
# Developing Autonomous Classification with EITS and AVEDac

**Gregor Bwye, University of Aberdeen, Scotland.**

***Mentors: Danelle Cline and Duane Edgington***

***Summer 2011***

**Keywords: AVEDac, Automation, Classification, Eye-in-the-Sea**

## ABSTRACT

Remotely operated vehicles (ROVs) have provided the marine science community with vast and invaluable volumes of raw video data. Manually annotating large quantities of this data is labor intensive and time-consuming, thus the induction of a tool to aid this process. AVEDac is a tool to analyze video data such as the Eye-in-the-Sea autonomous deep-sea viewing platform, which adopts a discreet approach in monitoring the behavior of deep-sea organisms. This study will look at calibrating the AVED classifier to detect deep-sea squid and fast-moving events within the sporadically instances of visual noise unique to the EITS data set. Critical analysis of the classifying performance will also scrutinized in the form of developing receiver operating characteristic (ROC) curves for each image class. Results showed that squid was incorrectly detected through miss-classification of another class. The results from the ROC curves indicate that further testing is needed to evaluate the accuracy of the classifier. Conclusively, AVEDac is still in its infancy in terms of achieving the desired goal of utilizing such automated applications to aid manual annotation.

## 1. INTRODUCTION

Remotely operated vehicles (ROVs) have provided the marine science community with vast and invaluable volumes of raw video data. Manually annotating large quantities of this data is labor intensive and time-consuming, thus the induction of a tool to aid this process.

The initial Automated Visual Event Detection and Classification (AVEDac) system was developed by the Monterey Bay Aquarium Research Institute (MBARI) and is a tool to analyze specially processed video data collected by the institutes' ROVs (Edgington *et al*, 2006). In addition, AVEDac has been modified to analyze data from autonomous video monitoring applications such as EITS - powered by underwater cabled-observatories such as the Monterey Accelerated Research System (MARS) at MBARI (Cline *et al*, 2009). The Eye-in-the-sea (EITS) experiment was designed by Ocean Research and Conservation Association's (ORCA) Dr. Edith Widder and was recently deployed at the MARS site in the fall of 2009.

EITS is an autonomous deep-sea viewing platform and adopts a discreet approach in monitoring the behavior of deep-sea organisms. In contrast to the widely accepted notion of disturbance caused by bright white lights from ROVs, EITS uses far-red illumination (695nm) that is considered to be beyond the visible spectrum of most deep-sea fishes – implementing a low profile method (Raymond & Widder, 2007). The viewing platform comprises of an intensified video camera in order to compensate for the rapidly absorbed red light that can be positioned in view of a bait box or an optical lure intended to attract deep-sea organisms. This report will only refer to data collected from the optical lure (termed electronic jelly) source.

The electronic jelly (e-jelly) was programmed to generate 3 distinctive display patterns based on deep-sea bioluminescent model organisms. The pinwheel display sequence mimics patterns produced by Scyphozoan *Atolla wyvillei* (Haeckel, 1880); the dim LED and single repetitive flash sequences simulate bacterial and Ostracod species respectively. Results from pinwheel display analysis show that the presence of deep-sea squid *Dosidicus gigas* (Orbigny, 1985) may exhibit distinct antagonistic behavioral responses to the e-jelly. Data was recorded using conventional methods of annotation through use

of a spreadsheet. There is insufficient evidence to suggest any significant ecological relationship between *D. gigas* and *A. wyvillei,* however recent documented behavioral studies in squid demonstrate similar aggressive tendencies towards blue-light sources *in situ* (Kubodera *et al,* 2007).

This study aims to train AVEDac to detect squid interactions within a sub-set of the EITS data with respect to the remaining 2 luminescent display patterns produced by the e-jelly. The primary objective of using AVEDac in this study involves analyzing large subsets of data in order to evaluate the current status of utilizing such automated annotation tools in comparison to human annotation. Studies have shown the benefits of using such machines to help identify and understand the biology of organisms that we may use for food resources or other commercial exploits (Aguzzi *et al*, 2009; Costa *et al*, 2008) and to aid in taxonomical classification (Culverhouse *et al*, 2003). Testing of the autonomous Batch Classification feature will also be included in the analyses. Results from this study will be used to document notable differences in the behavior of squid under 2 different artificial luminescent displays *in situ*.

## 2. MATERIALS AND METHODS

### 2.1 Workflow for EITS data

Digital video collected from the EITS camera system is transmitted back to shore via high-powered Ethernet connectivity and stored as 5-minute individual clips on the RAID server. The individual clips undergo a continual processing sequence before a designated user can utilize the data within the AVEDac user interface.

