Detection of Visual Events in Underwater Video Using a Neuromorphic Saliency-based Attention System

H11F-0912

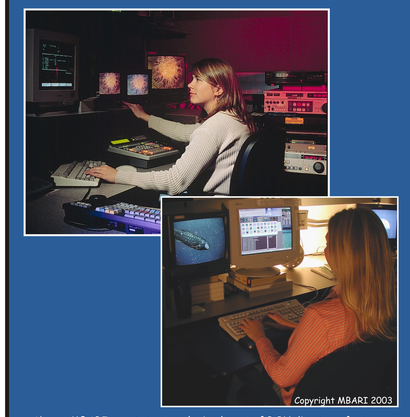
Duane R. Edgington[1], Dirk Walther[2], Danelle E. Cline[1], Robert Sherlock[1], Karen A. Salamy[1], Alexis Wilson[1], and Christof Koch[2]

Monterey Bay Aquarium Research Institute (MBARI), 7700 Sandholdt Rd., Moss Landing, CA 95039; [2]California Institute of Technology, Computation & Neural Systems Program, Pasadena, CA 91125

Abstract

The Monterey Bay Aquarium Research Institute (MBARI) uses high-resolution video equipment on remotely operated vehicles (ROV) to obtain quantitative data on the distribution and abundance of oceanic animals. High-quality video data supplants the traditional approach of assessing the kinds and numbers of animals in the oceanic water column through towing collection nets behind ships. Tow nets are limited in spatial resolution, and often destroy abundant gelatinous animals resulting in species undersampling. Video camera-based quantitative video transects (QVT) are taken through the ocean midwater, from 50m to 4000m, and provide high-resolution data at the scale of the individual animals and their natural aggregation patterns. However, the current manual method of analyzing QVT video by trained scientists is labor intensive and poses a serious limitation to the amount of information that can be analyzed from ROV dives.

Presented here is an automated system for detecting marine animals (events) visible in the videos. Automated detection is difficult due to the low contrast of many translucent animals and due to debris ("marine snow") cluttering the scene. Video frames are processed with an artificial intelligence attention selection algorithm that has proven a robust means of target detection in a variety of natural terrestrial scenes. The candidate locations identified by the attention selection module are tracked across video frames using linear Kalman filters. Typically, the occurrence of visible animals in the video footage is sparse in space and time. A notion of "boring" video frames is developed by detecting whether or not there is an interesting candidate object for an animal present in a particular sequence of underwater video frames that do not contain any "interesting" events. 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. Presented here is performance data that compare the automated detection method with that of human annotators. The system enhances the productivity of human video annotators and/or cues a subsequent object classification module by omitting "boring" frames and marking candidate objects.



Above: MBARI annotators analyzing hours of ROV dive tape footage.

Video Collection and Annotation

MBARI collects ~300 days per year of video from its ROVs. This equates to:

- **14,000+ tapes**
- **10,000+ hours of undersea video**

Each video is professionally annotated to feed the MBARI Video Annotation and Reference System (VARS) database which enables integration of annotation results and linking them to environmental data over many dives and over many years.

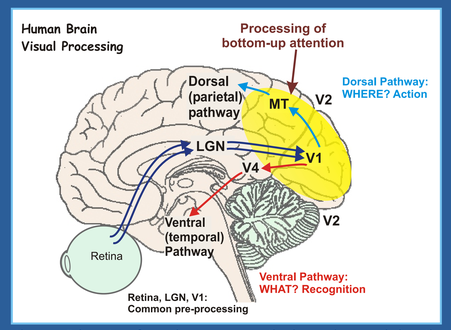
- **~1,000,000 individual observations in the MBARI annotation database**

Annotating video is time-consuming and tedious.

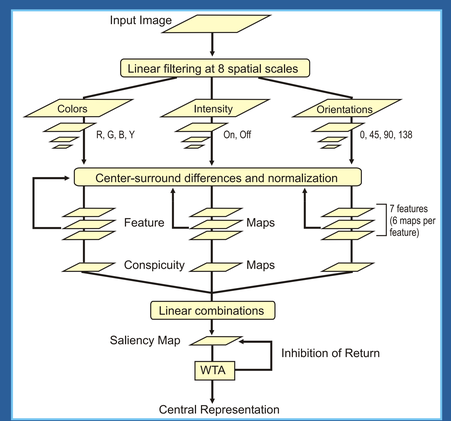
Can we supply tools to make the analysts more productive and efficient?

