# Detecting, Tracking and Classifying Animals in Underwater Video

Duane R. Edgington
Ishbel Kerkez
Danelle E. Cline
Jerome Mariette
Monterey Bay Aquarium Research Institute,
7700 Sandholdt Road,
Moss Landing, CA 95039, USA
{duane, dcline}@mbari.org
{ishbel.kerkez,jerome.mariette}@gmail.com

Marc'Aurelio Ranzato[1]
Pietro Perona[2]

[1]New York University, NY, NY, USA
(ranzato@cs.nyu.edu)
[2]California Institute of Technology,
Computations and Neural Systems Program,
Pasadena, CA 91125, USA (perona@caltech.edu)

## Abstract

*We demonstrate an attentional event selection, tracking, and classification system for processing video streams from remotely operated underwater vehicles (ROVs) that enables automated annotation of underwater video transects for quantitative studies of ocean ecology. The system identifies and tracks potentially interesting visual events spanning multiple frames based on low-level properties of salient objects. "Interesting" events are passed to a Bayesian classifier utilizing a Gaussian mixture model to determine the abundance and distribution of a representative benthic species. Presented data details the comparison between automated detection of organisms and program classification of deep-sea ocean bottom echinoderm (sea star) Rathbunaster californicus in video footage with professional annotations.*

## 1. Background and summary of work

For more than a century, the traditional approach for assessing the kinds and numbers of animals in the oceanic water column was to tow collection nets behind ships. This method is limited in its spatial resolution. Today, ROVs provide an excellent alternative to nets for obtaining quantitative data on the distribution and abundance of oceanic animals [1]. Using video cameras, it is possible to make quantitative video transects (QVT) through the water, providing high-resolution data at the scale of the individual animals and their natural aggregation patterns that advance studies in animal diversity, distribution and abundance. However, the current manual method of analyzing QVT video by trained scientists is very labor intensive and poses a serious limitation to the amount of information that can be obtained from ROV dives.

To overcome this bottleneck in analyzing underwater videos we have developed an automated system for detecting and classifying animals (events) visible in the video stream. This task is difficult due to the low contrast of many translucent animals and due to debris ("marine snow") cluttering the scene. We process video frames with an attentional selection algorithm [2] that has been shown to work robustly for target detection in a variety of natural scenes [3].

The candidate locations identified by the attentional selection module are tracked across video frames using linear Kalman Filters [4]. 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 with pan/tilt/zoom control and automated processing of video from cameras on autonomous underwater vehicles.

## 2. Explanation of the algorithms

Four main processing steps are involved in our video event detection analysis procedure after the video has been captured from the digital BetaCam video deck used for recording the HDTV signal from the ROVs. Initially, some generic pre-processing is performed for each frame of the input video stream (background subtraction, scan line smoothing, global contrast enhancement); then the vicinity of locations are scanned for the occurrence of animals at which the Kalman Filter trackers predict them; thirdly, every five frames the image is processed to find salient objects that are not yet tracked; and in the last step, visual events are classified into "interesting" or "boring" according to low-level properties of the tokens involved.

Tracking is achieved with two linear Kalman Filters for the $x$ and $y$ coordinates of each tracked object, assuming that the motion of the image of the object in the camera plane has constant acceleration. This is a good assumption for constant speed heading motion of ROVs obtaining QVTs. Data assignment for our multiple target tracking (MTT) system is done by a nearest-neighbor rule [5]. The Kalman trackers are initiated with salient objects detected by a system mimicking saliency-based bottom-up attention in humans and other primates [2, 6].

For this saliency-based detection system, each input 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”. 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) [2]. Each scanned location is compared with the events that are already being tracked. If it does not belong to any of these events, a new tracker for the detected object is initiated.

For each tracked object, we obtain a binary mask, which allows us to extract a number of low-level properties such as the object size, the second moments with respect to the centroid, the maximum luminance intensity, the average luminance intensity over the shape of the object, and its aspect ratio. We use the mask to extract a subframe of the event for further processing.

For classification, we pass these segmented square subframes to a Bayesian classifier [7] that extracts features using local jets from a labeled training set, and classifies with a Gaussian mixture model to determine the abundance and distribution of a representative benthic organisms.

## 3. Results

In NTSC single frames obtained from midwater dives of ROVs (ROV *Tiburon* and *Ventana*) 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. The processing of midwater video clips shows similarly promising results[8]. For epibenthic (near bottom) transects, the system detects 97% of the deep-sea benthic echinoderm *Rathbunaster californicus* in video footage, and correctly classifies those events as *R. californicus* in 90% of the events.

In our demonstration, we show unprocessed video clips to illustrate the nature of the task, and we show processed video clips with marked “interesting” events concurrently. In addition, we will have one computer displaying the processing steps in real time while running the algorithm. We will display a poster with explanations of the background and significance of our research, details of the processing steps, and performance data that compare the automated detection and classification method with human annotators. Presented data details the comparison between automated program classifications of *R. californicus* in video footage with professional annotations. We will present results from adapting our automated system to fixed observatory seafloor video.

Our demonstration is an example of applying computer vision algorithms to real-world video data, of broad interest since our work can be adapted and extended to other areas such as surveillance and robotic control.

## Acknowledgements

This project originated at the 2002 Workshop for Neuromorphic Engineering in Telluride, CO, USA in collaboration with Dirk Walther, California Institute of Technology, Pasadena, CA, USA. We thank the David and Lucile Packard Foundation for their generous support. At MBARI, Karen A. Salamy and Dorothy Oliver provided technical assistance, and Michael Risi provided engineering assistance.

## *References*

1. Robison, B.H., *The coevolution of undersea vehicles and deep-sea research.* Marine Technology Society Journal, 2000. **33**: p. 69-73.
2. 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.
3. 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.
4. Kalman, R.E. and R.S. Bucy, *New Results in Linear Filtering and Prediction Theory.* Journal of Basic Engineering, 1961. **83**(3): p. 95-108.
5. Kirubarajan, T., Y. Bar-Shalom, and K.R. Pattitipati, *Multiassignment for Tracking a Large Number of Overlapping Objects.* IEEE Transactions on Aerospace and Electronic Systems, 2001. **37**(1): p. 2-21.
6. Itti, L. and C. Koch, *Computational modelling of visual attention.* Nature Reviews Neuroscience, 2001. **2**(3): p. 194-203.
7. Ranzato, M.A., http://www.vision.caltech.edu/ranzato/. 2004, California Institute of Technology: Pasadena, CA.
8. Walther, D., D.R. Edgington, and C. Koch. *Detection and Tracking of Objects in Underwater Video.* in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR).* 2004. Washington, D.C.