First steps towards autonomous recognition of Monterey Bay's most common mid-water organisms: *Mining the ROV video database on behalf of the Automated Visual Event Detection (AVED) system.*

Alexis Wilson, Middlebury College
Mentor: Duane Edgington
Summer 2003

Keywords: AVED, ROV, QVT, VIMS, VARS, video, salience, pelagic organisms, winner-take-all, inhibition-of-return, bottom-up attention, recognition

ABSTRACT

The development of remotely operated vehicles (ROVs) has revolutionized the world of marine science by allowing quantitative video transects (QVTs) to be recorded underwater. This non-invasive technique allows previously unavailable information to be obtained concerning organism diversity, distribution, and abundance. Unfortunately, processing of these QVTs at the Monterey Bay Aquarium Research Institute (MBARI) is a lengthy and tedious process that prevents rapid data analysis. The development of a neuromorphic computer vision system is currently in progress to solve this problem of the data processing bottleneck. The proposed computer vision system combines a saliency-based attentional module and a recognition module, both designed after the human visual system. The "saliency" model mimics the largely unconscious "bottom-up" search mechanism of primates, responding specifically to visual cues in the surrounding environment. The recognition module involves learning of specific targets and requires "top-down" biasing based on learned traits. The attentional module at MBARI can already determine areas of potential interest in the marine environment. However, in order to begin training the computer system for the recognition module, a test database of representative organisms must first be established. The extensive MBARI database was mined to reveal the relative abundance and frequency of annotations of all organisms in the Monterey Bay. Based on the rankings, representative video clips were compiled of the top 25 most common mid-water organisms as a preliminary test database for the recognition module.

INTRODUCTION

In the past, it was common practice to tow collection nets behind ships to assess organism diversity, distribution, and abundance in the ocean. This method was not only destructive to the delicate gelatinous animals in the water column; it also integrated data over the entire length of the tow, making spatial resolution practically impossible [20]. Today, the use of remotely operated vehicles (ROVs) has revolutionized the field of ocean ecology by enabling underwater communities to be observed non-invasively. Video cameras are now employed as an alternative to nets to collect data on the distribution and abundance of pelagic organisms [17]. Quantitative video transects (QVTs) provide high-resolution data on individual organisms and their natural aggregation patterns [18]. At the Monterey Bay Aquarium Research Institute (MBARI) in Moss Landing, California, dive videos are collected approximately 300 days out of the year. More than 14,000 tapes containing over 10,000 hours of video footage have been recorded to date. Each video is professionally annotated and stored in the video annotation reference system (VARS). Unfortunately, QVT analysis is an extremely labor-intensive process that slows data filtration. Highly skilled scientists spend an average of four hours identifying the organisms in a single transect video (and 6-7 videos are collected per dive) [3].

To bypass this hurdle in data processing, scientists and engineers at MBARI are currently working to develop an automated visual event detection (AVED) system. Ultimately this trainable machine vision system is intended to autonomously detect, recognize and classify organisms from the ROV video database in real-time [3]. Videos are being processed using an attentional selection algorithm [12] that is known to be robust for target detection in a variety of natural scenes [10]. As the AVED project is now in its first of an estimated three years of development, preliminary steps towards organism recognition are being taken.

BACKGROUND:

While many computer vision systems are designed in an ad hoc, task-specific manner, the AVED computer vision system is modeled directly after the evolutionarily proven process of biological vision. In order to understand the basis of computer vision, primate visual search techniques and neural processes must first be considered. Though it is difficult to dissect and analyze all processes involved with biological vision, low level vision can be broken down into two general categories: attention and recognition.

1.1 PRIMATE VISION

Primates have the unique ability to perceive their surrounding visual environments and to rapidly filter large amounts of incoming ocular stimuli [1,12]. Though vision is often taken for granted, it is a computationally expensive task. More than one million distinct nerve fibers comprise each optic nerve and approximately half of the mammalian brain is dedicated to vision [7]. Massive amounts of incoming information (approximately 10^8 bits per second) must be processed to allow nearly immediate reactions [9]. An intricate network of brain areas from the primary visual cortex to the prefrontal cortex controls focal visual attention [6].

In order for "normal" vision to occur, cooperation between two hierarchical brain streams must take place to delegate attention, recognize visual stimuli, and trigger the corresponding eye movements [6]. The dorsal visual processing stream, or the "where/how" pathway is comprised of the central cortical areas and the posterior parietal cortex. This pathway determines where visual focus should be directed purely based on the stimuli in the visual environment. The ventral visual processing stream, or the "what" pathway consists of the central cortical areas and the posterior parietal cortex [5]. This stream recognizes and identifies objects in the visual scene (Figure 1).