To manage the workflow for EITS, a mass workload management system called Condor is used. Condor provides scheduled queuing for mass volumes of data to be processed. This is achieved by submitting the clips into an 8-node, 16 CPU Beowulf cluster (Cline *et al*, 2008). Once processed, the individual clips are saved and stored as metadata XML files which can then be imported into the AVEDac user interface for annotation. This workflow is represented in Figure 1.

Due to time constraints, only 0-40 minutes of each hour from the sub-set data 16th – 24th March 2010 was processed through AVEDac in single 5-minute clips (n=1292). This excluded day's 19th, 21st and 23rd where only 0-20 minutes was processed (when e-jelly was switched off). Due to technical errors, the first and last day of the data set (16th and 24th) have only a partially completed data set.

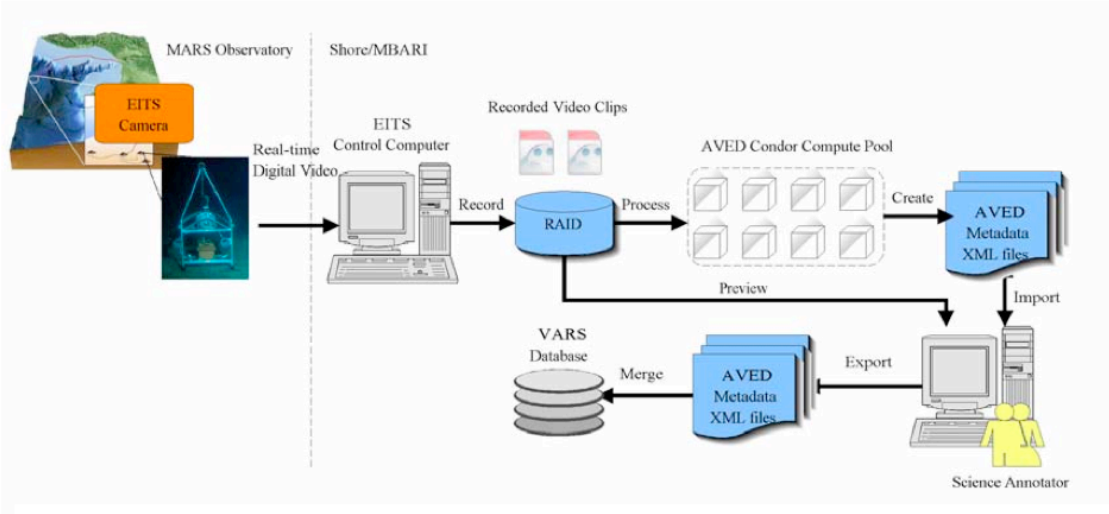



*Figure 1. Eye-in-the-Sea/AVED workflow (taken from Cline et al, 2007) from collection of raw data from autonomous lander system to user importing processed XML files.*

## 2.2 Basic Fundamental Parameters of AVEDac

AVEDac uses a neuromorphic vision algorithm (based on human vision) to detect events within the raw video data collected by ROVs. Events that may be regarded as "interesting" are determined by the potential event being tracked across several frames using linear Kalman filters (Cline *et al*, 2008).

For the purpose of this study, AVEDac was specifically calibrated to regard likely events with rapid motion by fast-moving organisms (such as deep-sea squid) as "interesting" to

maximize detection of species of interest. An event can vary in duration from a single frame to the entire length of the processed clip.

### 2.2.1 Image/Training libraries

Image libraries for organisms and classes of interest were predominantly created using previously analyzed images from the EITS pinwheel display sequence from February 2010. In this study, 5 initial distinct classes were used to classify the un-annotated March 2010 sub-set. The image classes were composed of Rattails, Rockfish, Squid, 'Noise' and 'Unknown organisms'. The latter class refers to organisms or organic material not of particular interest for the purpose of this investigation (such as micro-fauna on the benthos and drifting POM).

The inclusion of 'Noise' as a class in the training library was used to test the plasticity of the AVEDac classification tool. The EITS data set has periodical instances of visual noise and 'drop-outs' (caused by technical errors with the digital video encoder) that may reduce the overall effectiveness of the classifier. To account for this, a separate 'Noise' class was included in the training library to act as a potential filter for later batch classifications of unanalyzed portions of data. This filtering concept however is largely dependent on the accuracy of the classifier.

