<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper published in *IEEE Journal of Oceanic Engineering*. This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

Citation for the original published paper (version of record):

Verikas, A., Gelzinis, A., Bacauskiene, M., Olenina, I., Vaiciukynas, E. (2015)
An Integrated Approach to Analysis of Phytoplankton Images.
IEEE Journal of Oceanic Engineering, 40(2): 315-326
<http://dx.doi.org/10.1109/JOE.2014.2317955>

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:hh:diva-26558>

An Integrated Approach to Analysis of Phytoplankton Images

Antanas Verikas, *Member, IEEE*, Adas Gelzinis, Marija Bacauskiene, Irina Olenina, and Evaldas Vaiciukynas

Abstract—The main objective of this paper is detection, recognition, and abundance estimation of objects representing the *Prorocentrum minimum* (Pavillard) Schiller (*P. minimum*) species in phytoplankton images. The species is known to cause harmful blooms in many estuarine and coastal environments. The proposed technique for solving the task exploits images of two types, namely, obtained using light and fluorescence microscopy. Various image preprocessing techniques are applied to extract a variety of features characterizing *P. minimum* cells and cell contours. Relevant feature subsets are then selected and used in support vector machine (SVM) as well as random forest (RF) classifiers to distinguish between *P. minimum* cells and other objects. To improve the cell abundance estimation accuracy, classification results are corrected based on probabilities of interclass misclassification. The developed algorithms were tested using 158 phytoplankton images. There were 920 *P. minimum* cells in the images in total. The algorithms detected 98.1% of *P. minimum* cells present in the images and correctly classified 98.09% of all detected objects. The classification accuracy of detected *P. minimum* cells was equal to 98.9%, yielding a 97.0% overall recognition rate of *P. minimum* cells. The feature set used in this work has shown considerable tolerance to out-of-focus distortions. Tests of the system by phytoplankton experts in the cell abundance estimation task of *P. minimum* species have shown that its performance is comparable or even better than performance of phytoplankton experts exhibited in manual counting of artificial microparticles, similar to *P. minimum* cells. The automated system detected and correctly recognized 308 (91.1%) of 338 *P. minimum* cells found by experts in 65 phytoplankton images taken from new phytoplankton samples and erroneously assigned to the *P. minimum* class 3% of other objects. Note that, due to large variations of texture and size of *P. minimum* cells as well as background, the task performed by the system was more complex than that performed by the experts when counting artificial microparticles similar to *P. minimum* cells.

Index Terms—Classification committee, feature extraction, feature selection, phytoplankton images, *Prorocentrum minimum*, random forests (RFs), support vector machine (SVM).

Manuscript received October 30, 2013; revised February 10, 2014; accepted April 11, 2014. This work was supported by the Research Council of Lithuania under Grant LEK-09/2012.

Associate Editor: J. Jaffe.

A. Verikas is with the Intelligent Systems Laboratory, Halmstad University, Halmstad S-30118, Sweden and also with the Department of Electrical and Control Equipment, Kaunas University of Technology, Kaunas LT-51368, Lithuania (e-mail: antanas.verikas@hh.se).

A. Gelzinis, M. Bacauskiene, and E. Vaiciukynas are with the Department of Electrical and Control Equipment, Kaunas University of Technology, Kaunas LT-51368, Lithuania (e-mail: adas.gelzinis@ktu.lt; marija.bacauskiene@ktu.lt; evaldas.vaiciukynas@stud.ktu.lt).

I. Olenina is with the Department of Marine Research of the Environmental Protection Agency, Klaipeda LT-91149, Lithuania, and also with the Marine Science and Technology Center, Klaipeda University, Klaipeda LT-92294, Lithuania (e-mail: i.olenina@aaa.am.lt; irina.olenina@corpi.ku.lt).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/JOE.2014.2317955

I. INTRODUCTION

STUDIES of long-term changes in aquatic ecosystems, assessment of water quality parameters, and monitoring of toxic algal blooms are some examples where identification and counting of plankton cells is being used. Despite significant achievements in automated image analysis in the last two decades, much work in this area still remains in the form of conventional microscope analysis and is very time consuming and labor intensive. For example, obtaining accurate quantitative abundance estimates requires recognizing and counting cells in thousands or even hundreds of thousands of microscopic views. A robust automated image-analysis-based system capable of recognizing and estimating abundance of different plankton species would be of great help and would enable analysis at much larger scales [1]. Benfield *et al.* present an overview of recent developments and challenges in this area [2]. Usually, distinction is made between analysis of phytoplankton and zooplankton images. Phytoplankton cells are usually smaller than zooplankton objects: while phytoplankton cells are all microscopic, ranging from few to several hundred micrometers, the zooplankton objects usually vary within larger size spectra, from hundred micrometers up to several centimeters. Therefore, different systems are usually designed for automated analysis of zooplankton and phytoplankton images. Morphology of both groups is extremely variable, however the shape of zooplankton organisms usually is more complex than phytoplankton [3].

A number of successful approaches have been proposed for automated analysis of zooplankton images. Hu and Davis [4] and Davis *et al.* [5] developed several techniques for analysis of plankton images obtained from a video plankton recorder. Image analysis procedures applied in [5] and [6] included in-focus object detection, object feature extraction, and object classification. In total, 237 features were extracted, including shape factor, seven invariant moments, Fourier boundary descriptors, granulometric curves [7], and ratios computed from dimensions of the bounding box. The first 20–30 principal components were then used for object classification based on learning vector quantization (LVQ) [8]. The average classification accuracy for seven classes was about 61%. Hu and Davis elaborated upon the system presented in [5]. Co-occurrence matrices [9], [10] computed for distances $d = 1, 4, 8, 16$ pixels were used to extract features and a support vector machine (SVM) [11] for object classification. A very large image set consisting of 20 000 plankton images was used to verify the system. The overall classification accuracy of 71% was achieved using seven classes. The ZooScan digital imaging

