

## **Development of Automated Image Analysis Software for Suspended Marine Particle Classification**

Scott Samson  
Center for Ocean Technology  
University of South Florida  
St.Petersburg, FL 33701-5016  
phone: (727) 423-4805 fax: (727) 553-3967 email: [samson@marine.usf.edu](mailto:samson@marine.usf.edu)

Dmitry Goldgof  
Computer Science and Engineering  
University of South Florida  
4202 E. Fowler Ave,  
Tampa, FL 33620  
Phone: (813) 974-4055 fax: (813) 974-5456 email: [goldgof@csee.usf.edu](mailto:goldgof@csee.usf.edu)

Thomas Hopkins  
College of Marine Science  
University of South Florida  
St.Petersburg, FL 33701-5016  
phone: (727) 553-1501 fax: (727) 553-3967 email: [thopkins@marine.usf.edu](mailto:thopkins@marine.usf.edu)

Lawrence Hall  
Computer Science and Engineering  
University of South Florida  
4202 E. Fowler Ave,  
Tampa, FL 33620  
Phone: (813) 974-4195 fax: (813) 974-5456 email: [hall@csee.usf.edu](mailto:hall@csee.usf.edu)

Award #: N00014-02-1-0266

### **LONG-TERM GOAL**

The goal of this project is to develop a broadly-capable software package, to automatically classify digital images of zooplankton. The software is being developed initially for classification of images from the SIPPER linescan imaging instrument, but will also be able to process images from imaging systems that produce high quality digital image. Towards this end, the software is being created to include both training and multi-stage classification portions. The input is digital images in standard computer format. The output is an Internet-addressable classified particle database, including pertinent particle hydrographic information and sorted images. The project has application in rapid classification of microscopic marine particles, which have an effect on the optical properties and long-term variation in the local and global water column.

## **OBJECTIVES**

The project's objective is to develop automated image analysis software to reduce the effort and time required for manual identification of plankton images. Automated analysis will reduce the need for expert identification, produce quick results, and improve the reliability and repeatability of results over extended deployments. The software includes a training phase, identification phase, and a database retrieval phase. The training phase allows a trained user to enter expert-classified particle images into the system. The output of the training stage is a set of distinctive features and a set of classification boundaries (support vector machine), which is used as control parameters during the identification phase. The identification phase includes pre-processing to eliminate noise, feature computation, and image classification. The database retrieval phase allows the user to track and efficiently retrieve information (particle identity, size, features, location, and other water properties) on the anticipated billions of particles. The software is being developed for use on a commercially available personal computer, to allow for widespread use and simple shipboard transport.

## **APPROACH**

The testing and development of automated plankton image recognition software relies on high-resolution grayscale and binary SIPPER images [1] obtained in the field in the eastern Gulf of Mexico, from another ONR project (N00014-96-1-5020). We use a broad range of manually identified plankton images for the development, characterization, and refinement of various algorithms; these images encompass a challenging, yet accurate representation of the diverse subtropical coastal ecosystem sampled. The software has been improved in the past year with respect to: handling unidentifiable particles in the water (noise), speed, and capability to classify the new grayscale images produced by the recent version of the SIPPER instrument. The approach can be broadly broken down into several tasks: image decompression, feature extraction, training, classification, and database access. The software is being developed so it will be field-tunable for application in differing ecosystems.

Graduate students Tong Luo, Kurt Kramer, and Padmanabhan Soundararajan are primarily involved in implementation of the software, with guidance provided by Dmitry Goldgof and Lawrence Hall. Xiaou Tang, from the Chinese University of Hong Kong is acting as a technical consultant on the project, primarily in the development of advanced features applicable to plankton. Marine science professor Tom Hopkins and graduate student Andrew Remsen are providing biological expertise and manually classified images, while Scott Samson is acting as project manager.

## **WORK COMPLETED**

We have continued the development of an automated plankton imaging system previously reported [2]. Updates to the SIPPER imaging instrument have allowed collection of grayscale images, versus the previous binary images. These grayscale images were more easily identifiable by a human expert, so time was devoted to develop additional grayscale software features to more accurately classify these images. We have developed a probabilistic model for determining a confidence level of the classification. This is helpful both in reporting as well as training the software. A more comprehensive user interface is being developed, which allows the user to more effectively train and classify at sea. We have also continued experiments using in-situ collected images.

