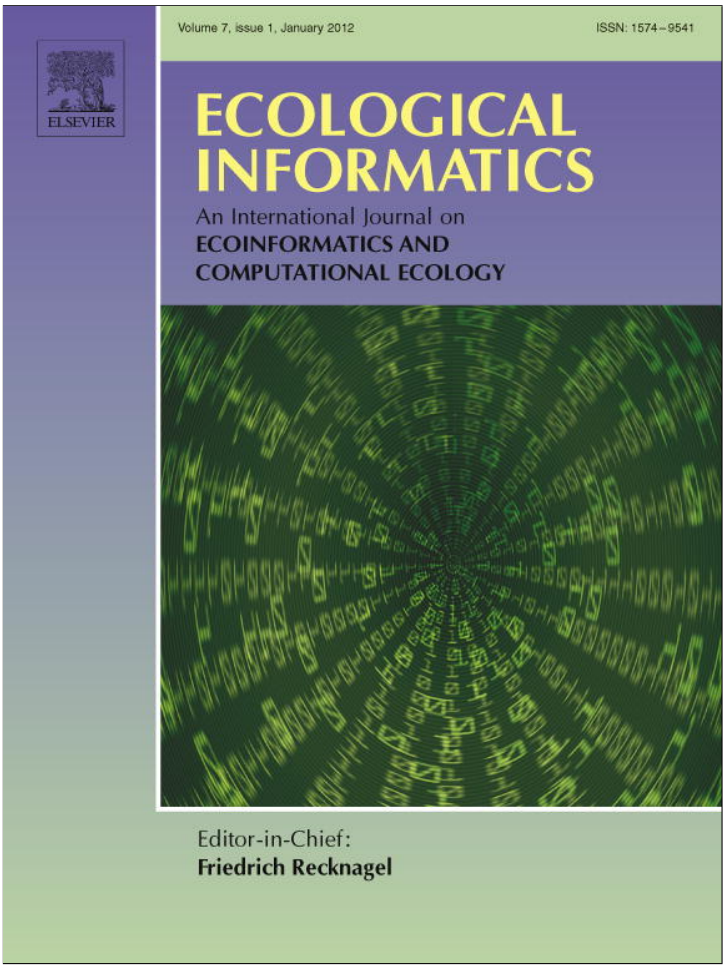

Ecological Informatics 7 (2012) 19–29

Contents lists available at SciVerse ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Hierarchical classification of diatom images using ensembles of predictive clustering trees

Ivica Dimitrovski [a,*], Dragi Kocev [b], Suzana Loskovska [a], Sašo Džeroski [b]

[a] Department of Computer Science and Computer Engineering, Faculty of Electrical Engineering and Information Technologies, Rugjer Boshkovik bb, 1000 Skopje, Republic of Macedonia
[b] Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 5 July 2011
Received in revised form 4 September 2011
Accepted 5 September 2011
Available online 10 September 2011

Keywords:
Diatoms
Automatic image annotation
Hierarchical classification
Predictive clustering trees
Feature extraction from images

ABSTRACT

This paper presents a hierarchical multi-label classification (HMC) system for diatom image classification. HMC is a variant of classification where an instance may belong to multiple classes at the same time and these classes/labels are organized in a hierarchy. Our approach to HMC exploits the classification hierarchy by building a single predictive clustering tree (PCT) that can simultaneously predict all different levels in the hierarchy of taxonomic ranks: genus, species, variety, and form. Hence, PCTs are very efficient: a single classifier is valid for the hierarchical classification scheme as a whole. To improve the predictive performance of the PCTs, we construct ensembles of PCTs. We evaluate our system on the ADIAC database of diatom images. We apply several feature extraction techniques that can be used in the context of diatom images. Moreover, we investigate whether the combination of these techniques increases predictive performance. The results show that ensembles of PCTs have better predictive performance and are more efficient than SVMs. Furthermore, the proposed system outperforms the most widely used approaches for image annotation. Finally, we demonstrate how the system can be used by taxonomists to annotate new diatom images.

1. Introduction

Diatoms are a large and ecologically important group of unicellular or colonial organisms (algae). They are characterized by their highly patterned cell wall composed mainly of hydrated amorphous silica. The cell wall can be divided into two halves. Each half of the cell wall consists of a valve and a number of girdle bands. One half is slightly larger than the other and overlaps it. Together, the halves make a cylinder, with the two valves at the ends. The cross section of the cylinder, and hence the outline of the valve, varies greatly in shape between species and genera. This, together with the pattern of pores and other markings on the valve, provides the information needed for species classification. Fig. 1 depicts three example images of diatoms.

In the variety of uses of diatoms, such as water quality monitoring, paleoecology and forensics, microscope slides must be first scanned for diatoms: if diatoms are present, they need to be classified. Most classifications are done using classification keys and/or comparing specimens using slides, photographs or drawings of diatoms in books and atlases (Stoermer and Smol, 2004). This is not a trivial task, taking into consideration that taxonomists estimate that there may be 200,000 different diatom species, half of them still undiscovered, and many of these extremely hard to distinguish on the basis of morphology (du Buf and Bayer, 2002). Furthermore, this is very tedious and repetitive work, thus any degree of automation can greatly help.

Having this in mind, we propose a system for automatic diatom classification. This system consists of the two standard parts of image annotation systems: image processing (feature extraction from images) and image classification. The image processing part converts an image to a set of numerical features that are extracted directly from the image pixels. The second part, image classification, labels and groups the images. The labels can be organized in a hierarchy and an image can be labeled with more than one label (can belong to more than one group).

For the image processing part, we have implemented two feature extraction techniques that are most commonly used in this context. The first technique produces descriptors (called Fourier descriptors) that contain information concerning the properties of the valve outline. The descriptors from the second technique, called Scale Invariant Feature Transform-SIFT histograms, contain information about the ornamentation of the valve face. We believe that the diatom images can be appropriately described with the combination of these two techniques.

Considering the image classification part, we will use the recently proposed method of building ensembles of PCTs, in particular, bagging and random forests of PCTs (Kocev et al.,2007; Kocev, 2011). We directly compare the predictive performance and the efficiency of the ensembles of PCTs to the one of SVMs trained with Gaussian kernels — the most widely used classifiers used in image annotation.

* Corresponding author. Tel.: +389 2 3099156.
E-mail addresses: ivicad@feit.ukim.edu.mk (I. Dimitrovski), Dragi.Kocev@ijs.si (D. Kocev), suze@feit.ukim.edu.mk (S. Loskovska), Saso.Dzeroski@ijs.si (S. Džeroski).