TABLE I
LIST OF TECHNIQUES DEVELOPED FOR AUTOMATED CLASSIFICATION OF PHYTOPLANKTON SPECIES

#	Author & Year	Technique	# Features	# Classes	Accuracy %
1	Gorsky et al. (1989) [24]	Minimum Distance	5	3	77
2	Sosik and Olson (2007) [25]	SVM	210	22	88
3	Embleton et al. (2003) [26]	MLP	74	4	71
4	Blaschko et al. (2005) [27]	DT, NB, LR, k -NN, SVM, Committee	780	13	70
5	Culverhouse et al. (2003) [28]	RBF Committee	59	6	72
6	Ellis et al. (1997) [29]	MLP Committee	32	6	89
7	Rodenacker et al. (2006) [30]	DT, LDA	204	5	76

system developed by Grosjean *et al.* [12] for automated analysis of zooplankton images was able to achieve about 75% classification accuracy in a task with 29 zooplankton species. A committee made of linear discriminant analysis (LDA), LVQ, and random forests (RFs) [13], [14] was used for the classification. Gorsky *et al.* have recently compared the performance of six classifiers, namely: multilayer perceptron (MLP), 5-nearest neighbor (5-NN), SVM with linear and radial basis function (RBF) kernels, RF, and C4.5 decision trees (DTs) [15] in the task of classification of zooplankton images obtained using the ZooScan system [16]. The number of categories of objects ranged from 5 to 35. RF provided the best overall performance followed by SVM with the linear kernel. Six classifiers, namely LDA, DTs, k -NN, LVQ, MLP, and RF, are implemented in the ZooImage software [17]. Irigoien *et al.* [18] investigated performance of the classifiers in a zooplankton classification task with 63 categories and found that RF provided the best performance. RF was also the best in the task with 17 categories. Hu and Davis proposed using a sequential classifier consisting of a neural network trained on shape-based features followed by an SVM exploiting texture features [19]. The issue of selection the appropriate number of classes (as a tradeoff between the number of classes identified and the accuracy) in automated zooplankton classification was studied by Fernandes *et al.* [20].

Several techniques for classification of binary zooplankton images have also been developed. Performance of several classifiers including SVM, RF, C4.5 trees, and the cascade correlation neural network in the task of categorizing binary zooplankton images into six and seven classes has been explored by Luo *et al.* [21]. SVM provided the highest classification accuracy equal to 90% and 75% for the six and seven classes, respectively. Moment invariants of contour and original images, granulometric features, and domain specific features, such as object size, convex ratio, and eigenvalue ratio, were used. Zhao *et al.* [22] augmented the aforementioned set of features with circular projections, boundary smoothness, object density, moment ratios, and some other geometric features. To reduce the dimensionality, principal component analysis was applied. A committee, designed using the random subspace approach, was used to make a decision. The classification accuracy of 93% was

achieved, when categorizing objects into seven classes. A very similar approach was taken by Tang *et al.* [23].

Developments in the field of automated analysis of phytoplankton images are less numerous compared to zooplankton images. Research conducted by Gorsky *et al.* is one of pioneering attempts in this area [24]. Using a set of simple geometric features, the authors were able to distinguish between three species of distinct size and shape. Sosik and Olson developed a system for automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry [25]. Embleton *et al.* [26] used an MLP to identify four species in lake water samples. A set of suitable features for training the MLP was selected from Fourier descriptors, geometrical features, and features characterizing the gray level distribution in an image region. Blaschko *et al.* [27] achieved 50%–70% classification accuracy in a task of phytoplankton categorization into 12 classes plus an “unknown” class. Various shape features, moments, texture features, and contour features (780 features in total) were used. Several classifiers, including DTs, naive Bayes (NB), ridge linear regression (LR), k -NN, SVM, and bagged as well as boosted ensembles were explored. SVM was the best classifier for the task. Culverhouse *et al.* [28] studied the classification accuracy achieved by the classification-committee-based (consisting of RBF networks) automated system DiCANN [29] and argued that accuracy of about 72% achieved by the system in a six-class phytoplankton categorization task was similar to the accuracy achieved by the trained personnel. Rodenacker *et al.* applied fluorescence imaging in their image acquisition system, to capture more information for discrimination between five classes of phytoplankton [30]. Sosik and Olson [25] presented perhaps the most elaborate study regarding multiclass phytoplankton categorization using data obtained from Imaging FlowCytobot [31]. Combination of video and flow cytometric technologies is used in the Imaging FlowCytobot. A set of 6600 visually identified and manually inspected images distributed across 22 categories was used in the study. In total, 210 features characterizing geometry, shape, symmetry, texture, and invariant moments of objects were extracted and 131 features were selected and used in an SVM for the categorization. The overall accuracy of 88% was reported on the test set. Table I presents a summary of techniques developed by different authors for automated classification of

Fig. 1. Example of phytoplankton image containing *P. minimum* cells (an example is enclosed in a rectangle), cells of native species (enclosed in a circle), and other objects (enclosed in a pentagon).

phytoplankton species, where “Technique” refers to the type of classifier used.

It is evident from the analysis that classification accuracy achieved when solving phytoplankton classification problems varies in a broad range depending on the task and the data. Due to variety of tasks considered, different resolution of images and different size of the data sets used as well as different procedures taken to assess performance of the techniques, comparison of developed approaches is a rather complicated matter. It is worth noting that one very important problem, namely object detection, is seldom addressed in the literature. Robust object detection, however, is a prerequisite step for obtaining an accurate images analysis tool, especially when rather simple imaging systems are used and objects in resulting pictures appear contiguous or overlapped. Touching or overlapping organisms cause difficulties in automated categorization and abundance estimation of the species. Irigoien *et al.* point out that a linear relation between the number of items and the automatic counting holds if the percentage of image area occupied by the items remains below 3%. Above this threshold, automatic counting underestimates abundance due to increased percentage of organisms touching each other [18].

This paper is limited to analysis of one invasive species, a dinoflagellate *Prorocentrum minimum* (Pavillard) Schiller (*P. minimum*), aiming to provide image analysis-based abundance estimation of *P. minimum* cells. The species is known to cause harmful blooms in many estuarine and coastal environments [32]–[34]. Fig. 1 presents an example of phytoplankton image containing *P. minimum* cells (an example is enclosed in a rectangle) as well as cells of native species (enclosed in a circle), and other objects (enclosed in a pentagon).

A new technique for automated analysis of phytoplankton images with high percentage of image area occupied by objects was developed. The main objective of the technique is detection, recognition, and abundance estimation of *P. minimum* cells. To our knowledge, there were no attempts, documented in the literature, to solve these specific tasks. The proposed technique ex-

ploits images of two types, namely, obtained using light and fluorescence microscopy. The fluorescence microscopy is of great help in finding light microscopy image areas containing *P. minimum* cells, thus, mitigating the object detection problem. Various image preprocessing techniques, including phase-congruency-based image preprocessing [35], [36], are applied to extract a variety of features characterizing cells and cell contours in light microscopy images. Relevant feature subsets are then selected and used in SVM as well as RF classifiers to distinguish between *P. minimum* cells and other objects. Classification-accuracy-based floating search and genetic search are two techniques applied to select features for the RF and SVM, respectively. To improve the cell abundance estimation accuracy, classification results are corrected based on class-confusion probabilities p_{jk} —the probability that the classifier assigns a j th class observation to class k .

