Automated Event Detection in Underwater Video

Duane R. Edgington
Karen A. Salamy
Michael Risi
R. E. Sherlock

Monterey Bay Aquarium Research Institute (MBARI)
7700 Sandholdt Road
Moss Landing, California 95039, USA
{duane, salamy, mrisi, robs}@mbari.org

Dirk Walther
Christof Koch

California Institute of Technology
Computation and Neural Systems Program
Pasadena, California 91125, USA
{walther, koch}@caltech.edu

Abstract – We present an attentional selection system for processing video streams from remotely operated underwater vehicles (ROVs). The system identifies potentially interesting visual events spanning multiple frames based on low-level spatial properties of salient tokens, which are associated with those events and tracked over time. If video frames contain interesting events, they are labeled "interesting", otherwise they are labeled "boring". By marking the "interesting" events and omitting "boring" frames in the output stream, we augment the productivity of human video annotators, or, alternatively, provide input for a subsequent object classification algorithm.

I. INTRODUCTION

For more than a century, the traditional approach for assessing the kinds and numbers of animals in the oceanic water column was to tow nets behind ships. However, tows have many limitations, including spatial resolution (averaging over the entire length of tow), and selectivity (e.g. gelatinous animals are typically destroyed and hence under-sampled). Today, remotely operated underwater vehicles (ROVs) provide an excellent alternative to trawls for obtaining quantitative data on the distribution and abundance of oceanic animals [1].

Using video cameras, it is possible to make quantitative video transects (QVTs) through the water, providing high-resolution data at the scale of the individual animals and their natural aggregation patterns [2, 3]. The great success of ROVs in collecting high-resolution video during dives encourages us to look for ways to increase the amount of information and knowledge we gather from underwater cameras. The current manual method of analyzing QVT video by trained scientists is very labor intensive and poses a serious limitation to the amount of data that can be obtained from ROV dives. The limiting factor in applying this methodology to marine ecological research is 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 future use of autonomous underwater vehicles (AUVs) for QVTs offer potential for even more data, but that potential is constrained by the time and effort currently necessary to view and quantify tape data. Hence, we see tremendous potential benefit in automating portions of the analysis.

To overcome the bottleneck in analyzing ROV dive videos we are developing an automated system for detecting animals (events) visible in the videos. This task is difficult due to the low contrast of many translucent animals and due to debris ("marine snow") cluttering the scene. We are processing the videos 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 (tokens) identified by the attentional selection module are combined across video frames using token tracking. If tokens 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.

Especially in deep water video (below 100 meters), visible animals are often sparse in space and time. By detecting whether or not there is an "interesting" candidate object for an animal present in a particular sequence of underwater video, we have developed a notion of "boring" video frames – video frames that do not contain any "interesting" events. By omitting "boring" frames and marking candidate objects, we aim to enhance the productivity of human video annotators and/or cue a subsequent object classification module.

II. METHODS

We use a variety of devices and algorithms in order to record and process videos of underwater scenes. In the following section we describe the hardware and the algorithms that we use.

A. Hardware

At MBARI, we use two ROVs for deep sea exploration, the *ROV Ventana* and the *ROV Tiburon* [8-10]. *ROV Ventana*, launched 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) onboard the

We thank the David and Lucile Packard Foundation for their generosity in funding work at MBARI. D.W. and C.K. are funded by the NSF Center for Neuromorphic Systems Engineering at Caltech. We thank the NSF Research Coordination Network (RCN) Institute for Neuromorphic Engineering (INE) for support of collaborative travel.

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 video editing card in a Pentium P4 1.7 GHz personal computer (PC) running the Windows 2000 operating system is 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 is performed on several state-of-the-art PCs running Red Hat Linux. All software development is done in C++ 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.

B. Pre-processing

For the video analysis, the frames first undergo a number of pre-processing steps. The frames contain visible horizontal scan lines, introduced by analog components in the chain of video processing equipment. To smooth the scan lines we convolve the red, green and blue channels of each frame with a one-dimensional Gaussian kernel of length 5 in the vertical direction.

The videos frequently contain man-made features such as the edge of the camera housing, parts of the ROV that may be visible, or even the brightness gradients introduced by the ROV's onboard lights. These features are usually constant in time over at least a few seconds. We estimate these constant background features for each frame by computing the sliding average over the ten preceding frames. We subtract the average from the frame, setting negative values to zero. An example for the effect of these pre-processing steps can be seen in Fig. 1b, compared with the original frame in Fig. 1a.

