LIMNOLOGY
and
OCEANOGRAPHY: METHODS

Limnol. Oceanogr.: Methods 10, 2012, 538–554

Semi-automated image analysis for the identification of bivalve larvae from a Cape Cod estuary

Christine M. Thompson[1], Matthew P. Hare[2], and Scott M. Gallager[1]*
[1]Biology Department, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA
[2]Department of Natural Resources, Cornell University, Ithaca, New York 14853, USA

Abstract

Machine-learning methods for identifying planktonic organisms are becoming well-established. Although similar morphologies among species make traditional image identification methods difficult for larval bivalves, species-specific shell birefringence patterns under polarized light permit identification by color and texture-based features. This approach uses cross-polarized images of bivalve larvae, extracts Gabor and color angle features from each image, and classifies images using a Support Vector Machine. We adapted this method, which was established on hatchery-reared larvae, to identify bivalve larvae from a series of field samples from a Cape Cod estuary in 2009. This method had 98% identification accuracy for four hatchery-reared species. We used a multiplex polymerase chain reaction (PCR) method to confirm field identifications and to compare accuracies to the software classifications. Image classification of larvae collected in the field had lower accuracies than both the classification of hatchery species and PCR-based identification due to error in visually classifying unknown larvae and variability in larval images from the field. A six-species field training set had the best correspondence to our visual classifications with 75% overall agreement and individual species agreements from 63% to 88%. Larval abundance estimates for a time-series of field samples showed good correspondence with visual methods after correction. Overall, this approach represents a cost- and time-saving alternative to molecular-based identifications and can produce sufficient results to address long-term abundance and transport-based questions on a species-specific level, a rarity in studies of bivalve larvae.

The larvae of many coastal benthic invertebrates have complex life cycles beginning with a pelagic larval stage lasting from a few days to weeks. During development, larvae are passively transported by ocean currents that determine their fates (Thorson 1950; Scheltema 1986). Studies of invertebrate larval dispersal have been met by challenges associated with small sizes of individuals, high mortality, and patchiness over large spatial scales (Boicourt 1988; Garland 2000; Pineda et al. 2007). Particularly for bivalve larvae, it is difficult to perform species-specific field studies because of an inability to accurately identify early stage larvae (Garland 2000; Garland and Zimmer 2002; Gregg 2002). Because bivalve larvae exhibit species-specific behaviors in the field (Shanks and Brink 2006), one cannot accurately assess transport without identifying species. This is especially important when considering populations of commercially important species, as an understanding of larval transport is necessary to address management questions concerning species productivity and decline, shellfish enhancement through seeding, and habitat restoration (Gregg 2002).

Once a bivalve larva begins shell mineralization (usually 20 h post-fertilization), most species proceed to a straight-hinge (veliger) stage followed by transformation to a more rounded, umbonate (pediveliger) stage after several days (Chanley and Andrews 1971). It is particularly difficult to distinguish species of straight-hinged larvae by morphological features alone, but as the larva develops, characteristic morphological changes can

*Corresponding author: E-mail: cmingione@whoi.edu;
present address: Horn Point Laboratory, University of Maryland Center for Environmental Science, Cambridge, MD.

Acknowledgments

Field assistance for sample collection was provided by R. York, E. Bonk, C. Weidman, and M. Mingione. R. Kraus from the Aquaculture Research Corporation in Dennis, MA and H. Lind of the Eastham Shellfish Hatchery provided larvae used in this study. Molecular work was performed with the support of H. Borchardt-Wier at Cornell University and T. Shank and W. Cho at WHOI. K. Bolles, E. North, J. Goodwin, and two anonymous reviewers provided helpful comments on this manuscript. This project was supported by an award to S. Gallager and C. Mingione Thompson from the Estuarine Reserves Division, Office of Ocean and Coastal Resource Management, National Ocean Service, National Oceanic and Atmospheric Administration and a grant from Woods Hole Oceanographic Institution's Coastal Ocean Institute.

DOI 10.4319/lom.2012.10.538

sometimes help distinguish species or genera. Photographs of cultured species are limited to those that can be reared in the laboratory and matching photographs to larvae from the field can result in misidentification (Loosanoff et al. 1951). Electronmicrographs of the larva's hinge structure have historically been the standard for larval species identification (Lutz et al. 1982), but the labor required to perform these identifications is unrealistic for field studies. The pros and cons of more recent species-specific identification methods have been reviewed by Garland and Zimmer (2002) and Hendriks et al. (2005). It has been a challenge to develop a reliable and cost-effective solution for larval identification to handle the large volume of samples required for many field studies. Current successful methods involve multiplex PCR (Hare et al. 2000; Larsen et al. 2005), quantitative PCR (Wight et al. 2009), and fluorescent in situ hybridization with DNA probes (Henzler et al. 2010), but each method has specific limitations on sample volume, specificity, and cost per sample.

Recent advances in imaging technology have allowed for greater spatial and temporal resolution of plankton studies through optical sampling methods (Benfield et al. 2007). Insitu optical sampling instruments such as the Video Plankton Recorder (Davis et al. 1992), benchtop equipment such as FLOW-CAM (Sieracki et al. 1998), and laboratory-based scanning methods such as ZOOSCAN (Grosjean et al. 2004) have created a need for image recognition software to identify plankton based on characteristic features that the computer reads from each image (Davis et al. 2004). Each class of organisms must have distinguishing characteristics (or features, i.e., shape, texture, color) for the computer to recognize and use for training. Not every statistical classifier is optimal for analysis of a given image set, so there can be a lengthy start-up time for optimizing image processing techniques (Grosjean et al. 2004; Lou et al. 2005; Gorsky et al. 2010). Furthermore, computer image analysis is not capable of discriminating images as exactly as humans and is generally assumed to be less accurate than having a human expert carefully analyze microscope samples (Culverhouse et al. 2003). Ultimately, image identification of plankton samples must balance accuracy, or how well the system compares with traditional methods, with efficiency and repeatability in order to handle large volumes of material.

Image-processing techniques can be used to address taxa-specific questions of abundance, spatial distribution, and biomass in zooplankton studies. Studies have demonstrated the use of these methods for observations of real-time zooplankton distribution through quantitative high-resolution maps (Gallager et al. 1996, Davis et al. 2004), seasonal zooplankton abundance and biomass estimates from preserved net samples (Bell and Hopcroft 2008; Gorsky et al. 2010), phytoplankton size and biomass (Sieracki et al. 1998), and taxa-specific phytoplankton distributions (Sosik and Olson 2007). More recent studies have employed these techniques to address large-scale biological questions, such as zooplankton biomass and spatial distribution in the Bay of Biscay over an eight-year period (Irigoien et al. 2009), spatial structure of zooplankton distribution in relation to oceanographic variables in an upwelling region in Chile (Manriquez et al. 2009; Manriquez et al. 2012), and association of zooplankton taxa with water mass types on the Western Antarctic Peninsula (Ashjian et al. 2008). As these techniques become more tested and available, the traditional taxonomic approach used for large-scale plankton studies should be adapted to include automated image processing (MacLeod et al. 2010).

The similar morphologies of veliger larvae make them less amenable to traditional identification methods using size features and black-and-white images (Hendriks et al. 2005), but color images of larvae under polarized light show distinct birefringence patterns (Tiwari and Gallager 2003a, 2003b). Once a larva begins shell formation, each species uses a specific protein matrix to control the orientation of the aragonite crystals forming the shell. Mineralization continues throughout the larval phase as the shell changes shape. Cross-polarized light and a full wavelength compensation plate create color patterns that reflect the crystal orientations. These color-patterns are species-specific and can be used in pattern-recognition software (Tiwari and Gallager 2003a, 2003b). Initial work using six species of preserved hatchery larvae showed accuracies between 80% to 90% (Tiwari and Gallager 2003b). As only color patterns are used as features, polarization techniques are insensitive to shell orientation, size, and morphology (Tiwari and Gallager 2003a, 2003b), eliminating many of the ambiguities involved in differentiating bivalve larvae (Perino et al. 2008).

Although hatchery-reared samples allow for definitive measures of identification accuracy, they are likely to represent a simplified sample set relative to field-caught larvae. Field larvae may appear different due to environmental heterogeneities and may contain more species than can be featured in a reference set of reared larvae. Using a reference set that doesn't accurately represent the field sample composition violates classification assumptions (Provost 2000) and could generate misclassifications, particularly false-positives (Gorsky et al. 2010).

The objective of this work was to develop a supervised image classification technique using shell birefringence patterns to distinguish species of bivalve larvae into a reproducible method that can be applied to field studies. Here we present the first application of this polarization technology to larval bivalves from field-collected samples. Our goal was to evaluate and optimize the identification accuracy of this technique and compare it to other available methods for bivalve larval identification. We employed visual identification as well as DNA identification methods using multiplex polymerase chain reaction (PCR) and genetic database searches (Hare et al. 2000). Finally, we present a species-specific assessment of larval bivalve abundance in weekly field samples taken from Waquoit Bay, MA, USA over a six-month period using computer classifications trained with field images and compare results to visual classifications. This research is the first systematic step needed to generate species-specific data to better address questions related to bivalve larval transport, dispersal, and survival, all of which are important for restoration and management efforts.

Materials and procedures

Our study employed four approaches for identifying larvae: (1) hatchery rearing, (2) genetic methods using multiplex PCR, (3) visual identification, and (4) supervised image classification. The first three approaches were necessary to set up the supervised image classification technique (Fig. 1A).