1574-9541/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecoinf.2011.09.001

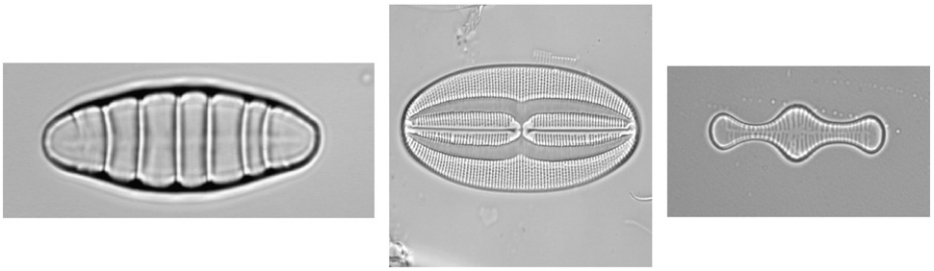

Fig. 1. Example images of diatoms. From left to right: *Diatoma mesodon*, *Fallacia sp.5* and *Tabellaria flocculosa*.

Moreover, we contrast the predictive performance of the proposed approach to the best reported results (du Buf and Bayer, 2002) on the used database of diatom images (ADIAC, 2011).

The goal of the complete system is to assist a taxonomist in identifying a wide range of different diatoms. To this end, we develop a web-based interface to the proposed system than can help taxonomists with the identification of diatom taxa in images. The user (i.e., taxonomist) loads an image to the system by using the interface. After that, the system recommends an annotation for the image, accompanied with a probability value for the prediction. The taxonomist can then select the proposed annotation and browse through the images from the same species to further check the validity of the annotation.

The remainder of the paper is organized as follows. In Section 2, we present the related work. Section 3 describes the system for annotation of diatom images, namely, image segmentation, techniques we use for feature extraction from images, and predictive clustering trees and their use for HMC. In Section 4, we explain the experimental setup. The obtained results and a discussion thereof are given in Section 5. Section 6 concludes the paper and points out some directions for further work.

2. Background and related work

The process of automatic diatom classification consists of three phases (du Buf and Bayer, 2002): image segmentation, feature extraction and image classification. The goal of image segmentation is to locate and obtain the contour of the diatom. Then, using these segmented images and extracted contours, the feature extraction algorithms generate image descriptors. At the end, machine learning algorithms are used to train a classifier that will perform the classification for previously unseen diatom images (e.g., provide the taxonomic rank). Here, we shortly describe each phase and the algorithms which are usually used in each of them.

2.1. Image segmentation

An ideal diatom image depicts only a single diatom shell. However, in reality, diatoms may lay on top of each other or very close to each other, the image may not be in proper focus, dust specks and background texture may be visible in some images etc. Fig. 3 (first row) shows diatom images that contain some of the aforementioned anomalies. Because of this, one needs to perform image segmentation before extracting features from the images.

The problem of image segmentation, i.e., contour extraction, of gray-scale diatom images can be solved mainly by applying four methods: threshold-based, boundary-based, region-based and hybrid methods (Jalba et al., 2004). *Threshold methods* assume that all pixels with gray-level values within a certain range belong to one class. They do not use any spatial information of the image, are sensitive to noise, and do not cope well with blurred edges. The *boundary-based methods* are local filtering techniques, such as edge detectors or active contour methods. Edge detectors usually cannot ensure continuous edge-detection and an edge-linking step must be used to produce closed contours. In contrast, active contour methods automatically produce closed contours and usually provide better edge localization, but are sensitive to noise and require an initialization step that is hard to automate. *Region-based methods* assume that neighboring pixels within the same region have similar values. Their main advantage is that they use and adapt the statistics inside the region, but they generate small holes and irregular boundaries. *Hybrid techniques* combine both boundary and region criteria. All in all, there is a variety of approaches that one can choose for the problem at hand. In our system, we are using marker-controlled watershed segmentation which has already been successfully applied for diatom image segmentation (Jalba et al., 2004).

2.2. Feature extraction

Once the segmentation and contour extraction are completed, different feature extraction techniques can be employed on the diatom images (Westenberg and Roerdink, 2002). The diatoms can be primarily distinguished by evaluating properties of the valve's outline. The contour features measure the symmetry, global and local shape characteristics, as well as geometric properties, such as length and width of the diatoms (Ciobanu and du Buf, 2002; Fischer and Bunke, 2002; Loke and du Buf, 2002).

An important characteristic of diatoms is also the ornamentation of the valve face, which is a specific type of texture (Wilkinson et al., 2002). There are several known visual descriptors able to measure these texture properties: features derived from gray level co-occurrence matrices, Gabor wavelets (Santos and du Buf, 2002), scale invariant feature transform (SIFT) (Lowe, 2004) and local binary patterns (LBP) (Ojala et al., 2002). To summarize, these features capture several aspects of an image. Depending on the application, one can choose to use some specific feature extraction technique or to combine several of them into a single, more complex set of features.

2.3. Image classification

The last phase of an automatic classification system is classification. In this phase, a machine learning algorithm is first employed to construct a classifier using the features extracted in the previous two stages and the annotations/labels of the images (taxonomic ranks). Then, the obtained classifier maps the images of unidentified specimens to annotations from the set observed during training, i.e., provides annotations for previously unseen images. In the context of diatom image classification, the most typically used classifiers are neural networks, naïve Bayes, support vector machines (SVMs) and decision trees.

Santos and Du Buf (Santos and du Buf, 2002) use a fully-connected neural network classifier with one hidden layer. The number of input units equals the number of features. The hidden layer has an equal number of units as the input layer, and the output layer has as many units as there are classes. The neural network is trained until the error rate on a validation set reaches a local minimum.

The naïve Bayes classifier estimates the probability density function of the features for each class (Zhang, 2004). It classifies an unseen image by first computing the conditional probabilities for each class, given the image's feature vector.Then, it assigns the image to the class with the highest probability.

The SVMs are most widely used classifiers for image annotation in general, current state-of-the-art results in image annotation are obtained using per-label-trained SVM classifiers (Mensink et al., 2010). There are also several studies concerning automated taxonomic classification that use SVMs as classifiers (MacLeod, 2008; Morris et al., 2001; Sosik and Olson, 2007).