Application of Biomimic Model to Detection of Visual Events

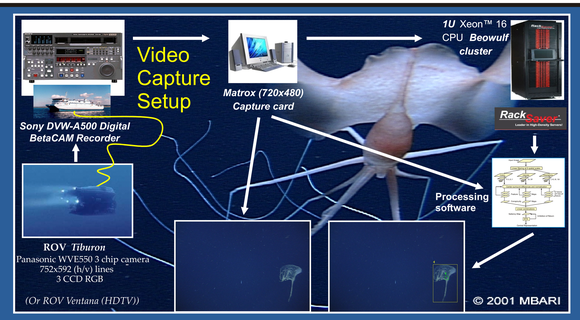
- Humans and many animals are extremely good at attending to novel features in a scene. Is it possible to model this ability?
- A model was developed in 1985 by Koch and Ullman (MIT). It was based on the biology of human perception and visual system (see image below).
- The model was implemented as a computer program by Itti in the Koch lab at Caltech as a Ph.D. Thesis in the late 1990's.
- This model encompassed a bleeding edge neuromorphic engineering approach to computer "Vision".
- The model has been applied to terrestrial surveillance, traffic surveillance and advertising copy. This research is the first application of the model to underwater video scenes.



Above: Schematic of visual processing in the human brain including the pathways of bottom-up attention modeled in our approach.



Above: Flow diagram of a typical model for the control of bottom-up attention. This diagram is based on Koch and Ullman's hypothesis that a centralized two-dimensional saliency map can provide an efficient control strategy for the deployment of attention on the basis of bottom-up cues. The input image is decomposed through several pre-attentive feature detection mechanisms (sensitive to color, intensity, etc), which operate in parallel over the entire visual scene. Neurons in the feature maps then encode for spatial contrast in each of those feature channels. In addition, neurons in each feature map spatially compete for salience, through long-range connections that extend far beyond the spatial range of the classical receptive field of each neuron (here shown for one channel; the others are similar). After competition, the feature maps are combined unto a unique saliency map, which topographically encodes for saliency irrespective of the feature channel in which stimuli appeared salient. The saliency map is sequentially scanned by attention through the interplay between a winner-take-all network (which detects the point of highest saliency at any given time) and inhibition of return (which suppresses the last attended location from the saliency map, so that attention can focus onto the next most salient location). Top-down attentional bias and training can modulate most stages of this bottom-up model.



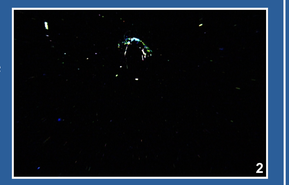
Processing

1. Record and Capture Video

- Video recorded by broadcast or HDTV Cameras on ROV (Digital BetaCam Recorder).
- On shore, video is captured into data files that are further processed.

2. Pre-process Frames

- Smooth to remove scan lines.
- Subtract the sliding average of the last 10 frames to remove constant background.

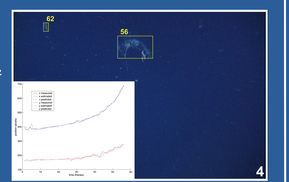


3. Identify Salient Locations

- Salience based on low-level properties such as luminance contrast, local orientation contrast and color contrast (red-green and blue-yellow).
- Salient points are scanned by the interaction of a Winner-Take-All (WTA) neural network and Inhibition-Of-Return (IOR).

4. Track Salient Objects

- Track the x and y coordinate of the centroid with linear Kalman Filters.
- Assume constant acceleration (good assumption for projection of constant velocity motion onto the camera plane).
- Data assignment for multiple target tracking made easy by sparseness of salient objects.
- Every 5 frames: Check for salient objects that are not yet tracked, and initialize new trackers.

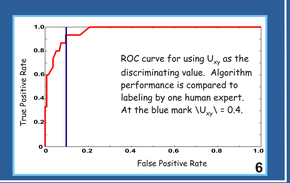


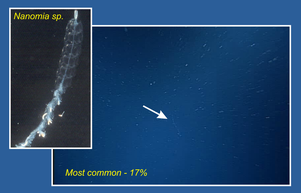
5. Extract Binary Objects at Salient Locations

- Segment objects using image flooding with fixed thresholding.
- Extract a number of intermediate-level properties for each object ("object-token")
  - Area, centroid
  - Second moments
  - Major and minor axes, elongation
  - Major axis orientation
  - Maximum, minimum & average image intensity within the object shape

6. Decide which events are "Interesting"

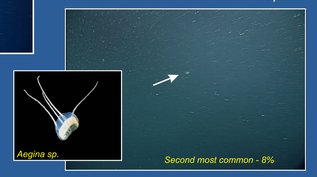
- Initial approach: Decision was based on area.
- Improved approach: Use the second momentum [U_{xx}] - a threshold of 0.4 gives **93.3% correct positive rate and 9.6% false positive rate**.
- Ongoing work: Using a support vector machine (SVM) to learn "Degree of Interesting" value from the features extracted from binary objects.



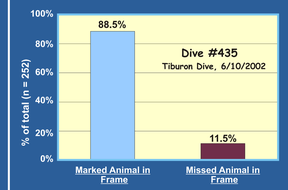


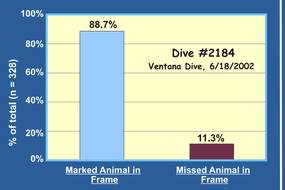
Above: Graph illustrating the mean percentage of annual annotations versus midwater organisms observed in Monterey Bay from 1997-2002. There were 325 different annotation classes during this time.

