

# 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) 553-3915 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)

Grant Number: N000140210266  
Project Completion: 31 July, 2003

## LONG-TERM GOAL

The goal of this project is to develop a broadly-capable software package, which can automatically classify digital images of zooplankton. The software is being developed initially for classification of images from the SIPPER linescan imaging instrument, but has a wider application to alternate imaging systems where high quality digital images are available. Toward this end, the software is being developed to include both training and multi-stage classification portions. The input is digital images in standard computer format. A database-addressable classified particle list, including pertinent particle information and sorted images is the output. The project has application in rapid classification of microscopic marine particles, which have an effect on the instant 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 by manual identification of plankton images. The software will include a training phase, identification phase, and a database retrieval phase. The training phase will allow a trained user to enter expert-classified particle images into the system. The output will be a set of distinctive features, and a classification scheme (neural net, support vector machine, etc.), which will be used as control parameters during the identification phase. The identification phase will include pre-processing to eliminate noise, feature computation, and image classification. The database retrieval phase will include a method to track and efficiently retrieve information (particle identity, size, features, location) on the anticipated billions of particles. The software is being developed for use on a commercially available personal computer, to allow possible widespread use.

## **APPROACH**

The testing and development of automated plankton image recognition software will rely on high-resolution binary SIPPER images [1] obtained in the field in the eastern Gulf of Mexico, from another ONR project (N00014-96-1-5020). The development will initially use a broad range of manually identified plankton images for the development, characterization, and comparison of various algorithms; the images encompass a challenging, yet accurate representation of the diverse subtropical coastal ecosystem sampled. The software will build on previously reported work that sorted images from a more restrictive sample-set [2]. The approach can be broadly broken into several tasks. The first is the extraction of image particle features. The features from the manually classified training data will act as inputs to a training set development. Several techniques (discussed below) will be attempted and their performance compared. Using accepted methods, the feature set and classification scheme will be optimized. Performance with the inclusion of the expected noise class (non-identifiable particles) will be studied. Using the aforementioned results, software will be developed for classification of the unidentified images. The developed software will be field-tunable for application in differing ecosystems. Finally, a database for efficient retrieval of the classified images and their pertinent information (geographic location, size, identification) will be developed.

Graduate students Tong Luo and Kurt Kramer are primarily involved in implementation of the software, with guidance provided by Drs. 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 graduate student Andrew Remsen, and Dr. Tom Hopkins are providing biological expertise and manually classified images, while Scott Samson is acting as project manager.

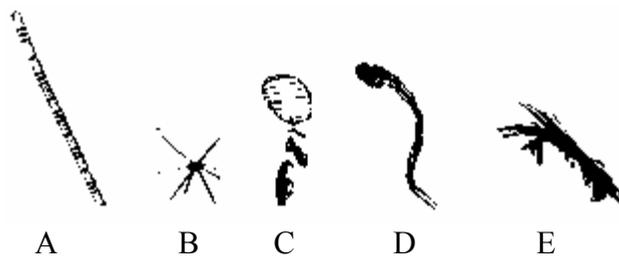
## **WORK COMPLETED**

Since the project inception, we have manually classified several thousand digital SIPPER images from Gulf-of-Mexico transects. The classes include a number of opaque and semi-transparent plankton: trichodesmium, siphonophores, shrimps, salps, protests, polychaetes, ostracods, mollusks, larvaceans, gelatinous plankton, doliolids, acantharians, diatoms, chateognaths, and several types of copepods. Additionally, images of unidentifiable particles have been found.

Software has been developed to exclude small noise pixels in the vicinity of the larger plankton images. Working with the biologists, the computer processing group has developed and computed a rich feature set. These features (numeric values derived from the input image) include: size, convex ratio (roughly-a ratio of particle bounding area to its footprint), object transparency, fourteen moments invariant to rotation, translation, and scaling, seven granulometric features, and a custom “head” feature. A total of 29 features extraction algorithms have been selected and implemented. These features, or a subset of them, act as input vectors to the classification stage.