Most state-of-the-art results in automatic diatom classification are achieved using decision trees and bagging thereof as classifiers (Fischer and Bunke, 2002). The decision trees do not make prior assumptions for the probability distribution of the dependent and the independent variables, they can use discrete and/or continuous independent variables, can handle missing values and the learning process is not influenced by redundant variables and noise. Furthermore, they are not computationally expensive and are easily interpretable. When the trees are combined into an ensemble, then very high predictive performance can be achieved (Breiman, 1996).

The aforementioned classification approaches however do not use the semantic knowledge about the inter-class relationships among the classes. The classes can be organized into different levels in the hierarchy of taxonomic ranks: genus, species, variety, and form. To this end, we propose to use the predictive clustering trees (PCTs) as classifiers. PCTs can exploit the hierarchical taxonomy and simultaneously predict all taxonomic ranks (Vens et al., 2008). This approach yields a very efficient classifier that offers high predictive performance.

3. System for automatic diatom classification

The architecture of the proposed system for annotation of diatom images is outlined in Fig. 2. As mentioned earlier, the system is composed of image segmentation part, feature extraction part and annotation part. Fig. 2 further shows the specific methods used as components of the proposed system. In this section, we further describe each part in more detail.

3.1. Image segmentation

The first step in automatic annotation of diatom images is applying image segmentation. The image segmentation concerns location of relevant objects, i.e., its goal is to locate the diatoms in the input microscopic images. The complete procedure for image segmentation is shown in Fig. 3. The procedure starts by applying watershed transform in order to separate the touching objects. The watershed transform finds catchment basins and watershed ridge lines in an image by treating it as a surface where light pixels are high and dark pixels are low (third row in Fig. 3).

Segmentation using the watershed transform works better if the foreground objects and the background locations are automatically identified at the beginning (second row in Fig. 3). This type of watershed segmentation is called *marker-controlled watershed segmentation*. To find the foreground markers, which must be connected blobs of pixels inside each of the foreground objects, a variety of procedures can be applied. In our system, we are using a morphological technique called *opening-by-reconstruction* to clean up the image. Opening-by-reconstruction is an erosion followed by a morphological reconstruction. This operation will create flat maxima inside each object that can be easily located and selected as foreground markers. This procedure tends to leave some isolated pixels that must be removed and not considered as marker regions. In our system, the size of the structuring element for the erosion was set to 3×3, and the threshold on area, for small region removal, was set to 200 pixels. The background markers were obtained with a thresholding operation on the resulting image.

The contours produced by the marker-controlled watershed segmentation method (fourth row in Fig. 3) are traced using a standard *contour following algorithm*. This procedure traverses the whole image pixel-by-pixel starting at the top-left corner and proceeding from left-to-right and top-to-bottom. It searches for a pixel from the 3×3 neighborhood with the same value as the current pixel. The search is continued until the starting point of the contour is reached again and it labels the encountered pixels as 'visited' (fifth row in Fig. 3).

All extracted contours are then filled at gray-level zero by a *flood-fill algorithm*, and all obtained regions are drawn in the same image. In a further post-processing step, an opening with a structuring element of size 3×3 is used to prune the thin structures, which may still be connected to diatom regions, due to debris or fragments of other diatoms. In this way, the union of all diatom and inner-diatom regions is created and all diatom contours can be found by tracing only one contour per region. The regions obtained after flood-filling the traced contours are shown in the sixth row in Fig. 3. Notice that the surviving inner-diatom regions, which were not removed by the marker selection procedure, are now merged into one large diatom region. In our system, contours which enclose regions of areas smaller than 4900 pixels are not considered as diatom regions and are rejected, i.e. this procedure rejects the contours that have fewer pixels than the minimum diatom size.

3.2. Feature extraction

We apply two feature extraction techniques on the regions identified by the image segmentation procedure. Given the specific problem of annotating microscopic diatom images, we apply two techniques

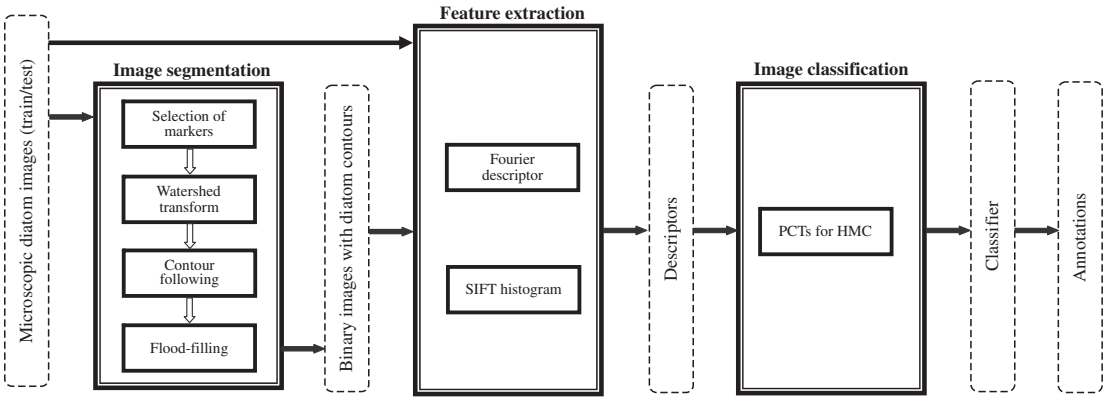

Fig. 2. Architecture of the proposed system for classification/identification/annotation of diatom images.