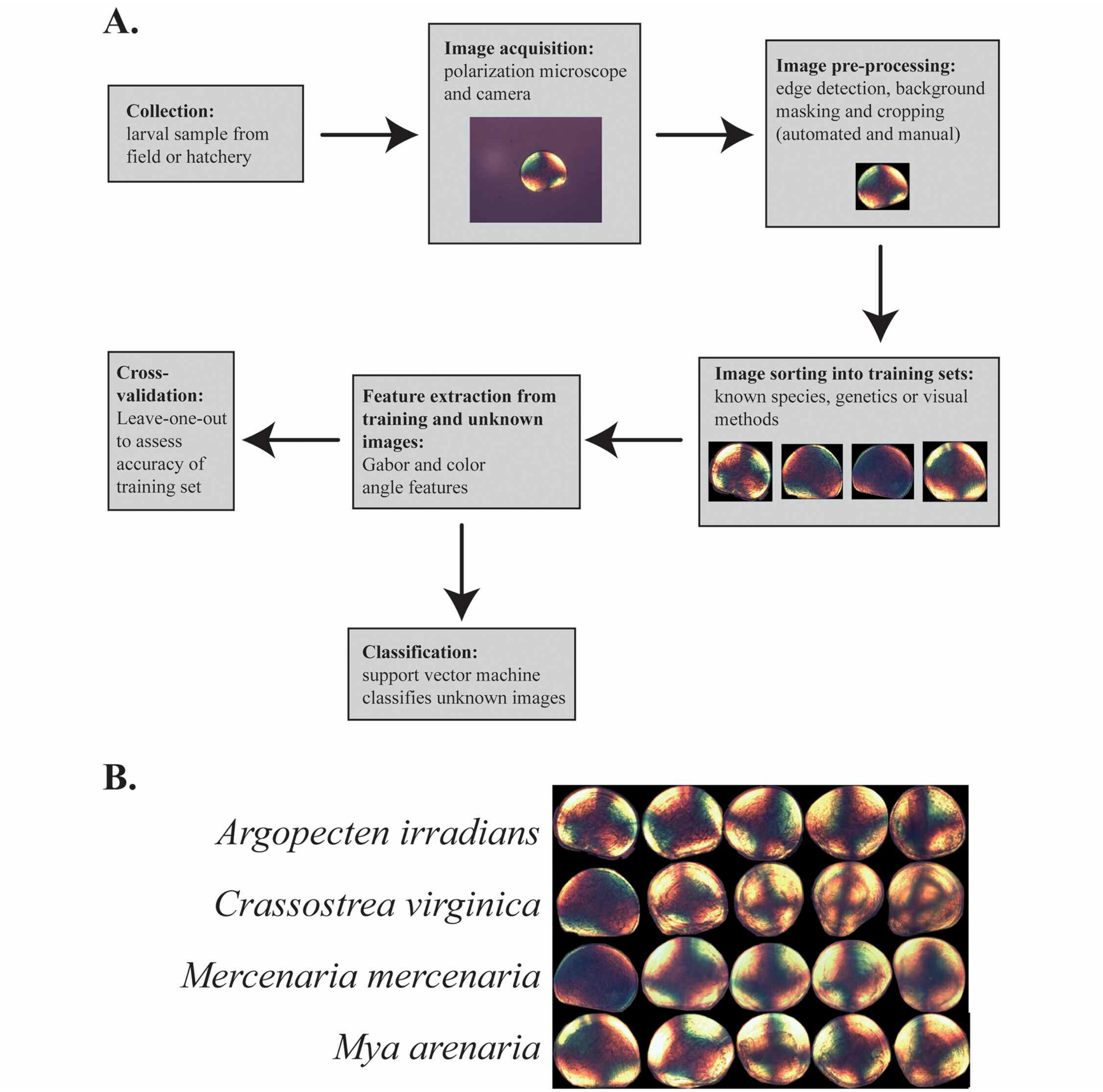
A.
Collection:
larval sample from field or hatchery
Image acquisition:
polarization microscope and camera
Image pre-processing:
edge detection, background masking and cropping (automated and manual)
Image sorting into training sets:
known species, genetics or visual methods
Feature extraction from training and unknown images:
Gabor and color angle features
Cross-validation:
Leave-one-out to assess accuracy of training set
Classification:
support vector machine classifies unknown images
B.
Argopecten irradians
Crassostrea virginica
Mercenaria mercenaria
Mya arenaria

Fig. 1. (A) Diagram of image processing technique from sample collection to classification of unknown images. (B) Sample images from the hatchery training set. Polarization images of larvae from four species throughout larval development with varying color patterns. Images are not to scale.

Hatchery rearing

Reference larvae were spawned from two Cape Cod aquaculture facilities between 2007 and 2010 and preserved in 80% ethanol. Larvae of four commercially important species, *Argopecten irradians* (bay scallop), *Crassostrea virginica* (eastern oyster), *Mercenaria mercenaria* (quahog), and *Mya arenaria* (soft-shell clam), were sampled from cultures every 1-2 d after spawning.

Images of hatchery larvae were taken using a Moticam 1000 4 megapixel camera mounted on a Zeiss IM 35 compound microscope fitted with a polarization filter and full wave compensation plate to achieve cross-polarization (Fig. 2). The optical path setup was similar to that used previously (Tiwari and Gallager 2003a, 2003b), but omitting bleaching and using a different wave compensation plate prohibited cross-comparisons with the 2003 images. Bleaching shells was not shown to affect classification accuracy (Thompson 2011). A 12V 100W halogen bulb was used as light source. Motic Images Plus (version 2.0; Motic China Group) captured JPEG images with color and exposure settings for the capture window matching the appearance of the larvae under the microscope. All larvae were imaged on a glass slide with coverslip in distilled water after rinsing off any fixative. For each species, 100 larvae were imaged from each sample to total between 500-3000 images representing different larval stages and orientations (Fig. 1B).

Genetic methods

A multiplex PCR method targeted to identify five species of bivalves from field samples was used for molecular identifications (Hare et al. 2000). Species targeted were *M. mercenaria*, *A. irradians*, *M. arenaria*, *Mulinia lateralis* (little surf clam), and *Spisula solidissima* (surf clam). DNA was extracted from ethanol preserved larvae after rinsing and used in multiplex PCR assays containing five species-specific primer pairs mapping to the cytochrome oxidase I (CO1) gene and a universal 18S-rRNA primer pair as a positive control. Each primer pair amplified a different length DNA fragment. Specific details of primer design, larval DNA extraction, and PCR assays can be found in Hare et al. (2000). Only reactions prepared from a master mix for which no amplification products appeared in the negative (no DNA) control reaction were used for comparison with images. Sequencing of the 18S region in reactions that did not produce a species-specific band provided further species identification through genetic database searches.

Visual identification

We set up reference image sets from both the hatchery (known species) and field sample (unknown) images for comparison. Subsamples of 100 larvae from each field sample were imaged using the setup described above resulting in a field set of over 7000 images. These field images were visually classified to species to form image groups from which we set up training sets for the supervised image classification. We used field identification guides of Chanley and Andrews (1971) and Loosanoff et al. (1966) for morphology and size criteria. Polarized images from hatchery and molecularly confirmed larvae were used to identify unknown larvae based on birefringence patterns.

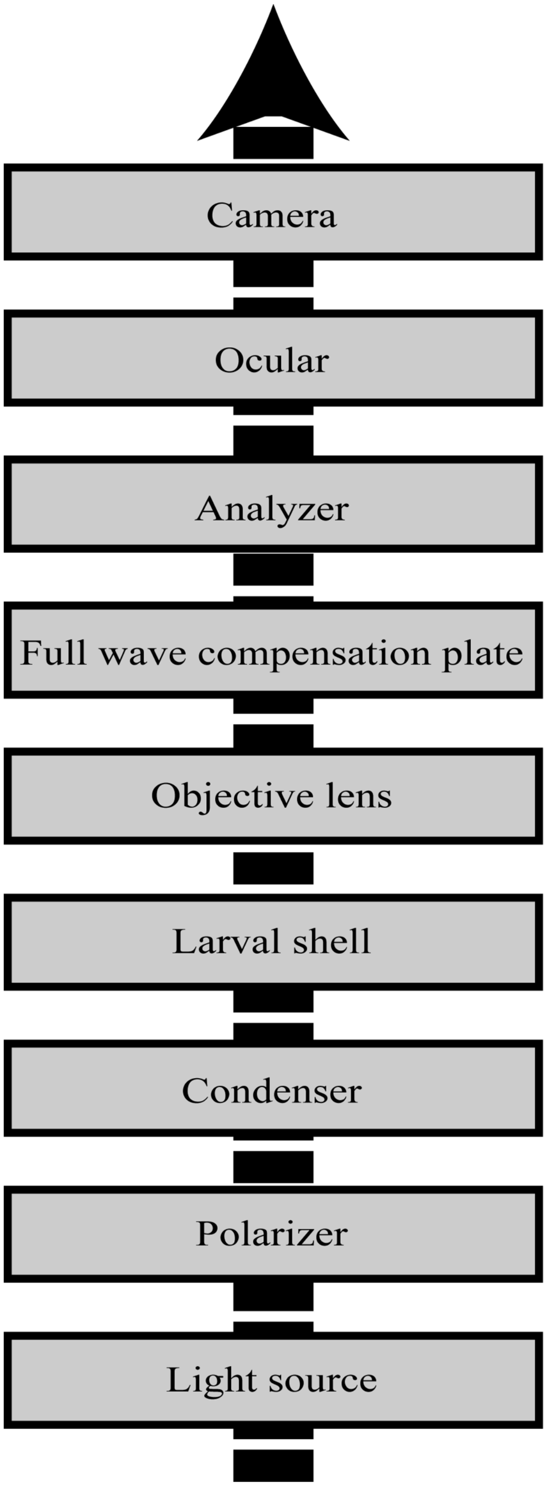

Fig. 2. Diagram of optical path for polarization setup of microscope for image acquisition. Black arrow represents path of light.

Supervised image classification

This supervised image classification technique requires three key steps after sample collection and imaging (Fig. 1A): (1) image preprocessing to remove background image "noise,"

(2) training set feature extraction and cross-validation, (3) classification of unknown images using a support vector machine (SVM).

Image preprocessing

Before images could be run through the classification software, a region of interest (ROI) had to be defined and distinguished from its background. All image analysis routines were run with the MATLAB software package (version R2009a; Mathworks) and its Image Processing Toolbox (version 6.3; Mathworks). Preprocessing was done through an automated Canny edge routine to detect the shell's edges, apply a binary mask, and crop the image to the area of interest. In a few cases where this routine failed (i.e., too much background or overlapping shapes with the larvae), the preprocessing was performed using a manual ROI masking routine in MATLAB.

Training set feature extraction and cross-validation

Reference sets of various sizes, or "training sets," were created by randomly selecting hatchery or visually classified larvae from each species category to train the classifier. Each image from our training sets was run through feature extraction software implemented in MATLAB and identical to that used in Tiwari and Gallager (2003b).