Four algorithms have been developed or tested to classify the plankton images, using the extracted features. A K-Nearest Neighborhood (KNN) method creates a matrix of features from each image in the training data; a classification decision is made on which training images the current image most closely matches. The decision tree (DT) method uses a divide-and-conquer strategy to split the features and build a decision tree. We have used software developed by Quinlan [3] for this work. A Cascade-Correlation Neural Network (CCNN) from CMU [4] has been used to test a neural network architecture. The CCNN differs from a back-propagation learning method in that it dynamically adds additional neurons to the classifier as needed. The final method, Support Vector Machine (SVM) [5], is used to map the input feature set into a multi-dimensional feature space. A set of hyper-planes is determined to minimize classification error for the training set images. These then become the decision boundaries for the unclassified images. We have used both Svmfu2.004 [6] and Libsvm [7] software.

Initial experiments have been performed using 15 of the 29 extracted features as inputs to the four different classification algorithms, using a randomly sampled subset of 1,285 images (64 diatoms, 100 acantharians, 321 doliolids, 366 larvaceans, 434 trichodesmium), which represented the real-world distribution of these particles. Example images are shown in Fig. 1. A 10-fold cross validation (using 90% of images as training set, and 10% as unknowns, running ten times, then averaging the classification accuracy) was run on each of the four classification schemes. The accuracies over 10-fold cross validation are: SVM = 88.25%, CCNN = 87.25%, KNN (k = 3) = 85.21%, and DT = 83.27%.

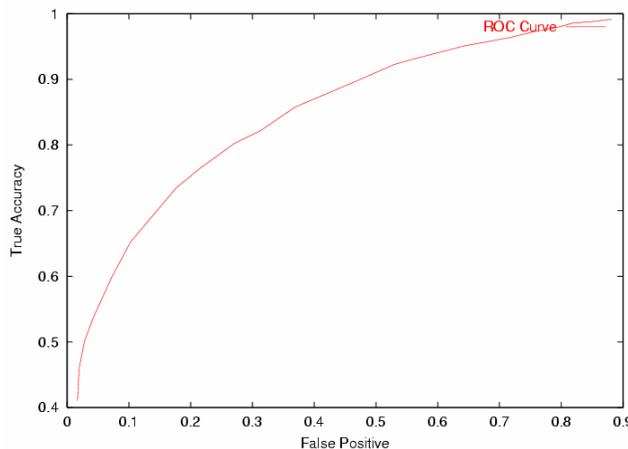


***Figure 1- Example binary images of the five classes used in the initial feature selection and classification performance experiments. A-diatom, B-acantharian, C-doliolid, D-larvacean, E-trichodesmium.***

Several features may be redundant or actually detrimental to classification accuracy; however, we don't know a priori which ones are helpful, so we performed a brute-force minimized feature set selection. This involved tests using feature sets of all possibilities in which one and/or two features were removed from an early 15-features set. In addition, tests utilizing all 3003 possible combinations

of just five out of the 15 features were used. The experiment showed that we could increase the accuracy by reducing the number of features. Some classifiers still perform well on a 5-feature set: CCNN=83.27% and KNN=78.52%. Therefore, feature selection will ultimately help in reducing computation time and also improve accuracy. Since enumerating all the combination of features is computational intractable, we used a greedy beam search in the later experiments on the 29 feature-set and another test image gallery. The search method gives us better accuracy. However, the greedy beam search is still computationally expensive when the feature dimension is very high. Other selection methods will be explored if many new features are added.

We have initiated work on determining the effect of and solutions to training sets which contain images belonging to a “noise” or untrained class of particles. The inclusion of untrained-for images into the data set will undoubtedly have an affect on any classifier’s performance. Images from a trained class may be incorrectly classified as noise or conversely, noise images may be misclassified as plankton. Different architectures to handle non-planktons have been explored. Initially, hierarchical architecture classification was used. We did non-plankton classification (plankton vs. non-plankton) in a first stage and then plankton classification (with 5 type of test plankton) in a second stage. In recent experiments we used the plankton images, which included noise images, from one single sea deployment as a training set. We chose copepods, protoctista, doliolids, larvaceans, trichodesmium and non-planktons to classify because most of the identifiable images in one run appeared to be of these types. In our training set, we used 1000 images from each type of planktons and 5000 images from non-plankton. We used the complete 29 feature set in the experiments. Figure 2 shows the receiver operating curve (ROC) for non-plankton classification. The second stage 5-plankton classifier had 82% accuracy. However, if we consider the loss of plankton in the first stage of non-plankton classification, the results were weaker. This is because the plankton images are of five types, and vary considerably in these real-world images, because of the three-dimensional nature of the particles; non-plankton have been treated as a single class, which makes the problem difficult. Therefore, we switched to one stage initial classification, where we initially included the non-plankton with their most similar plankton class. This gave us from 81.6% to 88.4% overall accuracy (depending on the training set mixture) and from 80.9% to 78.9% corresponding accuracy for the 5-type of planktons by varying the number of the non-plankton in our training set (i.e., altering the bias for the non-plankton class).