II. DATA

P. minimum cells vary in shape from triangular- to oval- and heart-shaped. The length of *P. minimum* cells varies from 14 to 22 μm , while the width ranges from 12 to 18 μm . Phytoplankton samples used for obtaining images were received from: 1) natural southeastern Baltic Sea phytoplankton samples containing *P. minimum* cells; 2) cultured *P. minimum*; and 3) natural phytoplankton mixed with cultured *P. minimum*. All samples were fixed with acetic Lugol’s solution (in proportion: 0.5 mL of solution for 100 mL of sample). Images for the analysis were obtained from a simple red–green–blue (RGB) color camera, providing photos of 3264×2448 pixels, attached to an inverted microscope with the magnification factor of $400\times$. Only one view, resulting from the G channel, has been used in this study, since the other two components do not provide much new information for detection and recognition of *P. minimum* cells. Therefore, a monochrome camera with a “green” filter could be used instead. Presence of chlorophyll in phytoplankton cells makes them glow under ultraviolet (UV) light, and this property becomes of immense help for detecting *P. minimum* cells. Therefore, two consecutive, with a delay of a few seconds, photos (using light and fluorescence microscopy) were obtained of the same location (see Fig. 2), presenting two phytoplankton images with *P. minimum* cells.

III. OBJECT DETECTION AND FEATURE EXTRACTION

A. Object Detection

The fluorescence microscopy is of great help in finding light microscopy image areas containing *P. minimum* cells. To solve this object detection task, we use a simple, histogram-based image binarization technique [37], applied to fluorescence microscopy images. A threshold value t^* needs to be found to accomplish the binarization.

Let L be the number of gray levels in the image and K_i the number of pixels at a gray level i . The total number of pixels in a given image is then equal to $K = \sum_{i=1}^L K_i$, and the probability of a gray level i is defined as $p_i = K_i/K$. The histogram of the image is then given by $h(i) = K_i$. The optimal t^* sought is such that the between-cluster variance $\sigma_B^2(t)$ is maximized

Fig. 2. Example of phytoplankton image obtained using (a) the light microscopy and (b) the fluorescence microscopy.

when dividing the histogram into two clusters C_0 and C_1 at the image gray value equal to t^*

$$\sigma_B^2(t^*) = \max_{t \in T} \sigma_B^2(t) \quad (1)$$

where T is a set of gray values of the histogram. The between-cluster variance is given by:

$$\sigma_B^2(t) = P_0(\mu_0 - \mu_T)^2 + P_1(\mu_1 - \mu_T)^2 \quad (2)$$

where P_0 and P_1 stand for the cluster occurrence probabilities, μ_0 and μ_1 are means of the clusters, and μ_T is the total mean.

A high $\sigma_B^2(t^*)$ value indicates reliable image binarization. We defined the following measure γ to assess the reliability of binarization result

$$\gamma = 1 - \exp \{-\vartheta \sigma_B^2(t^*)\} \quad (3)$$

where the experimentally chosen parameter ϑ determines sensitivity of the measure. Value of $\vartheta = 10$ was a good choice for the task. Using such value of ϑ , values of $\gamma > 0.6$ indicated a reliable segmentation result. Images exhibiting low γ values can be left aside from the analysis.

The obtained binary objects are then superimposed on light microscopy images and regions (slightly larger than the binary objects), cropped, and used to extract features.

B. Feature Extraction

To obtain comprehensive characterization of *P. minimum* cells, a relatively large number of features of various types were used. We categorize the features into the following five groups.

- 1) object geometry;
- 2) Fourier descriptors;
- 3) contour curvature;
- 4) image properties near the contour;
- 5) cell image properties.

Compared to our previous work [36], the feature set used in this study was extended by groups 2), 3), and 4), and by adding several features to group 1).

1) *Object Geometry*: Area, perimeter, major axis, circularity, eccentricity, roundness, and compactness are the features used in this study to reflect object geometry.

2) *Fourier Descriptors*: A K -point digital boundary in the xy -plane can be represented as the sequence of coordinates $c(k) = [x(k), y(k)]$, for $k = 0, 1, \dots, K - 1$. Each coordinate pair can be treated as a complex number so that

$$c(k) = x(k) + jy(k). \quad (4)$$

The discrete Fourier transform of $c(k)$ is

$$a(u) = \frac{1}{K} \sum_{k=0}^{K-1} c(k) e^{-j2\pi uk/K} \quad (5)$$

for $u = 0, 1, \dots, K - 1$. The complex coefficients $a(u)$ are called the Fourier descriptors of the boundary. Instead of using all the Fourier descriptors, only the absolute values of the first $P < K$ coefficients were used.

3) *Curvature Features*: For a plane curve given as $c(k) = [x(k), y(k)]$, the curvature is given by

$$\kappa(k) = \frac{|x'y'' - x''y'|}{(x'^2 + y'^2)^{3/2}} \quad (6)$$

where primes refer to derivatives with respect to k .

Based on the curvature assessment at each contour point, we compute four curvature related features: a) $\text{mean}(\kappa)$; b) $\text{max}(\kappa)$; c) $\text{std}(\kappa)$; and d) upper quartile value of κ .

4) *Image Properties in the Vicinity of the Contour*: For each contour $c(k)$ point $k = 1, 2, \dots, K$, one interior $c_{\text{int}}(k)$ and one exterior $c_{\text{ext}}(k)$ points are chosen perpendicular to the contour at a distance δ . Various image intensity characteristics on both sides of the contour are computed based on positions of $c_{\text{int}}(k)$ and $c_{\text{ext}}(k)$. Average image gradient as well as average image gradient near the contour is also used as image features. Features from the first four groups are given in Table II.

5) *Cell Image Properties*: A variety of features are used to characterize cell image properties. Among others, we also use features computed from the so-called M , m , and C images obtained by applying the phase-congruency-based image preprocessing [35], [36]. A complete list of features used to characterize cell image properties is given in Table III.