We calculated both Gabor and color-angle features to represent the texture and color of each polarized image. Gabor fast Fourier transform were generated from the spatial domain of Gabor wavelets from 4 scales and 90 rotations of the original image using parameters as described in Tiwari and Gallager (2003b). Rotation and size invariant Gabor features were calculated from the magnitudes of discrete Fourier transform of the Gabor feature matrix. This resulted in 184 values of the mean and standard deviations for the magnitude of the transform coefficients, which were used to represent the image for each RGB (Red, Green, Blue) color channel. This achieved a total of 184 × 3 × 2 = 1104 Gabor texture features. Nine color edge and distribution angles were calculated from HSV (Hue, Saturation, and Value) components of the image as defined in Tiwari and Gallager (2003b) and converted to true angles. Nine invariants of the color image matrix were included in the feature space. A Principle Component Analysis (PCA) was run on the Gabor features to isolate 10-40 of the most significant features and remove redundancy and noise from the 1113 dimensional vector (Zhao et al. 2010).

We used a Support Vector Machine (SVM) classifier toolbox for both cross-validation (CV) and classification implemented in MATLAB (Cawley 2000; http://theoval.cmp.uea.ac.uk/svm/toolbox/). The SVM sorted the feature data from each species, mapped each species to a multi-dimensional space equal to the number of features used, and created decision boundaries for each species group. We combined an SMO (Sequential Minimal Optimization) training algorithm with a DAG-SVM (Directed Acyclic Graph Support Vector Machine) algorithm to form a multi-class neural network for a one-to-one SVM classifier. A one-to-one SVM works with multiple categories by comparing each class to each other, and the image is identified as the class with the highest probability of classification (Lou et al. 2003). We used an SVM with a Gaussian Radial Basis Function (RBF) Kernel of $\gamma = 2$ and regularization parameter, $C = 70$ as in Tiwari and Gallager (2003a). The SVM was chosen because of its ability to operate in a high-dimensional feature space and its history of use in color pattern-recognition algorithms on which the feature extraction software is based (Daugman 2001; Tiwari and Gallager 2003a, 2003b). Initial tests of larval images with Linear Discriminate Analysis (LDA) were not as accurate (Tiwari and Gallager unpub. data).

A leave-one-out (LOO, Fukunaga and Hummels 1989) CV method using the SVM output was then run on every image in the training set to ensure that it was set up to accurately classify unknown images. The LOO method iterates through each image in the training set, trains the SVM classifier with every image except the current image, and uses those boundaries to make the decision to classify the left-out image. The result is an accuracy based on how many images fall into the correct category (from which the image was removed) and how many fall into an "unknown" category. While this is absolute for the hatchery training sets, the accuracy for the visually classified training sets includes a portion of human classification error, and for the purposes of this article, is reported as "agreement."

Classification: unknown images

Once the training set was created and the SVM was trained and cross-validated, we could classify unknown images from the sample set. The process works by loading images from samples and extracting the same texture and color features as the labeled training images. Because any unknown set may contain new species that are indistinguishable using the features previously defined as informative for the training set, the entire PCA to SVM procedure is repeated for the unknowns plus training sets. After this procedure, recognizable false positives were manually removed from classified groups to improve accuracy.

Assessment

The performance, accuracy, and versatility of the polarized image analysis method was assessed in four ways (Table 1): (1) assessing optimal conditions for feature selection and training set formation using images of hatchery-reared larvae; (2) measuring error rates for genetic, visual, and computer classification using hatchery-reared larvae; (3) using genetic methods to identify field larvae; and (4) assessing the supervised computer identification technique to identify species of bivalve larvae collected in the field. Each of these tests were not previously performed by Tiwari and Gallager (2003a, 2003b) and represent important optimization and assessment of this method for bivalve larval identification from field samples.

Optimizing training sets using images from hatchery-reared larvae

We performed several iterations of training and CV using the hatchery larvae as a model for how our method works

Table 1. List of assessment tests, training image and classification detail, and sources of error. KEY: LOO = leave-one-out cross-validation, SVM = Support Vector Machine classifier, AI = *Argopecten irradians*, CV = *Crassostrea virginica*, MA = *Mya arenaria*, MM = *Mercenaria mercenaria*, GD = *Geukensia demissa*, AO = *Arca* sp., AS = *Anomia simplex*, ED = *Ensis directus*, MB = *Macoma balthica*, SS = *Spisula solidissima*, UA = Unknown A.

Assessment test	Reference set	Species	Number of images	Classification method	Sources of error
Feature selection	Hatchery	AI, CV, MA, MM	500 per species	LOO, SVM	computer
Training set size	Hatchery	AI, CV, MA, MM	400/300/200/100/ per species	Hold-out × 5, SVM	computer
Age classes	Hatchery, ages 2,5,7 d	AI, CV, MA, MM	100/species for each age class	5-fold CV, SVM	computer
Computer control	Hatchery	AI, CV, MA, MM	500/species	5 fold CV × 5, SVM	computer
Visual control	Hatchery	AI, CV, MA, MM	398	Visual sorting	human
Molecular control	Hatchery	AI, CV, MA, MM	20/species	Multiplex PCR (Hare et al. 2000)	PCR
Field training set assessment	Visually sorted field images	AO, AS, ED*, GD, MA,MB, MM, SS, UA*	250/9 species, 250/6 species, 400-500/6 species, Unbal./6 species	LOO, 10-fold CV, SVM	human, computer
Field data classification	Visually sorted field images	AO, AS, GD, MA, MB, MM	3250 balanced, 3250 unbalanced	LOO, SVM	human, computer

*Species not verified by DNA.

under ideal conditions. This is because the larval species were known, larvae of each species were grown under relatively uniform hatchery conditions, and the image sets contain equal representation of all age classes and thus birefringence patterns for the larvae.

First, we determined the optimal number of features to extract from images. To assess classification error with varying number of Gabor features, we used a LOO CV analysis from a training set of 500 images of each hatchery species (Fig. 3). Only the principal components (features) with the highest eigenvectors were used, and classification errors decreased as the number of features increased from 10 to 35, but increased with 40. The balance of error and processing time was determined to be optimal with 25 Gabor features. The highest loadings from each principal component were 18 red Gabor features, 6 green Gabor features, and 1 blue Gabor feature. All subsequent classifications were performed by creating new feature sets with 25 PCA-transformed Gabor features and 9 color angle features, for a total feature vector of length 34.

We also determined the optimal number of images to include in training sets by comparing classification accuracies of different sized training sets. We created five different test sets of 100 images from each category, randomly sampled without replacement, to act as unknown images. From the remaining images after each test set was sampled, we created training sets of 100, 200, 300, and 400 images per species. A training set of at least 100 images is necessary for this method to encompass various sizes and orientations of larval shells for each species. We calculated the true accuracies for each species as the number of images that were classified into the correct category divided by total images for that species in the test set (100) and then averaged the values for each test set (Fig. 4A). This was to prevent bias that may result from resampling images for training sets (Bouckaert 2008).

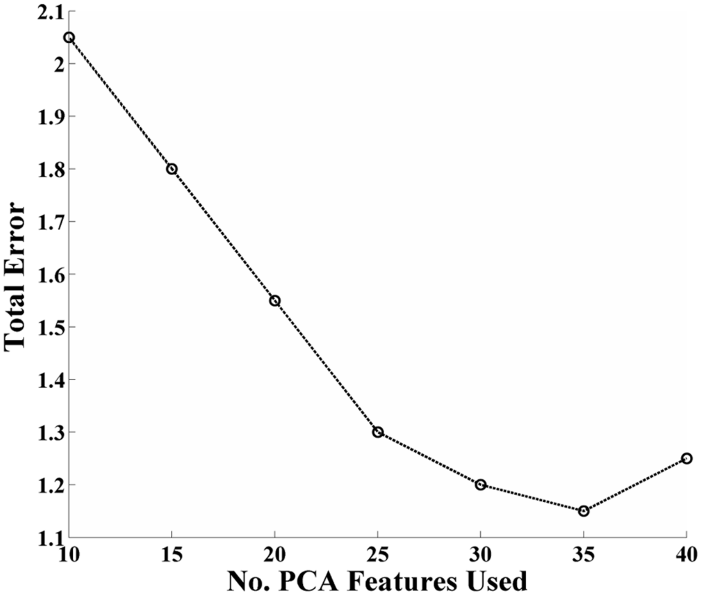

Fig. 3. Error analysis for varying numbers of Gabor features. Total error (as percentage of misclassified images) for hatchery species in a 500 image per species training set is shown versus number of selected variables after Principal Components Analysis on all Gabor features. Errors were calculated as percent misclassified images in a leave-one-out cross-validation analysis, and Principle Components with the highest eigenvectors were selected.

The total accuracies showed general improvement with larger training set size, however, each species behaved differently. *M. arenaria* had greater accuracies with more images, whereas *A. irradians* and *M. mercenaria* did not change much, and *C. virginica* got slightly worse with larger training set size.

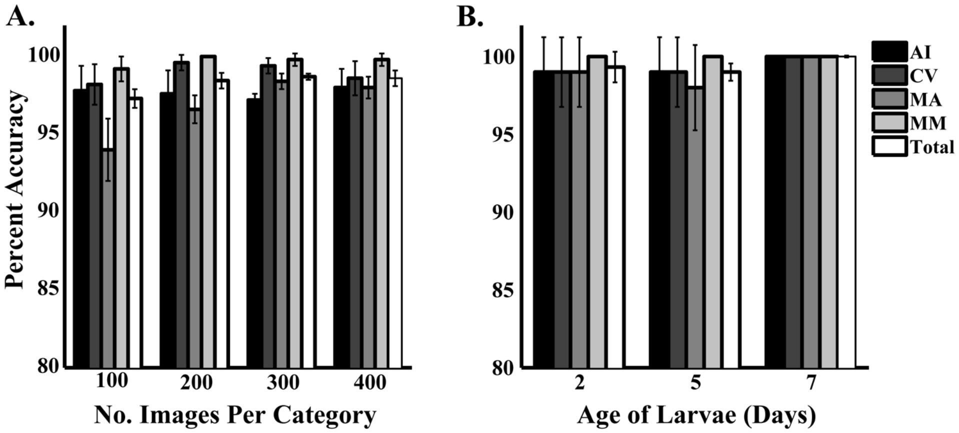

Fig. 4. Accuracy test for hatchery training set with (A) size of training categories and (B) age of larvae. (A) Percent accuracy and standard deviations for classifying 100 test images of each hatchery species (repeated 5-fold with no replacement) with training sets of 100, 200, 300 and 400 images per species category. (B) Accuracies and standard deviations from 5-fold cross validation of training sets containing 100 images of 2, 5, and 7 day old larvae for each hatchery species. Accuracies are the percentage of test images classified correctly (true positives) and then averaged across folds. Accuracies are shown for individual species and combined for the full training set. AI = *Argopecten irradians*, CV = *Crassostrea virginica*, MA = *Mya arenaria*, MM = *Mercenaria mercenaria*.

