

# Detecting and Tracking Animals in Underwater Video

Duane R. Edgington  
Danelle E. Cline  
R.E. Sherlock

Monterey Bay Aquarium Research Institute,  
7700 Sandholdt Road,  
Moss Landing, CA 95039, USA  
{dedgington, robs}@mbari.org

Dirk Walther  
Christof Koch

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

## Abstract

*We demonstrate an attentional selection and tracking system for processing video streams from remotely operated underwater vehicles (ROVs). The system identifies and tracks potentially interesting visual events spanning multiple frames based on low-level properties of salient objects, which are associated with those events. If video frames contain interesting frames, 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.*

## 1. Background

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, and because of the design of the nets, gelatinous animals are destroyed and hence under-sampled. 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. 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 data that can be obtained from ROV dives.

## 2. Summary of work

To overcome this bottleneck in analyzing ROV dive videos we have developed 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 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.

Typically, the occurrence of visible animals in the video footage is 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.

The demonstrated attentional selection and tracking module is only the first step towards an integrated automated video annotation system that will consist of an object classification module and control modules for pan/tilt/zoom cameras and autonomous underwater vehicles (AUVs) in addition to the attentional module.

## 3. Explanation of the algorithms

Four main processing steps are involved in our video 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 (subtracting of the background, smoothing of scan lines, 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 saliency 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.

#### 4. Results

Our attentional selection shows very promising results for single image NTSC frame-grabs obtained from midwater dives of ROVs (ROV *Tiburón* 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. The processing of video clips shows similarly promising results[7].

In our demonstration, we show unprocessed video clips to illustrate the difficulty of the task, and the processed video clips with marked “interesting” events and omitted “boring” frames concurrently. In addition, we 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; we will show details of the processing steps, as well as performance data that

compare the automated detection method with human annotators.

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.

#### Acknowledgement

This project originated at the 2002 Workshop for Neuromorphic Engineering in Telluride, CO, USA. We thank the David and Lucile Packard Foundation for their generous support of research and development 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. At MBARI, Karen A. Salamy provided technical assistance, and Michael Risi engineered our video capture system.

#### 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. 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.