LIMNOLOGY
and
OCEANOGRAPHY: METHODS

Limnol. Oceanogr.: Methods 5, 2007, 204–216

Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry

Heidi M. Sosik and Robert J. Olson
Biology Department, MS 32, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts, USA

Abstract

High-resolution photomicrographs of phytoplankton cells and chains can now be acquired with imaging-in-flow systems at rates that make manual identification impractical for many applications. To address the challenge for automated taxonomic identification of images generated by our custom-built submersible Imaging FlowCytobot, we developed an approach that relies on extraction of image features, which are then presented to a machine learning algorithm for classification. Our approach uses a combination of image feature types including size, shape, symmetry, and texture characteristics, plus orientation invariant moments, diffraction pattern sampling, and co-occurrence matrix statistics. Some of these features required preprocessing with image analysis techniques including edge detection after phase congruency calculations, morphological operations, boundary representation and simplification, and rotation. For the machine learning strategy, we developed an approach that combines a feature selection algorithm and use of a support vector machine specified with a rigorous parameter selection and training approach. After training, a 22-category classifier provides 88% overall accuracy for an independent test set, with individual category accuracies ranging from 68% to 99%. We demonstrate application of this classifier to a nearly uninterrupted 2-month time series of images acquired in Woods Hole Harbor, including use of statistical error correction to derive quantitative concentration estimates, which are shown to be unbiased with respect to manual estimates for random subsamples. Our approach, which provides taxonomically resolved estimates of phytoplankton abundance with fine temporal resolution (hours for many species), permits access to scales of variability from tidal to seasonal and longer.

Introduction

Many aspects of how phytoplankton communities are regulated remain poorly understood, in large part because we lack critical observational tools. Traditional organism-level sampling strategies are not amenable to high-frequency, long-duration implementations. Methods such as conventional microscopic analysis, for instance, are prohibitively labor intensive and time consuming, whereas newer and more rapid approaches, such as bulk water optical measurements (e.g., chlorophyll fluorescence or light absorption) provide little or no information about taxonomic composition and other details that are critical for ecological studies.

Working to overcome aspects of this limitation, we have developed a series of automated submersible flow cytometers capable of rapid, unattended analysis of individual plankton cells (and other particles) for long periods of time. The first such instrument, FlowCytobot, has proven capable of multimonth deployments (Olson et al. 2003) that provide new insights (e.g., Sosik et al. 2003). FlowCytobot, now in its fourth year of long-term deployment at the Martha's Vineyard Coastal Observatory (http://www.whoi.edu/mvco), is optimized for analysis of pico- and small nanoplankton (~1 to 10 μm). To complement FlowCytobot, we have now developed Imaging FlowCytobot (Olson and Sosik 2007), designed to sample natural assemblages of phytoplankton (and microzooplankton) in the size range ~10 to 100 μm. This is a critical development because phytoplankton in this size range, which include many diatoms and dinoflagellates, can be especially important in a variety of bloom conditions and as sources of new and export production.

The advent of instruments that permit rapid and automated microscopic analysis of natural waters, such as the laboratory-based FlowCam (Sieracki et al. 1998) and our

Acknowledgments

This research was supported by grants from NSF (Biocomplexity IDEA program and Ocean Technology and Interdisciplinary Coordination program; OCE-0119915 and OCE-0525700) and by funds from the Woods Hole Oceanographic Institution (Ocean Life Institute, Coastal Ocean Institute, Access to the Sea Fund, and the Bigelow Chair). We are indebted to Alexi Shalapyonok for expert assistance in the lab and field; to Melissa Patrician for hours of manual image classification; to Cabell Davis, Qiao Hu, Kacey Li, Mike Neubert, and Andy Solow for insights into image processing, machine learning, and statistical problems; and to the Martha's Vineyard Coastal Observatory operations team, especially Janet Fredericks, for logistical support.

submersible Imaging FlowCytobot (Olson and Sosik 2007), promise to revolutionize the ability to sample phytoplankton communities at ecologically relevant scales. They also, however, present new challenges for data analysis and interpretation. For instance, Imaging FlowCytobot can generate more than 10 000 high-quality plankton (and/or detritus) images every hour, and it can do so every day for months. This volume of data precludes manual inspection for cell identification as a feasible tool for many applications.

If adequate analysis techniques can be developed for large data sets of plankton images, the results will bring new insight into a range of ecological phenomena including bloom dynamics, species succession, and spatial and temporal patchiness. With these applications in mind, an initial goal for analysis of image datasets is to quantify abundance accurately for a wide range of taxa present in mixed assemblages. To do this requires efficient and accurate identification of individual plankton images.

This kind of classification problem has been addressed previously in particular applications involving plankton images. An important area of focus has arisen in response to availability of imaging systems optimized for observations of metazooplankton (> ~0.1 mm); these include systems designed for underwater measurements of live organisms, such as the Video Plankton Recorder (VPR) (Davis et al. 1992) and the Shadow Image Particle Profiling Evaluation Recorder (SIPPER) (Samson et al. 2001), as well the ZOOSCAN system for automated measurement of preserved samples (Grosjean et al. 2004). Davis and co-workers (Tang et al. 1998; Davis et al. 2004; Hu and Davis 2005) have made important contributions in developing several approaches for rapid analysis of plankton images generated by the VPR. This group has explored use of image characteristics (or features) such as invariant moments, granulometry, and co-occurrence matrices and use of machine-learning methods including learning vector quantization neural networks and support vector machines. In another approach, Luo et al. (2004), working with SIPPER-generated images, also addressed some of the challenges in including image features (such as area and transparency) that require accurate detection of the organism boundary within an image.

Compared to the case for zooplankton, efforts to automatically analyze and identify phytoplankton images have been more limited, although some recent progress suggests that new developments are likely to be productive. In an early demonstration example, Gorsky et al. (1989) showed that simple geometric properties were sufficient to reliably distinguish 3 species with distinct size and shape. In a similar study, Embleton et al. (2003) were able to define a neural network to identify 4 very distinct species from microscopic images of lake water samples, with accuracy sufficient to resolve seasonal patterns in total cell volume. In another example involving several dinoflagellate species from the same genus, Culverhouse et al. (2003) argued that a neural network approach can achieve accuracy similar to manual identification by trained personnel. Culverhouse et al. (2006) have proposed that this be implemented for detection of harmful algal species, although the ability of their HAB Buoy system to acquire cell images of sufficient quality remains to be demonstrated. There has also been considerable effort to develop species-level automated classification techniques for diatoms from ornamentation and shape details of cleaned frustules (du Buf and Bayer 2002 and chapters therein, e.g., Fischer and Bunke 2002). Most recently, for the special case of *Trichodesmium* spp. present in colonies large enough for detection with the VPR, automated analysis has provided striking ecological and biogeochemical insights (Davis and McGillicuddy 2006). These examples from previous work point to the utility of automated image processing and classification techniques for problems in phytoplankton identification, but they all address a relatively narrow scope in terms of taxonomic range or image type (e.g., cleaned frustules). Blaschko et al. (2005) highlighted the challenges of moving beyond this level by presenting results with ~50% to 70% accuracy for a 12-category (plus "unknown") image classification problem involving a variety of phytoplankton groups.

For adequate ecological characterization of many natural marine phytoplankton assemblages, the relevant image analysis and classification problem is broad (taxonomically diverse) and must accommodate many categories (10-20, or more). Taxonomic breadth necessarily means a wide range of cell sizes and relevant identifying characters. Moreover, for images collected automatically over long periods of time, such as from Imaging FlowCytobot, it is critical that techniques are robust to a range of sampling conditions (e.g., changes in co-occurring taxa and variations in image quality related to lighting and focus).

Here we describe a technique to address these challenges by combining selected image processing methods, machine-learning based classification, and statistical error correction to estimate taxonomically resolved phytoplankton abundance from high-resolution (~1 μm) images. Whereas the general approach is independent of the particular image acquisition system, we focus on data collected with Imaging FlowCytobot. Our approach builds on previous efforts in image classification for plankton, as well as some other image processing and classification applications such as face recognition and fingerprint recognition, while addressing the particular combination of image characteristics and identification markers relevant for Imaging FlowCytobot measurements of nano- and microphytoplankton in assemblages of highly mixed taxonomy. By characterizing temporal variability in a natural plankton community, we demonstrate that our approach achieves the overall goal of automatic classification of a wide variety of image types, with emphasis on morphologically distinct taxonomic groupings and accurate estimation of group abundance.

Materials and procedures

Our approach involves 5 main steps: 1) image processing and extraction of features (characteristics or properties), 2) feature selection to identify an optimal subset of characteristics for multi-

category discrimination, 3) design, training, and testing of a machine learning algorithm for classification (on the basis of selected features as input), 4) statistical analyses to estimate category-specific misclassification probabilities for accurate abundance estimates and for quantification of uncertainties in abundance estimates following the approach of Solow et al. (2001), and 5) application of the resulting feature extraction, classifier algorithm, and statistical correction sequence to sets of unknown images.