A repeated measures ANOVA was run to test whether there were any significant differences in accuracies between each training set. This test was used despite the violation of the independence assumption due to repeating images in the training sets (Demsar 2006), and therefore these statistical results should be interpreted with care (Pizarro et al. 2002). The ANOVA showed significant differences with training set size ($F_{3,4}$ = 10.89; P = 0.001), and the 100 images per category training set was significantly different from the rest after a Tukey HSD multiple comparison test. Thus, we suggest that training sets with 200 images per species should provide sufficient accuracy. Including more than 200 images would essentially increase processing time with no significant gain in overall accuracy. In training sets with higher error, more images may be necessary to increase accuracy. In other plankton image analysis, a visually sorted training set with 200-300 images per category was recommended with 60% accuracy (Gorsky et al. 2010), but in other analyses acceptable training sets have contained 100 images or less (Bell and Hopcroft 2008; Gislason and Silva 2009; Fernandes et al. 2009).

The final element that we tested was classifier performance on different age classes of larvae, because each species category contains combinations of images representing different larval stages with different birefringence patterns. We made training sets of images from each species for days 2, 5, and 7, as these ages were present in samples of all four species. Each category contained approximately 100 images per species. A 5-fold cross validation was run on each training set. This works by splitting the images into five equal test groups and training the SVM with the remaining images for each iteration or 'fold.' Accuracies for each class were determined the same way as for the training set size tests (4B).

Results of these tests show this method was highly accurate across age groups and was perfect for day 7 larvae. A nonparametric Friedman's test was performed on the accuracies for total larvae across folds due to unequal variances and confirmed results were not significant between age groups ($\chi_{2,5}$ = 4.77, P = 0.092). It should be noted that this statistical test does not exactly conform to those presented for model selection in the literature (Vazquez et al. 2001; Pizarro et al. 2002). Those examples performed analyses on test data from the same source, and our test data were composed of different images across folds for each training set. We conclude that the classifier does not seem to favor one size class over another, but we recommend including as many different size classes as possible within training sets to encompass the changing shell birefringence patterns that occur with growth, especially if larval size distribution is not known a priori. With all age classes included in a training set, CV accuracies were slightly lower at 92% to 96% of all species because images within each class are less homogenous (Thompson 2011). In other plankton imaging methods, copepods had different error rates between size classes and were found to be better classified if separated by size (Bell and Hopcroft 2008).

Measuring error rates for genetic identification, visual sorting, and computer classification using hatchery-reared larvae

To test the error for the genetic method, DNA was extracted from 20 hatchery larvae of *A. irradians*, *M. mercenaria*, and *M. arenaria* and amplified using the multiplex PCR method described earlier. No false positives (the case of a wrong CO1 band amplification) were reported for the hatchery-reared larvae, but 15% to 35% of the samples were false negatives (the case of an 18S band, but no CO1 band when it was expected, Table 2), resulting in accuracies between 65% to 85%. This

Table 2. Comparisons of visual sorting, molecular identification, and computer image analysis to identify hatchery larvae. Visual sorting was performed by a double-blind classification test, the computer test was performed using a 5-fold cross-validation technique (equivalent to the 400 image/species training set test in Fig. 4), and the multiplex PCR was performed using the protocol of Hare et al. (2000). Accuracies are defined as follows—visual test: the number of images from each species sorted correctly, divided by the total number of images from each species; computer test: the number of true positive classifications divided by the total images in each species category and averaged for each fold; molecular method: the number of correct species-specific amplifications divided by total DNA amplifications from the 18S primer for each species. Only three of four species were analyzed using the molecular method as no primers for *Crassostrea virginica* were used. Total = total images classified, SD = standard deviation of accuracies across species categories. (AI = *Argopecten irradians*, CV = *Crassostrea virginica*, MA = *Mya arenaria*, MM = *Mercenaria mercenaria*).

	AI	CV	MA	MM	Total	Accuracy	SD
Visual	93.6%	100.0%	85.1%	91.0%	398	92.7%	6.2%
Computer	98.0%	98.6%	98.0%	99.8%	1200	98.6%	0.8%
PCR	65.0%	n/a	85.0%	70.0%	60	73.3%	10.4%

indicates that the multiplex method by itself is not always informative for species-specific identifications based on CO1 results if proper amplification does not occur.

To test accuracies for visual classification, we had an outside assistant randomly select 100 images from each of the four hatchery species while maintaining even age class representation. The four image groups were then randomized across species and renamed to make a double-blind test. Each group was then visually classified to species by CMT. Results for the visual classifications produced accuracies ranging from 85% to 100% for each species, with overall accuracy of 92% (Table 2). In visually classifying phytoplankton images, Culverhouse et al. (2003) showed that human performance can vary between 67% to 83%, which alone could introduce substantial variability into a visually classified training set. Sorting accuracies are highest for species like *C. virginica* that have distinct morphologies, and these accuracies represent a minimum estimate for human sorting error as we only performed this test on four species when all possible categories were known.

We used the average accuracies from the 400 images per species training sets on the five test set splits from the analysis in section 1 (Fig. 4A) as our test for computer classification accuracy. This test was equivalent to a 5-fold CV, which gives a less biased estimate for classifier accuracy (Bengio and Grandvalet 2004) and is comparable to the visual and molecular tests. The classification accuracies for these "unknown" images ranged from 98-99.8% for each age group (Table 2), thus demonstrating strong performance of the classifier using error-free training sets and unknown samples with low diversity. Based on overall accuracies, the computer-based classification method was the most accurate of the three for the hatchery larvae.

Genetic identification of field larvae

The multiplex PCR method described above was applied to larvae collected in the field to validate our visual identifications of field larvae. Live (unpreserved) plankton samples were collected from Waquoit Bay, a National Estuarine Research Reserve site on the south shore of Cape Cod, MA, on three dates in June and July 2008 and five dates from May-September 2009. Individual larvae were isolated and placed into separate wells in 1.5 mL 24-well glass-bottom plates to culture in the laboratory. A total of 24 larvae were isolated from three sites in 2008 and four sites in 2009. Every 3 d, larvae were fed algae and imaged live on the polarization microscope. This resulted in a series of images for each larva depicting morphological changes over 12 d (our expected time to metamorphosis for most species) to compare to the molecular IDs. Larvae that survived were washed into 8 mL vials and preserved in 70% ethanol for molecular analysis.

A total of 31 larvae from 2008 and 50 larvae from 2009 were analyzed for a combined total of 81 samples corresponding to 355 images (Table 3). About half of the field PCR samples only amplified at the 18S locus, and those reactions were re-amplified with only the 18S primers. Sequencing this 430 base-pair band provided an alternative means of identifying field larvae that were not targeted by multiplex CO1 primers, as 18S can be diagnostic for some bivalve families and genera (Bell and Grassle 1998). This step was not performed in the above test with the hatchery species but may have led to increased accuracy for that method by eliminating the false negatives resulting from no CO1 amplification (Table 2). Successful PCR products from the 18S re-amplification were purified using the QIAquick PCR Purification kit (Qiagen) and used in one-eighth format sequencing reactions in 96-well plates using Big Dye terminators (version 3, Perkin-Elmer). Samples were purified by isopropanol precipitation and sequenced bi-directionally on an ABI 3700 Capillary Sequencer. Sequences were edited in Sequencher 4.8 (Gene Codes Corporation Inc.) and compared with the GenBank universal database for species identification using BLAST searches (National Center for Biotechnology Information database). A few sequences that did not match with species located in the Cape Cod region were assumed to be from species not represented in GenBank and left out of further analysis.

To verify that our 18S DNA sequence identifications from the BLAST searches correctly corresponded to known Cape Cod species, we extracted DNA from five adult bivalve species

Table 3. Multiplex PCR and 18S sequencing identifications for larvae from live field samples of 2008 and 2009. Eighty-one larvae were used in this analysis, corresponding to 360 images. About half of the samples were re-amplified for 18S sequencing. Total identified from the multiplex and sequencing are shown in the bottom rows.

	Samples		Images		Samples	Images
	2008	2009	2008	2009	Totals	Totals
Guekensia demissa	4	22	19	92	26	111
Macoma balthica	1	0	5	0	1	5
Mercenaria mercenaria	22	1	87	5	23	92
Mya arenaria	3	23	14	114	26	128
Petricola pholadiformis	0	1	0	5	1	5
Spisula solidissima	0	4	0	20	4	20
Total CO1 Multiplex	20	23	82	113	43	195
Total 18S Sequencing	11	27	43	122	38	165
Total amplified	31	50	125	235	81	360

to compare with our larval sequences (for analytical methods, see Thompson 2011). Based on the sequence divergence for different bivalve families for this region of the gene (Bell and Grassle 1998), we can conclude that these identifications using the 18S rRNA are accurate and any disagreement between the image and the sequence identification would thus be a result of human misclassification or error in sample preparation. We also used our molecularly identified field images of larvae to compare the computer and molecular methods (Web Appendix A, Table A1). Overall, the molecular method enabled us to get positive identifications on field larvae from several species that had corresponding images throughout the larval period. This helped us reduce classification error for our visually sorted field training sets.

Assessing the supervised computer identification technique to identify species of bivalve larvae collected in the field

The final assessment was testing classifier performance on unknown field samples using training sets based on visually sorted images of field larvae (referred to as 'visually sorted training sets'). We tested training set size and class numbers, balanced and unbalanced categories within training sets, and compared computer and manual classifications in a larval concentration time series for four species.

Sample collection and training set creation

Samples were taken at four locations throughout Waquoit Bay on a weekly basis from May–Oct 2009. Volumes of 100-200 L were collected in a 53 µm screen and preserved in 4% buffered formalin. Samples were processed by counting total bivalve larvae using a dissecting microscope and imaging a subset on the polarized microscope. See Thompson (2011) for more details on the field sampling procedure.