The software functions in the following method. Grayscale 4096 pixel-wide continuous linescan camera images are available in compressed run-length encoded format from the SIPPER instrument. This continuously-compressed image data is first decompressed and converted to frame images. The vertical dimension of the frame is expanded slightly in instances where a particle is cut by the artificially set boundary. This dynamic frame boundary reduces partial images of plankton. Plankton subimages are located and cleaned of any outlying tiny noise particles within each subimage area. Image features are then calculated for input into the developed Support Vector Machine (SVM) classifier. In addition to the 29 binary image features previously reported (size, shape, contour, Fourier descriptors, etc.), twenty additional grayscale features have been developed. These include grayscale contours, moments, weighted moments, and Fourier descriptors. Classification accuracy using the best binary only feature classifier is near 79%. Adding the 20 features enhances classification accuracy significantly (90.7%) on the same dataset.

The enhanced feature set can be beneficially reduced to a smaller number. This has two advantages over the full 49-feature set- it reduces the computation time, and it improves classification accuracy. To help select the important features, we experimented with 4965 gray-level images from five types of common plankton (copepods, trichodesmium, larvaceans, medusae, and protists). The 10-fold cross validation accuracy with all 49 features is 90.7%. Our feature selection approach (using a greedy beam search) successfully reduced the features to just 13, while achieving an improved 92% 10-fold cross validation accuracy (Table 1). A comparison of the *SVM* classifier to two other classification methods also implemented- *Decision Tree* and a *Cascade Correlation Neural Network* is shown in Figure 2. The SVM classifier produces the best accuracy, as was found in the more limited 29-feature binary work reported last year. The time required to decompress images, compute all 49 features, classify and sort images, and write out information file is around 25 minutes for ten minutes of collected grayscale image data on a Windows 2000 Pentium 4, 2.4GHz computer. Optimizing the feature set reduces computation time additionally.

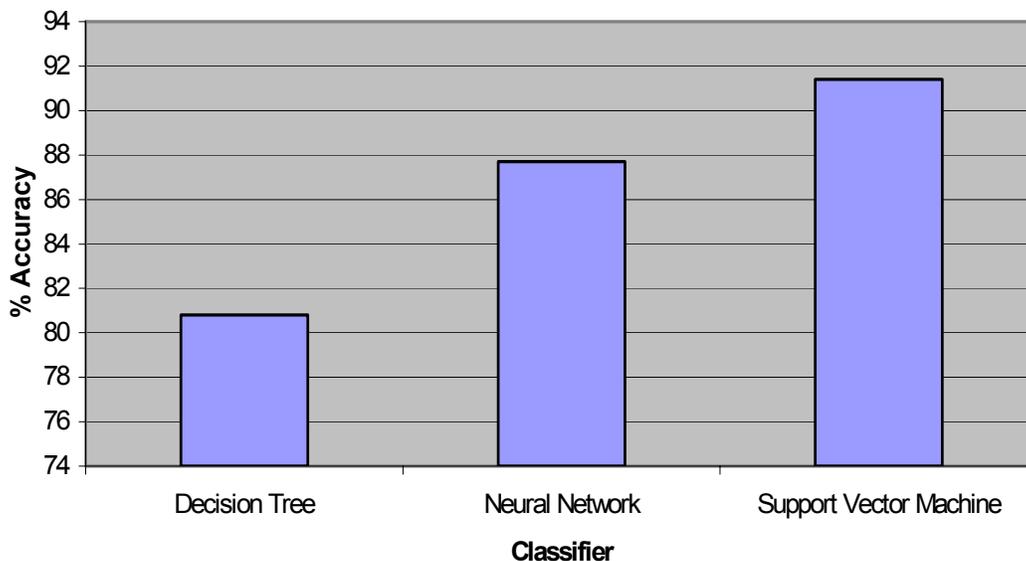
**Table 1- Classification accuracy for only 13 best (of 49 total) binary and grayscale features for five most abundant classes in dataset. Worst confusion occurs between Trichodesmium (Tricho) and Larvaceans, with less than 6% of these images incorrectly classified.**