IV. CLASSIFICATION

A Gaussian kernel SVM, an RF classifier [13], [14], and a committee made of these two are used to make a decision in

TABLE II
CELL FEATURES

#	Features	Comment
<i>Object geometry</i>		
1	Area	
2	Perimeter	
3	Major Axis	
4	Circularity	$\frac{4\pi \cdot \text{Area}}{\text{Perimeter}^2}$
5	Eccentricity	The ratio between the length of the short axis to the long axis.
6	Roundness	$\frac{4 \cdot \text{Area}}{\pi \cdot \text{Major Axis}^2}$
7	Compactness	$\frac{\text{Perimeter}^2}{\text{Area}}$
<i>Fourier descriptors</i>		
8-15	Fourier Descriptors	
<i>Contour curvature</i>		
16	mean(κ)	
17	max(κ)	
18	std(κ)	
19	upper quartile value of κ	
<i>Image properties in the vicinity of the contour</i>		
20	std($I_{c_{int}} - I_{c_{ext}}$)	Standard deviation of the intensity difference between the interior and exterior points.
21-22	mean($I_{c_{int}}$); std($I_{c_{int}}$)	Mean and standard deviation of intensity of the interior and exterior points.
23-24	mean(I_c); std(I_c)	Mean intensity and standard deviation of contour points.
25-26	mean($I_{c_{ext}}$); std($I_{c_{ext}}$)	Intensity mean and standard deviation of the exterior points.
27	mean($I_{c_{int}}$) - mean($I_{c_{ext}}$)	Difference between mean intensity of the interior and exterior points.
28	mean($\ \nabla I\ $)	Average gradient magnitude.
29	mean($\ \nabla I_{c_{int}, c, c_{ext}}\ $)	Average gradient magnitude in the vicinity of the contour.
30-39		Features 20-29, obtained using image (G) of the local intensity gradient in the direction of cell center, instead of I.

this work. We combine outputs of SVM and RF by applying weighted averaging to the *a posteriori* probabilities obtained from these two classifiers. The *a posteriori* probability from a trained RF is estimated as

$$p(\{t_1, \dots, t_L\}, \mathbf{r}, q) = \frac{\sum_{i=1}^L f(t_i, \mathbf{r}, q)}{L} \quad (7)$$

where L is the number of trees $t_1, \dots, t_L$ in the RF, $\mathbf{r}$ is the object being classified, q is a class label, and $f(t_i, \mathbf{r}, q)$ stands for the q th class frequency in the leaf node, into which $\mathbf{r}$ falls in the i th tree t_i of the forest

$$f(t_i, \mathbf{r}, q) = \frac{n(t_i, \mathbf{r}, q)}{\sum_{j=1}^Q n(t_i, \mathbf{r}, q_j)} \quad (8)$$

where Q is the number of classes and $n(t_i, \mathbf{r}, q)$ is the number of training data coming from the class q and falling into the same leaf node of t_i as $\mathbf{r}$.

V. FEATURE SELECTION

Feature selection for committees is a difficult problem, since an accurate committee is obtained by combining members, which are not only accurate but also diverse. Committee-targeted feature selection can help creating diverse members of a committee. Since a committee in this work is made of only two members, of completely different kind, we perform feature selection separately for SVM and RF. Completely different methodologies used to design SVM and RF usually lead to some diversity of the models.

Feature selection for RF is usually based on feature importance evaluations obtained from the RF software [39]. For an

TABLE III
CELL FEATURES (CONTINUATION)

#	Features
<i>Cell image properties</i>	
40-46	Hu 7 invariant moments [38].
47	Standard deviation (Std) of object grey levels.
48	Entropy of object grey levels.
49-50	Mean and std of local entropy of object grey levels. A 9×9 window is used for the computation.
51-52	Mean and std of local std. A 3×3 window is used for the computation.
53-66	14 Haralick's coefficients [9], computed from the averaged (over 4 directions) co-occurrence matrices estimated using the distance parameter $d = 5$.
67-69	Mean intensity of image filtered by Gabor filters of scales: 7, 10, and 15. Orientations 30° , 60° , 90° , 120° , 150° , and 180° are used and the results are averaged.
70-72	Std of intensity of image filtered by Gabor filters of the same four scales and orientations.
73-74	Mean and std of M image pixels corresponding to the object.
75-76	Mean and std of m image pixels corresponding to the object.
77-78	Mean and std of C image pixels corresponding to the object.
79-90	Mean and std of object pixels of the following six images obtained by applying different image preprocessing techniques: a) binary object image obtained from the FCM clustering; b) binary object perimeter image obtained from the FCM clustering; c) Sobel gradient magnitude image; d) discrete Laplacian image; e) image of differences between adjacent elements of object grey levels; f) image of local standard deviations computed using window of 7×7 size.
91-100	Features computed from the local std image to characterize average intensity variation when going from object's exterior towards center of the object (30 steps are used), where the average intensity is represented by the average intensity of points on the shrinking object perimeter line. In this way a vector $\mathbf{z}$ of 30 average intensity values is obtained and the following features are then computed: (91-92) mean and std of the intensities; (93-100) the first eight coefficients of the Discrete Cosine Transform of $\mathbf{z}$.
101-110	As 91-100, except that M image is used instead of local std image.

RF tree grown on a bootstrap sample, the out-of-bag (OOB) data (data not used to grow the tree) are used to estimate feature importance. The importance measure $\overline{D}_j$ for feature x_j is given by [39]

$$\overline{D}_j = \frac{1}{L} \sum_{b=1}^L (R_b^{\text{obj}} - R_{b,j}^{\text{obj}}) \quad (9)$$

where L is the number of bootstrap samples (trees in the forest), R_b^{obj} is the number of correct classifications of the OOB data by the tree T_b , and $R_{b,j}^{\text{obj}}$ is the number of correct classifications of the OOB data when values of x_j in the OOB set were randomly permuted. It is expected that, for an important feature, a large decrease in the number of correct classifications will be observed when the feature values are randomly permuted.

Sequential backward feature elimination based on feature importance evaluations given by (9) is usually applied to select features for RF. A recent study has shown, however, that better performance can be achieved by using classification-accuracy-

based feature selection instead of feature importance evaluations provided by the RF software [14]. Therefore, we applied classification-accuracy-based floating search to select features for RF. The process starts with a best pair of features. In the next step, one feature, exhibiting the least decrease in classification accuracy, is eliminated from the current feature subset and a pair of features providing the highest increase in classification accuracy is added. The process of eliminating one and adding two features is continued until the maximum classification accuracy of OOB data is determined.

We used the same genetic algorithm as in [36], to select features for SVM and to determine the optimal values of SVM hyperparameters.