We determined the hatchery training set would not be an accurate representation of the larvae in our samples and would lead to a disproportionate amount of false positives. We expected the classification accuracies of the visually sorted training set to be lower than our hatchery training set because 1) these images were manually sorted and thus subject to human error and bias, 2) the quality and appearance of field-preserved larvae in images is slightly less than those from the hatcheries because of fungal and other particulate matter that sometimes clouded the image of the shell, and 3) not all growth stages would be equally represented due to higher larval mortalities seen in the field. Of the species in the hatchery training set, only *M. arenaria* and *M. mercenaria* were present in large quantities in the field samples, and *A. irradians* and *C. virginica* composed only about 2% of the total images. Initial classifications of field images using the hatchery training set falsely classified all images as either *A. irradians* or *M. mercenaria*.

Our visually sorted training sets consisted of good quality images of species that were most abundant in the field samples based on the visual identification method. The true ability of a classifier to provide proper estimates of community composition relies on how accurately it represents the sample to be analyzed (Embleton et al. 2003; Bell and Hopcroft 2008). Groups we could not identify with certainty were left out of the training sets, as these showed poor CV results. We created four training sets to compare number of categories, number of images, and whether image numbers were even or unbalanced, with the unbalanced training image numbers proportional to each species' abundance in the field samples. For class selection in plankton identification methods, one must consider the tradeoff between incorporating high taxonomic resolution and achieving highest accuracies, as well as the manual labor it takes to establish larger sets (Gislason and Silva 2009).

Training set size and number of classes

We conducted several tests on field images to determine the appropriate number of images and species classes to include. We used both LOO and a 10-fold CV and employed a corrected resampled t test similar to that in Nadeau and Bengio (2003) and Bouckaert and Frank (2004) to test for significance. This is different from the corrected paired t test reported in the above works, but it includes the same variance correction necessary due to the random partitioning with the CV procedure.

A paired *t* test was not appropriate in this instance as each fold in our CV was subsampled from a different training set. Additionally, each calculation is not completely independent as many images were resampled between training sets. Due to this apparent violation, the statistics reported in these next sections should again be interpreted with care.

We compared a small training set with 250 images for each of nine species (representing 83% of all larvae) to a small training set with 250 images of six species representing 71% of the total larval abundance (Table 4). The six category training set was better at classifying larvae than the nine category training set (t = 2.41, df = 20, *P* = 0.027), with agreements between individual species all above 50% and an overall agreement of 70.8% compared to 65.6%. Thus, a gain of an extra 12% of species resolution by using the nine-category training set results in a 5% decline in classification accuracy, which may be acceptable for some cases. Increasing the number of categories in our training sets can make identifications more difficult as the decision boundaries between species categories are more likely to overlap. Going from six to nine categories increased the chances for misclassification to another species by about 10%. Highest accuracies are often observed with fewer categories (Fernandes et al. 2009).

Next we compared our small six-species training set of 250 images per species to a large six-species training set with 500 images per species (Table 4). Visually sorted images are more sensitive to training set size than our hatchery training set. This comparison was not significant at our α level of 0.05 (t = 2.04, df = 20, *P* = 0.056), but the low *P* value suggests a larger training set could be significantly better in some cases. Doubling the training set size increased overall agreements by 4% with individual species agreements improving by as much as 9%.

Overall, these training set results suggest classification accuracy increases with increasing number of training images, but can decrease as the number of categories increases. This result has been demonstrated in other plankton image processing methods (Davis et al. 2004; Grosjean et al. 2004). Agreement with visual classifications was lower than the hatchery training sets as expected, although size of the categories still did not significantly affect accuracy. Sorting field images presents more challenges as field samples contain mostly smaller, straight-hinged veligers. Not only are these larvae difficult to classify, but they could also bias the classifier by overtraining it with smaller larvae.

Balanced versus unbalanced training sets

We tested whether a training set that better reflected the distribution of our samples would have better accuracy. A common assumption of decision algorithms is that the classifier will operate on data drawn from the same distribution as the training sets (Provost 2000; Lin et al. 2002). Since creating the previous training sets involved balancing the training sets so each category contained equal membership, we may have violated this assumption. In some cases, rebalancing a training set by over- or under- sampling categories can improve training accuracy (Japkowitz and Stephen 2002; Sun et al. 2007). In some cases, Support Vector Machines have been shown to be resistant to some levels of imbalance, and over- or undersampling either does not help or hurts performance (Japkowitz and Stephen 2002; Akbani et al. 2004). To test this assumption, we created an unbalanced training set with the same

Table 4. Results of field training set assessments for varying numbers of species classes and number of images per class. Leave-one-out cross-validations were made for four training sets: a small training set (250 images per category) with nine species, a small training set with six species, a large training set (400-500 images per category) with six species, and a large training set with six species but unequal numbers per category. Total agreement was determined from the total images and false negatives summed over each category. Highest percentage agreement for each species is shown in bold. KEY: No. = number of images, FN = false negative, AG = percent agreement (1-FN/No. Images), AI = *Argopecten irradians*, CV = *Crassostrea virginica*, MA = *Mya arenaria*, MM = *Mercenaria mercenaria*, GD = *Geukensia demissa*, AO = *Arca* sp., AS = *Anomia simplex*, ED = *Ensis directus*, MB = *Macoma balthica*, SS = *Spisula solidissima*, UA = Unknown A.

	small/9 species			small/6 species			large/6 species			unbal./6 species		
Species	**No.**	**FN**	**AG**	**No.**	**FN**	**AG**	**No.**	**FN**	**AG**	**No.**	**FN**	**AG**
AO	250	79	68.4%	250	54	**78.4%**	427	121	71.6%	261	96	63.2%
AS	250	53	78.8%	250	44	82.4%	500	58	**88.4%**	358	65	81.8%
ED*	250	88	64.8%									
GD	250	59	76.4%	250	45	82.0%	500	79	84.2%	531	75	**85.9%**
MA	250	111	55.6%	250	89	64.4%	500	153	69.4%	914	110	**85.7%**
MB	250	104	58.4%	250	90	64.0%	500	152	**69.6%**	442	155	64.9%
MM	250	117	53.2%	250	116	53.6%	500	186	**62.8%**	421	209	50.4%
SS	250	130	48.0%									
UA*	250	33	86.8%									
Total	2250	774	65.6%	1500	438	70.8%	2927	749	74.4%	2927	710	**75.7%**

*Larvae not confirmed by DNA.

total images as the 6-category training set, but with category sizes proportional to each species' abundance in the visually sorted images.

In our approach to balance the training set, only some categories showed improvement (Table 4). Neither balanced or unbalanced performed significantly better (t = 0.039, df = 20, *P* = 0.955). The training set that performed the best for each species was always the one that contained more images. With unequal classes, SVMs will favor those with more examples (Lou et al. 2003). There are other algorithm-level approaches to this problem that could be explored, such as adjusting the cost function or incorporating a Random Forest algorithm, which is more equipped to handle unbalanced data (Lin et al. 2002; Tao et al. 2005; Eitrich and Lang 2006; Sun et al. 2007), but these are beyond the scope of this current study.

We then tested both these training sets on a set of field larvae to determine if a balanced or unbalanced training scheme had better agreement with unknown images. We chose images from one sampling site in the middle of Waquoit Bay as unknowns. Because it is important to represent all types of images in a training set (Gorsky et al. 2010), we added an "other" category composed of 322 images of rarer species that were not represented in the training sets. Although this is a common method for eliminating some false positive classifications (Davis et al. 2004), it can also result in lower classification accuracies between species categories (Thompson 2011). Confusion matrices (CMs) from classifications of this field set for the balanced and unbalanced training sets are shown in Table 5. Overall agreements were similar for both training sets at 63.5% and 64%, however, the highest agreements for five of seven categories were seen with the unbalanced training set. The choice of the best training can be subject to needs and purposes of a particular study (Gislason and Silva 2009). Despite the unbalanced set having higher agreement with most categories, we found the balanced training had higher or similar agreements for the target species in our field study (Thompson 2011). In particular, *M. mercenaria*, a commercially important clam, had only 50% agreement in the unbalanced set compared with 78% agreement in the balanced set. This species was particularly sensitive to training set size.

Table 5. Confusion matrix comparing visual and computer classifications for the balanced and unbalanced 6 category training sets classifying unknown field larvae. Results of the visually classified species are summed up in the rows, whereas results of the computer identifications are summed up in the columns. Diagonals correspond to agreements. Cell colors represent percentages of visually classified larvae classified into each category by the computer (dark red = 75% to 100%, red = 25% to 75%, orange = 10% to 25%, beige = > 0% to 10%). PA = percent agreement or how many larvae were classified the same by both methods (true positives), AO = *Anadara* sp., AS = *Anomia simplex*, GD = *Geukensia demissa*, MA = *Mya arenaria*, MB = *Macoma balthica*, MM = *Mercenaria mercenaria*.

BALANCED									
	AO	AS	GD	MA	MB	MM	OT	TOT	PA
AO	70	1	26	2	6	6	18	129	54.3%
AS	0	177	0	20	5	4	10	216	81.9%
GD	9	1	110	0	3	2	12	137	**80.3%**
MA	0	5	0	273	8	66	27	379	72.0%
MB	1	0	4	16	115	49	27	212	54.2%
MM	0	1	0	24	6	185	21	237	**78.1%**
OT	24	33	32	116	64	88	352	709	49.6%
TOT	104	218	172	451	207	400	467	2019	63.5%

UNBALANCED									
	AO	AS	GD	MA	MB	MM	OT	TOT	PA
AO	87	3	11	2	3	0	23	129	**67.4%**
AS	1	200	0	7	3	0	5	216	**92.6%**
GD	19	0	103	0	2	0	13	137	75.2%
MA	1	30	0	277	17	17	37	379	**73.1%**
MB	6	5	10	9	146	7	29	212	**68.9%**
MM	2	4	1	45	23	119	43	237	50.2%
OT	49	92	20	110	43	34	361	709	**50.9%**
TOT	165	334	145	450	237	177	511	2019	**64.0%**

Classification agreements with time-series

We compared classification results for two species with high classification accuracies (>80%, *Anomia simplex* or jingle clam and *G. demissa*), and two species with lower accuracies (<80% *M. arenaria* and *M. mercenaria*) as a time-series of total larval concentration as estimated from species' abundance in 100 image subsamples (Fig. 5). We compared our visual classifications with the supervised image classification results using the balanced training set and the same computer results after a final manual correction by removing false positive images. This is a common method of improving agreements (Davis et al. 2004; Bell and Hopcroft 2008; Gorsky et al. 2010). Because the computer software does not use size or shape as a distinguishing feature, many false-positive images have distinct morphologies from the target species and can be removed manually. This correction procedure works well because morphology (i.e., size, shape of umbo) is a much better criterion for excluding nontarget species than it is for positively identifying a species (Perino et al. 2008). Any larvae removed manually were not re-sorted into other categories.