Left and Below: Images and frame grabs of the two most common species. The frame grabs show how difficult these animals are to see and identify.

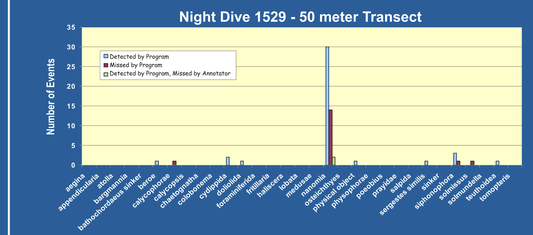


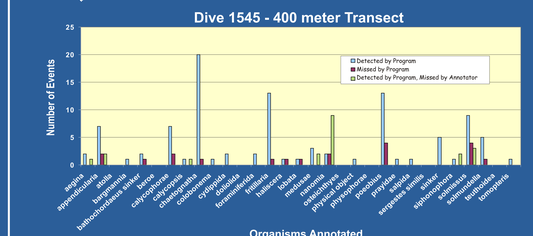
Vision Results





Above: The attentional selection algorithm shows very promising results for single image NTSC frame-grabs obtained from midwater dives of MBARI's two 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 an animal in about 90% of the cases.

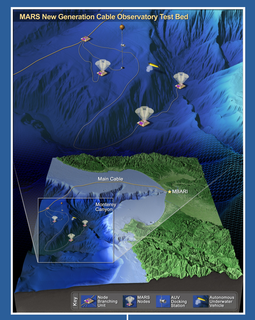


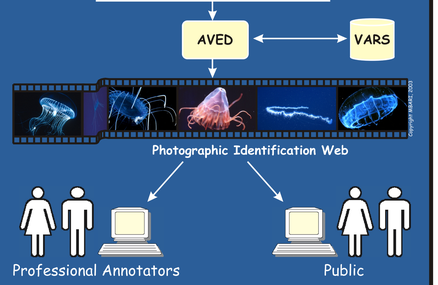


Above: Comparison of human annotation versus computer analysis for two midwater dives. Each 10 minute transect video was annotated in detail by a professional annotator using the MBARI VARS database system. All organisms recognized by professional annotators were classified into standard categories. These same transect videos were analyzed by the program. For each human annotation, we compared whether the program detected and tracked that organism or if the program missed the annotation event. "Detected by Program" is each case where the program detected and tracked the same organism as a human annotator. "Missed by Program" is each case where the program missed the human annotation. "Detected by Program, Missed by Annotator" were the very interesting cases where the professional annotator missed the organism, but the program clearly found an organism. The annotation categories shown are the combined categories for these two transects.

Future Applications

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 benefit in automating portions of the analysis. We also see great benefit in automating analysis of video from fixed ocean observatory cameras, where autonomous response to potential events (pan/zoom to events), and automated processing of largely "boring" video streams from 10s or 100s or even 1000s of network cameras could be key to those cameras being useful practical scientific instruments.





Above: AVED technology for real-time analysis of underwater video on cabled observatories. Results from real-time analysis of stationary cameras and mobile cameras mounted on ROVs and AUVs are transmitted to shore through a high bandwidth communication link. Results are stored in the VARS database. Still images and short video clips from cabled observatories are then made available for both professional annotators and the public to help in the identification of undersea organisms through a web interface.

Future Work

- Continued detailed comparison of the algorithm with video lab Annotation.
- Processing more videos with large set of more species.
- Recognition of selected organisms.
- Use of motion as a feature.
- Rapid processing of streamed video using a Beowulf computer cluster.

Acknowledgements

We would like to thank the David and Lucile Packard Foundation for their generosity in funding work at MBARI. D. Walther and C. Koch are sponsored by the NSF Center for Neuromorphic Systems Engineering at Caltech. This project was initiated at the 2002 Neuromorphic Engineering Workshop, Telluride, Colorado sponsored by NSF. We would like to thank the NSF Research Coordination Network (RCN) Institute for Neuromorphic Engineering (INE) for support of collaborative travel. We would also like to thank the staff and management of the MBARI video lab and Dr. Bruce Robison for their support, help, and interest in our project.

Citations

L. Itti and C. Koch, (2001). " Computational Modeling of Visual Attention". Nature Reviews Neuroscience, 2: pp. 194-203.

L. Itti, C. Koch, et al. (1998). "A model of saliency-based visual attention for rapid scene analysis". IEEE PAMI, 20(11): pp. 1254-1259.

C. Koch and S. Ullman. (1985). "Shifts in selective visual attention: towards the underlying neural circuitry". Human Neurobiology, 4: pp. 219-227.

Compiled by Karen A. Salamy for MBARI, December 2003