VI. ESTIMATING ABUNDANCE OF *P. MINIMUM* CELLS

Results of automatic classification can be directly used to estimate the abundance of *P. minimum* cells. However, such an estimate is not accurate, unless the classification error is equal to zero. Therefore, correction of results of automatic classification

Fig. 3. Two examples of object detection results, where detected objects are shown by contours, while arrows point to undetected *P. minimum* cells.

is usually required. Let us assume that $\mathbf{n} = (n_1, n_2, \dots, n_Q)^T$ denotes correct, but unknown, number of observations from different classes $j = 1, 2, \dots, Q$ in a given data set. The total number of observations $N = \sum_{j=1}^Q n_j$ in the set is known. Let p_{jk} denote the probability that the classifier assigns a j th class observation to class k and let $\mathbf{P} = [p_{jk}]$ be a $Q \times Q$ matrix of these probabilities. Let $m_j = \sum_{i=1}^Q m_{ij}$ be the number of observations assigned to class j by the classifier, where m_{ij} is the number of i th class observations assigned to class j . The expected m_{ij} value is equal to $n_i p_{ij}$. The expected m_j value is then equal to $E(m_j) = \sum_{i=1}^Q n_i p_{ij}$. In the matrix form, $E(\mathbf{m}) = \mathbf{P}^T \mathbf{n}$, where $\mathbf{m} = (m_1, \dots, m_Q)^T$ and E denotes expectation. The unbiased estimate of $\mathbf{n} = (n_1, n_2, \dots, n_Q)^T$ can then be computed as [40]

$$\hat{\mathbf{n}} = (\mathbf{P}^T)^{-1} \mathbf{m} \quad (10)$$

where $\mathbf{m} = (m_1, \dots, m_Q)^T$ is given by the classification results.

We use classification results of the validation set data to estimate the matrix $\mathbf{P}$. Classifiers were trained 20 times using different random splits of the data set into training and validation subsets, and the obtained classification results were averaged. According to Yuan [41], the matrix $\mathbf{P}$ is nonnegative and strictly diagonally dominant if $p_{jj} > 0.5, \forall j, j = 1, 2, \dots, Q$. Nonnegative and strictly diagonally dominant matrices are invertible [42]. It was also shown that inverse correction of classification results provides reliable and stable estimates if $p_{jj} \geq 0.7$ [41]. These conditions are clearly satisfied by our classifiers.

VII. EXPERIMENTAL INVESTIGATIONS

A. Parameters of the Algorithms and Experimental Setup

There are several parameters governing behavior of the algorithms. The appropriate values of the parameters are chosen experimentally. The number of randomly selected features used to split a tree node in the RF design process was set to $\sqrt{\eta}$, where η is the total number of variables used (values ranging from about 0.05η to 0.3η gave good performance). The number of trees in the RF was set to 1000 (RFs consisting of 200 trees and larger worked well).

We used Breiman's implementation of RFs [39] and the LIBSVM software [43]. Observe that to evaluate fitness of each

chromosome during the genetic search process, the SVM was trained 20 times using different random splits of the data set into training and validation subsets and results were averaged.

Two criteria were used to assess performance of the algorithms: the percentage of detected *P. minimum* cells (if compared to all *P. minimum* cells present in the images) and the test data set classification accuracy obtained when classifying detected objects into *P. minimum* cells and other objects (since objects not belonging to the class of *P. minimum* cells are also detected). The accuracy reported here is the average accuracy obtained from 20 trials using random split of the data set into learning and test subsets.

B. Cell Detection Results

In total, 158 images were processed and results of automated analysis were verified by manual inspection. The manual inspection has shown that there were 920 *P. minimum* cells in these images in total. The object detection algorithms based on the fluorescence microscopy have found 903 *P. minimum* cells. Thus, 98.1% of *P. minimum* cells present in the images were detected.

Fig. 3 presents two examples of object detection results without using the fluorescence microscopy. Detected objects are shown by contours, while arrows point to undetected *P. minimum* cells. As can be seen from Fig. 3, majority of small "uninteresting" objects were eliminated and all *P. minimum* cells, except the two marked by arrows and those with center points too close to the image boundaries, were detected. Note that objects with center points too close to image boundaries were left out of the analysis, due to constraints imposed by the phase-congruency-based image preprocessing [36]. When using fluorescence, all erroneous detections due to small "uninteresting" objects [clearly seen in the upper left and bottom right parts of the image shown in Fig. 3(b)] disappear, since these objects do not fluoresce.

C. Cell Classification Results

Fig. 4 plots OOB data classification accuracy as a function of the number of features used to design RF. Results for two strategies to select features for RF, namely add a pair and remove one (+2-1) and add one (+1), are shown in Fig. 4. As can be seen from Fig. 4, the (+2-1) strategy clearly outperforms the (+1)

Fig. 4. Classification accuracy as a function of the number of variables used.

Fig. 5. Variable importance values [based on (9)] obtained from the RF software.

one. For RF, the best classification accuracy was achieved using 47 features. We also computed importance [based on (9)] of all 110 features, using the RF software. The importance values are shown in Fig. 5 along with triangles identifying features, which have been included into the final subset of 47 features using the $(+2 - 1)$ feature selection strategy. Feature numbering on the x -axis of Fig. 5 corresponds to the numbers given in Tables II and III. As can be seen from Fig. 5, some features exhibiting relatively high importance values in Fig. 5 were eliminated, while some features with relatively low importance values in Fig. 5 were included into the final feature subset used to design the final RF classifier. Thus, classification-accuracy-based feature selection is more appropriate than the one based on feature importance evaluations obtained from the RF software. Features of various types were found to be salient for the RF.

The optimal feature subset, selected by the genetic algorithm for the SVM classifier, was found to consist of 20 features identified by circles in Fig. 5. The average test data set classification accuracy obtained from the SVM, RF, and the committee was equal to 98.04%, 97.37%, and 98.09%, respectively. Note that the test data set contained both *P. minimum* cells and other objects. The classification accuracy of detected *P. minimum* cells was equal to 98.9%. Since the algorithms detected 98.1% of *P.*

minimum cells and correctly classified 98.9% of detected cells, a 97% overall recognition rate of *P. minimum* cells was achieved. On average, the SVM was more accurate than the RF on the 95% significance level. Since in many cases of classification errors both SVM and RF erroneously classified the same samples, the committee was not so effective.

A receiver operating characteristic (ROC) illustrates the performance of a binary classifier when its decision making threshold is varied and enables selection of possibly optimal models. Fig. 6 presents ROC curves computed for the SVM, RF, and the committee. As can be seen from Fig. 6, the ROC curves are very similar and are almost identical for the SVM and the committee. Thus, SVM can be considered as a preferred classifier for this particular task.

To correct classification results, based on (10), entries of the probability matrix $\mathbf{P}$ were estimated and are provided as follows:

$$\mathbf{P} = \begin{pmatrix} 0.9893 & 0.0309 \\ 0.0107 & 0.9691 \end{pmatrix}. \quad (11)$$

The RF software also produces a data proximity matrix $\mathbf{\Pi}$. To obtain the matrix, for each tree grown, the data are run down the tree. If two observations $\mathbf{x}_i$ and $\mathbf{x}_j$ occupy the same terminal

Fig. 6. ROC curves computed for the RF, SVM, and the committee.