Image data sets—The images used to develop, assess, and demonstrate our methods were collected with a custom-built imaging-in-flow cytometer (Imaging FlowCytobot) analyzing water from Woods Hole Harbor. All sampling was done between late fall and early spring in 2004 and 2005. Here we provide a brief summary of Imaging FlowCytobot design and image characteristics; details are available elsewhere (Olson and Sosik 2007).

Imaging FlowCytobot uses a combination of flow cytometric and video technology to both capture images of organisms for identification and measure chlorophyll fluorescence and scattered light associated with each imaged particle. Its submersible and autonomous aspects were patterned after successes with the original FlowCytobot (Olson et al. 2003), while the addition of cell imaging capability and a design with higher sample volumes are critical for the application to microplankton. Imaging FlowCytobot uses a customized quartz flow cell (800 by 180 μm channel), with hydrodynamic focusing of a seawater sample stream in a sheath flow of filtered seawater to carry cells in single file through a red (635 nm) diode laser beam. Each cell passing through the laser beam scatters laser light, and chlorophyll-containing cells emit red (680 nm) fluorescence. Fluorescence signals are then used to trigger a xenon flashlamp strobe to emit a 1-μs flash of light, which illuminates the flow cell after passing through a green bandpass filter (514 nm). A monochrome CCD camera (1380 by 1034 pixels) and a frame grabber board are used to capture an 8-bit grayscale image of the corresponding cell. A 10× microscope objective focused on the flow cell is used to collect the images, as well as the scattered light and fluorescence from cells as they traverse the laser beam. This combination of fluidics and optical configuration provides images with target objects in consistent focus and typically having their major axis oriented with the longer axis of the camera field (i.e., along laminar flow lines). As described in Olson and Sosik (2007), the resulting images (considering the effects of magnification, camera resolution, and cell motion during flash exposure) can be resolved to approximately 1 μm, with the full camera field spanning ~300 by 400 μm. In real time, binary thresholding and a "blob" analysis algorithm (ActiveMIL 7.5, Matrox Electronic Systems Ltd.) are used to record only rectangular subregions of the camera field that contain cells or other objects (along with some adjacent background).

Manual inspection of many images from our Woods Hole Harbor data set led us to define 22 explicit categories that represent subjective consideration of taxonomic knowledge, ecological perspective, and practical issues regarding groupings that can be feasibly distinguished from morphology visible in the images (Fig. 1; see also Appendix A). Many of the categories correspond to phytoplankton taxa at the genus level or groups of a few morphologically similar genera. Diatoms account for most of these categories: 1) *Asterionellopsis* spp.; 2) *Chaetoceros* spp.; 3) *Cylindrotheca* spp.; 4) *Cerataulina* spp. plus the morphologically similar species of *Dactyliosolen* such as *D. fragilissimus* (all having many small distributed chloroplasts; category labeled *DactFragCeratul* in figures and tables); 5) other species of *Dactyliosolen* morphologically similar to *D. blavyanus* (with chloroplasts typically concentrated in a small area within the frustule); 6) *Ditylum* spp.; 7) *Guinardia* spp. plus occasional representatives of *Hemialus* spp.; 8) *Licmophora* spp.; 9) *Pleurosigma* spp.; 10) *Pseudonitzschia* spp.; 11) *Rhizosolenia* spp. plus rare occurrences of *Proboscia* spp.; 12) *Skeletonema* spp.; and 13) *Thalassiosira* spp. plus similar centric diatoms. Nondiatom genera are 14) *Dinobryon* spp.; 15) *Euglena* spp., plus other euglenoid genera; and 16) *Phaeocystis* spp. In addition to the genus-level categories, we defined several mixtures of morphologically similar particles and cell types: 17) various forms of ciliates; 18) various genera of dinoflagellates > ~10 μm in width; 19) a mixed group of nanoflagellates; 20) single-celled pennate diatoms (not belonging to any of the other diatom groups); 21) other cells < ~20 μm that cannot be taxonomically identified from the images; plus 22) a category for "detritus," noncellular material of various shapes and sizes.

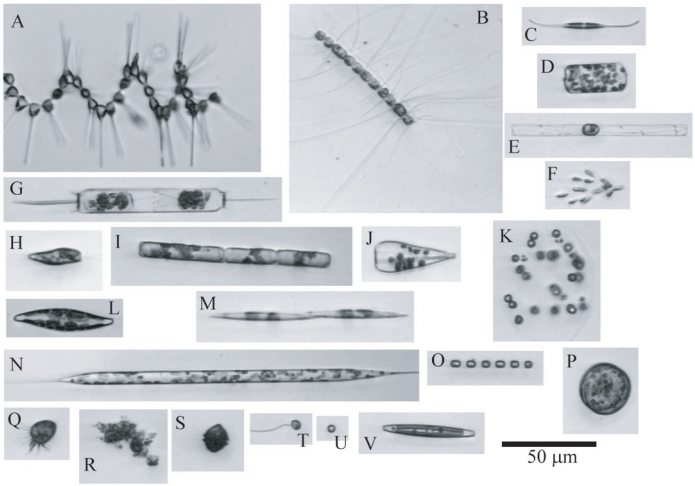


Fig. 1. Example images from 22 categories identified from Woods Hole Harbor water. Most categories are phytoplankton taxa at the genus level: *Asterionellopsis* spp. (A); *Chaetoceros* spp. (B); *Cylindrotheca* spp. (C); *Cerataulina* spp. plus the morphologically similar species of *Dactyliosolen* such as *D. fragilissimus* (D); other species of *Dactyliosolen* morphologically similar to *D. blavyanus* (E); *Dinobryon* spp. (F); *Ditylum* spp. (G); *Euglena* spp. plus other euglenoids (H); *Guinardia* spp. (I); *Licmophora* spp. (J); *Phaeocystis* spp. (K); *Pleurosigma* spp. (L); *Pseudonitzschia* spp. (M); *Rhizosolenia* spp. and rare cases of *Proboscia* spp. (N); *Skeletonema* spp. (O); *Thalassiosira* spp. and similar centric diatoms (P). The remaining categories are mixtures of morphologically similar particles and cell types: ciliates (Q); detritus (R); dinoflagellates > ~20 μm (S); nanoflagellates (T); other cells <20 μm (U); and other single-celled pennate diatoms (V).

Table 1. Summary of different features types determined for each image, specifying algorithm source and the stage of image processing at which the features are calculated.

Feature type	Algorithm or code source	Image processing stage	No. features	No. selected
Simple geometry	MATLAB Image Processing Toolbox	Blob image	18	17
Shape & symmetry	DIPUM and custom	Simplified boundary	16	16
Texture	DIPUM Toolbox	Original image (blob pixels only)	6	6
Invariant moments (standard and affine)	DIPUM & custom	Original image, blob image, filled simplified boundary	22	12
Diffraction pattern (ring/wedge)	Custom	Simplified boundary	100	41
Co-occurrence matrix statistics	MATLAB Image Processing Toolbox	Original image	48	39
		Total:	210	131

For each type, the total number of features originally calculated is indicated, along with the final number selected for use with the classifier (see text for details)

For development and testing of the analysis and classification approach, we compiled a set of 6600 images that were visually inspected and manually identified, with even distribution across the 22 categories described above (i.e., 300 images per category). These identified images were randomly split into "training" and "test" sets, each containing 150 images from each category (see Appendix A for full image sets, provided here to facilitate future comparison with other methods applicable to this problem). Independent of the training and test sets, we also inspected *every* image acquired during randomly selected periods of natural sample analysis (~27 000 images in sample volumes ranging from 5 to 50 mL and measured spanning the period February to April 2005) for manual identification; this allowed specification of misclassification probabilities under real sampling conditions and evaluation of error correction procedures (described below) for accurate abundance estimates.

Image processing and feature extraction—Our first objective was to produce, for each image, a standard set of feature values (characteristics or properties) which might be useful for discriminating among the 22 categories. We specified the standard feature set by considering characteristics that seem important for identification of images by human observers and on the basis of previous successes in related image classification problems. All imaging processing and feature extraction was done with the MATLAB software package (version 7.2; Mathworks, Inc.), including the associated Image Processing Toolbox (version 5.2; Mathworks, Inc.). We also incorporated algorithms described in Gonzalez et al. (2004) and implemented in the accompanying toolbox Digital Image Processing for MATLAB (DIPUM) (version 1.1.3; imageprocessingplace.com). For each image, the result of all feature extraction is a 210-element vector containing values that reflect various aspects of object size, shape, and texture, as described in more detail below (see Table 1).