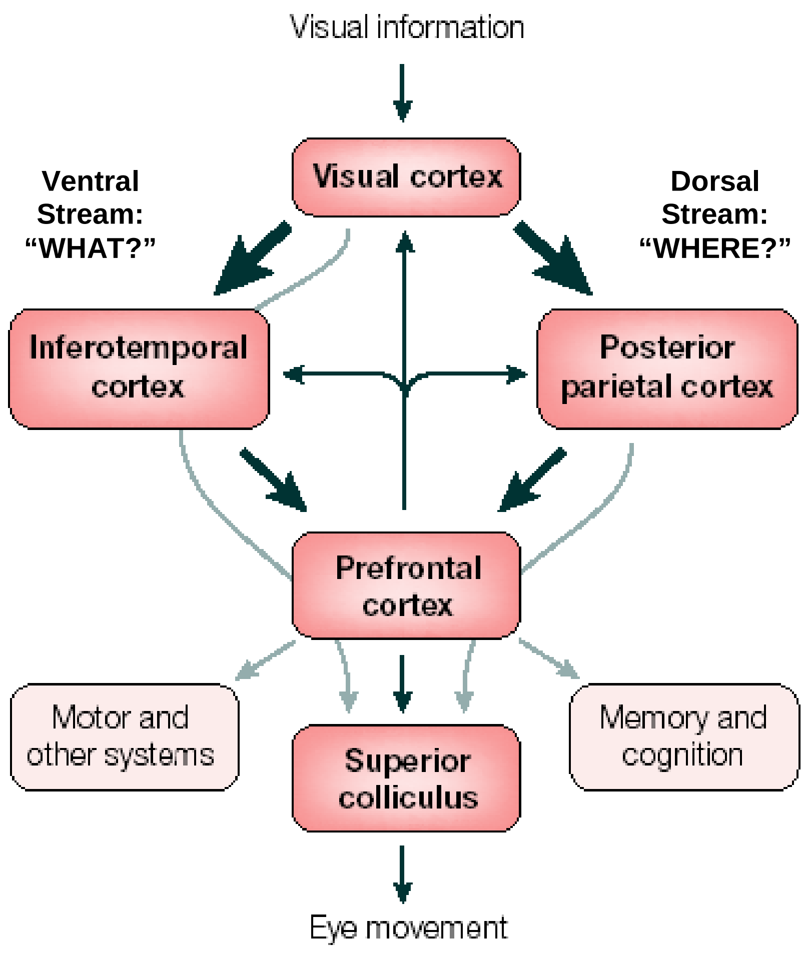

Figure 1: A simplified overview of the main brain areas involved in the control of visual attention. Visual information enters the primary visual cortex through the lateral geniculate nucleus. The information then moves through the dorsal and ventral streams. The control of attention is believed to take place mostly in the dorsal stream, which includes the posterior parietal cortex. Consisting of the Inferotemporal cortex, the ventral stream is responsible for object recognition and identification. The dorsal and ventral streams must interact to allow recognition and spatial allocation of attention. This interaction takes place in the prefrontal cortex, which then transfers the signal to the superior colliculus allowing eye movement. Image adapted from Itti and Koch (2001) [11].

The mechanism by which humans and primates rapidly direct their gaze towards objects of interest in the visual environment is known as "selective visual attention." This is a dual component mechanism and can act from the bottom-up (image-based attention), or from the top-down (task-dependent attention) [5,9,13]. Bottom-up control is largely unconscious and is driven purely by the specific aspects of the stimuli present in the visual environment, while top-down attention acts under cognitive, or "volitional control" [9] and depends on the exact task being performed. In the absence of task specification, visual attention appears to be guided purely by bottom-up processes and can be studied using simple visual search techniques. This saliency-based process is commonly modeled with computer software systems [15].

Modeling of primate vision with machines has long been of interest to scientists and engineers; however the realization of this task has been less than optimal due to the lack of computer processing power [1]. Additionally, computational modeling often is too task-specific or goal-oriented and is not readily transferable to novel environments, targets, or tasks [7]. Yet, recently the development of low-cost supercomputers such as the "Beowulf" cluster is making machine vision a reality and a potentially very powerful tool for myriad diverse fields of science [1,7,8]. Neuromorphic engineering promises to bridge the gap between computational neuroscience and machine vision [7]. Imitation of the primate visual system and ocular search mechanisms has already yielded fairly accurate software models and the potential for versatility, advancement, and adaptability is great [1].

1.2 COMPUTER VISION: "THE ATTENTIONAL MODULE"

Recent progress in the field of visual neuroscience (single neuron electrophysiology, psychophysics, and functional neuroimaging) has laid the foundation for understanding biological vision [1]. Experiments measuring eye motions suggest that primates use a non-sequential, though non-random, and largely reproducible scanning strategy to explore new environments [13]. As a primate scans its visual surroundings, attention is drawn to the most interesting, or "salient" features in the scene. Development of computer vision systems based on primate visual techniques depends entirely on the ability to predict the most "salient" regions in a scene. A probability of fixation can be estimated for each object in the visual environment based on certain characteristics of contrast, color, motion, intensity, orientation, and the similarity to the background [2].