Fig. 7. Data proximity matrix mapped onto the 2-D space by multidimensional scaling.

node of the tree, $\pi(i, j)$ is increased by one. When an RF is grown, the proximities are divided by the number of trees in the RF. Data proximities possess an important property, meaning that only variables used by the forest contribute to the proximity values.

Insights into similarity of objects coming from the two classes can be obtained by exploring the data proximity matrix. The matrix can be mapped on the 2-D space. Fig. 7 presents such mapping obtained using multidimensional scaling. *P. minimum* cells are denoted by rectangles and other objects by triangles in the figure. Due to simplicity of the imaging system used, there were images containing rather unfocused regions. *P. minimum* cells extracted from such regions are labeled by filled-in markers in Fig. 7. There were 20.2% of such *P. minimum* cells in total, while the overall error rate is about 5.7%. Thus, most of *P. minimum* cells extracted from unfocused regions were recognized correctly. Moreover, Fig. 7 shows that *P. minimum* cells extracted from unfocused regions overlap, to a great extent, with other *P. minimum* cells. Thus, the feature system used is rather robust to such type of distortions.

Fig. 8. Example of the first type error. The arrow points to an erroneously classified *P. minimum* cell lying on their narrower side. Correctly classified *P. minimum* cells are marked by white and other objects by black contours. Fig. 9. Examples of the first type error. The arrows point to erroneously classified *P. minimum* cells, coming from an unfocused image area. Correctly classified *P. minimum* cells are marked by white and other objects by black contours. Fig. 10. Example of the second type error. The arrow points to a phytoplankton cell of other type, which was erroneously assigned to the *P. minimum* class. Correctly classified *P. minimum* cells are marked by white and other objects by black contours.

Fig. 11. Results of the interexpert calibration obtained from ten phytoplankton experts.

D. Comparison of Results With Assessments Obtained by Experts

The developed system has been tested by experts of the Coastal Research and Planning Institute of Klaipeda University (Klaipeda, Lithuania). To test the software, new 65 phytoplankton images were taken from different areas of the camera view. Experts of the Institute found 338 *P. minimum* cells in these images. The number of cells in one image ranged from 0 to 11, with five cells on average. The automated system detected and correctly recognized 308 (91.1%) of all *P. minimum* cells present in the images.

When investigating results of the automated cell detection and classification, the experts found the following.

- 1) The automated system erroneously classified 8.9% of all *P. minimum* cells present in the images (the first type error). The first type error is characteristic for cells lying on their narrower side (see Fig. 8 for an example) and/or cells coming from unfocused image areas (see Fig. 9).
- 2) The system erroneously assigned to the *P. minimum* class 3% of other objects present in the images (the second type error). An example of the second type error can be found in Fig. 10.

To compare performance of the automated system with performance of phytoplankton experts exhibited in manual counting of *P. minimum* cells, we used results of interexpert calibration presented in [44]. Each of ten experts (H1–H10) counted artificial microparticles of 20 μm in diameter, similar to *P. minimum* cells, in two chambers (chamber 1 and chamber 2). Density of particles in the specimen was equal to $200\,000 \pm 20\,000$ microparticles per liter. Results of the interexpert calibration are shown in Fig. 11. As can be seen from Fig. 11, the number of particles found by the experts ranged from 175 000 to 220 000 particles per liter. Thus, the counting accuracy achieved by the experts was 88%–90%. Therefore, the 91.1% accuracy obtained by the automated system in the invasive *P. minimum* species recognition task is rather encouraging.

It is worth mentioning that when preparing specimens for the automated system, care should be taken that *P. minimum* cells do not stick together into large clusters, as the one seen in the upper part of Fig. 12. *P. minimum* cells correctly recognized by the automated system are labeled by white contours in Fig. 12.

Fig. 12. Image containing a cluster of *P. minimum* cells in the upper part of the image. *P. minimum* cells correctly recognized by the automated system are labeled by white contours, objects assigned to the “other objects” class are marked by black contours, and the arrow points to a *P. minimum* cell erroneously assigned to the “other objects” class.

Objects assigned to the “other objects” class are marked by black contours, while the arrow points to a *P. minimum* cell erroneously assigned to the “other objects” class. As can be seen, some *P. minimum* cells have not been identified in the upper part of Fig. 12. Sticking together objects and rather unfocused image areas are two main reasons causing this kind of errors.

Experience with the system indicates that areas of specimens with clusters of objects usually result into unfocused image areas obtained from such clusters. Even if the system exhibits considerable tolerance to this kind of distortion, recognition accuracy deteriorates when analyzing a cluster of objects in an unfocused image area. The analysis results presented in Fig. 12 illustrate the case.

VIII. DISCUSSION AND CONCLUSION

A technique for automated detection, recognition, and derivation of abundance estimates of *P. minimum* invasive species, which is known to cause harmful blooms in many estuarine and coastal environments, was presented in this work. The technique exploits images obtained using both light and fluorescence microscopy. A variety of features characterizing *P. minimum* cells and cell contours are extracted using various image preprocessing techniques and used in SVM as well as RF classifiers to distinguish between *P. minimum* cells and other objects. To improve cell abundance estimation accuracy, classification results are corrected using probabilities of interclass misclassification.

Experimental tests have shown that robust object detection is possible even in images where image area occupied by objects was much larger than 3%. Combination of light and fluorescence microscopy allowed obtaining a relatively high cell detection accuracy. The algorithms detected 98.1% of *P. minimum* cells present in the images and correctly classified 98.09% of all detected objects. The classification accuracy of detected *P. minimum* cells was equal to 98.9%. Therefore, a 97% overall recognition accuracy of *P. minimum* cells was achieved. It is worth

mentioning that this accuracy was reached including objects extracted from relatively unfocused images. Thus, the feature set used in this work has shown considerable tolerance to blurring and out-of-focus distortions. The SVM classifier showed superior performance to that of the RF, although these two classifiers agreed on many of the erroneously classified data. Therefore, only insignificant improvement in accuracy was obtained from the committee.

The developed system has been tested by phytoplankton experts in a cell abundance estimation task of *P. minimum* species. New phytoplankton samples were used to accomplish these tests, and a new set 65 phytoplankton images were collected from different areas of the camera view. The automated system detected and correctly recognized 308 (91.1%) of 338 *P. minimum* cells found by experts in the images. The system erroneously assigned to the *P. minimum* class 3% of other objects present in the images. The average counting accuracy achieved by experts, when counting artificial microparticles of 20 μm in diameter similar to *P. minimum* cells, was 88%–90%. Thus, performance achieved by the developed system is comparable or even better than performance of phytoplankton experts exhibited in manual counting of artificial microparticles, similar to *P. minimum* cells. It is worth mentioning that, due to large variations of texture and size of *P. minimum* cells and background as well, the task performed by the system was more complex than that performed by the experts.