The original grayscale image is used to derive some features, but various stages of image processing are required for others (Table 1). As a first step, many of the features we calculate require information about the boundary of the targets of interest (or "blobs") within an image, so preliminary image processing is critical for edge detection and boundary segmentation. We found that conventional edge detection algorithms were inadequate for reliable automated boundary determination over the range of image characteristics and plankton morphologies that we encounter with Imaging FlowCytobot. Approaches relying on intensity gradients, such as the commonly used Canny algorithm, could not be optimized to avoid artifacts from noise and illumination variations while reliably detecting challenging cell features such as spines, flagella, and localized areas that vary from brighter to darker than the background. For this reason, we turned to a computationally intensive but effective approach based on calculation of the noise-compensated phase congruency in an image (Kovesi 1999), as implemented in MATLAB by Kovesi (2005). Phase congruency is independent of contrast and illumination, and we found that simple threshold-based edge detection applied to phase congruency images provides excellent results for a wide range of phytoplankton cell characteristics (Fig. 2A–C).

After edge detection, we used standard MATLAB functions for morphological processing (closing, dilation, thinning) and for segmentation algorithms to define blobs or connected regions (Fig. 2D). For some feature calculations (e.g., symmetry measures), images were also rotated about the centroid of the largest blob to align the longest axis horizontally (compare Fig. 2C and D). Finally, we used DIPUM toolbox functions to reconstruct a simplified boundary of the largest blob in each image on the basis of the first 10% of the Fourier descriptors (Fig. 2E) (Gonzalez et al. 2004).

For the largest blob in each field (Fig. 2D), we calculate a set of relatively common geometric features such as major and minor axis length, area and filled area, perimeter, equivalent spherical diameter, eccentricity, and solidity (MATLAB Image Processing Toolbox, regionprops function), as well as several simple shape indicators (e.g., ratio of major to minor axis lengths, ratio of area to squared perimeter). For more detailed shape and symmetry measures, calculations were done on the

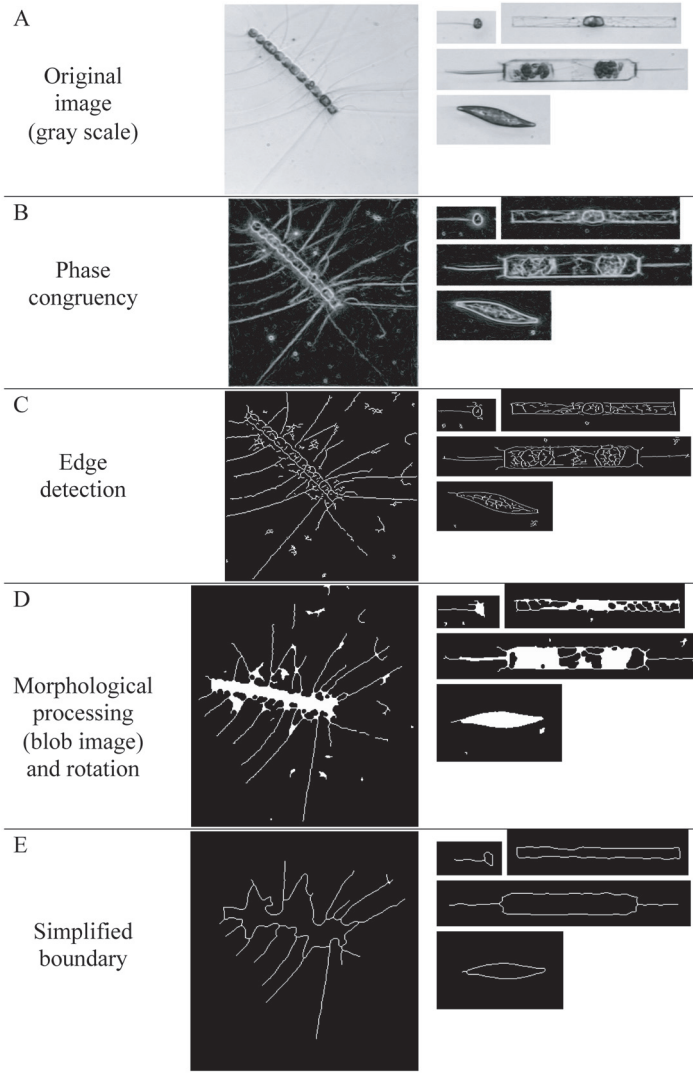


Fig. 2. Image processing stages for example images from several categories. The original grayscale images (A) are used for determining some classification features, but different preprocessing is required for others (see Table 1). Calculation of phase congruency (B) in the original images is a critical step to produce robust edge detection (C). Through morphological processing, edge images are converted into blob images (black and white) and then rotated (D). Finally, the first 10% of the Fourier descriptors are used to reconstruct a simplified blob boundary (E). Both the blobs (D) and the simplified boundaries (E) are used directly for feature calculations. Each panel shows corresponding results for the same set of 5 images.

simplified boundary with a combination of DIPUM functions and custom algorithms. These detailed features include 1) the number of line segment ends on the blob perimeter (potentially indicative of spines, for instance); 2) relative cell width near the left and right edges compared to the midpoint (a possible indicator for cells with distinctive ends such *Ditylum* spp. and *Rhizosolenia* spp.); 3) mean (absolute and relative to equivalent spherical diameter) and standard deviation of the distances between the blob centroid and points along its perimeter; 4) the number of concave and convex segments along the perimeter [see Loke and du Buf (2002) for development of this idea applied to diatom frustule characterization]; 5) symmetry metrics based on the Hausdorff distance, a measure of how much 2 shapes overlap, as applied to blobs compared with themselves rotated 90 and 180 degrees and reflected along the longitudinal centerline [e.g., see Fischer and Bunke (2002) for application to diatom frustules]; and 6) triangularity and ellipticity metrics specified by Rosin (2003) on the basis of the first affine moment invariant of Flusser and Suk (1993).

Various texture properties [e.g., contrast, smoothness, uniformity, and entropy as specified by Gonzalez et al. (2004) and implemented in the DIPUM toolbox] were determined on original grayscale images, but only considering the pixels within the largest blob determined as described above. In addition, following the success of Hu and Davis (2005) with this technique for zooplankton images, more detailed characterization of texture was included through calculation of gray-level co-occurrence matrices (MATLAB Image Processing Toolbox functions) for the original images. As individual features, we used statistics (mean and range of 4 properties: contrast, correlation, energy, and homogeneity) of 6 different gray-level co-occurrence matrixes (pixel offsets of 1, 2, 4, 16, 32, and 64, each averaged for 4 angles, 0, 45, 90, and 135 degrees).

As indicators of geometric pattern, we also used the 7 invariant moments described by Hu (1962). These are independent of object position, size, and orientation and were determined with DIPUM algorithms. In the absence of evidence suggesting the most appropriate image processing stage for these features, we chose to calculate them for the original image, the blob image, and the simplified boundary (filled solid) image.

For the final features, we used digital diffraction pattern sampling (custom MATLAB code), previously shown to be effective for fingerprint and other pattern recognition problems (Berfanger and George 1999). We implemented a modified version of the method developed by George and Wang (1994), applied to the simplified boundary images. The approach involves calculation of the 2-dimensional power spectrum for an image, and then sampling it to determine the energy distribution across a pattern of wedges and rings radiating from the origin. We used 48 rings and 50 wedges, each evenly distributed around one-half of the power spectrum. To prevent low frequencies from dominating wedge signals, the portion of each wedge near the origin (within an equivalent 15-pixel radius) was removed. Energy in each ring or wedge was normalized by the total energy in the image to specify 98 features; 2 additional values were included as features: the total energy in the power spectrum and the ratio of energy near the center (within the low frequency band eliminated from wedges) to the total.

As mentioned earlier, the final standard feature set for each image corresponds to a 210-element vector. The features for the 3300-image training set then comprise a 210-by-3300 element matrix. Before proceeding with further steps involved with classifier development or application, all features were

transformed to have mean = 0 and standard deviation = 1 in the training set (i.e., each of the 210 rows have mean = 0 and std = 1). The untransformed mean and standard deviation values for the training set features are later used for all other feature transformations (i.e., for test set and unknown image features before they are presented to the classifier).

Feature selection—Because inclusion of redundant or uninformative features can compromise overall classifier performance, feature selection algorithms can be useful to choose the best features for presentation to a machine learning algorithm. We have used the Greedy Feature Flip Algorithm (G-flip) as described by Gilad-Bachrach et al. (2004b) and available in a MATLAB implementation (Gilad-Bachrach et al. 2004a). G-flip, developed specifically for multicategory classification problems, is an iterative search approach for maximizing a margin-based evaluation function, where margin refers to a distance metric between training set instances and decision boundaries between categories. By selecting a small set of features with large margins, the G-flip algorithm helps to increase classification generality (i.e., avoids a classifier overly fitted to training data). For our 22-category training set with 210 input features, G-flip typically converges in less than 10 iterations (passes over the training data), although it does converge at local maxima, so we used 10 random initial points and picked the solution with the overall maximum of the evaluation function. With our current 22-category problem applied to the manually identified training set (150 images from each category), G-flip selection reduces our feature set from the original 210 elements down to 131 (Table 1). Only these 131 features are then presented to the classifier for training, testing, and classification of unknowns.