Today, most models of visual search are based on the concept of a "saliency map" [2,9,6,11-13]. A saliency map is a two-dimensional topographic map that encodes for the "saliency" or conspicuity of objects in the visual environment. This map receives input from early visual processing and provides a reliable control strategy to focus attention by scanning the map in order of decreasing saliency. Input images are decomposed into different features based on color, motion, orientation, intensity, velocity, contrast, etc. [2, 14]. Ranking of saliency is designated by a combination of "winner-take-all" and "inhibition-of-return" mechanisms [6]. Competition among neurons in the map determines a single winning location that is the most salient position in the scene [3].

This "winner-take-all" neural network determines the next target "attended" by the system [2,5]. To allow attention to shift to the next most conspicuous location, the first most salient position must be transiently inhibited using an "inhibition-of-return" mechanism, such that the winner-take-all network can once again converge on the next most salient location in the scene [5] (Figures 2 and 3). It is difficult to determine the single most interesting location in a scene because human vision may be influenced by many environmental factors [2]. However, the purely bottom-up model reliably provides an accurate prediction of the most probable areas of focal attention.

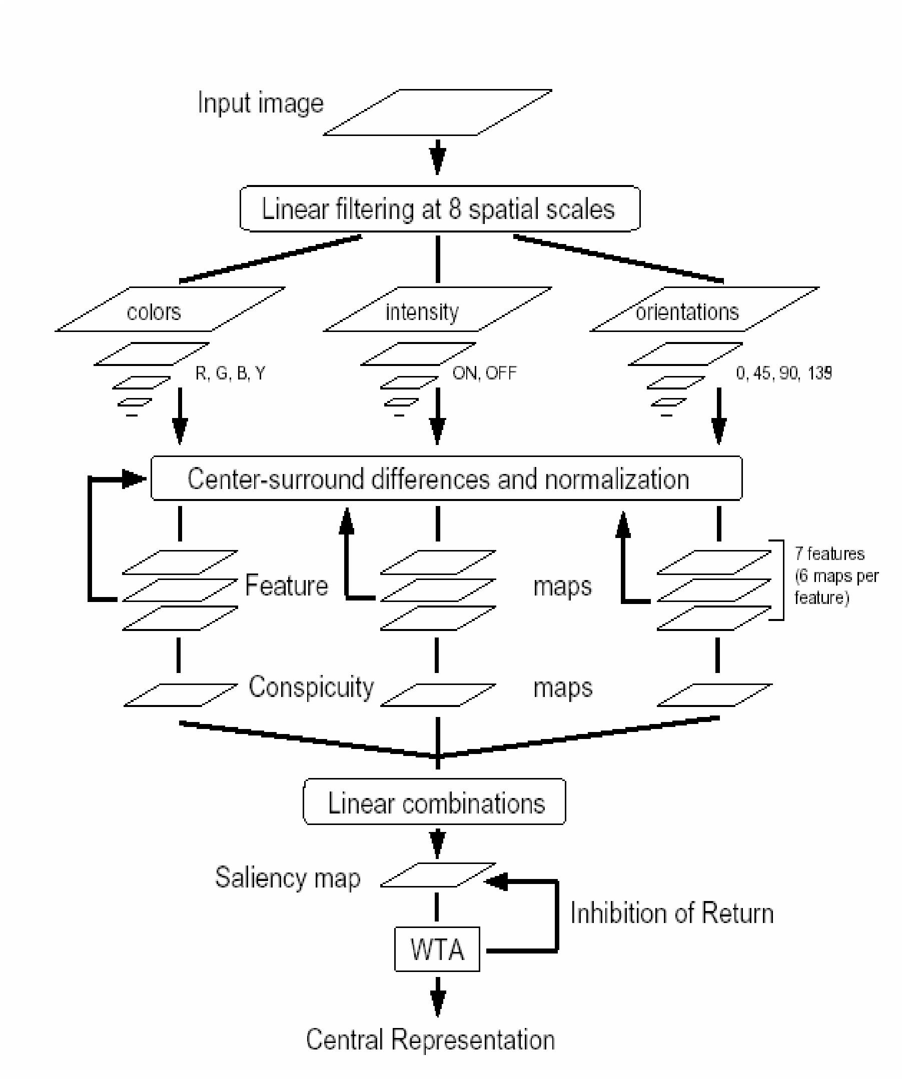

Figure 2: Simplified architecture of a typical bottom-up visual attention model. The input image is analyzed by several visual filters that are sensitive to properties such as color, intensity, and orientation at different spatial scales. These feature maps are combined to yield three conspicuity maps that then feed into a single saliency map. Attention is directed to the area in the saliency map with the maximum activity (as detected by the winner-take-all network). The transient inhibition-of-return at this location then allows the system to delegate attention to the next most salient location. Figure reprinted from Itti (2003) [5].