Training libraries are composed of groups of user-defined image classes for AVEDac to relay against un-annotated data. The training library compiles every image from the assigned classes and is compared to an un-annotated event by the classifier. In this study 2 training libraries were tested, one containing all 5 initial classes ('Dominator') and the second with 6 classes ('Re-vamp'). Re-vamp is a modified version of Dominator and includes the insertion of a Sea anemone class. In principal the two libraries share the same images, however images deemed potentially poor quality or 'destitute' in the Re-vamp library were removed. It is important that the annotator carried out this procedure.

In order to verify the quality of the image classes, each class is tested against a training library. This preliminary analysis tests against a 10% sample of randomly selected

images within the class and results are compiled in a confusion matrix (later described). In this study all 7 classes were tested and demonstrated 90-100% accuracy.

### 2.3 Batch Classifier

Each training library was created using the appropriate image classes and was run against the March data set (per day-directory). Each directory was imported from the EITS-processed server and run against one training library at a time. The probability threshold for this experiment was set to 80%. The autonomous batch classifier has an in-built feature that exports the output results in an Excel table format, stating the number of predicted classes per 5-minute clip.

In addition to the user defined image classes, the batch classifier automatically inserts a default 'Unknown' column into the results table. This functions as a bin for events that are regarded as unidentifiable by the classifier with respect to the pre-determined group of images within the training library.

### 2.4 Classifier Accuracy

In order to assess the validity of the output results generated by the batch classifier, human annotation was utilized as comparative measure. An annotator labeled a randomized portion (n=60) of the total data set and then allowed commencement of the AVEDac classifier. Only the Revamp training library was used for this part of the study. Thus, the construction of a confusion matrix table was an appropriate visualization tool for comparing the two methods of annotation. The classification probability thresholds were altered to determine the variance in accuracy under different conditions (20%, 50%, 80% and 95%). The same labeled data set was used at all times.

### 2.5 Data analyses

Evaluating the output of results required comparison of total events correctly/incorrectly labeled by the classifier. These values were prepared in Excel (the default format output

of the batch classifier) and run through a confusion matrix function in MATLAB. A Receiver operating characteristic (ROC) curve was determined per class (using the values from a confusion matrix) to illustrate the varied discriminatory thresholds of the classifier. ROC curves graphically plot the true positive rate (sensitivity) against the false positive rate (1-specificity) as outlined below:

- True Positive Rate (TPR) = TP/(TP+FN)
- Specificity = TN/(TN+FP)
- False Positive Rate (FPR) = 1-Specificity

Where TP, FP, TN and FN represent True positive, False positive, True negative and False negative respectively. Original confusion matrices are not included in this study.

## 3. RESULTS

### 3.1 Batch classification

The results from the batch classifier show occurrences of all 5/6 classes within the March data set (n= >31,000) across both training libraries. The classifier detected 548 events of predicted squid when trained with the Dominator library and 317 events with Re-vamp (see Table 1). In comparison to the large number of detections of Noise, it is apparent that the predicted occurrences of squid are notably lower than other classes. Both training libraries predict a large percentage of the data set to be Noise (Dominator – 40% and Re-vamp – 64%, see Table 2).

There is little difference in the number of 'Unknown' and 'Squid' classified events in between two training libraries. However the results also indicate a clear discrepancy in values upon alteration of a training library. There is a reduction in percentage values for events per class when trained with the Re-vamp library such as Rattail and Rockfish (see Table 2).

*Table 1. Results of the classifier showing the sum of all classified events assigned to appropriate classes for March data set between two training libraries 'Dominator' and 'Re-vamp'. Total N = >31,000.*

| Training library | *Unknown* | *Noise* | *Rattail* | *Rockfish* | *Squid* | *Unknown org.* | *Sea anemone* |
|---|---|---|---|---|---|---|---|
| Dominator | 6293 | 12628 | 5945 | 1016 | 548 | 4777 | |
| Re-vamp | 6509 | 19923 | 764 | 380 | 317 | 2299 | 1038 |

*Table 2. Percentage composition of events per class. Values taken from Table 1.*

| Training library | *Unknown* | *Noise* | *Rattail* | *Rockfish* | *Squid* | *Unknown org.* | *Sea anemone* |
|---|---|---|---|---|---|---|---|
| Dominator | 20.3% | 40.7% | 19.1% | 3.3% | 1.3% | 15.4% | |
| Re-vamp | 20.9% | 64.3% | 2.5% | 1.2% | 1% | 7.4% | 3.3% |