Classifier design and training—For our multicategory classification problem, we use a support vector machine (SVM), a supervised learning method that is typically easier to use than neural networks and is proving popular for a variety of classification problems, including those involving plankton (Luo et al. 2004; Blaschko et al. 2005; Hu and Davis 2005). SVM algorithms are based on maximizing margins separating categories in multidimensional feature space. The algorithms we use have been implemented with a MATLAB interface as the LIBSVM package (Chang and Lin 2001). LIBSVM uses a one-against-one approach to the multi-category problem, as justified by Hsu and Lin (2002). We selected this implementation over others because of its ease of use in the MATLAB environment and its full development for multiclass applications. An additional consideration is that the LIBSVM package includes an extension of the SVM framework to provide probability estimates for each classification (p_c), according to Wu et al. (2004). We use these probabilities for accurate abundance estimates (see details below).

We used a radial basis function kernel, which means the overall SVM requires specification of 2 parameters, 1 kernel parameter and 1 for the cost function (penalty parameter for errors). The optimal values of these parameters cannot be specified a priori, so we used 10-fold cross-validation on the training set (G-flip selected features only) for parameter selection, maximizing overall classification accuracy of the SVM. The cross-validation approach involves random splits of the training data into 10 subsets, one of which is used to test accuracy after training with the other 9; this is implemented as a standard option in LIBSVM and minimizes effects of overfitting to the training data during parameter selection. We used a simple brute-force nested search approach over a wide range of parameter combinations to find the global maximum cross-validation accuracy.

After parameter selection, we trained the SVM (fixed with the best model parameters) with the entire training set (all 3300 entries without cross-validation, G-flip selected features only). The results of this training step determine the final SVM classifier, which is specific to the selected feature set and the feature transformation statistics described above.

Statistical error correction for abundance estimates—To extend our automated classifier to ecological applications that require quantitative determination of group-specific abundances, we followed the approach of Solow et al. (2001). This involves statistical correction on the basis of empirically determined misclassification probabilities and permits not only improved abundance estimates (especially for rare groups that may be subject to large errors from false-positive identifications), but also estimation of uncertainties (standard errors) for abundance.

We used manual analysis of all images in randomly selected field samples (not used as part of the training and test sets), combined with the automated SVM classifier, to produce a matrix of classification probabilities, where the diagonal elements represent the probability of detection for each category and the off-diagonal elements are misclassification probabilities for each possible combination of categories. This is a 23-by-23 element matrix: 22 categories plus 1 for "other" images, i.e., unidentifiable images or species not represented in the 22 explicit categories. We then used this information to correct abundance estimates for expected misclassification errors, considering the complete mix of identifications in the sample, and to calculate approximate standard errors for the abundance estimates as described in Solow et al. (2001). In applying this approach, we include 1 modification to the example classification application described by Solow et al. We take advantage of the probability estimates available from the LIBSVM classification results and use only SVM classification results with relatively high certainty, $p_c > 0.65$; identifications with lower probabilities are initially placed in the "other" category. This leads to lower values of detection probability (diagonal element of the matrix) for some categories, but gives better overall performance (lower residuals relative to manual results) for corrected abundance estimates (see details below). We selected the threshold value of $p_c = 0.65$, by searching for the value that provided the lowest overall relative residuals between manual and classifier-based abundance estimates.

Assessment

Classifier performance—We evaluated overall performance of the final SVM classifier by applying it to the independent test

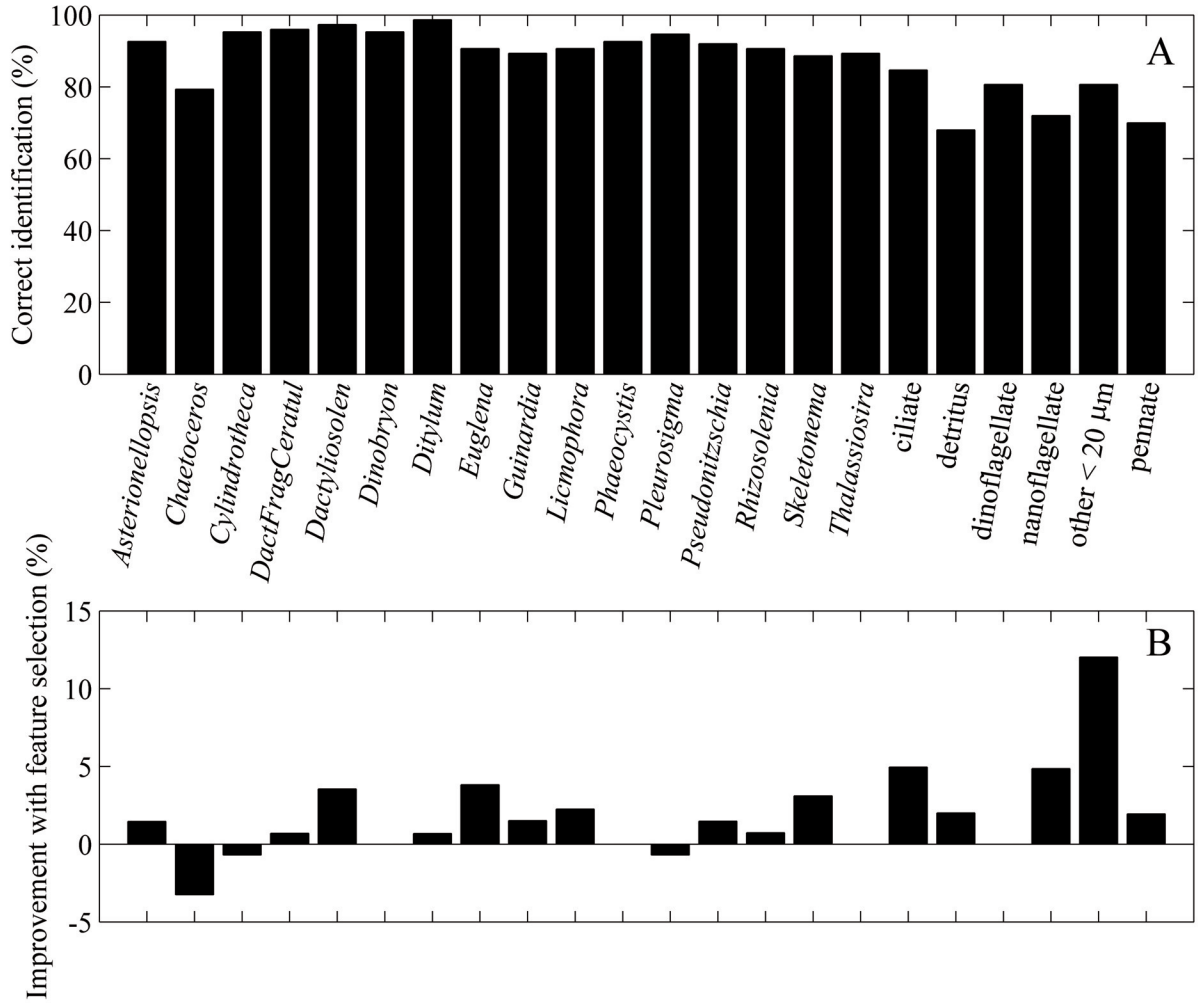


Fig. 3. Automated classification results for 22 categories in the independent test set of images (i.e., images not used for classifier development) (A). Values shown here represent the percentage of images manually placed in each category that were also placed there by the SVM classifier. Percent improvement in classification rate due to feature selection (B) was determined by comparing test results in (A) with those from a separate classifier trained with the complete 210 feature set. Categories appear in the same order as images labeled A–V in Fig. 1; see text for detailed explanation of category labels.

set of 3330 manually identified images (i.e., those not used in features selection, parameter selection, or training). This evaluates the full analysis and classification scheme encompassing image processing, feature extraction, feature selection, parameter selection, and SVM training (including effectiveness of training set). The classifier provides excellent results for many genera of phytoplankton (>90% correct identifications for 12 categories), and overall classification accuracy across all 22 categories in the test set is 88% (Fig. 3A). Only 4 categories have accuracies <80%: 1 phytoplankton genus, *Chaetoceros* (79%), which is challenging because of its morphological diversity; and 3 relatively nonspecific categories: detritus (68%), nanoflagellates (72%), and pennate diatoms (70%). Both specificity (true positives/classifier total) and probability of detection (true positives/manual total) for each class follow the same pattern: 80% to 100% for the phytoplankton genera and somewhat lower for the less precise categories (Table 2, Test set columns).