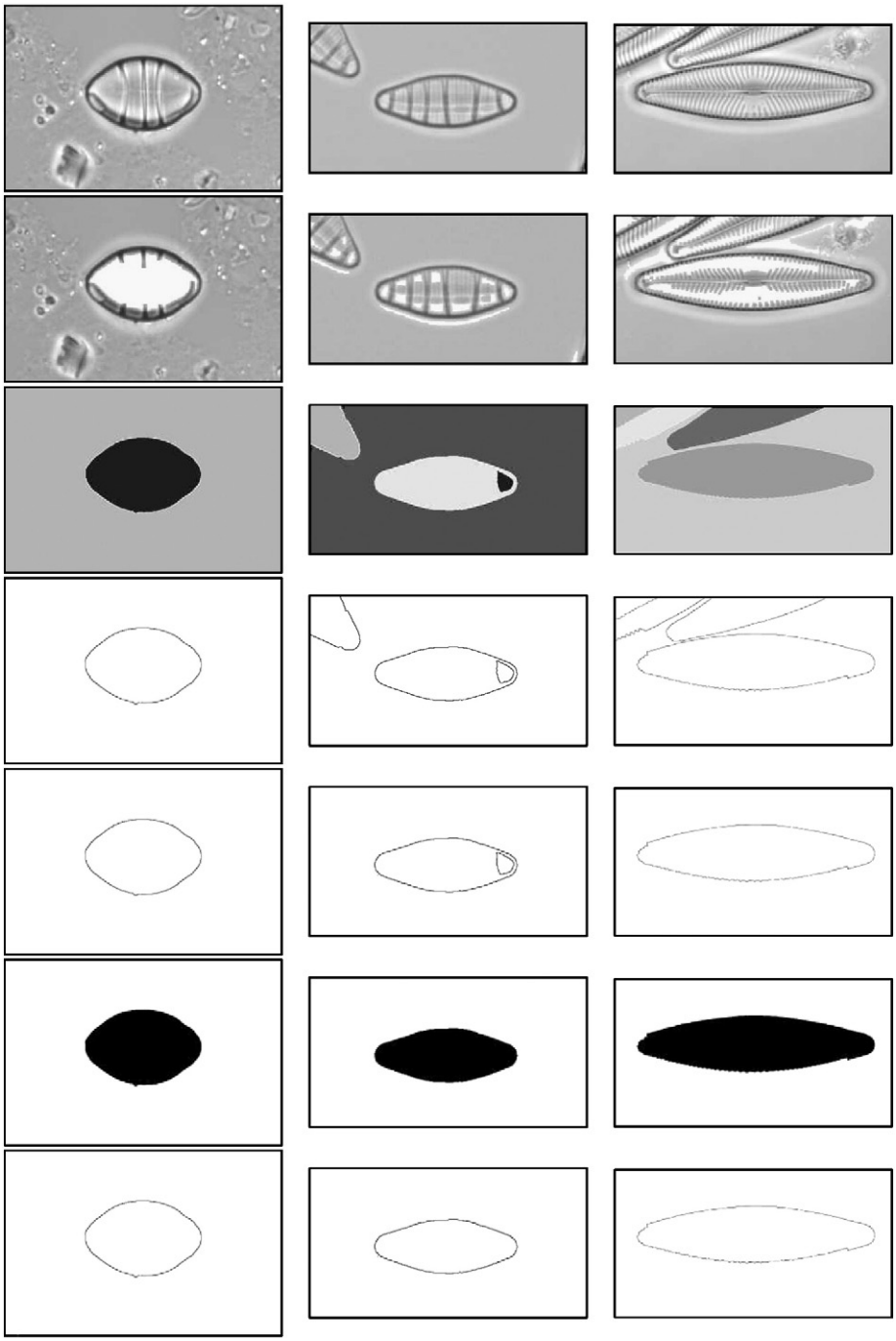

Fig. 3. Illustration of the diatom segmentation procedure. First row: example images, from left to right: *Diatoma mesodon*, *Denticula tenuis* and *Navicula lanceolata*. Second row: images with markers. Third row: images with watershed regions. Fourth row: images with watershed lines. Fifth row: images with contours after applying contour following. Sixth row: regions obtained after flood-filling the contours. Seventh row: images with final diatom contours.

that have been used in this context: Fourier descriptors and histograms of local SIFT descriptors. The Fourier descriptors characterize the shape of the diatom contour, while SIFT histograms describe the texture of the diatom. In the following, we briefly describe these two techniques.

3.2.1. Fourier descriptors

The Fourier descriptors view a closed curve (e.g., a diatom contour) as a periodic function and represent it by a set of Fourier coefficients/descriptors. They are calculated by approximating the best-fitting ellipse over the examined shape and are obtained through a Fourier transform using the coordinates of the shape boundary. The magnitudes of the obtained Fourier coefficients are normalized by the magnitude of the first coefficient. The Fourier coefficients are invariant to translation, rotation and scaling.

The high frequency noise can be significantly reduced by limiting the number of coefficients c (the effect of low pass filtering). At the same time, this will preserve the main patterns in the contour. On the other hand, this reduction can lead to the loss of spatial information in terms of fine detail. Following the recommendations by Fischer and Bunke (Fischer and Bunke, 2002), we consider 30 coefficients as sufficient to distinguish between most shapes.

3.2.2. SIFT histograms

An important property of the diatoms is the ornamentation of the valve face, which is a specific type of texture. This means that descriptors for local regions of an image can provide significant information for distinguishing and discrimination of the images. To this end, we use the Scale Invariant Feature Transform (SIFT), which is reasonably invariant to changes in illumination, image noise, rotation, scaling, and small changes in viewpoint (Lowe, 2004).

The SIFT method extracts and describes local key-points. The number of descriptors obtained with SIFT is large because an image may contain many key-points and each key-point is described by 128 numerical values. Using histograms of local features, the amount of data is reduced by estimating the distribution of local feature values for every image.

The creation of these histograms is a three step procedure. First, key-points detected inside the diatom contour are extracted from all database images using the default parameters proposed by Lowe (2004). Then, the key-points are clustered into 200 clusters using *k*-means. For each key-point, all information is discarded except the identifier of the most similar cluster center. We then create for each image a histogram of the occurring patch-cluster identifiers. To be independent of the total number of key-points in an image, the histogram bins are normalized to sum up to 1. This results in a 200 dimensional histogram for each image.

3.3. Ensembles of PCTs for HMC

The descriptors that were obtained using the procedures described above, combined together with the annotations of the images, are used to train a classifier. The annotations of the images can be unstructured or structured. In the first case, the annotations are a simple vector of binary variables meaning that an image is or is not labeled with a given label. In the second case, the labels can be organized in some kind of taxonomy (e.g., hierarchy or directed acyclic graph). The problem of annotation of microscopic diatom images belongs to the second case, since the diatoms can be described by their taxonomic rank. So, we use classifiers that are able to exploit the information about the structure of the annotations, namely, we use predictive clustering trees (PCTs) (Blockeel et al., 1998) for hierarchical multi-label classification (HMC) (Vens et al., 2008). Moreover, to increase their predictive performance, we use ensemble methods, such as bagging and random forests. In the following, we first define the task of hierarchical multi-label classification. We then present the approaches to learning PCTs for HMC and ensembles of PCTs for HMC.