How do research results achieved in this work compare to results obtained by other researchers when solving similar tasks? This is not an easy question to answer. The recognition accuracy of cells achieved in this work is much higher than that shown in Table I. However, the works summarized in Table I consider more than two classes. Moreover, due to different resolution of images and different size of the data sets used as well as different procedures taken to assess performance of the techniques, comparison of developed approaches is a rather complicated matter even in the case of the same number and kind of species. A considerable improvement in recognition accuracy of *P. minimum* cells was achieved in this work compared to the results presented in [36]. It is well known in the pattern recognition community that significant recognition accuracy improvements over 95% are rather difficult to achieve.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewer for valuable comments on the manuscript.

REFERENCES

- [1] P. F. Culverhouse, R. Williams, M. Benfield, P. R. Flood, A. F. Sell, M. G. Mazzocchi, I. Buttino, and M. Sieracki, "Automatic image analysis of plankton: Future perspectives," *Mar. Ecol. Progr. Ser.*, vol. 312, pp. 297–309, 2006.
- [2] M. C. Benfield, 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, "RAPID research on automated plankton identification," *Oceanography*, vol. 20, no. 2, pp. 172–187, 2007.
- [3] I. Olenina, S. Hajdu, L. Edler, A. Andersson, N. Wasmund, S. Busch, J. Gobel, S. Gromisz, S. Huseby, M. Huttunen, A. Jaanus, P. Kokkonen, I. Ledaine, and E. Niemkiewicz, "Biovolumes and size-classes of phytoplankton in the Baltic Sea Helsinki Commission (HELCOM), Baltic Marine Environment Protection Commission, 2006.
- [4] Q. Hu and C. Davis, "Automatic plankton image recognition with co-occurrence matrices and support vector machine," *Mar. Ecol. Progr. Ser.*, vol. 295, pp. 21–31, 2005.
- [5] C. S. Davis, Q. Hu, S. M. Gallager, X. Tang, and C. J. Ashjian, "Real-time observation of taxa-specific plankton distributions: An optical sampling method," *Mar. Ecol. Progr. Ser.*, vol. 284, pp. 77–96, 2004.
- [6] X. Tang, W. K. Stewart, L. Vincent, H. Huang, M. Marra, S. M. Gallager, and C. S. Davis, "Image features from phase congruency," *Artif. Intell. Rev.*, vol. 12, pp. 177–199, 1998.
- [7] L. Vincent, "Morphological grayscale reconstruction in image analysis: Application and efficient algorithm," *IEEE Trans. Image Process.*, vol. 2, no. 2, pp. 176–201, Apr. 1993.
- [8] T. Kohonen, "The self-organizing map," *Proc. IEEE*, vol. 78, no. 9, pp. 1461–1480, Sep. 1990.
- [9] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst. Man Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.
- [10] A. Gelzinis, A. Verikas, and M. Bacauskiene, "Increasing the discrimination power of the co-occurrence matrix-based features," *Pattern Recognit.*, vol. 40, no. 9, pp. 2367–2372, 2007.
- [11] V. N. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998, pp. 493–520.
- [12] P. Grosjean, M. Picheral, C. Warembourg, and G. Gorsky, "Enumeration, measurement, and identification of net zooplankton samples using the ZOOSCAN digital imaging system," *ICES J. Mar. Sci.*, vol. 61, pp. 518–525, 2004.
- [13] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.
- [14] A. Verikas, A. Gelzinis, and M. Bacauskiene, "Mining data with random forests: A survey and results of new tests," *Pattern Recognit.*, vol. 44, no. 2, pp. 330–349, 2011.
- [15] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. London, U.K.: Chapman & Hall, 1993, pp. 297–317.
- [16] G. Gorsky, M. D. Ohman, M. Picheral, S. Gasparini, L. Stemmann, J. Romagnan, A. Cawood, S. Pesant, C. Garcia-Comas, and F. Prejger, "Digital zooplankton image analysis using the ZooScan integrated system," *J. Plankton Res.*, vol. 32, no. 3, pp. 285–303, 2010.
- [17] J. L. Bell and R. R. Hopcroft, "Assessment of ZoolImage as a tool for the classification of zooplankton," *J. Plankton Res.*, vol. 30, no. 12, pp. 1351–1367, 2008.
- [18] X. Irigoien, J. A. Fernandes, P. Grosjean, K. Denis, A. Albaina, and M. Santos, "Spring zooplankton distribution in the Bay of Biscay from 1998 to 2006 in relation with anchovy recruitment," *J. Plankton Res.*, vol. 31, no. 1, pp. 1–17, 2009.
- [19] Q. Hu and C. Davis, "Accurate automatic quantification of taxa-specific plankton abundance using dual classification with correction," *Mar. Ecol. Progr. Ser.*, vol. 306, pp. 51–61, 2006.
- [20] J. A. Fernandes, X. Irigoien, G. Boyra, J. A. Lozano, and I. Inza, "Optimizing the number of classes in automated zooplankton classification," *J. Plankton Res.*, vol. 31, no. 1, pp. 19–29, 2009.
- [21] T. Luo, K. Kramer, D. B. Goldgof, L. O. Hall, S. Samson, A. Remsen, and T. Hopkins, "Recognizing plankton images from the shadow image particle profiling evaluation recorder," *IEEE Trans. Syst. Man Cybern. B, Cybern.*, vol. 34, no. 4, pp. 1753–1762, Aug. 2004.
- [22] F. Zhao, F. Lin, and H. S. Seah, "Binary SIPPER plankton image classification using random subspace," *Neurocomputing*, vol. 73, pp. 1853–1860, 2010.
- [23] X. Tang, F. Lin, S. Samson, and A. Remsen, "Binary plankton image classification," *IEEE J. Ocean. Eng.*, vol. 31, no. 3, pp. 728–735, Jul. 2006.
- [24] G. Gorsky, P. Guilbert, and E. Valenta, "The autonomous image analyzer: Enumeration, measurement and identification of marine phytoplankton," *Mar. Ecol. Progr. Ser.*, vol. 58, pp. 133–142, 1989.
- [25] H. M. Sosik and R. J. Olson, "Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry," *Limnol. Oceanogr., Methods*, vol. 5, pp. 204–216, 2007.
- [26] K. V. Embleton, C. E. Gibson, and S. I. Heaney, "Automated counting of phytoplankton by pattern recognition: A comparison with a manual counting method," *J. Plankton Res.*, vol. 25, no. 6, pp. 669–681, 2003.
- [27] M. B. Blaschko, G. Holness, M. A. Mattar, D. Lisin, P. E. Utgoff, A. R. Hanson, H. Schultz, E. M. Riseman, M. E. Sieracki, W. M. Balch, and B. Tupper, "Automatic in situ identification of plankton," in *Proc. 7th IEEE Workshop Appl. Comput. Vis.*, Breckenridge, CO, USA, 2005, pp. 79–86.