Overall, our supervised classification method was able to capture seasonal trends in larval abundance of our four target species. The time-series for *A. simplex* and *G. demissa* show strong correspondence between computer and visual classification (Fig. 5A, B). For *M. arenaria* and *M. mercenaria*, correspondence was not as strong (Fig. 5C, D), possibly a result of misclassifications between the two species (Table 5). Thus, 80% agreement or higher should be strived for when evaluating field training sets to estimate trends in species abundance. False positive classifications for *M. arenaria* that occurred during a peak

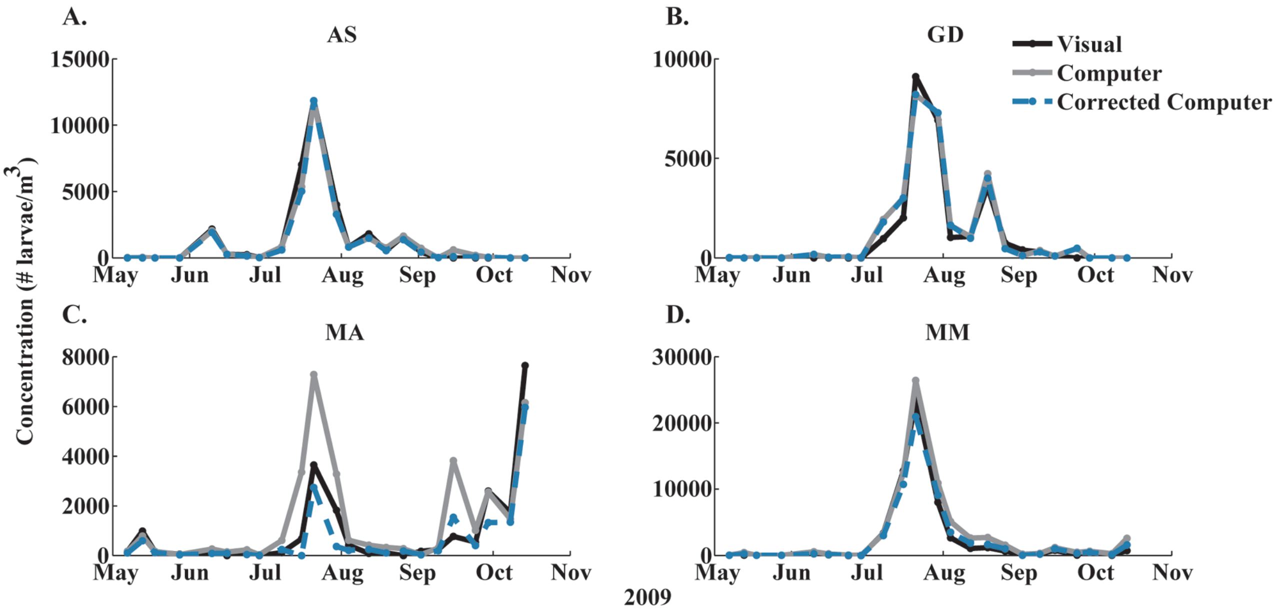

Fig. 5. Time-series of four species' concentrations classified by visual and supervised image classification methods. Samples were collected from a site in Waquoit Bay, MA from May-October 2009. Concentrations were calculated from the total number of bivalve larvae in each sample multiplied by the percentage of each species in a subsample classified by each method. Black solid line corresponds to visual classifications, and the gray and light blue dashed lines correspond to computer and manually corrected computer classifications, respectively. The balanced training set with 6 species and one "other" category was used for computer classifications. (A) *Anomia simplex* (AS), (B) *Geukensia demissa* (GD), (C) *Mya arenaria* (MA), and (D) *Mercenaria mercenaria* (MM).

period of larval abundance resulted in a significant overestimation of abundance (Fig. 5C), although manual correction was able to resolve this error. For *M. mercenaria*, overestimation of abundance by the computer and manually corrected methods in August was small relative to the range of larval concentration for the full series (Fig. 5D), but it was in excess of 100% for the uncorrected images, and up to 76% for the corrected images. This could be significant error if high-frequency samples were taken during this period, as trends may be missed. In a shorter-term study, one should focus a training set on species abundant at the time period of interest. For longer time-series, it may be helpful to change training sets based on species composition at a given period (Gorsky et al. 2010).

We tested the agreement between visual identifications counts and manually validated computer classifications using the Bland-Altman method. This method compares agreement between two methods of measurement subject to error by comparing the residuals of both estimates (Bland and Altman 1986). Plots of residuals show the relationship between the mean of both estimates and the difference observed for each sample (Fig. 6). A perfect correspondence would have points falling on the y = 0 line. Most samples fell within 95% confidence limits for estimates. A slight downward slope for some species indicates the computer may underestimate large sample sizes. Confidence limits were widest for *M. arenaria* (Fig. 6C), indicating that this species has the weakest agreement in estimates. The narrowest limits were observed for *A. simplex* and *G. demissa*. Most estimates differed by less than 10% of the sample. Disparities between training sets and preserved samples have been observed in other plankton identification studies and were attributed to lack of representation in the training sets, human error, presence of false positives, and low numbers of training images (Embleton et al. 2003; Grosjean et al. 2004; Bell and Hopcroft 2008; Gislason and Silva 2009). Although no supervised image analysis method is devoid of error, this polarized image classification method shows potential for estimating species abundance of bivalve larvae in field samples.

Discussion

Our goal was to convert an image processing technique using shell birefringence patterns to distinguish species of bivalve larvae into a reproducible method that can be applied to field studies. The true strength of this method lies in its ability to inexpensively and accurately handle large amounts of samples in a short amount of time. This method works best when known or genetically verified larvae are used to create the training sets to eliminate human misclassification error. The assessment tests confirmed that the classifier performs well on training sets as small as 100 images per species and is consistent at identifying larvae of all ages and morphologies. Using training sets created from sorted field images introduces more error, but this step may be necessary to achieve the best results for field studies. We showed that a few simple correction methods can achieve results consistent with visual identification of larval images but with less overall effort.

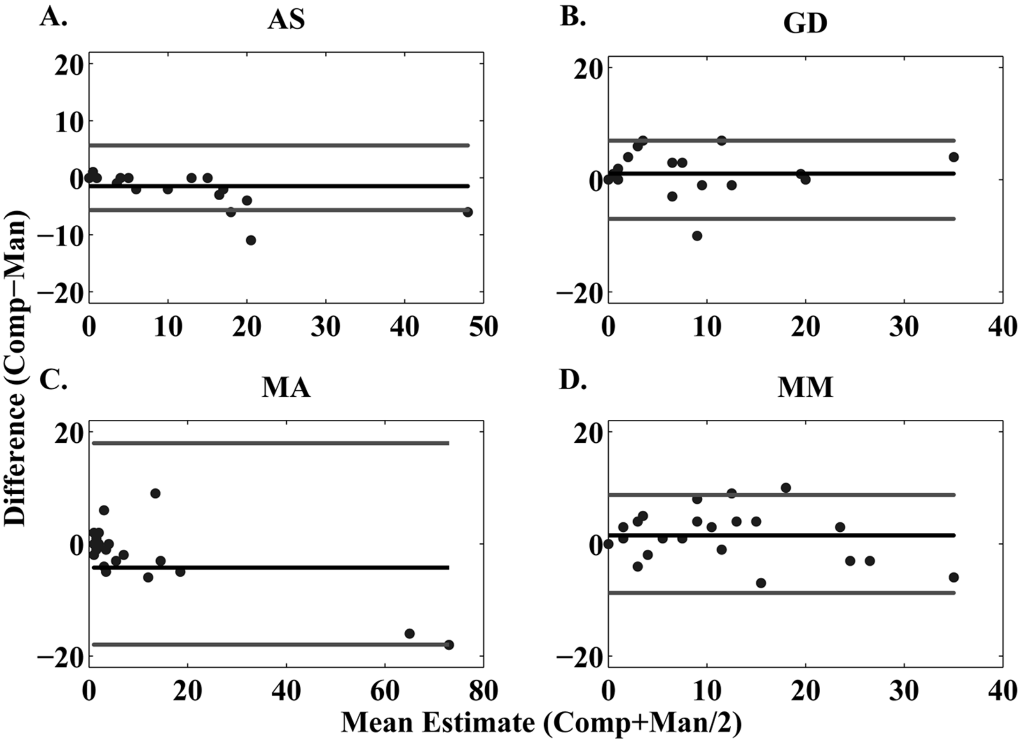

Fig. 6. Bland-Altman plots of residuals for the classification with manual correction. The difference between the number of computer classified and visually classified images of each species are plotted against the mean value of both estimates for (A) *Anomia simplex* (AS), (B) *Geukensia demissa* (GD), (C) *Mya arenaria* (MA), and (D) *Mercenaria mercenaria* (MM). Mean difference (dark line) and 95% confidence intervals (light lines) for estimates are shown for each species. A perfect correspondence would have all points on the y = 0 line.

When compared with the accuracies of other methods of automated plankton identification, these results fall in the middle. For our hatchery training sets, our accuracies fall under the high end of image analysis capabilities (with up to 100% accuracy), but for our field training sets our accuracies are lower, but still acceptable (62% to 88% for the large six species field training set). The video plankton recorder group found that their plankton classification method had higher accuracies for more abundant taxa and lower accuracies for rare taxa, with an overall accuracy range of 45% to 91% (Davis et al. 2004), which was later improved with a dual-classification method (Hu and Davis 2006). Plankton recognition software for the SIPPER II underwater camera was improved from 77% to 90% by adding an active-learning approach with the SVM (Lou et al. 2005). ZooScan users found accuracy was highest for a simplified approach using 8 groups and a random-forest classification method (83.9%) and were able to improve total accuracy after manually reclassifying "suspect" images identified by the computer and merging categories of similar image types (Grosjean et al. 2004; Fernandes et al. 2009; Gorsky et al. 2010). Phytoplankton are traditionally difficult to identify, but automated classification for the in-situ imaging flow cytometer, FlowCytobot, achieved between 68% to 99% accuracy among 22 categories (Sosik and Olson 2007). The DiCANN machine-learning system for dinoflagellates categorized six species with accuracies ranging between 41% to 100% (Culverhouse et al. 2003). Our system has a more limited set of reference images available as compared to some of these methods, as each image had to be manually captured on the microscope. In the future, automation of this process may allow for collection of a greater number of images and the ability to further increase accuracy through active learning and error correction.