C. Finding salient objects

After pre-processing, each frame is scanned for salient locations using the model for saliency-based attention in humans by Itti & Koch [6]. For 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 different spatial scales. We compute center-surround contrasts across scales – similar to those found in the human retina and primary visual cortex, yielding six "feature maps" for each of the seven features. In a series of non-linear normalization and summation steps, the feature maps are combined into one "saliency map", in which only a sparse number of salient locations remain active (Fig. 1c). In a winner-take-all neural network these locations compete for saliency, yielding one winner. This winning location is then inhibited (inhibition of return), and the competition continues, thus producing a sequence of the most salient locations in the frame. An example for the resulting scan path is shown in Fig. 1d.

Naturally, we are not only interested in where salient objects are in the image, but also how large they are, their aspect ratios, and other properties of the objects. We can extract this information from a binary version of the image. We first apply binary morphology filters to the image, removing artifacts, which are difficult to avoid when thresholding the original image. We then identify the largest clusters of pixels at the salient locations, yielding a mask for the location, size and extend of the objects there (Fig. 1e).

From these masks we can derive a number of object properties, namely the area occupied by the object, the maximum, minimum and average intensity in the original image within the object mask, the object's centroid and bounding box. We can also derive the second moments u_{xx}, u_{yy} and u_{xy}, which are measures for the variance of the pixels around the x and y axes that intersect in the centroid, and for the covariance of the pixels with respect to the centroid. By approximating the object with an ellipse, we derive the length and orientation of the major and the minor axes of the object from the second moments. From the length of the major and minor axes we derive an estimate for the aspect ratio of the object. This set of properties will later serve as a basis for object classification.

D. Tracking and discriminating 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 does find 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). Naturally, it depends on the objective of the observer whether a particular event is interesting or not.

We are learning from the human annotators with their long experience and training what they consider interesting and what not. We ask annotators to view video clips in which the algorithm has marked all potentially "interesting" events. Using the annotators' feedback we can learn by example a notion of how "interesting" an event is, using the medium-level object properties as explained in Section IIC. At the time of writing this paper, this part of our work is still in

progress. As a preliminary version of the discrimination process we use a method of simply thresholding property values that are extracted from the binary object masks.

Once we have established the visual events and made the decision which ones are "interesting", we can mark them in the video by drawing the boundary box for the objects into the frames (Fig. 1f). Occasionally, video has several seconds or even minutes in which nothing salient appears. We now also have the option of omitting those long stretches of "boring" video from being passed on to the annotators.