3.3.1. The task of HMC

Hierarchical multi-label classification is a variant of classification where (1) a single example may belong to multiple classes at the same time and (2) the possible classes are organized in a hierarchy. An example that belongs to some class *c* automatically belongs to all super-classes of *c*: This is called the hierarchical constraint. Problems of this kind can be found in many domains including text classification, functional genomics, and object/scene classification. For a more detailed overview of the possible application areas we refer the reader to Silla and Freitas (Silla and Freitas, 2011).

In diatom classification, the diatom species are separated/classified using a logical system of categories with a hierarchical structure (taxonomic rank). Fig. 4 depicts such a system where the top level category is *genus*. Within each *genus*, there are many *species* that are further divided into *subspecies*, *varieties*, *forms*, *morphotypes*, etc.

The data, as presented in the table in the left-hand side of Fig. 4, constitute a training data set for HMC. Each image is represented with: (1) a set of descriptors (in this example, the descriptors are the first five Fourier coefficients) and (2) a set of classes/annotations. A single image can belong to multiple classes at different levels of the predefined hierarchy of taxonomic ranks. For example, the image in the second row of the table in Fig. 4 belongs to the class *minutissimum*. This image also belongs to the classes *Olivaceum* and *Gomphonema*. The testing set of images contains only the set of descriptors and has no a priori annotations.

3.3.2. PCTs for hierarchical-multi label classification

In the PCT framework (Blockeel et al., 1998), a tree is viewed as a hierarchy of clusters: the top-node corresponds to one cluster containing all data, which is recursively partitioned into smaller clusters while moving down the tree. Note that the hierarchical structure of the PCT does not necessarily reflect the hierarchical structure of the annotations (Fig. 4). PCTs are constructed with a standard "top-down induction of decision trees" (TDIDT) algorithm. The heuristic for selecting the tests is the reduction in variance caused by partitioning the instances. Maximizing the variance reduction maximizes cluster homogeneity and improves predictive performance. A leaf of a PCT predicts the prototype of the set of examples belonging to it. A detailed description of the PCT framework can be found in (Blockeel et al., 1998). The PCT framework is implemented in the CLUS system, which is available at http://sourceforge.net/projects/clus/.

To apply PCTs to the task of HMC, the variance of a set of examples (S) is defined as the average squared distance between each example's label v_i and the mean label $\bar{v}$ of the set ($Var(S) = \sum_i d(v_i, \bar{v})^2/|S|$). Considering that an error at the upper levels of the hierarchy costs more than an error at the lower levels, a weighted Euclidean distance is used: $d(v_1, v_2) = \sqrt{\sum_i w(c_i)(v_{1,i} - v_{2,i})^2}$, where $v_{k,i}$ is the ith component of the class vector v_k of an instance x_k, and the class weights $w(c_i)$. The class weights decrease with the depth of the class in the hierarchy, $w(c_i) = w_0 \cdot w(c_j)$, where c_j is the parent of c_i.

Each leaf in the tree stores the mean $\bar{v}$ of the vectors of the examples that are sorted in that leaf. Each component of $\bar{v}$ is the proportion of examples $\bar{v}_i$ in the leaf that belong to class c_i. An example

image	features/descriptors						taxonomy
	Fourier coefficients						
	48	24	59	66	37	…	olivaceum
	36	25	53	45	15	…	minutissimum
	35	25	56	52	19		exigua
…	…	…	…	…	…	…	…

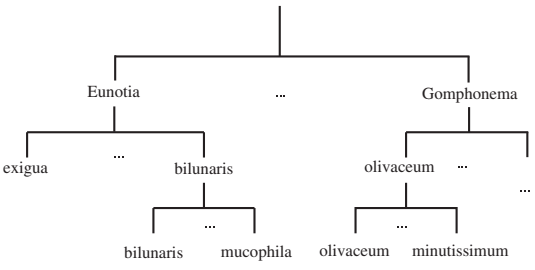

Fig. 4. An example task of HMC in diatom image classification. The table (on the left-hand side) contains a set of images with their visual descriptors and annotations. The annotations are part of the taxonomic rank with hierarchical structure (of which a small part is shown on right hand side).

arriving in the leaf can be predicted to belong to class c_i if $\bar{v}_i$ is above some threshold t_i. The threshold can be chosen by a domain expert. For a detailed description of PCTs for HMC, we refer the reader to Vens et al. (2008). Next, we explain how PCTs are used in the context of an ensemble classifier, in order to further improve the performance of PCTs.

3.3.3. Ensemble methods

An ensemble classifier is a set of (base) classifiers. A new example is classified by the ensemble by combining the predictions of the member classifiers. The predictions can be combined by taking the average (for regression tasks), the majority vote (for classification tasks) (Breiman, 1996; Breiman, 2001), or more complex combinations.

We use PCTs for HMC as base classifiers. Average is applied to combine the predictions of the different trees: the leaf's prototype is the proportion of examples of different classes that belong to it. Just like for the base classifiers, a threshold should be specified to make a prediction. However, each image from the database contains a single diatom. Thus, we select as prototype the species with the highest proportion of examples averaged across all base classifiers.

We consider two ensemble learning techniques that have primarily been used in the context of decision trees: bagging and random forests. It was previously shown that in the domain of functional genomics (Schietgat et al., 2010) and annotation ofmedical X-ray images (Dimitrovski et al., 2011), both random forests and bagging of PCTs, outperform a single PCT.

Breiman (1996) constructs the different classifiers by making bootstrap replicates of the training set and using each of these replicates to construct one classifier. Each bootstrap sample is obtained by randomly sampling training instances, with replacement, from the original training set, until a number of instances are obtained equal to the size of the training set. Bagging is applicable to any type of learning algorithm.

A random forest (Breiman, 2001) is an ensemble of trees, obtained both by bootstrap sampling, and by randomly changing the feature set during learning. More precisely, at each node in the decision tree, a random subset of the input attributes is taken, and the best feature is selected from this subset (instead of the set of all attributes). The number of attributes that are retained is given by a function f of the total number of input attributes x (e.g., $f(x)=x$, $f(x)=\sqrt{x}$, $f(x)=\lfloor log_2 x \rfloor+1$, …). By setting $f(x)=x$, we obtain the bagging procedure.

4. Experimental design

In this section, we present the experimental setup used to evaluate the proposed system and compare it to other state-of-the-art approaches for image annotation. First, we state the experimental questions that we investigate in this study. Next, we present the databases of images that we use and the design of the experiments. We then specify the parameter instantiations for the algorithms and describe the evaluation measures used to assess the predictive performance of the classifiers.