Image processing—We did not undertake any quantitative assessment of the effectiveness of the image processing methods we used, except as evident indirectly through performance of the final classifier scheme. An important aspect of our method development, however, involved visual examination of the results of image processing stages as applied to thousands of example images drawn from the range of categories in our training and test sets. These inspections were used subjectively to optimize details of the processing scheme, for example, the choice to use phase congruency calculations for acceptable edge detection results, use of the first 10% of the Fourier descriptors for boundary reconstruction, and selecting the size of structuring elements (2 to 5 pixels) used for morphological processing.

Feature and parameter selection—We assessed the importance of our feature selection step by comparing classification test results (Fig. 3A) to those achieved with a scheme that omits

Table 2. Specificity (Sp) and probability of detection (Pd) for each of the 22 categories during application of the SVM classification scheme to the image test set, and also to a series of field samples for which every acquired image was included in the analysis.

	Test set		Complete field samples			
	all *P*		all *P*		*P* > 0.65 only	
	Sp	Pd	Sp	Pd	Sp	Pd
Asterionellopsis	0.93	0.91	0.28	0.75	0.54	0.59
Chaetoceros	0.85	0.82	0.93	0.70	0.98	0.56
Cylindrotheca	0.91	0.96	0.77	0.81	0.94	0.70
DactFragCeratul	0.97	0.95	0.69	0.99	0.91	0.96
Dactyliosolen	0.97	0.94	0.97	0.94	1.00	0.90
Dinobryon	0.97	0.95	0.74	0.97	0.95	0.96
Ditylum	1.00	0.98	0.71	0.83	1.00	0.67
Euglena	0.79	0.87	0.21	0.88	0.44	0.75
Guinardia	0.84	0.88	0.97	0.75	1.00	0.59
Licmophora	0.94	0.89	0.34	0.75	0.95	0.66
Phaeocystis	0.95	0.93	0.35	0.79	0.75	0.69
Pleurosigma	0.90	0.95	0.61	1.00	0.89	0.94
Pseudonitzschia	0.91	0.91	0.13	0.67	0.27	0.50
Rhizosolenia	0.81	0.90	0.64	0.80	0.98	0.60
Skeletonema	0.91	0.86	0.14	0.60	0.23	0.43
Thalassiosira	0.81	0.89	0.43	0.82	0.76	0.63
ciliate	0.81	0.81	0.42	0.79	0.63	0.62
detritus	0.75	0.67	0.59	0.42	0.78	0.18
dino	0.79	0.81	0.65	0.79	0.79	0.50
flagellate	0.76	0.69	0.19	0.66	0.34	0.45
other < 20 μm	0.66	0.72	0.92	0.73	0.92	0.73
pennate	0.77	0.69	0.24	0.70	0.41	0.43

Results for the field samples are shown both for the case where all identifications are included, regardless of the maximum category probability, and for the case where only instances with $p_c > 0.65$ are considered.

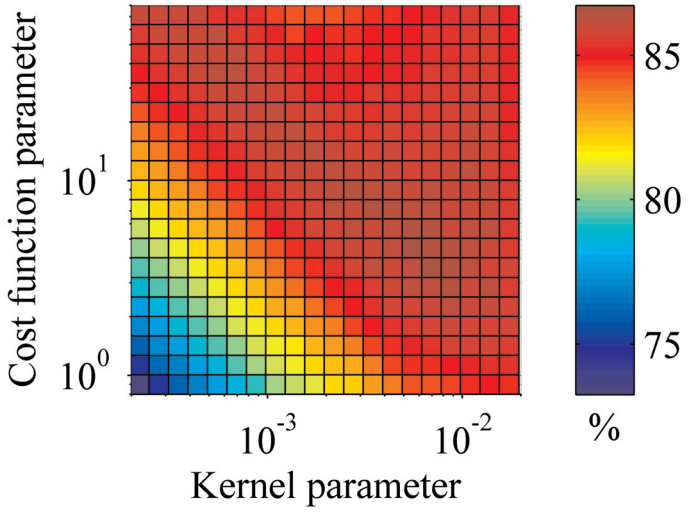


Fig. 4. Grid search results showing 10-fold cross-validation accuracy (%) for various combinations of SVM parameters, emphasizing the importance of optimal parameter selection for classifier performance.

feature selection (Fig. 3B). In other words, the same SVM details and the same training images were used, but the SVM training was conducted with all 210 features instead of the reduced set of 131. The overall correct classification rate on the test set was only 2% better with feature selection (88% vs. 86% accurate identification); however, some category-level accuracies were substantially better with feature selection (Fig. 3B), most notably "other cells <20 μm," for which the rate increased from 72% to 80%. Although the overall advantage of feature selection on correct identification rate is relatively modest, it does provide improvement in performance for almost all categories and adds only modest computational cost.

Another potential advantage of feature selection is reduction in the number of features that must be calculated for each unknown image. For large datasets, this can affect overall computation time and provide some insights into which feature types may be worth further investigation or refinement. Feature selection for our 22-category training set showed that all the types of features in our full 210-element set were useful for some aspect of classification, but that within certain feature types not all the elements were needed (Table 1). For instance, all but the fifth invariant moment was retained for the original image, but only the first was selected for the case of the simplified boundary image, and moments 4, 5, and 7 were eliminated for the blob image. For co-occurrence matrix statistics, contrast and correlation values were consistently chosen for all pixel offsets, but only about half of the energy and homogeneity values were needed; and for the ring-wedge diffraction pattern sampling, just over half the wedges were chosen, with these spread over the full pattern, but only 13 (of 48) rings were retained, with these concentrated near the center. Exact details of which features are chosen change slightly with different realizations of the G-flip algorithm, but these general trends are persistent, suggesting that in our initial feature set we have oversampled the diffraction patterns and co-occurrence matrix properties.

Compared with feature selection, SVM parameter selection had a larger impact on accuracy of the final trained SVM classifier. Results of grid search near the accuracy maximum show that >10% changes in cross-validation accuracy occur with 100-fold changes in kernel parameter and cost function parameter (Fig. 4). Because there is no a priori way to choose these parameters, the parameter selection step is critical for optimal results. There may be more elegant and faster search approaches for parameter selection, but we opted for a nested brute force search because of its simplicity, near guarantee of locating the global maximum, and because the added computational time is relatively minor (since this search need only be done once, after feature selection and before training).

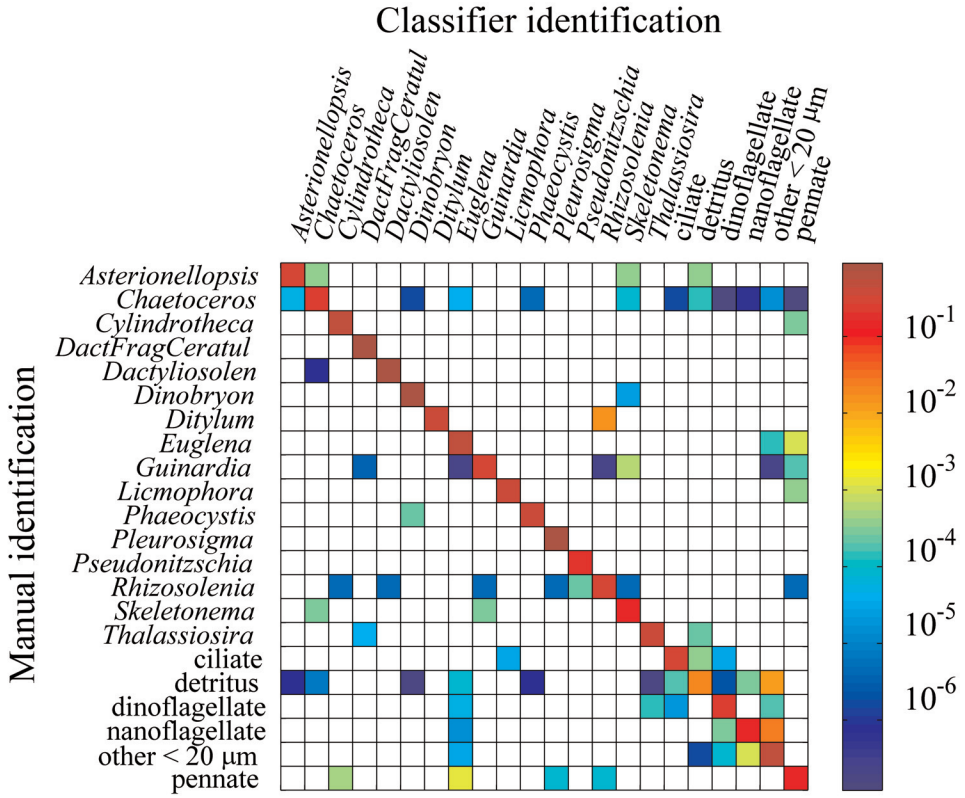


Fig. 5. Matrix of classification probabilities for the 22 image categories, derived from analysis of all images in a series of natural samples (~8600 images). Probability values range from 0 to nearly 1 and are colored with a logarithmic mapping to emphasize both the high values along the diagonal (probability of detection for each category) and the low values off diagonal (category-specific misclassification probabilities). White elements correspond to $P = 0$.