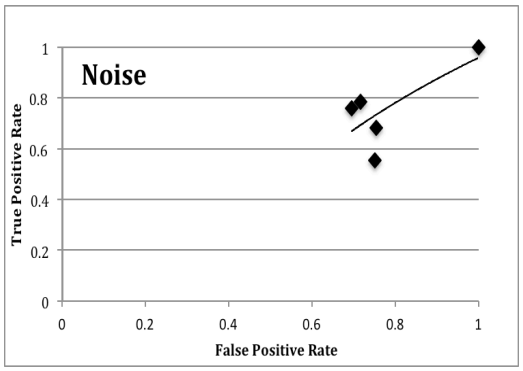
### 3.2 Accuracy of Classification

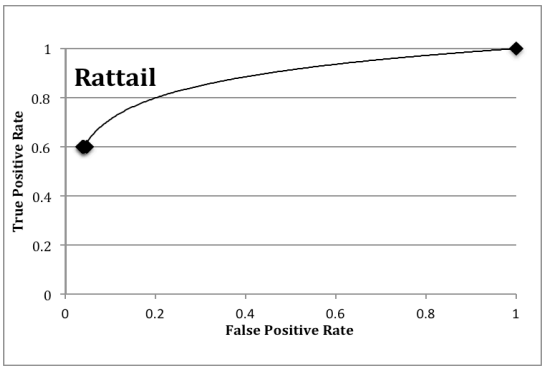
The randomized data set was found to have 7 clips with no detected events, thus reducing the sub-set sample (N=53). There were annotator-labeled events containing classes which were not constituted in the training library. As this section of the study focuses on testing the accuracy of the classifier, it was deemed appropriate to remove these 'alien classed' events from the data set, further reducing the total number of events (N=1066) between the 53 clips.

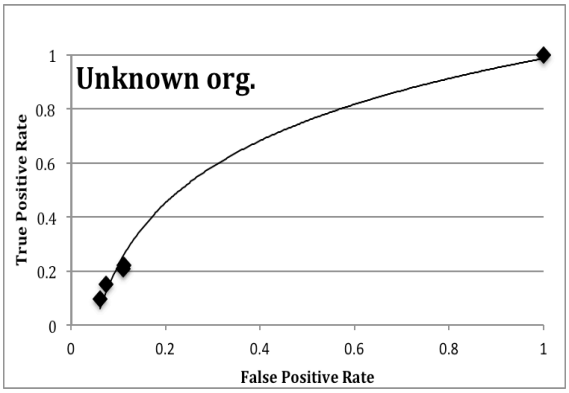
*Table 3. Percentage accuracy of classifier under varying probability thresholds (20, 50, 80 and 95) of labeled event sub-set (N=1066) using Re-vamp training library.*

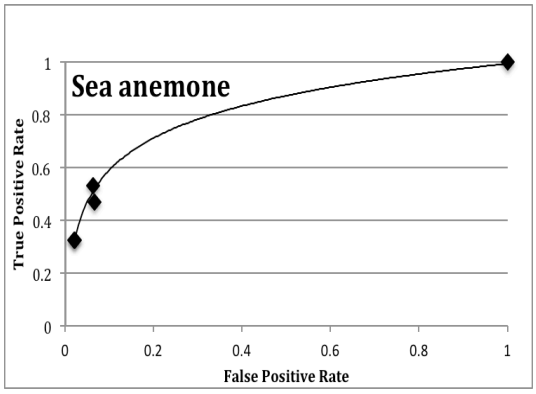
| | Unknown | Noise | Rattail | Rockfish | Squid | Unknown org. | Sea anemone |
|---|---|---|---|---|---|---|---|
| PT20 | N/A | 78.21% | 60% | 20% | N/A | 21.94% | 32% |
| PT50 | N/A | 75.62% | 60% | 20% | N/A | 20.38% | 52% |
| PT80 | N/A | 68.09% | 60% | 20% | N/A | 15.05% | 32% |
| PT95 | N/A | 55.51% | 60% | 6.67% | N/A | 9.40% | 46% |

There were no detections of 'squid' or 'unknown' classes within the sub-set sample therefore no critical analysis on these classes can be drawn. 'Noise' was classified with the highest accuracies (55-78% accuracy) and 'Rattail' and 'Rockfish' indicating consistent classified accuracy (60-20% respectively) under increasing probability thresholds. With the exception of 'Sea anemone' and 'Rattail', all classes show to have reducing accuracies under increasing probability thresholds.





