detected several other objects that the scientists had not annotated. However, we did not consider these cases as false positives because a classification system capable of recognizing species would also be able to distinguish target animals from salient debris.

Table 2: Missed species.

Annotation	# of misses	Remark
Nanomia	14	sub-class Siphonophora
Siphonophora	1	faint, string-like animal
Calycophora	1	sub-class Siphonophora
Solmissus	1	Narcomedusa: small jellyfish

Our system missed 17 of the 57 annotated objects in the video clip. Table 2 shows details about the species that were missed. As can be seen, almost all misses are of the subclass Siphonophora, which are long string-like colonial animals that yield very low contrast in the video frames. In further improvements of the system we plan to specifically address detection of these objects in the video stream.

4 Discussion

We have presented a new method for processing video streams from ROVs automatically. This technology has potentially significant impact on the daily work of video annotators by aiding their analysis of noteworthy events in the videos. After continued refinement, 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.

Beyond its applications to ROV videos, our method for automated underwater video analysis may potentially have a larger impact by enabling Autonomous Underwater Vehicles (AUVs) to collect and analyze quantitative video transects, with the potential to sample more frequently and at an ecologically significant finer spatial resolution and greater spatial range than is practical and economical for ROVs [14]. We also see 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 very sparse video streams from 100s or 1000s of network cameras could be key to those cameras being practical scientific instruments.

Acknowledgments

Omitted for the review process.

References

[1] T. Clarke. Robots in the deep. *Nature*, 421(30):468–470, 2003.

[2] B.H. Robison. The coevolution of undersea vehicles and deep-sea research. *Marine Technology Society Journal*, 33:69–73, 2000.

[3] B.H. Robison, K.R. Reisenbichler, R.E. Sherlock, J.M.B. Silguero, and F.P. Chavez. Seasonal abundance of the siphonophore, nanomia bijuga, in monterey bay. *Deep-Sea Research II*, 45:17411752, 1998.

[4] J.M.B. Silguero and B.H. Robison. Seasonal abundance and vertical distribution of mesopelagic calycophoran siphonophores in monterey bay, ca. *Journal of Plankton Research*, 22:1139–1153, 2000.

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

[6] L. Itti and C. Koch. Target detection using saliency-based attention. In *Proc. RTO/SCI-12 Workshop on Search and Target Acquisition (NATO Unclassified)*, pages 3.1–3.10, Utrecht, The Netherlands, 1999.

[7] F. Miau, C. Papageorgiou, and L. Itti. Neuromorphic algorithms for computer vision and attention. In *SPIE 46 Annual International Symposium on Optical Science and Technology*, volume 4479, pages 12–23, 2001.

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

[9] R.E. Kalman and R.S. Bucy. New results in linear filtering and prediction theory. *Journal of Basic Engineering*, 83(3):95–108, 1961.

[10] P. Zarchan and H. Musoff. *Fundamentals of Kalman filtering: a practical approach*. Progress in astronautics and aeronautics. American Institute of Aeronautics and Astronautics, Inc., 2000.

[11] T. Kirubarajan, Y. Bar-Shalom, and K.R. Pattipati. Multiassignment for tracking a large number of overlapping objects. *IEEE Transactions on Aerospace and Electronic Systems*, 37(1):2–21, 2001.

[12] J. B. Newman and D. Stakes. Tiburon, development of an rov for ocean science research. In *Proceedings MTS/IEEE Oceans*, Brest, France, 1994.

[13] E. Mellinger, A. Pearce, and M. Chaffey. Distributed multiplexers for an ROV control and data system. In *Proceedings MTS/IEEE Oceans*, Brest, France, 1994.

[14] K. Smith. NEPTUNE science white paper #6: Deep-sea ecology. `http://www.neptune.washington.edu/science_wrkgp/127_deep_sea_final.html`, June 18, 2002 2002.