**Figure 2- Receiver Operating Curve for non-plankton classification showing tradeoff between accuracy and false positive identification.**

The SVM classification scheme performed very well, in terms of computation time for training and classification, and produced high accuracy (over 80% in our real-world test images). Using this algorithm, we have developed a pair of software applications that will enable shipboard training and classification. The input to the training application is a text-based file that defines the training images (their names and locations on a disk drive), and a variety of parameters (which features to use, number of images, kernel parameters). This allows maximum flexibility, which will be especially useful when new ecosystems are sampled. The application performs noise reduction, computes the selected features, and creates training models, which are saved to files for use by the classification application. The classification application extracts the features for each unclassified image in the source directory, and uses the training models to classify the image. The image is then moved to the appropriate class subdirectory within the destination directory.

## **RESULTS**

Significant progress has been made toward producing a shipboard application for use in classification of marine plankton images. Twenty-nine features have been created and implemented, though for fastest computation time and highest accuracy, a reduced subset is preferred. Removing features has been shown to improve accuracy by several percentage points. For high speed, and only slightly reduced accuracy, the five best features can be chosen. Size, moment1, convex ratio, and transparency features are some of the most important features. The Support Vector Machine algorithm has had the best combination of computation speed and accuracy. The initial software for shipboard classification, using the SVM algorithm and full feature set, operating on a 2.0 GHz Pentium 4 desktop PC can extract features for and classify approximately 730 images per minute. This compares favorably to a human expert, who can sort approximately 1,000 images per *hour*.

The introduction of noise images is a challenging aspect in the development of the software, and will need additional development, especially when operating nearshore, where significant amounts of detritus and sediment may be present in the water. The software should be flexible enough to allow tuning of the bias for including plankton and excluding noise.

## **IMPACT/APPLICATIONS**

This project has the potential to reduce the turnaround time for evaluating plankton images from months to days. 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. In-situ imaging techniques have been shown to have a scientific benefit [8], especially when compared to net collection of fragile plankton. 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.

## **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>.

We have met with Dr. Horst Bunke from the Institute of Computer Science and Applied Mathematics, University Bern, who has worked with consortium of universities on an EU-funded project (ADIAC) for the specific classification of diatom images.

## REFERENCES

1. Scott Samson, Thomas Hopkins, Andrew Remsen, Lawrence Langebrake, Tracey Sutton, and Jim Patten, “A System for High-Resolution Zooplankton Imaging,” *IEEE J. Oceanic Engineering*, v. 26, p. 671-676 (2001).
2. X. Tang, W. K. Stewart, L. Vincent, H. Huang, M. Marra, S. Gallager and C. Davis, “Automatic Plankton Image Recognition,” *Artificial Intelligence Review*, 12, p. 177-199 (1998).
3. J.R. Quinlan, *C4.5: Programs From Empirical Learning*, Morgan Kaufmann, San Francisco, CA.
4. S. E. Fahlman and C. Lebiere, “The Cascade-Correlation Learning Architecture,” *Advances in Neural Information Processing Systems*, 1991.
5. V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 2000.
6. Ryan Rifkin, <http://fpn.mit.edu/SvmFu/index.html>, 2002.
7. C. C. Chang and C.J. Lin (2002), “LIBSVM: A Library for Support Vector Machines,” <http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf>, 2002.
8. Andrew Remsen, Scott Samson, Thomas Hopkins, “What you see is not always what you get: Comparison of a Zooplankton-Imaging Sensor (SIPPER) With Concurrent Optical Plankton Counter and net Data From the Gulf of Mexico”, AGU-ASLO Ocean Sciences Meeting, Talk OS-41M-09, Honolulu, HI, Feb. 11-17, 2002.

## PUBLICATIONS AND PRESENTATIONS

<currently none related to this work>