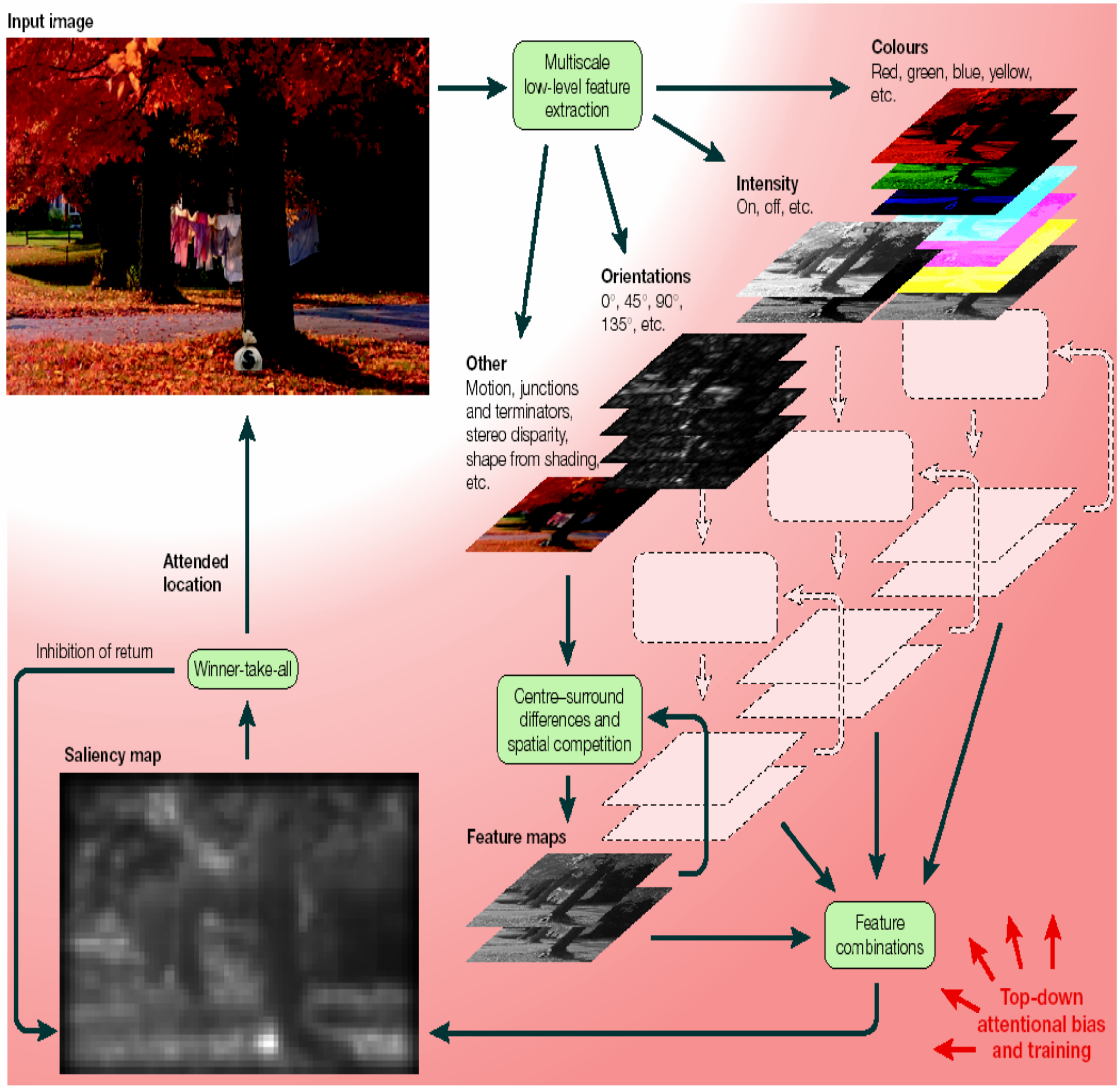

Figure 3: A flow diagram for a typical bottom-up control model. The allocation of attention can be determined based on bottom-up cues. The input image from the natural scene is decomposed into specific feature maps that eventually are combined into a unique saliency map. The winner-take-all network then scans the resulting map to determine the most salient location and to designate the focus of attention. The inhibition-of-return mechanism then suppresses the last attended location from the saliency map so attention can be focused on the next most "interesting" location. Figure taken from Itti and Koch (2001)[11].

1.3 THE RECOGNITION MODULE:

Generally speaking, saliency is robustly computed in most situations using bottom-up models; however object recognition remains more problematic, especially in scenes when background noise/clutter is present [1]. Organism recognition in marine environments promises to be a particularly arduous task, as "marine snow" habitually clutters the visual scene, while the low contrast of many of the gelatinous animals makes identification difficult even for professional video annotators [4]. Only recently have researchers begun to integrate recognition modules with the attention systems. Currently, Navalpakkam and Itti (in press) are using learned representation of target objects to perform top-down biasing on the basic bottom-up attention model. This renders specific objects more salient by enhancing the characteristics of the desired target [16]. Another study indicates that attention improves visual recognition in the presence of multiple salient objects [19]. A combined saliency and recognition module will be implemented in the MBARI automated visual event detection system (AVED) to ultimately recognize and identify organisms in the water column. The attentional module will be used to determine potentially interesting objects that will then be analyzed by the recognition system [3]. In order to make this goal a reality, a test database of representative organisms must be established to teach the system the characteristic features (i.e. shape, contrast, orientation, color, and motion) of each target organism.

