

# Automated Video Analysis for Oceanographic Research

Dirk Walther

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

Christof Koch

California Institute of Technology,  
Computation and Neural Systems  
Program, Mail Code 139-74,  
Pasadena, CA 91125, USA  
walther@caltech.edu

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

California Institute of Technology,  
Division of Biology,  
Mail Code 139-74,  
Pasadena, CA 91125, USA  
koch@klab.caltech.edu

## Abstract

*We demonstrate 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 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 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 are processing the videos 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 (tokens) identified by the attentional selection module are combined across video frames using a token tracking algorithm [4]. 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.

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 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 salient tokens are selected; thirdly, each of the tokens is associated with a visual event that is tracked across frames; and in the last step, visual events are classified into “interesting” or “boring” according to low-level properties of the tokens involved.

In order to detect salient tokens in the video stream, 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, 3].

Each scanned location (token) is inspected for its compatibility with the existing events tracked in previous frames [4]. The compatibility is assessed by determining the distance of the token location from its expected location extrapolated linearly from the last two tokens of the event. Finally, the token is stored for the event for which the expected and the actual token locations match best, provided that their distance is smaller than some upper threshold. If a token cannot be associated with any of the existing events, a new event is created with this token as its initial location. Events to which no token is associated in two successive frames are declared “closed”, and no subsequent association of tokens to these events is possible. Closed events that could not be tracked for more than a preset number (typically 7) of frames are discarded as noise.

For each token, we obtain a binary mask of the object, which allows us to extract a number of low-level properties such as the object size, the maximum luminance intensity, the average luminance intensity over the shape of the object, and its aspect ratio. Among all the tokens stored for one event, the token with the largest object size is selected as the representative of the event for the subsequent classification. A threshold classifier decides which of the events are labeled as “interesting” or “boring” based on these sets of low-level properties of the

representative tokens. Interesting events are marked in the video frames. Frames that do not contain any interesting events can be omitted from being written to the output video stream.

### 4. Results

Our attentional selection algorithm 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. We describe our implementation of a protocol for quantitative comparison of the detection performance of the algorithm with human annotators, and report on the results of that quantitative comparison in a poster at the conference.

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. In order to demonstrate the attentional selection algorithm, we show a pan/tilt/zoom camera that is controlled by the attention algorithm, acting as an artificial eyeball executing eye movements.

### Acknowledgement

This project originated at the 2002 Workshop for Neuromorphic Engineering in Telluride, CO. We wish to acknowledge financial support from the Institute for Neuromorphic Engineering, NSF-ITR, and the David and Lucile Packard Foundation as part of their generous support of research and development at MBARI.

### 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. Zhang, Z.Y., *Token Tracking in a Cluttered Scene*. Image and Vision Computing, 1994. **12**(2): p. 110-120.