4.1. Experimental questions

The goal of this study is to answer the following questions:

1. Does the use of the hierarchy (in ensembles of PCTs) improve the predictive performance over flat classification (SVMs) for the task of diatom image annotation?
2. Is the proposed system with ensembles of PCTs for HMC scalable and efficient?

For the first question, we compare the performance of the ensemble classifiers (bagging and random forests) of PCTs for HMC with SVMs for flat classification—the most widely used classifiers for image annotation. Since the most elaborate work so far on the problem of diatom identification is presented by Du Buf and Bayer (du Buf and Bayer, 2002), we compare our results to the ones presented there.

Considering the second question, we compare the execution times of the different classifiers to assess the efficiency and scalability of the system. We measure the time needed to train the classifiers (for SVMs this includes also the time needed to optimize the parameters), and the time needed to obtain annotation (taxonomic rank) of new and unseen images.

4.2. Diatom image database

The presented system for hierarchical multi-label classification of diatom images was evaluated on the ADIAC diatom image database (ADIAC, 2011). In our experiments, we used a subset of 1098 microscopic images that are classified using the taxonomic rank of the diatoms. The diatoms from the images belong to 55 different taxa. For each taxon, there are at least 10 images available, up to a maximum of 29 images (the number of images per diatom taxa is shown in Table 4). The diatoms in this set vary in shape but also in ornamentation (three examples are shown in Fig. 1, and the first row in Fig. 3).

An extensive analysis of this database was performed in (du Buf and Bayer, 2002), where two other versions of this dataset were used. The first one consists of 38 taxa, for which at least 20 images per taxa are available (837 images in total). The second one consists of 48 taxa, for which at least 15 images per taxa are available (1019 images in total). For comparability, we conduct experiments on the complete database as well as on the two additional variants of the database.

4.3. Experimental setup

We evaluate the proposed system along three different evaluation measures: overall recognition rate, precision and recall. The overall recognition rate is the fraction of the diatom images where the complete taxonomic rank was predicted correctly. The precision and recall were measured for each part of the taxonomic rank (for genus, species, variety and form). Precision measures the proportion of diatom images belonging to a given genus (or other part of the taxonomic rank) that were correctly labeled by the classifier ($Precision = \frac{TruePositives}{TruePositives+FalsePositives}$). Recall, on the other hand, measures the proportion of diatom images labeled by the classifier with a given genus (or other part of the taxonomic rank) that truly belong to the given genus ($Recall = \frac{TruePositives}{TruePositives+FalseNegatives}$). Precision can be seen as a measure of exactness, while recall as a measure of completeness. These three evaluation measures are widely used by the image annotation community. We estimate the predictive performance of the classifiers on unseen cases, for each of the three measures, by using 10-fold cross validation.

For training the SVMs, we used a custom developed application. This application uses the Lib SVM library (Chang and Lin, 2001). We apply the One-against-All (OvA) approach to solve the partial binary classification problems. Each of the SVMs was trained with a Gaussian kernel. We optimize the values of the kernel parameter σ and cost parameter C of the SVMs by using an automated parameter search procedure.

The algorithm for learning PCTs requires as input the weight of the depth in the hierarchy (the weight used in the Euclidean distance). We set w_0 to 0.75 (as recommended in (Vens et al., 2008)) thus penalizing the algorithm more if it makes an error in the upper levels of the hierarchy than if it makes an error in the lower levels of the hierarchy. We constructed ensembles of 100 un-pruned trees (PCTs) (Bauer and Kohavi, 1999; Kocev et al., 2007). The size of the feature subset that is retained at each node, when training a random forest,

Table 1
Predictive performance of the feature extraction algorithms and their combination evaluated using overall recognition rate (boldface indicates the best performance).

Classifier	Descriptors	# features	Overall recognition rate		
			55 diatom taxa	48 diaton taxa	38 diatom taxa
Bagging	Fourier descriptors	30	87.61	88.42	89.00
	SIFT histograms	200	87.89	88.91	90.92
	Fourier desc. + hist.	230	**95.45**	**96.27**	**97.49**
Random forests	Fourier descriptors	30	87.61	88.52	89.25
	SIFT histograms	200	89.44	90.87	91.64
	Fourier desc. + SIFT hist.	230	**96.17**	**97.15**	**97.97**
SVM	Fourier descriptors	30	83.97	86.75	87.93
	SIFT histograms	200	85.25	88.32	89.96
	Fourier desc. + IFT hist.	230	**92.35**	**94.80**	**96.54**

Table 2
Time needed to construct the classifier.

Classifier	Descriptors	# features	Total training time [sec]		
			55 diatom taxa	48 diatom taxa	38 diatom taxa
Bagging	Fourier descriptors	30	79.900	67.427	42.272
	SIFT histograms	200	633.896	513.962	284.269
	SIFT desc. + SIFT hist.	230	701.230	589.346	340.123
Random forests	Fourier descriptors	30	12.316	9.680	6.448
	SIFT histograms	200	60.730	49.037	30.927
	SIFT desc. + SIFT hist.	230	74.890	53.340	37.510
SVM	Fourier descriptors	30	20.404	12.920	8.309
	SIFT histograms	200	114.267	86.133	55.353
	SIFT desc. + SIFT hist.	230	140.900	98.078	65.780

was set to 10% of the number of descriptive attributes. Remember that the output of the classifier is a probability that a given example is annotated with a given label. If the probability is higher than a given threshold (obtained during the training of the classifier), then the example is annotated with the given label.

5. Results and discussion

In this section, we present the results from the experimental evaluation of the proposed system for automatic diatom identification. We first compare the performance of the ensembles of PCTs for HMC and the SVMs. The performance is assessed in terms of predictive power through the overall recognition rate (Table 1) and in terms of efficiency through the training (Table 2) and testing times. Next, we compare the aforementioned approaches to the results presented in an extensive study (du Buf and Bayer, 2002) concerning the ADIAC diatom database. We then focus the analysis on the precision and recall values for each diatom species obtained with random forests of PCTs for HMC. Finally, we present a web-based interface that uses the results of the annotation system.

5.1. Performance of the annotation system

Table 1 summarizes the performance of the three machine learning algorithms (SVMs, random forests and bagging of PCTs for HMC) in terms of the overall recognition rate. The predictive performance is compared on the three variants of the image database. Overall, random forests of PCTs for HMC perform best, followed by bagging of PCTs for HMC and SVMs. The random forests of PCTs for HMC have overall recognition rates of 96.17%, 97.15% and 97.97%, respectively, for the three variants of the image database. We can thus note that the random forest method is better than bagging and SVMs over the three variants of the database and the three types of descriptors.