1.4 SPECIFIC AIMS:

As the AVED project nears the end of its first year of development, an attentional selection module is already in place to identify potentially "interesting" or salient objects in the visual scene. A token tracking mechanism [21] is successfully being used to track and label these "interesting" objects across several frames of video footage [3,4]. Analysis has already been done to assess the program's ability to discriminate boring from interesting events. In a test of 10sec. video clips taken from the *ROV Ventana,* a true positive rate of 93.3% was obtained to identify objects deemed "interesting" or "salient" [3]. The next task is to create a video clip test database of Monterey Bay's most common pelagic organisms as a preliminary step towards autonomous organism recognition and identification. This portion of the AVED project primarily aims to:

1. Rank Monterey Bay's most common mid-water organisms according to relative abundance in the water column, as determined from the frequency of annotation.

2. Create representative video clips of the most common animals to train the AVED system to recognize objects present in QVTs based on color, shape, contrast, orientation, and eventually motion.

3. Finesse the process of capturing short (2-10sec) video clips in the proper format to build the preliminary test library for the AVED system.

MATERIALS AND METHODS

2.1 HARDWARE AND PROCESSING

At MBARI, two remotely operated vehicles are used in deep-sea exploration for QVT collection. The *ROV Ventana* is launched from the *R/V Point Lobos* and uses a Sony HDC-750 HDTV (1035i30, 1920x1035 pixels) camera for video data acquisition. The

ROV Tiburon is launched from the *R/V Western Flyer* and uses a Panasonic WVE550 3-chip CCD (625i50, 752x582 pixels) camera. For both ROVs, video data are recorded on a DVW-A500 Digital BetaCam video tape recorder (VCR).

On shore, the resources in MBARI's digital video lab were used to create short (2-10sec) digital video clips for each of the organisms from the Digital beta tapes. A Matrox DigiSuite LE digital video editing card in a Pentium P4 computer, running the Windows 2000 operating system and Adobe® Premiere®6.5 was used to capture video as AVI movie files at a resolution of 720 x 480 pixels and 30 frames per second. To ensure compatibility with the AVED processing system, all videos were converted to Quicktime movies and then were completely uncompressed using the video software program "Bink and Smacker" from RAD Video tools. Finally, Roxio Easy DVD creator was used to burn the uncompressed video clips on 4.7GB DVD's. These final, uncompressed AVI movie clips are decompressed into single frames that can be processed by the video analysis software.

Before being tested for saliency, the individual video frames must undergo a series of pre-processing steps. Most notably, the videos often contain features resulting from the ROV itself (i.e. portions of the ROV, bright spots introduced by the lights, etc.) These features are usually constant in time for at least several seconds. By estimating the constant background features for each frame by computing the sliding average over the ten preceding frames. The average is then subtracted from each frame, setting negative values to zero. For a complete description of pre-processing steps refer to Edgington et al. (in press) [4]. After processing is complete, each frame is scanned for salient locations based on the attentional model proposed by Itti et al. (1998) [12]. To date, these analyses have only been performed on random video clips to locate potentially "interesting" regions in the frames. Thus far, no recognition of specific organisms has been introduced. The following steps are intended to lay the preliminary foundation for future organism recognition work.

2.2 DIVULGING MONTEREY BAY'S MOST COMMON MARINE ORGANISMS

The video information management system (VIMS) at MBARI contains information on all organisms annotated in transect and non-transect videos from 1989 through 2003. The system was queried to reveal a complete set of annotations recorded from 1997 to 2002 in the top 1000m of the water column. Only these six years of data were considered in the analysis of organism frequency because the MBARI team of video annotators changed in 1997. Combining the work of current and past video annotators may have introduced undesirable bias attributed to variation in viewing/identification/labeling techniques. For each entry in the output, a complete, detailed set of returns was obtained from the VIMS system; however, for the purposes of the AVED project, only a few select returns were considered (Figure 4). Also, because this project is primarily concerned with mid-water transect data, only returns from the MBARI mid-water scientists were considered (Robison, Matsumoto, Sherlock, Hamner, Raskoff, and Reisenbichler).

All returns from the VIMS queries were imported into a Microsoft Excel spreadsheet and data was sorted based on depth and organism type. Using PivotTables, yearly observed frequency was calculated for all organisms. The data was averaged over all seven years and the organisms were ranked based on their relative abundance in the non-transect clips. Organisms were considered based on their mean percentage (± standard error) of the total yearly number of annotations. Transect annotations were not included in this calculation because transects are often selectively annotated for specific target organisms and do not contain an accurate representation of the actual range of organisms present.

Because of the enormity of the resulting dataset, 10 occurrences for each of the top 25 most common organisms were randomly selected and AVI video clips were created to span the corresponding timecode.