We compared our supervised image classification software to a multiplex PCR method. Molecular methods are commonly used to identify species for which distinguishing morphological features are absent. As this method is based on DNA, it gave us a higher level of certainty for many of our field identifications, however, this method was more time consuming and expensive than the image analysis method, which limited the number of larvae we could test. Sequencing adult DNA and measuring sequence divergences confirmed our identifications of four species, but for some rarer species that are not in GenBank, false positive BLAST searches can occur. In addition, larval DNA can be difficult to extract from preserved samples (Larsen et al. 2005). Our 7000 images of field larvae were only a subsample of the total field larvae collected, and performing PCR on this quantity of larvae would be daunting.

The ultimate goal of automated and semi-automated image analysis is to produce useful measures of species abundance, biomass and size for ecological purposes (Gorsky et al. 2010). Our proof-of-concept application involved a series of weekly plankton samples taken from Waquoit Bay, MA in the summer of 2009 (Thompson 2011). Each of the four species had different periods of abundance throughout the time series, and species-specific data such as these can be useful to identify spawns, uncover transport patterns, and track larval distribution and survival over time. Further applications using this method include studies of larval transport patterns in Waquoit Bay (Thompson 2011). Additional applications for this method could involve relating larval supply to juvenile and adult recruitment, revisiting information from archived samples, identifying larvae to analyze gene frequency patterns, and validating transport models with information on species-specific distributions.

The polarization method can be easily adapted for use in other geographical areas, requiring only a polarization microscope with digital camera, computer with at least 2 GB of RAM, and software package for MATLAB. Many molecular-based methods require significant start-up for including a new species. Primer or antibody design both require significant knowledge, data collection, cost, and time to perform. Ideally, our image-analysis method involves imaging a collection of hatchery reared larvae, but if this is not possible, a field-identified training set may suffice. Thus this method could potentially be applied to bivalve larval samples from any location where they could be imaged using polarization as the characteristics of shell mineralization for all bivalves are species-specific (Gallager et al. 1988). Preliminary studies of polarized image analysis on six larval bivalves from the Chesapeake Bay have showed promising results for identification (J. Goodwin unpub. data). We have yet to analyze birefringence patterns for closely related species (ie, the same genus), but previous comparison of the bay scallop, *A. irradians*, and sea scallop, *Placopecten magellanicus* showed distinctive shell birefringence patterns (Tiwari and Gallager 2003a).

Overall, we conclude that a minor sacrifice to accuracy biased by human sorting is worth the ability to handle a large amount of field samples. Automation of image collection could enable larger spatial and temporal coverage than by any published bivalve larval identification method to date by eliminating time-consuming sample processing. Currently, few species-specific field studies of bivalve larvae exist, which limits our understanding of their larval ecology compared with other larval groups. The ability to estimate species-specific abundance from studies with large spatial and temporal coverage in relatively short time periods will greatly increase our understanding of bivalve larvae abundance, distribution, transport, and how species might be responding to climate change. This will have lasting implications in the fields of larval ecology, biological-physical processes, and shellfish restoration and management.

Comments and recommendations

Our method presents a versatile and cost-effective alternative for bivalve larval species identification that compares well with other methods for bivalve larval identification and image analysis techniques for plankton. A main drawback to this method is that it is unknown as to how much variability is present in shell polarization patterns due to environmental conditions that could affect growth and mineralization. Bleaching the samples, which removes tissue and cleans shells, may remove some variability but inhibits DNA analysis (Thompson 2011). Microscope settings may also affect these patterns, which could affect the performance of the classifier if they are not kept standard. This method is similar to that of ZooScan and ZooProcess as it doesn't require a specific instrument to sample images, and thus image quality and type may differ between users (Grosjean et al. 2004; Gorsky et al. 2010). Differences between training and sample images may lead to weaker classifications (Bell and Hopcroft 2008), and many software identifications are sensitive to image quality and illumination (Sieracki et al. 1998). Because our image collections spanned several months to years, we suspect that variations in microscope settings over time may have affected color patterns. Our trials found that training sets cannot be used to classify images taken with different microscope settings unless those images are also represented in a training set.

Another issue with the method we used is that training sets must accurately represent the species composition of the sample set, or many false-positive classifications will occur. Based on our results, training sets should contain more abundant species that together represent at least 50% of the entire sample composition. Samples that contain large numbers of different species may be difficult to use due to the reduced classifier performance with more categories. If species composition of the field samples is not known a priori, it may be difficult to set up a training set using known, lab-reared species. Sorting field larvae can add more error to human classification which is then reflected in classifier performance.

We recommend that further applications of this method take careful examination of species composition of each sample to create a training set that accurately represents the sample to reduce error. If keys or cultured individuals are not available, genetic information can provide a reasonable background for some species identifications. If a known training set cannot be established to accurately represent field larvae, we recommend the following protocol when creating a field-training set:

1. Classify 1000 randomly selected images to the most accurate number of species categories (based on genetics and key information, or preferably cultured individuals).

2. Evaluate which species are most abundant, based on these categories (at least 50% of the entire sample).

3. From the rest of the images (leaving out the ones that were classified), create training sets, starting at 200 images per

category, representing different sizes or morphologies of the species.

4. Evaluate accuracy using these training sets to classify the visually sorted images. Compare to the visual sorted images and adjust species categories and/or number of images per category until the best agreement is reached.

5. Once the training set is optimized, use it to classify all images from the sample set.

For our field image classifications, it was necessary to add a category for images not represented in training sets and perform manual corrections to achieve better correspondence with our visual counts (Thompson 2011). We recommend this if initial agreement of both methods shows many false-positives with unlabeled images, although this may increase manual-processing efforts. More sophisticated methods of feature selection (Lou et al. 2003; Sosik and Olson 2007), dual-classification (Hu and Davis 2006), or active-learning approaches for classifiers (Lou et al. 2005) may help with misclassifications between species categories. Other classifiers such as the Random Forest, which has demonstrated to be superior to SVMs in cases with many zooplankton categories and unbalanced data (Grosjean et al. 2004; Gislason and Silva 2009), could also be investigated, but this classifier has no record of performance on color images of plankton. We found that our simple correction methods provided sufficient agreement to our visual counts when considering the error present in both methods.

The shell birefringence method has been applied to a field transport study of bivalve larvae on Cape Cod, but it can be applied to other environments. Our image analysis method can be applied from both manually extracted images (as in this study) or from optically sampled images from a machine (future studies). The next step for this method is to integrate it into an automated image collection and analysis routine. The Larval Identification and Hydrographic Data Telemetry System (or LIHDAT) is being tested in laboratory settings for analysis of bivalve larvae from plankton samples (Gallager and Tiwari 2008) with the goal of being field-operational. In addition, this software could be appended to other image analysis systems to identify polarized color images of bivalve larvae. The requirements for expanding this method to other environments are minimal, and the software is available by contacting S. Gallager.

References

Akbani, R., S. Kwek, and N. Jackowicz. 2004. Applying support vector machines to imbalaced data set, p. 39-50. *In* Proceedings of European Conference on Machine Learning, Pisa, Italy. September 2004, LNCS.

Ashjian, C. J., C. S. Davis, S. M. Gallager, P. H. Wiebe, and G. L. Lawson. 2008. Distribution of larval krill and zooplankton in association with hydrography in Marguerite Bay, Antarctic Peninsula, in austral fall and winter 2001 described using the video plankton recorder. Deep-Sea Res. II 55:455-471 [doi:10.1016/j.dsr2.2007.11.016].

Bell, J. L., and J. P. Grassle. 1998. A DNA probe for identification of the larvae of the commercial surf clam (*Spisula solidissima*). Molec. Mar. Biol. Biotechnol. 2:129-136.

———, and R. R. Hopcroft. 2008. Assessment of ZooImage as a tool for the classification of zooplankton. J. Plankton Res. 30(12):1351-1367 [doi:10.1093/plankt/fbn092].

Benfield, M.C. , P. Grosjean, P.F. Culverhouse, X. Irigoien, M.E. Sieracki, A. Lopez-Urrutia, H.G. Dam, Q. Hu, C.S. Davis, A. Hansen, C.H. Pilskaln, E.M. Riseman, H. Schultz, P.E. Utgoff, and G. Gorsky. 2007. RAPID: Research on Automated Plankton Identification. Oceanography 20(2):13-26 [doi:10.5670/oceanog.2007.63].

Bengio, Y., and Y. Granvalet. 2004. No unbiased estimator of the variance of K-fold cross-validation. J. Mach. Learn. Res. 5:1089-1105.

Bland, J. M., and D. G. Altman. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet i: 307-310 [doi:10.1016/S0140-6736(86)90837-8].

Boicourt, W. C. 1988. Recruitment dependence on planktonic transport in coastal waters, p. 183-202. *In* B. J. Rothschild (ed.), Toward a theory on biological-physical interactions in the world ocean. Kluwer Academic Publishers.

Bouckaert, R. R. 2008. Practical bias variance decomposition. *In* AI 2008: Advances in artificial intelligence: Lecture notes computer science 5360:247-257 [doi:10.1007/978-3-540-89378-3_24].

———, and E. Frank. (2004). Evaluating the replicability of significance tests for comparing learning algorithms. *In* H. Dai, R. Srikant, & C. Zhang (Eds.), Proceedings 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 26-28, 2004 (pp. 3-12). Berlin: Springer. [doi:10.1007/978-3-540-24775-3_3]

Cawley, G. C. 2000. (MATLAB) Support vector machine toolbox (v0.55). <http://theoval.cmp.uea.ac.uk/svm/toolbox>.

Chanley, P., and J. D. Andrews. 1971. Aids for identification of bivalve larvae of Virginia. Malacologia 11:45-119.

Culverhouse, P. F., R. Williams, B. Reguera, V. Herry, and S. Gonzalez-Gil. 2003. Do experts make mistakes? A comparision of human and machine identification of dinoflagellates. Mar. Ecol. Prog. Ser. 247:17-25 [doi:10.3354/meps247017].

Daugman, J. G. 2001. Statistical richness of visual information: update on recognizing persons by iris patterns. Int. J. Comp. Vision 45(1):25-38 [doi:10.1023/A:1012365806338].

Davis, C. S., S. M. Gallager, and A. R. Solow. 1992. Microaggregations of oceanic plankton observed by towed video microscopy. Science 257:230-232 [doi:10.1126/science.257.5067.230].