Abundance estimation—Although the classifier performance on the test set was excellent, some errors or misclassifications are unavoidable. These errors can be significant for abundance estimates in natural samples, both because error rates tend to be higher when considering all images (not just those that manual inspectors find readily identifiable) and because abundance estimates for rare groups are very sensitive to effects of even low rates of false-positive classification associated with images from more abundant categories. Our evaluation of the classifier considering manual identification of all images (~8600) in a randomly selected set of natural samples showed that specificity and probability of detection decreased for almost all categories, in some cases dramatically, compared with the test set results (Table 2). As expected, consideration of only classifications with category probabilities above the threshold $p_c = 0.65$ (i.e., ignoring relatively uncertain identifications) resulted in decreased probability of detection for all categories but increased specificity, in many cases to levels near those achieved with the test set (Table 2).

We examined misclassification errors in more detail through the classification probability matrix calculated from the 8600-image field data set (Fig. 5). As expected given the high overall performance of the classifier, misclassification rates (off-diagonal elements) were always much lower than correct classification rates (diagonal in Fig. 5). This analysis emphasizes that certain types of misclassification are more common than others, such as between detritus and ciliates or nanoflagellates and other cells <20 μm or *Asterionellopsis* and *Chaetoceros* (both chain-forming diatoms with spines). Many elements of the matrix are zero, indicating that those types of misclassification did not occur in the analysis with this data set.

As described by Solow et al. (2001), the classification probability matrix is a property of the classifier and not dependent on properties of unknown samples (such as relative abundance in different categories), so once the matrix is determined with sufficient accuracy it can be used in a straightforward manner to correct initial classifier-predicted abundances. We evaluated this approach at 2 levels. First, we used the 8600-image field set to compare category-specific abundance estimates determined manually and with the error-corrected classifier. Then, we applied the same approach to a separate field data set of randomly selected samples containing nearly 19 000 images. The initial 8600-image set was used to calculate the probability matrix, but this latter set was not used in any aspect of the classifier development, training, or correction scheme, thus ensuring a completely independent test. Both field data sets span a range of sampling conditions over a 2-month period in February to April 2005.

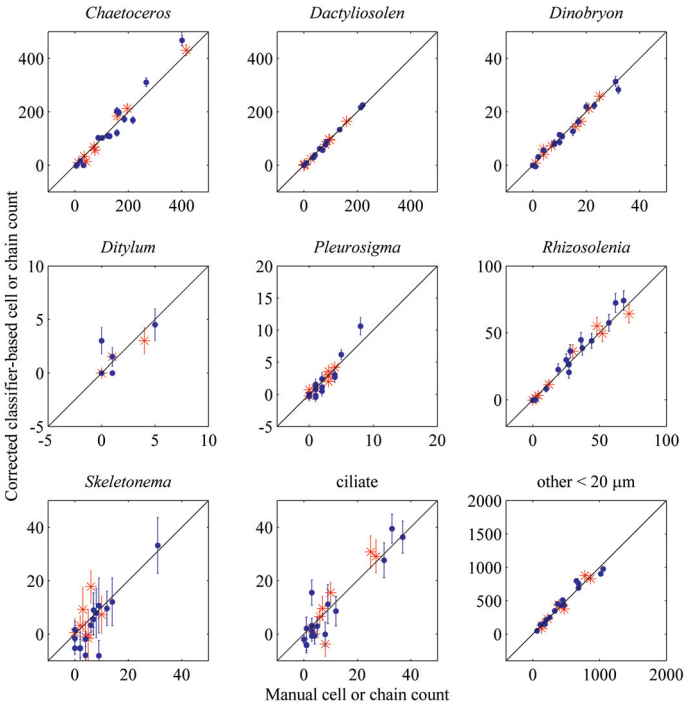


Fig. 6. Comparison between classifier-based counts (after correction for classification probabilities) and manual counts (from visual inspection of images) for selected categories chosen to span the range of abundances and classification uncertainties evident across all 22 categories and across the range of natural sample conditions encountered during the time series shown in Fig. 7. All points are shown with error bars indicating ± 1 standard error for the classifier estimates (in some cases these are as small as the plot symbols). Red points (asterisks) indicate samples used in the generation of the classification probability matrix (see text for details), and blue points (solid circles) are completely independent samples, each selected randomly from within week-long intervals of the full time series. The sample volumes examined for these comparisons ranged from 5 to 50 mL, providing a range of abundances and associated relative errors.

Use of the probability matrix shown in Fig. 5 results in 80% of the corrected classifier-based abundance estimates falling within 2 standard errors of the manual results and no evident biases between the manual and classifier-based results for any categories, in either field data set (Fig. 6). The few cases with differences outside 2 standard errors still show no bias and are concentrated in 1 phytoplankton genus (*Chaetoceros*, which is challenging because its morphological diversity) and in several relatively nonspecific categories: "detritus," "nanoflagellates," and "other <20 µm." As expected, categories with very low abundance in our samples (e.g., *Ditylum*) have higher relative errors than some of the more abundant categories; relative standard errors are also higher for categories that tend to get confused with more abundant ones, such as *Skeletonema*, which has modest misclassification rates with the more abundant (in these samples) *Guinardia* and *Chaetoceros* (see Fig. 5). Even when standard errors are high, however, the estimates are all without bias compared to manual counts (Fig. 6). If we compare these overall results to the case with uncorrected classifier abundance, summed squared residuals between classifier and manual estimates (i.e., residuals about the 1:1 lines in comparisons similar to those shown in Fig. 6) increase by 3-fold or more for most categories (7-fold mean across all categories), emphasizing the importance of the error correction step for abundance.

Application to time series studies—The field data used in the assessments discussed above were randomly selected from a much larger data set from trial deployment of Imaging FlowCytobot at the Woods Hole Oceanographic Institution dock during February to April of 2005. Because there were more than 1.5 million images collected over the 8-week period, the complete data set provides an opportunity to assess the potential ecological applications of our automated classification method. Imaging FlowCytobot was connected to power and data communication systems analogous to those at the Martha's Vineyard Coastal Observatory, and all control and data acquisition were fully automated. Images were processed and classified as described above, and category-specific concentrations (and associated standard errors) were determined with 2-h resolution. Averaged over the full data set, computational time (on a single 3.2-GHz Pentium-based computer) was roughly equal to the duration of the time series, with image processing and feature extraction dominating.

Historical observations in waters near Woods Hole point to late winter/early spring as a period of transition in the phytoplankton community. Blooms of large-celled species and chain-forming diatoms are more commonly found in fall and winter than at other times of year (e.g., Lillick 1937; Riley 1947; Glibert et al. 1985). When analyzed with the approach described in this article, our Imaging FlowCytobot observations capture this seasonal transition in unprecedented detail. In late February, the most abundant nano- and microphytoplankton (besides the mixed class of ~10- to 20-µm rounded cells that cannot be taxonomically discriminated from our images) were chain-forming diatom species, especially *Chaetoceros* spp., *Dactyliosolen* spp., and *Guinardia* spp., which were present at approximately 20 chains mL^{-1}, 15 chains mL^{-1}, and 10 chains mL^{-1}, respectively (other taxa were at levels of 3 cells mL^{-1} or less). By mid-March, the previously abundant diatom genera had declined by ~1 to 3 orders of magnitude, to near undetectable levels. The full 2-h resolution time series emphasize the power of these observations for exploring detailed ecological phenomena, such as species succession (Fig. 7). The dominant diatoms all declined over the 2-month sampling period, but they responded with very different temporal patterns. For example, *Dactyliosolen* spp. and *Guinardia* spp. started at similar concentrations, but *Dactyliosolen* spp. declined roughly exponentially over the entire period, whereas *Guinardia* spp. persisted longest and then declined most abruptly in early April (Fig. 7). The full time series are also rich with even higher frequency detail, such as fluctuations associated with the complexity of water masses and tidal currents in Woods Hole Harbor (e.g., Fig. 8).

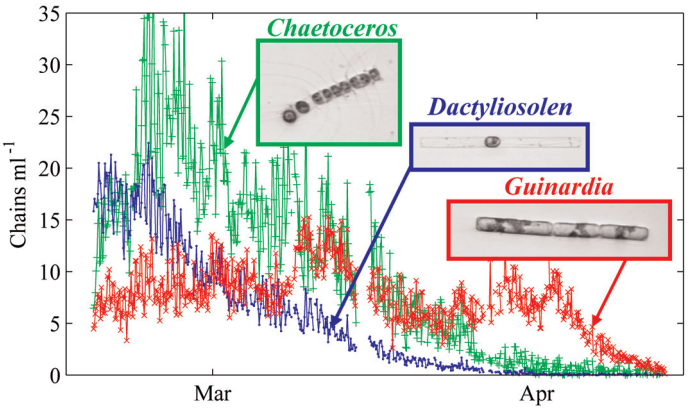


Fig. 7. Time series of diatom chain abundances observed during a 2-month deployment of Imaging FlowCytobot in Woods Hole Harbor during 2005. Automated image classification was used to separate contributions at the genus level; shown here are *Chaetoceros* spp. (green +), *Dactyliosolen* spp. (blue ●), and *Guinardia* spp. (red x). For each abundance estimate, we have an associated standard error, although these are not shown here for clarity; see Fig. 8 for representative values.