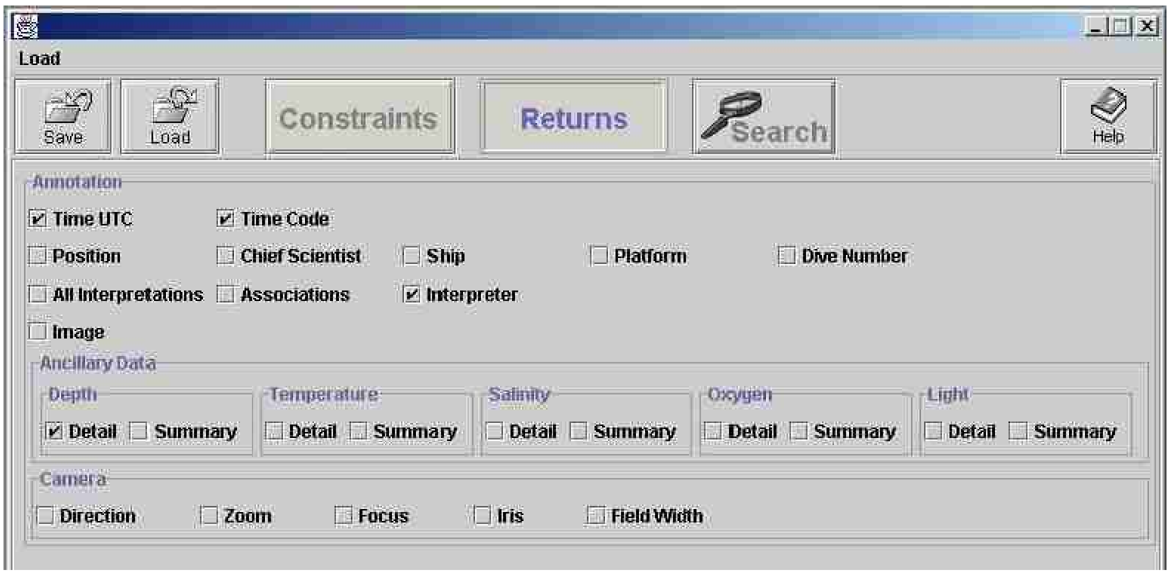

Figure 4: Specific returns needed from the MBARI web-based VIMS query system. Output from these queries can then be imported as a Microsoft Excel file to analyze the frequency and distribution of the observed marine organisms.

RESULTS

MONTEREY BAY'S MOST COMMON ORGANISMS:

A total of 236,186 annotations and 341 different types of organisms were returned from 1997 to 2002 (mean number of annotations = 39,400 ±16,000 s.e.). Though there was some seasonal and annual variability between organism abundance and depth distribution, the most common organisms were apparent, comprising an average of 0.73% to 17.3% of all yearly annotations. Uncommon organisms generally accounted for less than 0.50% of the yearly annotation total and often were observed on solitary occasions. Averaging 17.3% of the total annual annotations, the gelatinous siphonophore, *Nanomia spp.* was conclusively Monterey Bay's most common organism (Figure 5).