———, Q. Hu, S. M. Gallager, X. Tang, and C. J. Ashjian. 2004. Real-time observation of taxa-specific plankton distributions: an optical sampling method. Mar. Ecol. Prog. Ser. 284:77-96 [doi:10.3354/meps284077].

Demsar, J. 2006. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7:1-30.

Eitrich, T., and B. Lang. 2006. Efficient optimization of support vector machine learning parameters for unbalanced datasets. J. Comp. App. Math. 196(2):425-436 [doi:10.1016/j.cam.2005.09.009].

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. Plankton Res. 25(6):669-681 [doi:10.1093/plankt/25.6.669].

Fernandes, J. A., X. Irigoien, G. Boyra, J. A. Lozano, and I. Inza. 2009. Optimizing the number of classes in automated zooplankton classification. J. Plankton Res. 31(1):19-29 [doi:10.1093/plankt/fbn098].

Fukunaga, K., and D. M. Hummels. 1989. Leave-one-out procedures for nonparametric error estimates. IEEE Trans. Pattern Anal. Mach. Intellig. 11:421-423 [doi:10.1109/34.19039].

Gallager, S. M., J. P. Bidwell and A. M. Kuzirian. 1988. Strontium is required in artificial seawater for embryonic shell formation in two species of bivalve molluscs. *In* R. Crick, [ed.], Origin, history and modern aspects of biomineralization in plants and animals. Proceedings of the Fifth International Symposium on Biomineralization, Arlington, Texas. Univ. of Chicago Press.

———, C. S. Davis, A. W. Epstein, A. Solow, and R. C. Beardsley. 1996. High-resolution observations of plankton spatial distributions correlated with hydrograpy in the Great South Channel Georges Bank. Deep-Sea Res. II 43:1627-1664 [doi:10.1016/S0967-0645(96)00058-6].

———, and S. Tiwari. 2008. Optical method and system for rapid identification of biological and inorganic materials using multiscale texture and color invariants. US. Patent No. 7415136.

Garland, E. D. 2000. Temporal variability and vertical structure in larval abundance: the potential roles of biological and physical processes. Doctoral dissertation, Massachusetts Institute of Technology/Woods Hole Oceanographic Institution.

———, and C.A. Zimmer. 2002. Techniques for the identification of bivalve larvae. Mar. Ecol. Prog. Ser. 225:299-310 [doi:10.3354/meps225299].

Gislason, A., and T. Silva. 2009. Comparison between automated analysis of zooplankton using ZooImage and traditional methodology. J. Plankton Res. 31(12):1505-1516 [doi:10.1093/plankt/fbp094].

Gorsky, G., and others. 2010. Digital zooplankton image analysis using the ZooScan integrated system. J. Plankton Res. 32(3):285-303 [doi:10.1093/plankt/fbp124].

Gregg, C. S. 2002. Effects of biological and physical processes on the vertical distribution and horizontal transport of bivalve larvae in an estuarine inlet. Doctoral dissertation, Rutgers Univ.

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

Hare, M. P., S. R. Palumbi, and C. A. Butman. 2000. Single-step species identification of bivalve larvae using multiplex polymerase chain reaction. Mar. Biol. 137:953-961 [doi:10.1007/s002270000402].

Hendriks, I. E., L. A. van Duren, and P. M. J. Herman. 2005. Image analysis techniques: a tool for the identification of bivalve larvae? J. Sea Res. 54:151-162 [doi:10.1016/j.seares.2005.03.001].

Henzler, C. M., E. A. Hoaglund, and S. D. Gaines. 2010. FISH-CS – a rapid method for counting and sorting species of marine zooplankton. Mar. Ecol. Prog. Ser. 410:1-11 [doi:10.3354/meps08654].

Hu, Q., and C. Davis. 2006. Accurate automatic quantification of taxa-specific plankton abundance using dual classification with correction. Mar. Ecol. Prog. Ser. 306:51-61 [doi:10.3354/meps306051].

Irigoien, X., J. A. Jernandes, P. Grosjean, K. Denis, A. Albania, and M. Santos. 2009. Spring zooplankton distribution in the Bay of Biscay from 1998 to 2006 in relation with anchovy recruitment. J. Plankton Res. 31(1):1-17 [doi:10.1093/plankt/fbn096].

Japkowitz, N., and S. Stephen. 2002. The class imbalance problem: a systematic study. Intell. Data Anal. 6:429-449.

Larsen, J. B., M. E. Frischer, L. J. Rasmussen, and B. W. Hansen. 2005. Single-step nested multiplex PCR to differentiate between various bivalve larvae. Mar. Biol. 146:1119-1129 [doi:10.1007/s00227-004-1524-2].

Lin, Y., Y. Lee and G. Wahba. 2002. Support Vector Machines for Classification in Nonstandard Situations. Mach. Learn. 46(1-3):191-202 [doi:10.1023/A:1012406528296].

Loosanoff, V. L., W. S. Miller, and P. B. Smith. 1951. Growth and setting of larvae of Venus mercenaria in relation to temperature. J. Mar. Res. 10:59-81.

———, H. C. Davis, and P. E. Chanley. 1966. Dimensions and shapes of larvae of some marine bivalve mollusks. Malacologia 4:351-435.

Lou, T., K. Kramer, D. Goldgof, L.O. Hall, S. Samson, A. Remsen, and T. Hopkins. 2003. Learning to recognize plankton, p. 888-893. *In* Proceedings of the IEEE International Conference on Systems, Man and Cybernetics. Washington DC, October 2003. IEEE.

———, ———, 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.

Lutz, R. J., and others. 1982. Preliminary observations on the usefulness of hinge structures for identification of bivalve larvae. J. Shell. Res. 2(1):65-70.

MacLeod, N., M. Benfield, and P. Culverhouse. 2010. Time to automate identification. Nature 467:154-155 [doi:10.1038/467154a].

Manriquez, K., R. Escribano, and P. Hidalgo. 2009. The influence of coastal upwelling on the mesozooplankton community structure in the coastal zone off Central/Southern

Chile as assessed by automated image analysis. J. Plankton Res. 31(9):1075-1088 [doi:10.1093/plankt/fbp053].

———, ———, and R. Riquelme-Bugeno. 2012. Spatial structure of the zooplankton community in the coastal upwelling system off central-southern Chile in spring 2004 as assessed by automated image analysis. Prog. Oceanogr. 92-95:121-133 [doi:10.1016/j.pocean.2011.07.020].

Nadeau, C., and Y. Bengio. 2003. Inference for the Generalization Error. Mach. Learn. 52(3):239-281 [doi: 10.1023/A:1024068626366].

Perino, L. L., D. K. Padilla, and M. H. Doall. 2008. Testing the accuracy of morphological identification of northern quahog larvae. J. Shellfish Res. 27(5):1081-1085 [doi:10.2983/0730-8000-27.5.1081].

Pineda, J., J. A. Hare, and S. Sponaugle. 2007. Larval transport and dispersal in the coastal ocean and consequences for population connectivity. Oceanography 20(3):22-39 [doi:10.5670/oceanog.2007.27].

Pizarro, J., E. Guerrero, and P. L. Galindo. 2002. Multiple comparison procedures applied to model selection. Neurocomputing 48:155-173 [doi:10.1016/S0925-2312(01)00653-1].

Provost, F. 2000. Machine learning from imbalanced data sets 101, p. 1-3. *In* Proceedings of the AAAI'00 workshop on learning from imbalanced data sets, Austin, TX. AAAI.

Scheltema, R. S. 1986. On dispersal and planktonic larvae of benthic invertebrates: an eclectic overview and summary of problems. Bull. Mar. Sci. 39(2):290-322.

Shanks, A. L., and L. Brink. 2005. Upwelling, downwelling, and cross-shelf transport of bivalve larvae: test of a hypothesis. Mar. Ecol. Prog. Ser. 302:1-12 [doi:10.3354/meps302001].

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 [doi:10.3354/meps168285].

Sosik, H. M., and R. J. Olson. 2007. Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. Limnol. Oceanogr. Methods 5:204-216 [doi:10.4319/lom.2007.5.204].

Sun, Y., M. S. Kamel, A. K. C. Wong, and Y. Wang. 2007. Cost-sensitive boosting for classification of imbalanced data. Pattern Recogn. 40:3358-3378 [doi:10.1016/j.patcog.2007.04.009].

Tao, Q., G. W. Wu, F. Y. Wang, and J. Wang. 2005. Posterior probability support vector machines for unbalanced data. IEEE Trans. Neural Netw. 16(6):1561-1573 [doi:10.1109/TNN.2005.857955].

Thompson, C.M. 2011. Species-specific patterns in bivalve larval supply to a coastal embayment. Doctoral dissertation, Massachusetts Institute of Technology/Woods Hole Oceanographic Institution.

Thorson, G. 1950. Reproductive and larval ecology of marine bottom invertebrates. Bio. Rev. 25:1-45 [doi:10.1111/j.1469-185X.1950.tb00585.x].

Tiwari, S., and S. M. Gallager. 2003a. Optimizing multiscale invariants for the identification of bivalve larvae. *In* Proceedings of the 2003 IEEE International Conference on Image Processing, Barcelona, Spain, September 14-17, 2003. IEEE.

———, and S. Gallager. 2003b. Machine learning and multiscale methods in the identification of bivalve larvae. *In* Proceedings of the Ninth IEEE International Conference on Computer Vision, Nice, France, October 14-17, 2003. IEEE [doi:10.1109/ICCV.2003.1238388].

Vazquez, E. G., A. Y. Escolano, P. G. Riano, and J. P. Junquera. 2001. Repeated measures multiple comparison procedures applied to model selection in neural networks, p. 88-95. *In* Proc. of the 6th Intl. Conf. on Artificial and Natural Neural Networks (IWANN 2001).

Wight, N. A., J. Suzuki, B. Vadopalas, and C. S. Friedman. 2009. Development and optimization of quantitative PCR assays to aid Ostrea lurida Carpenter 1984 restoration efforts. J. Shellfish Res. 28(1):33-41 [doi:10.2983/035.028.0108].

Zhao, F., F. Lin, H. S. Sea. 2010. Binary SIPPER plankton image classification using random subspace. Neurocomp. 73:1853-1860 [doi:10.1016/j.neucom.2009.12.033].

Submitted 1 September 2011
Revised 1 April 2012
Accepted 17 May 2012