Class Names =====	Copepod =====	Larvacean =====	Meduae =====	Protist =====	Tricho =====
Copepod (1000)	<b>936</b>	15	16	11	22
Larvacean (1000)	24	<b>910</b>	7	11	48
Medusae (1000)	25	10	<b>920</b>	25	20
Protist (965)	17	5	25	<b>907</b>	11
Tricho (1000)	23	54	14	15	<b>894</b>
<b>Totals(4965)</b>	<b>1041</b>	<b>1019</b>	<b>972</b>	<b>968</b>	<b>965</b>
Copepod (20.1%)	<b>93.600%</b>	1.500%	1.600%	1.100%	2.200%
Larvacean(20.1%)	2.400%	<b>91.000%</b>	0.700%	1.100%	<b>4.800%</b>
Medusae (20.1%)	2.500%	1.000%	<b>92.000%</b>	2.500%	2.000%
Protist (19.4%)	1.762%	0.518%	2.591%	<b>93.990%</b>	1.140%
Tricho (20.1%)	2.300%	<b>5.400%</b>	1.400%	1.500%	<b>89.400%</b>

10 Fold Cross Validation Accuracy **91.984%**

We have implemented and tested shipboard a Windows-based interactive software application for managing image extraction, automated classification, and active learning for the classifier. A trained biologist operates the software in the learning and maintenance modes, and a low level operator in the classification mode. The application consists of the following features:

- a. *Class Maintenance Function*: allows the user to specify the various plankton classes that are to be classified.
- b. *Classification Model Maintenance*: allows the creation and maintenance of models to be used for image classification. These models specify the training classes that are to be used, and their location on the hard disk.
- c. *Image Extraction*: prompts user for run-time parameters such as a SIPPER data file, the destination for decompressed images, and the minimum image size to consider for classifying. It starts execution of the Image Extraction application while displaying a status screen.
- d. *Image Classification*: prompts the user for run-time parameters such as the source images to be classified, training model to be used (models maintained in feature *b*, above), and the destination directory for classified images.
- e. *Active Learning Function*: this is a work in progress; it will allow the user to select a set of classified images, have them presented in a selected order such as by probability, distance from the classifier decision border, or randomly. The user then views and specifies their proper classification while the computer places them in the appropriate training library.



**Figure 2- Comparison of typical accuracy for Decision Tree, Neural Network, and Support Vector Machine classifiers using 49 grayscale features. Accuracy is relative to expert human identified images.**

The function on active learning (*e*) was initiated due to the long time and large effort it takes to identify images, even during training of the system. The concept is to use coarse classification based on an unoptimized classifier developed from previous data. This will speed, simplify, and reduce strain on the expert biologists during the training and retraining of the system. This training and retraining may be done at sea when other shipboard tasks need to be performed. This has eventual significance in military applications, where a particular type of plankton may be of special interest, and the person training the system may not an expert biologist. The goal is to minimize the amount of work that a user

has to do to update existing training libraries with more relevant images. Consider the situation when classifying images from a current cruise. The training library that you start with would initially consist of images from a previous cruise. The training data will should be updated with the newer images being retrieved, to enhance accuracy and account for new classes that may exist presently in the water.

Using the original training data we select candidates to be classified by the user to enhance the current training set. This way the user would not have to deal with searching through millions of images for good candidates. The question then arises: How do you select which images to identify to retrain the system? To answer this, we performed experiments using three active learning approaches: lowest probability, highest probability (reversed probability), and random selection. The *lowest probability* approach selects example images with the lowest classification accuracy probability (e.g. not very similar to any single existing class) to be labeled by the expert. The *highest probability* approach labels examples with highest classification probability (e.g. almost certainly a copepod). *Random selection* means randomly selecting examples to re-label. To test, 13,282 images were divided into 10 groups with five classes of particles evenly distributed amongst the 10 groups. There were two sets of tests run, each using the three different sorting methods for selecting candidate images. The first set selected 20 candidate images from each predicted class to be added to the training libraries, the second set selected 100 images from the results of each test group without regard to class. We compared the three approaches in the two sets of experiments. In both cases, actively selecting examples with lowest probability provided better classification accuracy than random selection. Note that for all three approaches the number of examples needed to be labeled by the expert reduced significantly. A more in-depth discussion will be presented [3].