We further analyze the results for the individual feature extraction algorithms. The SIFT histograms have better predictive performance than the Fourier descriptors over all three different machine learning algorithms and all three databases. The differences are small for bagging and more pronounced for random forests and SVMs. Next, the results show that combining the two feature sets improves the predictive performance of the machine learning algorithms. This comes as a result of the better representation of the visual content in the images and the orthogonal information offered to the classifiers. This implies that no single set of features allows to discriminate completely between the different taxa. Two types of descriptors are needed to successfully annotate diatom images: contour and texture descriptors. The contour descriptors contain information concerning the properties of the valve outline, while the texture descriptors concern the ornamentation of the valve face. To sum up, the best annotation results are obtained by training random forests of PCTs for HMC using the combination of the two feature sets.

We also assess the efficiency of the algorithms by measuring the time needed to learn the classifier (Table 2) and the time needed to produce a taxonomic rank for a new and unseen image. Considering the time needed to learn the classifier, random forests are the fastest

Table 3
Comparison of the performance of the random forests of PCTs for HMC (given in italic typeface) to the performance of the approaches from Du Buf and Meyer (du Buf and Bayer, 2002). For each approach, we present the number of images, number of different taxa, used feature extraction techniques and classifiers, the approach to evaluating the performance, and the reported overall recognition rate.

Data		Descriptors	Classifier	Evaluation	Recognition rate [%]
# images	# taxa				
1098	*55*	*Fourier; SIFT*	*Random forest of predictive clustering trees*	*10-fold cross-validation*	*96.17*
1019	*48*	*Fourier; SIFT*	*Random forest of predictive clustering trees*	*10-fold cross-validation*	*97.15*
1009	48	Contour profiling; Legendre polynomials	Decision trees; Neural networks; syntactical classifier	Random separation (50/50) to train and test set	82.00
808	38	Geometric; shape; Fourier; image moments; ornamentation and morphological	Bagging of decision trees	Leave One Out	94.90
837	*38*	*Fourier; SIFT*	*Random forest of predictive clustering trees*	*10-fold cross-validation*	*97.97*
781	37	Contour; segment; global	Nearest-mean classifier	Set swaping (complex pseudo cross-validation)	82.90
781	37	Gabor; Legendre polynomials; ornamentation	Decision trees; Bayesian classifier	Random separation (50/50) to train and test set	88.00
781	37	Contour; ornamentation	Bagging of decision trees	10 times random separation (75/25) train and test	89.60
781	37	Gabor; Legendre polynomials; ornamentation; contour; global; geometric; shape; Fourier; image moments; morphological	Bagging of decision trees	10 times random separation (75/25) train and test	96.90

Table 4
The recognition rate per taxa obtained with the combined feature sets and the approach of learning random forests of PCTs for HMC. The *N/A* value for the precision and recall is used for the species and genera that were not present in the respective database.

Taxon	#images	55 diatom taxa		48 diatom taxa		38 diatom taxa	
		Precision	Recall	Precision	Recall	Precision	Recall
Achnanthes	22	0.79	0.92	N/A	N/A	N/A	N/A
Achnanthes/minutissima	10	0.83	0.71	N/A	N/A	N/A	N/A
Achnanthes/oblongella	12	0.67	1	N/A	N/A	N/A	N/A
Caloneis	18	1	1	1	1	N/A	N/A
Caloneis/amphisbaena	18	1	1	1	1	N/A	N/A
Cocconeis	62	1	1	1	1	1	1
Cocconeis/placentula	19	1	1	1	1	N/A	N/A
Cocconeis/neodiminuta	20	1	1	1	1	1	1
Cocconeis/stauroneiformis	23	1	1	1	1	1	1
Cymbella	67	0.97	1	0.97	1	0.99	1
Cymbella/helvetica	26	1	1	1	1	1	1
Cymbella/hybrida	20	1	1	1	1	1	1
Cymbella/subequalis	21	0.9	1	0.89	1	0.94	1
Diatoma	46	1	1	1	1	1	1
Denticula	22	1	1	1	1	1	1
Denticula/tenuis	22	1	1	1	1	1	1
Diatoma/mesodon	26	1	1	1	1	1	1
Diatoma/moniliformis	20	1	1	1	1	1	1
Encyonema	35	1	1	1	1	1	1
Encyonema/neogracile	10	1	1	N/A	N/A	N/A	N/A
Encyonema/silesiacum	25	1	1	1	1	1	1
Epithemia	19	1	1	1	1	N/A	N/A
Epithemia/sorex	19	1	1	1	1	N/A	N/A
Eunotia	75	1	0.99	0.98	1	1	0.98
Eunotia/bilunaris	12	0.8	1	N/A	N/A	N/A	N/A
Eunotia/denticulata	22	1	1	1	1	1	1
Eunotia/incisa	20	1	1	0.95	0.95	1	1
Eunotia/tenella	21	1	0.88	0.96	1	1	0.96
Fallacia	43	1	1	1	1	1	1
Fallacia/forcipata	26	1	1	1	1	1	1
Fallacia/sp.5	17	1	1	1	1	N/A	N/A
Fragilariforma	20	1	1	1	1	1	1
Fragilariforma/bicapitata	20	1	1	1	1	1	1
Gomphonema	64	1	0.95	0.98	0.97	0.99	0.99
Gomphonema/augur	20	0.91	0.88	1	0.95	0.95	0.95
Gomphonema/minutum	24	1	1	1	1	1	1
Gomphonema/sp.1	20	0.94	0.84	0.95	0.95	1	1
Gyrosigma	20	1	1	1	1	1	1
Gyrosigma/acuminatum	20	1	1	1	1	1	1
Meridion	20	1	1	1	1	1	1
Meridion/circulare	20	1	1	1	1	1	1
Navicula	210	0.96	0.99	0.95	0.99	0.97	0.99
Navicula/capitata	20	1	0.86	0.95	0.86	1	0.9
Navicula/constans	22	1	1	1	1	1	1
Navicula/gregaria	11	0.88	0.78	N/A	N/A	N/A	N/A
Navicula/lanceolata	27	1	1	1	1	1	0.97
Navicula/menisculus	18	1	1	1	1	N/A	N/A
Navicula/radiosa	21	1	1	1	1	1	1
Navicula/reinhardtii	29	1	1	1	1	1	1
Navicula/rhynchocephala	19	1	1	1	1	N/A	N/A
Navicula/viridula	19	1	1	1	1	N/A	N/A
Nitzschia	87	1	0.95	1	0.95	1	0.99
Nitzschia/dissipata	20	1	0.9	1	0.83	1	0.95
Nitzschia/hantzschiana	20	0.56	0.83	0.68	0.93	0.74	0.78
Nitzschia/sinuata	20	1	0.89	1	1	1	1
Nitzschia/sp.2	27	0.93	0.79	0.96	0.79	0.86	0.83
Opephora	20	0.84	0.89	0.95	0.95	1	1
Opephora/olsenii	20	0.84	0.89	0.95	0.95	1	1
Parlibellus	20	1	0.95	1	0.96	1	1
Parlibellus/delognei	20	1	0.95	1	0.96	1	1
Petroneis	20	1	1	1	1	1	1
Petroneis/humerosa	20	1	1	1	1	1	1
Pinnularia	45	0.98	1	1	1	1	1
Pinnularia/kuetzingii	21	1	1	1	1	1	1
Pinnularia/silvatica	10	0.78	1	N/A	N/A	N/A	N/A
Pinnularia/subcapitata	14	0.92	0.86	N/A	N/A	N/A	N/A
Sellaphora	18	1	1	1	1	N/A	N/A
Sellaphora/bacillum	18	1	1	1	1	N/A	N/A
Stauroneis	19	0.86	0.9	0.94	0.8	N/A	N/A
Stauroneis/smithii	19	0.86	0.9	0.94	0.8	N/A	N/A
Staurosirella	16	0.94	0.88	0.82	0.93	N/A	N/A
Staurosirella/pinnata	16	0.94	0.88	0.82	0.93	N/A	N/A
Surirella	26	1	0.91	1	0.88	1	0.93