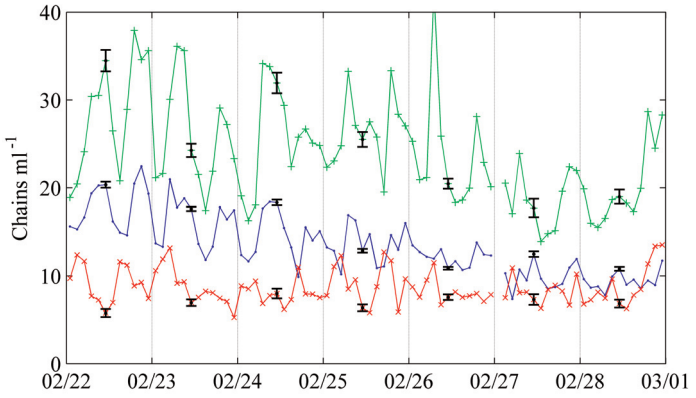


Fig. 8. Expanded view of a single week (end of February) of the time series in Fig. 7, emphasizing the ability of automatically classified Imaging FlowCytobot observations to capture variability related to water mass changes with semidiurnal tidal flows in Woods Hole Harbor. Standard errors for the abundance estimates (shown only once per day for clarity) are small compared to the variations evident at tidal frequencies. As in Fig. 7, green (+), blue (●), and red (x) correspond to abundances of *Chaetoceros* spp., *Dactyliosolen* spp., and *Guinardia* spp. chains, respectively.

Discussion

Our goal was to develop an analysis and classification approach to substantially increase ecological insight that can be achieved from rapid automated microplankton imaging systems. Our experiments and analyses are extensive enough to show that our approach meets the challenge of many-category (>20) classification, with unbiased abundance estimates for both abundant and rare categories. Furthermore, the extent of our training, testing, and field evaluation data ensures that the approach is robust and reliable across a range of conditions (i.e., changes in taxonomic composition and variations in image quality related to lighting and focus). The training and test sets were developed from many different sampling dates over more than a year, and the example field data span 8 weeks of continuous sampling when species composition was changing dramatically (Fig. 7).

The performance of our automated classifier exceeds that expected for consistency between manual microscopists (Culverhouse et al. 2003) or achieved with other automated applications to plankton images (e.g., Culverhouse et al. 2003; Grosjean et al. 2004; Blaschko et al. 2005; Hu and Davis 2005; Luo et al. 2005). The approach provides unbiased quantitative abundance estimates (and associated uncertainty estimates) with taxonomic resolution similar to many applications of manual microscopic analysis to plankton samples. When coupled with an automated image acquisition system, such as Imaging FlowCytobot, the advantages of our approach over manual identification are striking. Not only can we extend interpretation to include many more images, but the effort can be sustained indefinitely, providing access to information at scales of variability (e.g., the full spectrum from hours to years) that have been inaccessible with traditional methods.

Regarding taxonomic resolution, the set of 22 categories we identified in this work is only one way to parse our image data set. Although there are physical limits to the morphological detail that can be resolved in the images we used, other groupings or finer taxonomic detail could still be considered in future work depending on the ecological context and the expertise of personnel developing the necessary training sets. For instance, more detailed investigation of wintertime diatom bloom dynamics at our study site may require that the *Chaetoceros* genus be subdivided according to species with characteristic spine morphology evident in the images. We also anticipate that it will be necessary to add new genera to the list of categories as we build up longer time series of observations in waters of the New England continental shelf. For other study sites, new categories will certainly be necessary. Because of the diversity of categories we have already incorporated, the image processing and feature calculation methods we have used will likely be adequate for a range of changes in categories to classify. Similarly, although new classifier training must be carried out with any change in categories, the classifier development framework we have described can be readily applied to new category sets, as well as to new feature sets if necessary. Addition of new categories or any other change in the classifier will require routine updating of the classification probability matrix.

Our time series during the late winter diatom bloom near Woods Hole emphasize the range of temporal scales accessible in our observations (Figs. 7 and 8). This is most striking for high-abundance cell types. It is important to note that, as with any counting method, abundance estimates for rare cell types will not be statistically robust at small sample volumes. For our system, this statistical limit can be overcome at the

expense of temporal resolution. In other words, we can provide abundance estimates for rare species, but not reliably at 2-h resolution. For many ecological challenges, this tradeoff is acceptable. For low-level detection of harmful algal species, for instance, it may be both practical and more than adequate to provide abundance estimates with daily resolution.

We have recently begun deployments of the existing Imaging FlowCytobot at the Martha's Vineyard Coastal Observatory, located in 15 m of water on the New England continental shelf near Woods Hole. This study site is where we have operated the original FlowCytobot (for pico- and small nanoplankton observations) for several years. With these two instruments, FlowCytobot and Imaging FlowCytobot, now side-by-side, we can make high temporal resolution observations of the entire phytoplankton community, ranging from picoplankton to chain-forming diatoms, and do so for extended periods (months to years). As exemplified by the time series presented in this article (Figs. 7 and 8), the automated classification procedure we have developed is critical for exploiting the full potential of these observations. The coupled instrument and analysis system can provide new insights into ecological processes and responses to environmental perturbations in natural plankton communities.

Comments and recommendations

When combined with sampling strategies enabled by instruments such as Imaging FlowCytobot, our automated image processing and classification approach is reliable and effective for characterizing phytoplankton community structure with high taxonomic and temporal resolution. We expect that critical elements of the approach are general enough for application to plankton images from other sources, but future work is needed to evaluate this prospect quantitatively and identify any aspects that may require adaptation.

As with any supervised machine learning method, expert specification of training and test data are a critical aspect of our approach. The investment required for expert identifications is also by far the limiting factor in applying our approach to other data sets; steps dealing with aspects such as feature selection and optimization and training of the Support Vector Machine are straightforward and take negligible time (typically a few hours of automated computations) in comparison. In this article, we describe application to one set of specified categories, acknowledging that other groupings or more taxonomically detailed categories can be justified for specific ecological questions or by more experienced taxonomists. Undoubtedly, different categories will be required for other study sites. Our experience with development of the framework we describe here suggests that it will be readily adaptable to specification of different image categories. It was not necessary to modify the approach, for instance, as we added categories to the training set over the period of its development. New features can also be readily added if new knowledge or expert advice recommends them.

Use of phase congruency calculations for edge detection is an aspect of the image processing sequence that is particularly important for reliability and generality of our approach. This step is also one of the most computationally demanding, so future efforts to make our automated classification approach truly real time (e.g., on board a remotely deployed instrument with limited communication bandwidth to shore) should focus on hardware advances or acceptable algorithm alternatives for this calculation.

For ecological problems that require quantifying abundance accurately for a wide range of taxa present in mixed assemblages (e.g., including rare categories), the investment in developing the full category-specific classification probability matrix is critical for unbiased results. Although our experiments show that it is acceptable to apply a single realization of the probability matrix over a large data set spanning 2 months of data collection, future work is needed to determine how general the probabilities are across changes in sampling conditions. It is possible, for instance, that the likelihood of the classifier confusing certain categories changes with performance of the image acquisition system (e.g., more likely if focus is poor). Implementing a strategy to update the probability matrix is simple in concept, requiring only periodic manual inspection of a small subset of images.

For the case of Imaging FlowCytobot deployed at a cabled coastal ocean observatory, the approach described here is ready for application to targeted ecological studies, such as investigation of bloom dynamics and patterns and causes of species succession on the New England continental shelf. As discussed above, expansion into other environments is not limited, but will require new training and test sets to accommodate taxonomic groups not included here.

References

Berfanger, D. M., and N. George. 1999. All-digital ring-wedge detector applied to fingerprint recognition. Appl. Opt. 38:357-369.

Blaschko, M. B., and others. 2005. Automatic in situ identification of plankton. Seventh IEEE Workshops on Application of Computer Vision (WACV/MOTION'05) 1:79-86.

Chang, C.-C., and C.-J. Lin. 2001. LIBSVM: A library for Support Vector Machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Culverhouse, P. F., R. Williams, B. Reguera, V. Herry, and S. González-Gil. 2003. Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. Mar. Ecol. Prog. Ser. 247:17-25.

——— and others. 2006. HAB Buoy: a new instrument for in situ monitoring and early warning of harmful algal bloom events. African J. Marine Sci. 28:245-250.