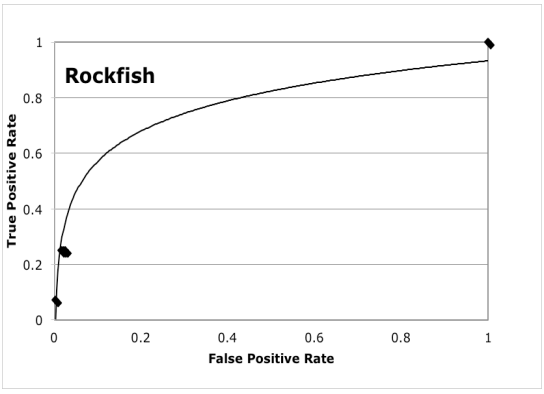




*Figure 2. ROC curves representing the accuracy of classification for each class under probability thresholds 20%, 50%, 80% and 95%. Data points were calculated using values from confusion matrices.*

The ROC curves in Figure 2 display trade-offs between true positive rates and false positive rates. With the exclusion of 'Noise', all classes show a similar trend with false positive rates less than 10%. 'Noise' displays high true positive values but also displays high false positive rates irrespective of varying probability thresholds. 'Rattail' and 'Sea anemone' show high true positive rates however, as already mentioned, 'Rattail' displays consistent 60% classification accuracy with little variance in false positive rate against increasing probability thresholds.

## 4. DISCUSSION

The AVED classifier detected few instances of squid within the time frame 16th-24th March 2010. However through critical analysis from the raw video data, it was apparent there were no squid interactions with the e-jelly irrespective of the results from the batch classifier. Consequently, there is no data to evaluate any variance in behavior under different luminescent displays. The high instances of events in Table 1 within the 8-9 day time frame would suggest an active period around the e-jelly in terms of regarded interesting events. Accordingly, this could imply reasoning into adopting such automation tools that would reduce the need for time-consuming manual annotations. However there are other factors that need to be addressed before vindicating this notion.

It was observed that there were large portions of the data that were composed of relatively short or single frame events as opposed to long, continuous events. With

respect to the Noise class that contributed around half of the total event detections, most if not all of the events classed as Noise were brief and fragmented. There were occasions where a potential continuous event displayed by a mobile organism would be broken into shorter fragments, which may be due to forced errors by Noise to the tracking filter. Conversely, the miss-classifications of Noise as other classes also significantly increase the number of events per class.

The use of two different training libraries illustrated a change in the number of predicted classes per class. 'Rattail' showed a decrease by 16.6% with the insertion of an additional class. It is difficult to suggest an appropriate solution to this change in class composition; other than the notion of providing an additional class allows more option for classifier. The increased percentage of Noise and decreased percentage of the other classes may verify this.

Another challenging concept faced in this study was evaluating the accuracy of the classifier. Table 3 indicates what may be regarded as good accuracy in classifying Noise, however the ROC curves (Figure 2) suggest that moreover there are high rates of false detections. It is difficult to quantify this with respect to the general low abundances of actual mobile organisms (e.g. Rattail) in comparison to the high counts of Noise in the particular sub-set of data used. This implies random, larger and multiple data sets should be used for future evaluation. This issue is also apparent in both Table 3 and Figure 2 where the classification accuracy for Rattail is saturated (60%) for all 4-probability thresholds tested.

Table 3 indicates that the increasing probability thresholds reduce the overall accuracy of classification (with the exception of Sea anemone which does not follow any trend). As this study is the first instance of testing such applications, it is unjust to provide any tangible explanation without comparison to further tests.

It important to consider the quality of images used for training AVEDac. While it is apparent that Noise had a significant influence in this data set, furthermore there were numerous misclassifications between the other faunal classes (Rockfish, Rattail, Squid etc). As the ultimate goal of this application focuses on the ability to distinguish between genus, class, species etc in a marine habitat, it is imperative to understand what common

feature an array of organisms may have that lead to high numbers of false positive classifications. In context with this study, one may presume the monochrome video quality may hinder the effectiveness of discriminating between classes. In addition, there were instances of the event area detected potentially being undersized therefore difficult for effective classification. A solution to this problem may lie within narrowing the parameters in the Gaussian algorithm further to regard certain sizeable areas to be interesting (if seeking to detect specific megafauna). However with many of the proposed solutions mentioned, they can only be justified upon further testing.