A web-accessible database is being developed to allow retrieval of classified images based on object class, properties, location, or other physical (e.g. CTD). The database uses PHP hypertext processing and a *mySQL* database to store and retrieve images and information based on queries input on a web page. Currently, the database allows searches on size and object type. We are working to expand the database and search engine to include concurrently sampled hydrographic data. Textual report generation will also be included to enable porting into other applications typically used by scientists.

## RESULTS

Progress has been made toward producing a user friendly and efficient shipboard application for use in classification of marine plankton images. Forty-nine features have been generated for enabling classification. Dependent on local waters, a different subset and weighting of the 49 features is optimum, both for time and classification accuracy. In performing feature selection, 9 of the 20 new grayscale features were almost always used in the optimum feature sets. Thus, grayscale information is extremely beneficial to accurate classification. This information has been used to direct technical development of the SIPPER instrument. The Support Vector Machine algorithm continued to have the best combination of computation speed and accuracy, compared to Cascade Correlation Neural Networks and Decision Tree methods. A new shipboard application using the SVM has been developed which enables training and modification of the system at sea using current images. Accuracy over 90 percent has been achieved, though results vary according to image variation and quality. Obtaining such high accuracy in the presence of noise (small fragments and unidentifiable images) is a current research problem and is being studied.

## **IMPACT/APPLICATIONS**

This project has the potential to reduce the turnaround time for evaluating large plankton image datasets, from months to minutes. Being able to quickly image and sort particles in the marine environment offers the possibility to react to or predict changes in optical or chemical properties of the water resulting from these particles. Consistency, speed, and lack of fatigue offered by computer processing versus human identification, and use by non-biologists are advantages to be seen by automated recognition. The software is being developed to be as generic as possible, to enable its use in a wide variety of marine ecosystems. Although the current approach uses SIPPER instrument images, the standard (bitmap) input format is amenable to other digital imaging systems that may be developed presently or in the future.

## **TRANSITIONS**

The developed software will be used in the near-term to classify millions of available SIPPER images, which will help in measuring the distribution of plankton in the Gulf of Mexico.

## **RELATED PROJECTS**

The SIPPER imaging instrument is being refined and deployed under the ONR-sponsored “Development and Field Application of Laser Particle Imagers” grant (ONR: N00014-96-1-5020-Hopkins et. al.). The field-collected images are being made available for the current project, and the developed software will have a direct and immediate use in the aforementioned grant. The SIPPER web site is <http://marine.usf.edu/sipper>.

## **REFERENCES**

- [1] Scott Samson, Thomas Hopkins, Andrew Remsen, Lawrence Langebrake, Tracey Sutton, and Jim Patten, “A System for High-Resolution Zooplankton Imaging,” IEEE Journal of Oceanic Engineering, v. 26, p. 671-676 (2001).
- [2] Scott Samson, Dmitry Goldgof, Thomas Hopkins, Lawrence Hall, “Development of Automated Image Analysis Software for Suspended Marine Particle Classification,” Office of Naval Research Ocean Atmosphere and Space Fiscal Year 2002 Annual Reports.
- [3] T. Luo, K. Kramer, D. Goldgof, L. Hall, S. Samson, A. Remsen, T. Hopkins, “Learning to recognize plankton”, to appear in IEEE International Conf on Systems, Man and Cybernetics, 2003.

## **PUBLICATIONS AND PRESENTATIONS**

X. Tang, F. Lin, S. Samson, and A. Remsen, “Feature extraction for binary plankton image classification,” Proc. of International Conference on Imaging Science, Systems, and Technology, Las Vegas, USA, June 2003. [published]

T. Luo, K. Kramer, D. Goldgof, L. Hall, S. Samson, A. Remsen, T. Hopkins, “Learning to recognize plankton”, IEEE International Conference on Systems, Man and Cybernetics, 2003. [to be presented, refereed].

T. Luo, K. Kramer, D. Goldgof, L. Hall, S. Samson, A. Remsen, T. Hopkins, “Recognizing Plankton from the Shadow Image Particle Profiling Evaluation Recorder”, submitted to IEEE Transactions on Systems, Man and Cybernetics, Part A [in review]

X. Tang, F. Lin, S. Samson, and A. Remsen, “Binary plankton image classification,” submitted to IEEE Journal of Oceanic Engineering. [in review]