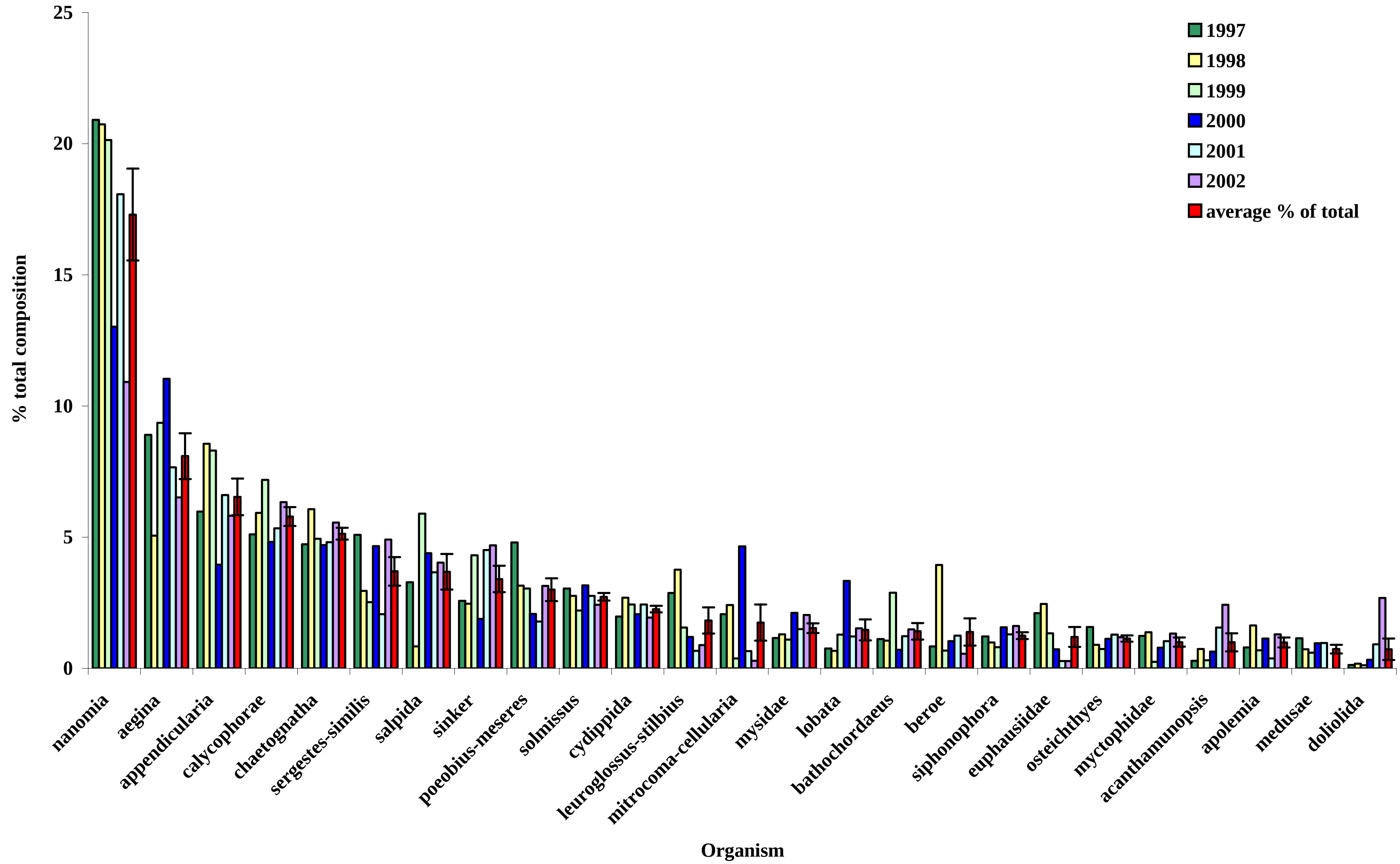

FIGURE 5: The top 25 most common organisms in the upper 1000m of the Monterey Bay water column, presented as the total percentage of yearly annotations. Since there was some variability in annual organism abundance, only the mean percentage (± s.e.) was considered to designate relative rankings.

DISCUSSION

Thus far, the AVED project operates purely from a bottom-up analysis; however, with the completion of this preliminary test database, experimentation with top-down, task-specific recognition and attention can begin. Ultimately the two-component framework of this software aims to mimic the human visual system, largely eliminating the need for human video annotators. If successful, the eventual integration of the attentional module and the recognition module could transform the world of marine ecology and ocean science.

Because the recognition model has never been tested in a marine environment, the success in training the software system to identify these 25 most common organisms remains uncertain. Certain large and conspicuous organisms should be easy for the AVED system to detect and identify. Yet, the natural environment is far from simplistic and the organisms present are not all salient. Though abundant, many of the most common organisms are extremely small, inconspicuous, and to the untrained eye, are entirely indistinguishable from marine snow. Additionally, many of the organisms (particularly the gelatinous animals) are extremely similar in appearance; real-time differentiation in QVT clips remains an often futile task for video annotators. Because of these frequent similarities in organism appearance and the inconspicuousness of other animals, it should be acknowledged that complete trainability of the AVED system will be difficult.

Though object recognition has long been of interest to computer software engineers and neuroscientists alike, realization of this task is exceedingly difficult. Previous successes with recognition have only been achieved in terrestrial environments. The impending computer learning process in the marine ecosystem will be largely experimental in nature. The next step in this lengthy process is to send the video clips to our collaborator's lab, Pietro Perona, at the California Institute of Technology. At the lab, defining characteristics of each test organism will be determined. Specific traits such as color, contrast, orientation, and motion will be averaged using the 10 representative clips and stored in the system. From the database, it will then be possible to test the AVED program with novel video clips and see if it can identify the organisms present as one of the top 25 most common organisms, or else label it as an unknown object. If these tests are successful, a larger test database will continue to be amassed. The goal is to eventually build a complete library database, containing 100 representative (transect) clips of the top 100 most common pelagic organisms.

Once the dual attentional/recognition module is in place, additional objectives for the AVED software can be tested. For the purpose of specific organism studies, the AVED system could be programmed to detect individual targets in the water column. Though this would narrow the focus of the project, specific search tasks promise to be an invaluable resource for many scientists.

CONCLUSIONS/RECOMMENDATIONS

Once operational, the automated visual event detection (AVED) system will provide great insight into the world of marine ecology and fishery research, especially by concentrating on the otherwise difficult question of gelatinous organism diversity, biomass, and abundance. New answers to the questions of organism succession, vertical and lateral migration, variability, community complexity, and food availability soon will be attainable. I summarize here the progress to date in the development of the automated visual event detection project and the future steps to be taken in the successful implementation of a computer vision system.

