



Detecting, Tracking and Classifying Animals in Underwater Video



Duane R. Edgington¹, I. Kerkez¹, D. Oliver¹, D. E. Cline¹, R. Sherlock¹, B. Robison¹, L. Kuhn², D. Walther², Marc'Aurelio Ranzato², and P. Perona²
¹Monterey Bay Aquarium Research Institute, 7700 Sandholdt Rd., Moss Landing, CA 95039; ²California Institute of Technology, Computation & Neural Systems Program, Pasadena, CA 91125

Abstract
We present a system that enables automated annotation of underwater video transects for quantitative studies of ocean ecology. Remotely operated vehicles (ROVs) equipped with high resolution video cameras enable quantitative video transects (QVTs) of the ocean midwater and benthos. QVTs provide data at the scale of the individual midwater macrofauna and epibenthic megafauna and their natural aggregation patterns that advance studies in animal diversity, distribution and abundance. Analysis of QVTs, however, is labor intensive and costly, reducing the amount of data analyzed from transects and thus limiting understanding of the factors regulating the abundance and distribution of marine populations. To address this problem we developed an automated system for detecting and classifying organisms, in which frames are processed with a neuromorphic selective-attention algorithm. The identified candidate locations are subject to a number of parameters and tracking, to mark detected events as "interesting" or not. The "interesting" events undergo further processing with a Bayesian classifier utilizing a Gaussian mixture model to determine the abundance and distribution of a selected organism category. Presented data detail the comparison between professional annotations and automated detection of organisms, and classification by the system of the deep-sea benthic echinoderm *Rathbunaster californicus* in video footage.



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

Video Collection and Annotation

MBARI collects ~300 days of video annually from its ROVs. This equals:

- 16,000+ tapes
- 12,000+ 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 MBARI annotation database
Annotating video is time-consuming and tedious.

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

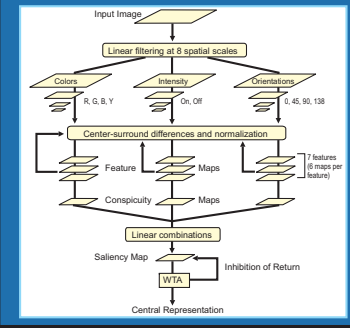
Application of Biomimic Models to Detection and Classification of Visual Events

Humans and many animals are extremely good at attending to novel features in a scene. A model of attention was developed in 1985 by Koch and Ullman (MIT). It was based on the biology of human perception and visual system.

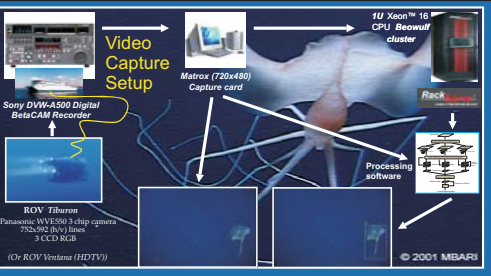
The model was implemented as a computer program by Itti in the Koch lab at Caltech as a Ph.D. Thesis in the late 1990s.

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.

The "interesting" events undergo further processing with a Bayesian classifier utilizing a Gaussian mixture model to determine the abundance and distribution of a selected organism category.



Left: 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 onto 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 HD-TV cameras on ROV (Digital BetaCam Recorder).
- On shore, video is captured into data files that are further processed.

2. Pre-process Frames & Identify Salient Locations

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

3. Track Salient Objects

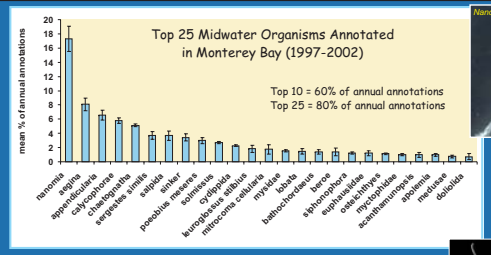
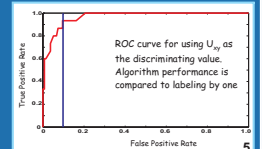
- Track the x and y coordinate of the centroid with linear Kalman Filters.
- Assume constant acceleration (good assumption for projection of constant velocity motion onto the camera plane).
- Data assignment for multiple target tracking made easy by sparseness of salient objects.
- Every 5 frames, check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

4. 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

5. Decide which events are "Interesting"

- Initial approach: Decision was based on area.
- Improved approach: Use the second momentum [U_{ij}] - a threshold of 0.4 gives 93.3% correct positive rate and 9.6% false positive rate.

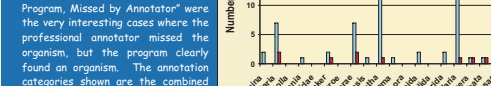
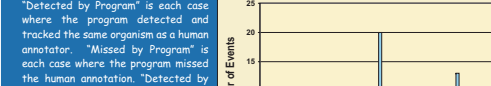


Above: Graph illustrating the mean percentage of annual annotations versus midwater organisms observed in Monterey Bay from 1997-2002. There were 325 different annotation classes during this time.

Right: Images and frame grabs of the two most common species (*Nanomia* sp. and *Aegina* sp.). The frame grabs show how difficult these animals are to see and identify.

Results

Comparison of human annotation versus computer analysis for two midwater dives. Each ten 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.

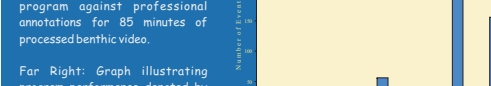


Right: Captured images depicting natural benthic scenes before and after AVED processing. Boxes drawn around i) *Rathbunaster californicus*, ii) *Parastichopus leukothele* and iii) *Microstomus pacificus* denote event detections.



Right: A comparison of event detections made by the AVED program against professional annotations for 85 minutes of processed benthic video.

Far Right: Graph illustrating program performance denoted by frequency of successful detection.



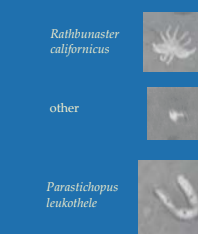
Acknowledgements

We would like to thank the David and Lucile Packard Foundation for their generosity in funding work at MBARI. D. Walther and P. Perona 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 for their support, help, and interest in our project. K. Solomy provided technical support.

Classification

- Used a classification program developed by Marc'Aurelio Ranzato at Caltech and Universita' degli studi di Padova
- Developed to analyze biological particles
- Classify by analysis of image features to recognize objects
- Based on extracting features using
 - local jets (Schmid et al. 1997) (convolution of the image with a derivative of Gaussian kernel)
 - image and power spectrum principal components (Torralba et al. 2003)
- Model training data with mixture of Gaussians (Choudrey and Roberts 2003)
- Implemented in Matlab
 - processes grayscale square subimages of the segmented scene containing the object to be classified

Sample images



Classification results to date

- Analyzed 7.5 minutes of benthic transect data at Smooth ridge
- Trained classifier with
 - ~8000 images, including ~2600 images of *Rathbunaster*
- Extracted 210 events (~7250 images) from transect data
- Program classified
 - 90% of the *Rathbunaster* events correctly
 - 10% of the *Rath* events incorrectly
 - 0% of non-*Rath* events incorrectly as *Rathbunaster*



Not just for animals... Manganese Nodules



Hawaii-2 Observatory (5000m) central North Pacific. JASON ROV. *Stace E. Beaulieu* WHOI

Next steps

- Evaluating and improving classification system.
- Evaluate automatically adjusting weights of low level detection feature maps from training images of target of interest (Navalpakkam & Itti, 2004).
- Evaluate using system for habitat classification