## CONCLUSIONS/RECOMMENDATIONS

The results of this study infer autonomous classification to show variance in accuracy, there is evidence to suggest tools such as AVEDac may be useful for certain data sets. The data used in this study showed to have excessive volumes of noise throughout and proved to be one of the greatest challenges. One may speculate the difference in results from both the batch classifier and accuracy testing if a cleaner, less-noisy data set was implemented. Changes in methods would provide a better platform to suitably assess the accuracy of the classifier. Testing with the addition/removal of images within class libraries would expose the features needed for optimum performance of Bayesian classifier. Increasing the randomized sample size for comparing human and classifier annotations would augment the feasibility of the end results. In addition, increasing repetitions with different randomized sub-sets would improve validity in testing for accuracy from a confusion matrix. In order to reduce bias during the human annotation step, the user should only refer to the event thumbnail image when labeling the event instead of viewing the entire event through the event player. The latter method gives the annotator the advantage of putting the event in context before making the final decision to classify the event. Testing with enlargement or reduction of training libraries with new or less classes would furthermore aid in understanding the optimal features that increase the classifiers performance.

In conclusion, the results of this study indicate there is no meticulous direct method to achieve reliable autonomous classification. Although AVEDac is still in its infancy in terms of achieving the desired objective, this study has exposed certain conceptual features that require further investigation.

**ACKNOWLEDGEMENTS**

This project would not have been possible without the help of my mentors Danelle Cline and Duane Edgington. I would also like give a special thanks to Erika Montague to whom I am indebted. The depth and breadth of their combined knowledge formed a remarkable foundation on which my internship was based, and they were always there when I needed advice. Given my limited background with software programming and artificial intelligence in general, I can only show enormous gratitude to Danelle and Erika for explaining new complex concepts in their simplest form. It was a challenge that I enjoyed every minute of throughout the 10-week internship. I wish them further future success in their work.

However, this summer was not only about my project work. I had the chance to meet with scientists and speakers I would otherwise never have met. The excursions to various locations with fellow interns within the area really made my time here in California unforgettable. It was this work/play ratio that rounded up an incredible internship. For this, I owe my thanks to George Matsumoto and Linda Kuhnz for providing the foundations of this internship. That said, I wish the best to my fellow interns in their future endeavors. I must also include a big thank you to the entire staff at MBARI.

Finally, I would like to say thank you to the David and Lucile Packard Foundation for funding my internship and giving me the opportunity to experience a taste of working in deep-sea research.

**References:**

Aguzzi, J., Costa, C., Menesatti, P., Garcia, J.A., Chiesa, J.J., Sarda, F. (2009). Monochromatic blue light entrains diel activity cycles in the Norway lobster, *Nephrops norvegicus* (L.) as measured by automated video-image analysis. *Scientia Marina*, 73(4): 773-783.

Costa, C., Aguzzi, J., Menesatti, P., Antonucci, F., Rimatori, V., Mattoccia, M. (2008). Shape analysis of different populations of clams in relation to their geographical structure. *Journal of Zoology*,276: 71-80.

Cline, D.E., Edgington, D.R., Mariette, J. An Automated Visual Event Detection System for Cabled Observatory Video. *3rd International Conference on Computer Vision Theory and Applications*, 22-25 January, 2008. Funchal, Madeira Portugal.

Cline, D.E., Edgington, D.R., Smith, K.L, Vardaro, M.F, and Kuhnz, L., An Automated Event Detection and Classification System for Abyssal Time-Series Images of Station M, NE Pacific. *MTS/IEEE Oceans 2009 Conference Proceedings*, October, 2009. Biloxi, Mississippi 2009.

Culverhouse, P.F., Williams, R., Beatriz, R., Herry, V, and Gonzalez-Gill, S. (2003). Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. *Marine Ecology Progress Series*, 247: 17-25.

Edgington, D.R., Cline, D.E., Davis, D., Kerkez, I., and Mariette, J., Detecting, Tracking and Classifying Animals in Underwater Video (060331-207). *MTS/IEEE Oceans 2006 Conference Proceedings*, Boston, MA September 2006. IEEE Press.

Kubodera, T., Koyama, Y, and Mori, K. (2007). Observations of wild hunting behaviour and bioluminescence of a large deep-sea, eight-armed squid, *Taningia danae*. *Proceedings of the Royal Society B,* 274: 1029-1034.

Raymond E,H and Widder, E.A. (2007). Behavioral responses of two deep-sea fish species to red, far-red, and white light. *Marine Ecology Progress Series,* 350: 291–298.