ACKNOWLEDGEMENTS (Normal, Times New Roman, 12 pt, bold)

I would like to thank Duane Edgington for his guidance on the project and for giving me a privileged glimpse into the world of software development and marine science. Thanks also to Danelle Cline, Rob Sherlock, Dirk Walther, and the MBARI video lab staff for their help on the computers and in general concept-clarification. I am extremely grateful to George Matsumoto for his tireless efforts to make the internship an enjoyable and rewarding experience for everyone and to the rest of MBARI for making the research possible. Cheers and best of luck to all the 2003 summer interns!

REFERENCES:

1. Chung, D., R. Hirata, T.N. Mundhenk, J. Mg, R.J. Peters, E. Pichon, A. Tsui, T. Ventrice, D. Walther, P. Williams, and L. Itti (2002). A new robotics platform for neuromorphic vision: Beobots. *Proc. 2nd Workshop on Biologically motivated computer vision (BMCV '02), Tuebingen, Germany,* pp. 558-566.

2. Dhavale, N. and L. Itti (in press). Saliency-based multi-foveated MPEG compression. *Proc. IEEE Seventh International Symposium on signal processing and its applications, Paris, France*.

3. Edgington, D.R. and AVED lab (2003). MBARI internal project proposal. *Moss Landing, California.*

4. Edgington, D.R., D. Walther, K.A. Salamy, M. Risi, R.E. Sherlock, and C. Koch (in press). Automated event detection in underwater video. *Proc. MTS/IEEE Oceans 2003 Conference, San Diego, California.*

5. Itti, L. (2003). Visual Attention. In: *The handbook of brain theory and neural networks, 2nd Ed.,* edited by M.A. Arbib, MIT Press, pp. 1196-1201.

6. Itti, L. (in press). Modeling primate visual attention. In: *Computational Neuroscience: A comprehensive approach,* edited by J. Feng, Boca Rato:CRC Press.

7. Itti, L. (in press). The beobot platform for embedded real-time neuromorphic vision. In: *Advances in neural information processing systems, Vol 15,* edited by T.G. Dietterich, S. Becker, and Z. Ghahramani, Cambridge, MA:MIT Press.

8. Itti, L. (in press). Toward highly capable neuromorphic autonomous robots: beobots. *Proc. SPIE 47 annual International symposium on optical science and technology,* **4787**.

9. Itti, L. and C. Koch (1999). Target detection using saliency-based attention. *Proc. RTO/SCI-12 Workshop on search and target acquisition (NATO unclassified), Utrecht, The Netherlands*: 3.1-3.10.

10. Itti, L., and C. Kock (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research* **40(10-12)**: 1489-1506.

11. Itti, L. and Koch (2001). Computational modeling of visual attention. *Nature reviews Neuroscience,* **2(3)**: 194-203.

12. Itti, L., C. Koch, and E. Niebur (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions of pattern analysis and machine intelligence*, **20(11)**: 1254-1259.

13. Miau, F., C. Papageorgiou, and L. Itti (2001). Neuromorphic algorithms for computer vision and attention. *Proc. SPIE 46 annual international symposium on optical science and technology*, **4479**: 12-32.

14. Mundhenk, T.N and L. Itti (2002). A model of contour integration in early visual cortex. *Proc. 2nd workshop on biologically motivated computer vision (BMCV'02), Tuebingen, Germany*, pp. 80-89.

15. Navalpakkam, V., and L. Itti (2002). A goal oriented attention guidance model. *Proc. 2nd workshop on biologically motivated computer vision (BMCV'02), Tuebingen, Germany*, pp. 453-461.

16. Navalpakkam, V. and L. Itti (in press). Sharing resources: Buy attention, get object recognition. *Proc. International workshop on attention and performance in computer vision (WAPCV'03), Graz, Austria*.

17. Robison, B.H. 2000. The coevolution of undersea vehicles and deep-sea research. *Marine Technology Society Journal*, **33**: 69-73.

18. Robison, B.H., K.R. Reisenbichler, R.E. Sherlock, J.M.B. Silguero, and F.P. Chavez (1998). Seasonal abundance of the siphonophore, *Nanomia bijuga*, in Monterey Bay. *Deep-Sea Research II*, **45**: 1741-1751.

19. Walther, D., L. Itti, M. Riesenhuber, T. Poggio, and C. Koch (2002). Attentional selection for object recognition – a gentle way. *Proc. 2nd workshop on biologically motivated computer vision (BMCV'02), Tuebingen, Germany*, pp. 472-479.

20. Walther, D., D.R. Edgington, K.A. Salamy, M. Risi, R.E. Sherlock, and C. Koch (2003). Automated video analysis of oceanographic research. *Proc. of the IEEE computer vision and pattern recognition conference, Madison, Wisconsin.*

21. Zhang, Z.Y. (1994). Token tracking in a cluttered scene. *Image and vision computing*, **12(2)**: 110-120.

ADDITIONAL RESOURCES:

VIMS query available at: http://mww.shore.mbari.org/itd/video/VIMS/vimsquery.htm

RAD game tools available at: http://www.radgametools.com/default.htm