III. RESULTS

In order to assess the suitability of the saliency-based de-

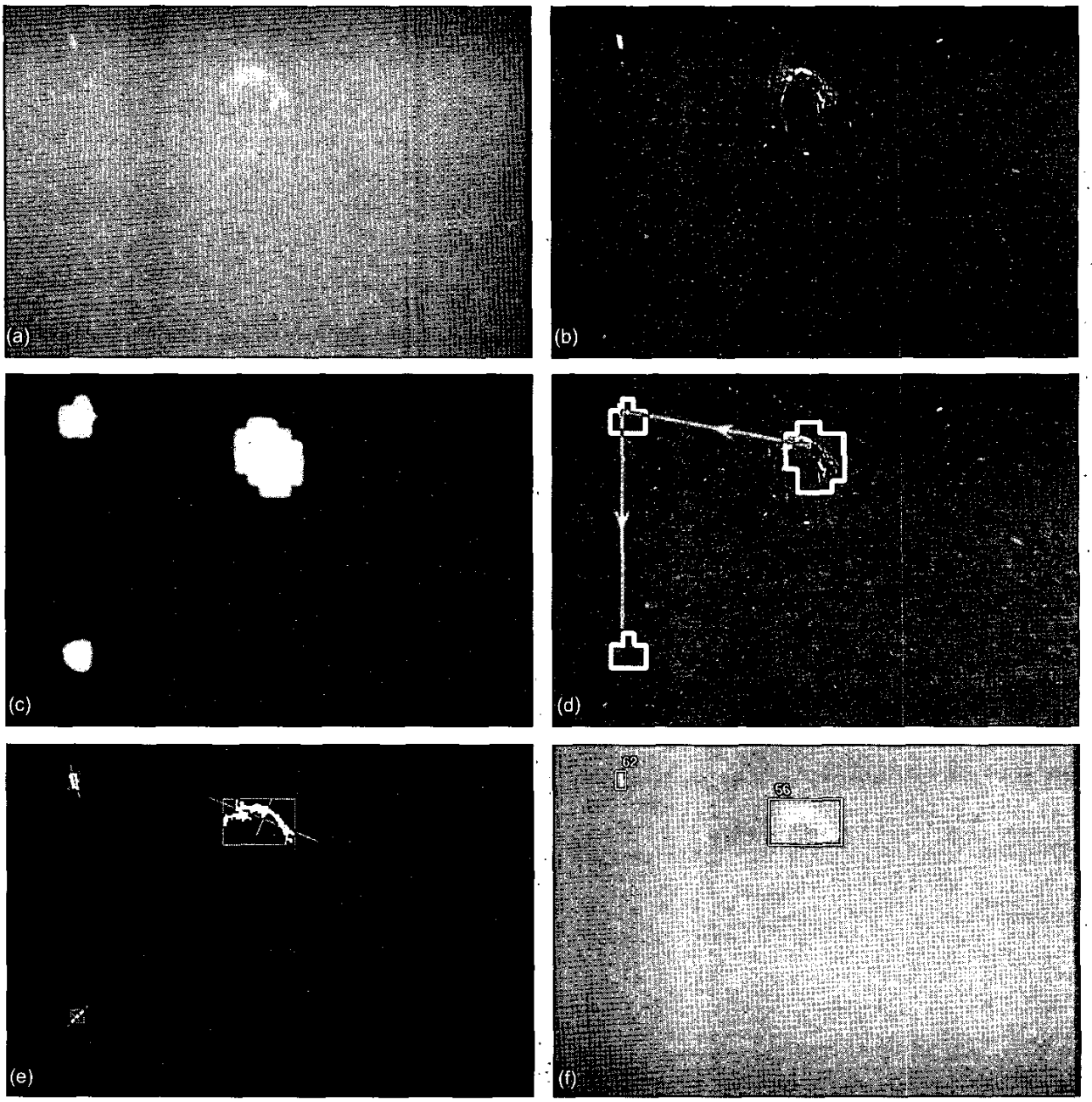

Fig. 1. Processing steps for the automated video analysis. (a) video frame captured by the PC with video editing card; (b) the result of the pre-processing. In this image, scan lines were smoothed out, and the average of the preceding ten frames was subtracted. The image contrast is enhanced for illustration purposes; (c) saliency map computed from image (b); (d) scan path of the saliency model overlaid on top of the image; (e) object masks that were extracted at the salient location with the boundary box (in gray) and the major (red) and minor (light blue) axes marked; (f) video frames with the objects marked that are part of "interesting" events.

tection of animals in video frames in the early stage of our project, we did an analysis of a number of single video frames. We further show results from processing video clips and compare our findings with the feedback from video annotators.

A. Single frame results

We obtained the single frames for this analysis as described in section II.A. Instead of saving AVI videos we captured single images from the video stream at random. We did this analysis for two images set – 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 contained animals, the others did not. We used the saliency-based detection program by Itti & Koch [6] 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 model, coincides with an animal. The results are displayed in table I.

TABLE I
SINGLE VIDEO FRAME ANALYSIS RESULTS

	Image set 1	Image set 2
Date of the dive	June 10, 2002	June 18, 2002
ROV used for the dive	ROV Tiburon	ROV 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 program identified other visual features as being the most salient ones, usually particles of marine snow. Originally, the system had no ability to distinguish "interesting" from "boring" images. We introduced this concept later in our work.

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 (Fig. 2). In 89% of all images in image set 1 that contained animals, the animals were identified as the most salient object in the image. In image set 2 this was the case for 88% of the images with animals.

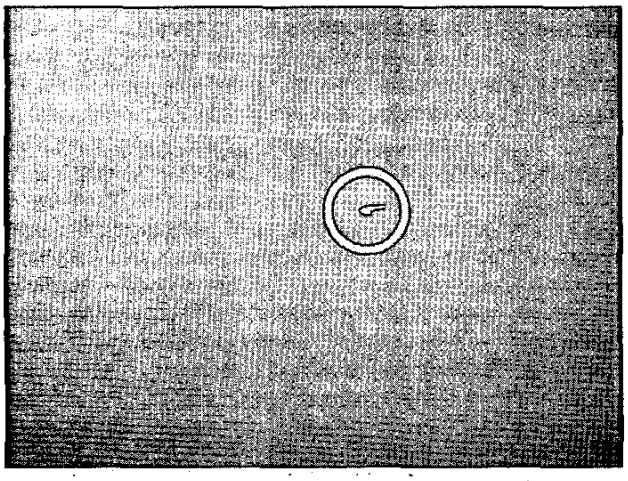

Fig. 2. Example of a single video image with the detected animal marked by the saliency program