Davis, C. S., S. M. Gallager, M. S. Berman, L. R. Haury, and J. R. Strickler. 1992. The video plankton recorder (VPR): design and initial results. Arch. Hydrobiol. Beih. Ergebn. Limnol. 36:67-81.

———, Q. Hu, S. M. Gallager, C. Tang, and C. J. Ashjian. 2004.

Real-time observation of taxa-specific plankton abundance: an optical sampling method. Mar. Ecol. Prog. Ser. 284:77-96.

——— and D. J. J. McGillicuddy. 2006. Transatlantic abundance of the N_2-fixing colonial cyanobacterium *Trichodesmium*. Science 312:1517-1520.

du Buf, H., and M. M. Bayer. 2002. Automatic Diatom Identification. World Scientific.

Embleton, K. V., C. E. Gibson, and S. I. Heaney. 2003. Automated counting of phytoplankton by pattern recognition: a comparison with a manual counting method. J. Plank. Res. 25:669-681.

Fischer, S., and H. Bunke. 2002. Identification using classical and new features in combination with decision tree ensembles, p. 109-140. *In* H. du Buf and M. M. Bayer [eds.], Automatic Diatom Identification. World Scientific.

Flusser, J., and T. Suk. 1993. Pattern recognition by affine moment invariants. Pattern Recognition 26:167-174.

George, N., and S. G. Wang. 1994. Neural networks applied to diffraction-pattern sampling. Appl. Opt. 33:3127-3134.

Gilad-Bachrach, R., A. Navot, and N. Tishby. 2004a. Large margin principals for feature selection, http://www.cs.huji.ac.il/labs/learning/code/feature_selection/.

———. 2004b. Margin based feature selection: theory and algorithms. ACM International Conference Proceeding Series, Proceedings of the 21st International Conference on Machine Learning 69:337-343.

Glibert, P. M., M. R. Dennett, and J. C. Goldman. 1985. Inorganic carbon uptake by phytoplankton in Vineyard Sound, Massachusetts. II. Comparative primary productivity and nutritional status of winter and summer assemblages. J. Exp. Mar. Biol. Ecol. 86:101-118.

Gonzalez, R. C., R. E. Woods, and S. L. Eddins. 2004. Digital Image Processing Using MATLAB. Prentice Hall.

Gorsky, G., P. Guilbert, and E. Valenta. 1989. The Autonomous Image Analyzer: enumeration, measurement and identification of marine phytoplankton. Mar. Ecol. Prog. Ser. 58:133-142.

Grosjean, P., M. Picheral, C. Warembourg, and G. Gorsky. 2004. Enumeration, measurement, and identification of net zooplankton samples using the ZOOSCAN digital imaging system. ICES J. Mar. Sci. 61:518-525.

Hsu, C.-W., and C.-J. Lin. 2002. A comparison of methods for multiclass support vector machines. IEEE Trans. Neural Networks 13:415-425.

Hu, M. K. 1962. Visual pattern recognition by moment invariants. IEEE Trans. Information Theory 8:179-187.

Hu, Q., and C. Davis. 2005. Automatic plankton image recognition with co-occurrence matrices and Support Vector Machine. Mar. Ecol. Prog. Ser. 295:21-31.

Kovesi, P. 1999. Image features from phase congruency. Videre: A Journal of Computer Vision Research. MIT Press 1:1-26.

Kovesi, P. D. 2005. MATLAB and Octave functions for computer vision and image processing. School of Computer Science & Software Engineering, The University of Western Australia., http://www.csse.uwa.edu.au/~pk/research/matlabfns/.

Lillick, L. C. 1937. Seasonal studies of the phytoplankton off Woods Hole, Massachusetts. Biol. Bull. Mar. Biol. Lab., Woods Hole 73:488-503.

Loke, R. E., and H. du Buf. 2002. Identification by curvature of convex and concave segments, p. 141-165. *In* H. du Buf and M. M. Bayer [eds.], Automatic Diatom Identification. World Scientific.

Luo, T., D. Kramer, D. B. Goldgof, L. O. Hall, S. Samson, A. Remsen, and T. Hopkins. 2005. Active learning to recognize multiple types of plankton. J. Mach. Learn. Res. 6:589-613.

Luo, T., K. Kramer, D. Goldgof, L. O. Hall, S. Samson, A. Remsen, and T. Hopkins. 2004. Recognizing plankton images from the shadow image particle profiling evaluation recorder. IEEE Trans. Syst. Man Cybern. B 34:1753-1762.

Olson, R. J., A. A. Shalapyonok, and H. M. Sosik. 2003. An automated submersible flow cytometer for pico- and nanophytoplankton: FlowCytobot. Deep-Sea Res. I 50:301-315.

——— and H. M. Sosik. 2007. A submersible imaging-in-flow instrument to analyze nano- and microplankton: Imaging FlowCytobot. Limnol. Oceanogr. Methods, in press.

Riley, G. A. 1947. Seasonal fluctuations of the phytoplankton population in New England Coastal Waters. J. Mar. Res. 6: 114-125.

Rosin, P. 2003. Measuring shape: ellipticity, rectangularity, and triangularity. Machine Vision Applications 14: 172-184.

Samson, S., T. Hopkins, A. Remsen, L. Langebrake, T. Sutton, and J. Patten. 2001. A system for high-resolution zooplankton imaging. IEEE J. Oceanic Eng. 26:671-676.

Sieracki, C. K., M. E. Sieracki, and C. S. Yentsch. 1998. An imaging-in-flow system for automated analysis of marine microplankton. Mar. Ecol. Prog. Ser. 168:285-296.

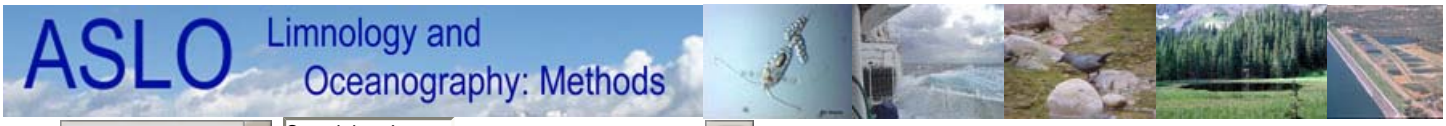
Solow, A., C. Davis, and Q. Hu. 2001. Estimating the taxonomic composition of a sample when individuals are classified with error. Mar. Ecol. Prog. Ser. 216:309-311.

Sosik, H. M., R. J. Olson, M. G. Neubert, A. A. Shalapyonok, and A. R. Solow. 2003. Growth rates of coastal phytoplankton from time-series measurements with a submersible flow cytometer. Limnol. Oceanogr. 48:1756-1765.

Tang, X., W. K. Stewart, H. Huang, S. M. Gallager, C. S. Davis, L. Vincent, and M. Marra. 1998. Automatic plankton image recognition. Artificial Intelligence Rev. 12:177-199.

Wu, T.-F., C.-J. Lin, and R. C. Weng. 2004. Probability estimates for multi-class classification by pairwise coupling. J. Mach. Learn. Res. 5:975-1005.

Submitted 5 October 2006
Revised 12 March 2007
Accepted 7 April 2007



Web Appendix

Heidi M. Sosik and Robert J. Olson
Automated taxonomic classification of phytoplankton sampled with ima cytometry
2007, 5:204-216

Electronic copy of image sets used for training and testing the classification s appendix contains 3300 manually categorized images of cells and other particles were collected with Imaging FlowCytobot from Woods Hole Harbor water. Th stored in tiff format and are organized at 2 levels. First, they are split between testing sets of equal size. Second, each set contains 22 categories with 150 indiv in each. The 22 categories are explained in more detail in the text. Category those described in the text except that the label "other_lt20" corresponds to "o in the text.

Images are available in compressed ZIP files for download. Because it can b download very large ZIP files, the images have been broken into two lar "training.zip" (128MB) and "testing.zip" (131 MB). If these are still too large for download, each of the 22 categories is available as a separate ZIP file from the li

	Testing images	Training images
All images	*download*	*download*
Asterionellopsis	*download*	*download*
Chaetoceros	*download*	*download*
ciliate	*download*	*download*
Cylindrotheca	*download*	*download*
DactFragCeratul	*download*	*download*
Dactyliosolen	*download*	*download*
detritus	*download*	*download*
Dinobryon	*download*	*download*
dinoflagellate	*download*	*download*
Ditylum	*download*	*download*
Euglena	*download*	*download*
Guinardia	*download*	*download*
Licmophora	*download*	*download*
nanoflagellate	*download*	*download*
other_lt20	*download*	*download*
pennate	*download*	*download*
Phaeocystis	*download*	*download*
Pleurosigma	*download*	*download*
Pseudonitzschia	*download*	*download*
Rhizosolenia	*download*	*download*
Skeletonema	*download*	*download*
Thalassiosira	*download*	*download*