- [28] P. F. Culverhouse, R. Williams, B. Reguera, V. Herry, and S. Gonzalez-Gil, "Do experts make mistakes? A comparison of human and machine identification of dinoflagellates," *Mar. Ecol. Progr. Ser.*, vol. 247, pp. 17–25, 2003.
- [29] R. Ellis, R. Simpson, P. F. Culverhouse, and T. Parisini, "Committees, collectives and individuals: Expert visual classification by neural network," *Neural Comput. Appl.*, vol. 5, no. 2, pp. 99–105, 1997.
- [30] K. Rodenacker, B. Hense, U. Tting, and P. Gais, "Automatic analysis of aqueous specimens for phytoplankton structure recognition and population estimation," *Microscopy Res. Tech.*, vol. 69, pp. 708–720, 2006.
- [31] R. J. Olson and H. M. Sosik, "A submersible imaging-in-flow instrument to analyze nanoand microplankton: Imaging FlowCytobot," *Limnol. Oceanogr., Methods*, vol. 5, pp. 195–203, 2007.
- [32] R. V. Azanza, Y. Fukuyob, L. G. Yapa, and H. Takayama, "Prorocentrum minimum bloom and its possible link to a massive fish kill in Bolinao, Pangasinan, Northern Philippines," *Harmful Algae*, vol. 4, pp. 519–524, 2005.
- [33] P. J. Tango, R. Magnien, W. Butler, R. Lacouture, K. Sellner, M. L. P. Glibert, C. Poukish, and C. Luckett, "Characterization of impacts and potential effects due to *Prorocentrum minimum* blooms in Chesapeake Bay," *Harmful Algae*, vol. 4, pp. 525–531, 2005.
- [34] I. Olenina, N. Wasmund, S. Hajdu, I. Jurgensone, S. Gromisz, J. Kownacka, K. Toming, D. Vaiciute, and S. Olenin, "Assessing impacts of invasive phytoplankton: The Baltic Sea case," *Mar. Pollution Bull.*, vol. 60, pp. 1691–1700, 2010.
- [35] P. Kovsi, "Image features from phase congruency," *Videre, J. Comput. Vis. Res.*, vol. 1, no. 3, pp. 1–26, 1999.
- [36] A. Verikas, A. Gelzinis, and M. Bacauskiene, "Phase congruency-based detection of circular objects applied to analysis of phytoplankton images," *Pattern Recognit.*, vol. 45, no. 4, pp. 1659–1670, 2012.
- [37] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.
- [38] M. K. Hu, "Visual pattern recognition by moment invariants," *IRE Trans. Inf. Theory*, vol. IT-8, no. 2, pp. 179–187, Feb. 1962.
- [39] L. Breiman, "RFtools—For predicting and understanding data," Berkeley, CA, USA, Tech. Rep., 2004 [Online]. Available: stat-www.berkeley.edu/users/breiman/RandomForests/cc_home.htm
- [40] A. Solow, C. Davis, and Q. Hu, "Estimating the taxonomic composition of a sample when individuals are classified with error," *Mar. Ecol. Progr. Ser.*, vol. 216, pp. 309–311, 2001.
- [41] D. Yuan, "Natural constraints for inverse area estimate corrections," *Photogram. Eng. Remote Sens.*, vol. 62, no. 4, pp. 413–417, 1996.
- [42] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985, p. 349.
- [43] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, 2011.
- [44] P. Schilling, "Phytoplankton Ring Test 2007: Identification, counting and biomass determination of synthetic and natural phytoplankton samples—Final report," Tech. Rep., 2010.

Antanas Verikas (M'96) received the M.S. degree in electrical engineering and the Ph.D. degree in pattern recognition from Kaunas Polytechnic Institute, Kaunas, Lithuania, in 1974 and 1981, respectively.

He is currently a Professor at both Halmstad University, Halmstad, Sweden and Kaunas University of Technology, Kaunas, Lithuania. His research interests include image processing, pattern recognition, artificial neural networks, fuzzy logic, and visual media technology.

Prof. Verikas is a member of the International Pattern Recognition Society, the European Neural Network Society, the

International Association of Science and Technology for Development, and the Swedish Society of Learning Systems.

Adas Gelzinis received the M.S. degree in electrical engineering and the Ph.D. degree in computer science from Kaunas University of Technology, Kaunas, Lithuania, in 1995 and 2000, respectively.

Currently, he is a Professor in the Department of Electrical and Control Equipment, Kaunas University of Technology. His research interests include artificial neural networks, kernel methods, pattern recognition, signal and image processing, and texture classification.

Marija Bacauskiene received the M.S. degree in applied mathematics from Vilnius University, Vilnius, Lithuania, in 1965.

Currently, she is a Senior Researcher in the Department of Electrical and Control Equipment, Kaunas University of Technology, Kaunas, Lithuania. Her research interests include artificial neural networks, image processing, pattern recognition, and fuzzy logic. She participated in various research projects and published numerous papers in these areas.

Irina Olenina received the M.Sc. degree in biology from Kazan State University, Kazan, Russia, in 1980 and the Ph.D. degree in botany from the Institute of Botany, Lithuanian Academy of Science, Vilnius, Lithuania, in 1997.

Currently, she is a Senior Specialist at the Department of Marine Research, Lithuanian Environmental Protection Agency, Klaipeda, Lithuania, and a Senior Researcher (part-time) at the Marine Science and Technology Center, Klaipeda University, Klaipeda, Lithuania. Her main research interests include phy-

toplankton biodiversity, long-term changes in phytoplankton community, and ecology of invasive phytoplankton species.

Dr. Olenina is a member of the HELCOM Phytoplankton Experts Group.

Evaldas Vaiciukynas received the M.S. degree in informatics engineering from Kaunas University of Technology, Kaunas, Lithuania, in 2006, where he is currently working toward the Ph.D. degree in computer science.

He is a Junior Research Assistant in the Department of Electrical and Control Equipment, Kaunas University of Technology. His research interests include computational intelligence, statistical learning, kernel methods, pattern recognition, and signal processing.