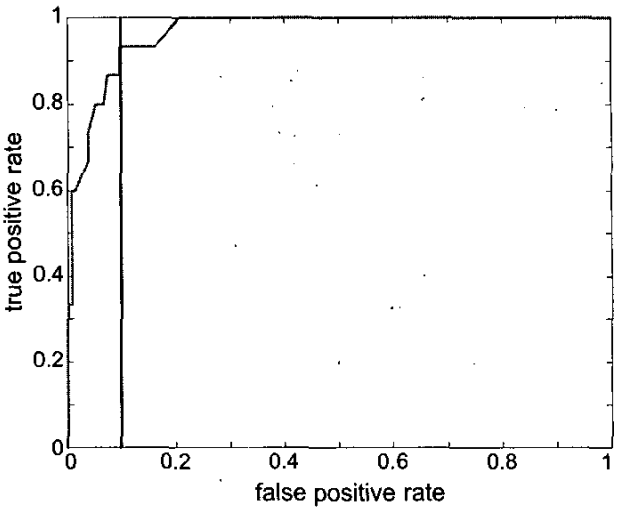

Fig. 3. Receiver Operating Characteristics (ROC) for the discrimination between "interesting" and "boring" events based on the covariance u_{xy}. The blue line marks the best performance at a threshold of $|u_{xy}| = 0.4$.

B. Video processing results

These promising results lead us to develop our video annotation software on the basis of the saliency-based attention model as described in section II.

A crucial part of our analysis is the discrimination between "interesting" and "boring" visual events in the underwater video. To evaluate our approach to this problem we captured ten video clips of ten seconds (300 frames) length each from video material recorded during a midwater dive by *ROV Ventana* on April 30, 1999 in the Monterey Bay[1]. We pre-processed the videos according to the procedure detailed in section II, marked all potentially "interesting" events in the video frames (see Fig. 1f), and stored the medium-level properties for the objects (see Fig. 1e). A video annotator was then asked to decide which of the events they would find "interesting", and which ones not.

We examined the object properties of the events that the annotator marked as "interesting" and found that the covariance of the object pixels with respect to the centroid u_{xy} is a good indicator for the how "interesting" an event is. In Fig. 3, we show the ROC (Receiver Operating Characteristics) for this discrimination. With a threshold of $|u_{xy}| = 0.4$ we could obtain a true positive rate of 93.3% with only 9.6% false positives. The unexpectedly good discrimination power with only one out of twelve properties that we extract from the

[1] Latitude 36.7, Longitude -122.0, Depth 299 meters.

objects makes us confident that we can improve the discrimination performance more by learning a measure for how "interesting" an event is from all available properties.

IV. DISCUSSION AND CONCLUSION

We report here our progress in developing a new method for processing video streams from ROVs 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 in the process of setting up the necessary hardware.

But even beyond its applications to ROV videos, our method for automated underwater video analysis may potentially have a larger impact by enabling 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 [11]. 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 largely "boring" video streams from 10s or even 1000s of network cameras could be key to those cameras being useful practical scientific instruments.

Acknowledgments

We thank the staff and management of the MBARI video lab for their support, help and interest in our project. This project was initiated at the 2002 Workshop for Neuromorphic Engineering in Telluride, Colorado.

REFERENCES

1. Robison, B.H., *The coevolution of undersea vehicles and deep-sea research.* Marine Technology Society Journal, 2000. **33**: p. 69-73.
2. Robison, B.H., 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, 1998. **45**: p. 1741–1752.
3. Silguero, J.M.B. and B.H. Robison, *Seasonal abundance and vertical distribution of mesopelagic Calycophoran siphonophores in Monterey Bay, CA.* Journal of Plankton Research, 2000. **22**: p. 1139-1153.
4. Estes, J.A. and C.H. Peterson, *Marine ecological research in seashores and seafloor systems: Accomplishments and future directions.* Marine Ecology Progress Series, 2000. **195**: p. 281-289.
5. Southward, A.J. *The importance of long time-series in understanding the variability of natural systems.* in *International Helgoland Symposium "The Challenge to Marine Biology in a Changing World.* 1992. Helgoland, Hamburg, Germany.
6. Itti, L., C. Koch, and E. Niebur, *A model of saliency-based visual attention for rapid scene analysis.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998. **20**(11): p. 1254-1259.
7. Itti, L. and C. Koch. *Target Detection using Saliency-Based Attention.* in *Proc. RTO/SCI-12 Workshop on Search and Target Acquisition (NATO Unclassified).* 1999. Utrecht, The Netherlands.
8. Robison, B.H., *Midwater research methods with MBARI's ROV.* Marine Technology Society Journal, 1993. **26**: p. 32-39.
9. Mellinger, E., A. Pearce, and M. Chaffey. *Distributed multiplexers for an ROV control and data system.* in *Proceedings MTS/IEEE Oceans 1994.* Brest, France.
10. Newman, J.B. and D. Stakes. *Tiburon, development of an ROV for Ocean Science Research.* in *Proceedings MTS/IEEE Oceans 1994.* Brest, France.
11. Smith, K., *NEPTUNE science white paper #6: Deep-sea ecology.* 2002.