Table 4 (*continued*)

Taxon	#images	55 diatom taxa		48 diatom taxa		38 diatom taxa	
		Precision	Recall	Precision	Recall	Precision	Recall
Surirella/brebissonii	26	1	0.91	1	0.88	1	0.93
Tabellaria	43	1	1	1	1	1	1
Tabellaria/flocculosa	20	1	1	1	1	1	1
Tabellaria/quadriseptata	23	1	1	1	1	1	1
Tabularia	41	1	1	1	1	1	1
Tabularia/investiens	21	1	1	1	1	1	1
Tabularia/sp.1	20	1	1	1	1	1	1

method: they are ~8.84 times faster than bagging and ~1.69 times faster than the SVMs (including also the optimization of the SVM parameters). Recall that the random forests are ensembles of PCTs for HMC predicting the complete hierarchy (a single model), while the SVMs construct a classifier for each node of the hierarchy separately. Hence, the increase of the hierarchy significantly increases the training time of SVMs (additional classifiers should be trained), while the training time for random forests will increase only slightly. For a detailed analysis of the computational complexity of the ensembles of PCTs, we refer the reader to Kocev (2011). Considering the time needed to annotate an unseen image, bagging and random forests of PCTs for HMC are faster than SVMs. Namely, to produce an annotation for unseen image, random forests and bagging need 1 ms for all datasets and descriptors, while SVMs need 1 ms for the datasets with Fourier descriptors, 2 ms for the SIFT descriptors and 3 ms for the combination. All in all, random forests of PCTs for HMC are more efficient than bagging of PCTs for HMC and SVMs.

du Buf and Bayer (2002) present a very extensive study on the problem of diatom identification. The ADIAC diatom image database was analyzed using a variety of feature extraction techniques (mainly based on contour and ornamentation) and their combinations, as well as a variety of machine learning algorithms. The results of this study together with the best results from the proposed system for diatom identification are summarized in Table 3. Since all systems used a similar variant of the database, we can relatively directly compare the performance of the used machine learning algorithms (i.e., the classifiers) and the feature extraction techniques (i.e., the descriptors). Note that in the other studies a purified version of the database was used: some images with bad quality were removed. In our study, however, we used all the images.

We focus the comparison on the approaches that use an equal number of diatom taxa. For the 38 taxa database, our approach has ~3% better recognition rate than the best performing previous method (du Buf and Bayer, 2002). This approach uses several types of descriptors (geometric, shape, Fourier, image moments, ornamentation and morphological) and bagging of decision trees. This means that our system uses a subset of the features used by the competing approach and a similar classifier. The difference is that we use predictive clustering trees as base classifiers in the ensembles as opposed to simple decision trees. Considering this, we can conclude that our system draws its predictive power from the specific classifier (i.e., predictive clustering trees), which predicts the whole hierarchy and exploits the dependencies between the taxa.

Next, we discuss the results for the 48 taxa database. The best reported recognition rate here is 82%, while our system achieves a 97.15% recognition rate — an absolute improvement of ~15%. The competing method uses a similar set of descriptors and a custom-made classifier for the task of diatom identification. As for the smaller database, we can conclude that the competitive advantage of our system is the usage of predictive clustering trees as base classifiers in the ensembles.

5.2. Annotation results per species

We now focus the discussion of the results on the precision and recall values (shown in Table 4) for each diatom species obtained by using random forests of PCTs for HMC. Moreover, we also show the results for each genus of diatoms present in the database. The results from Table 4 show that our system achieves the perfect score (both precision and recall are 1) on 34 species and 14 genera for the largest database (55 taxa), 33 species and 15 genera for the middle sized database (48 taxa) and 28 species and 14 genera for the smallest database (38 taxa).

We further discuss the species for which we achieve poor annotation results. In particular, Fig. 5 depicts images from four species with bad annotation results. The worst precision of 0.56 is obtained for *Nitzchia hantzschiana*. We believe that this is because of the similarity with the other species from the same genus (see for example the images for *Nitzchia sp.2* shown in Fig. 5) and the fact that the images are not clean and contain other artifacts than diatoms of the given species (see Fig. 5). The first claim is also confirmed by the recall values for the other three species from the *Nitzchia* genus: images from the species *Nitzchia sp.2*, *Nitzchia sinuata* and *Nitzchia dissipata* were incorrectly annotated as *Nitzchia hantzschiana*. Also, the low precision value for *Achnanthes oblongella* and the low recall value for *Achnanthes minutissima* is due to the fact that images from the latter species were annotated as images from the former species. The same holds for *Pinnularia silvatica* and *Pinnularia subcapitata*.

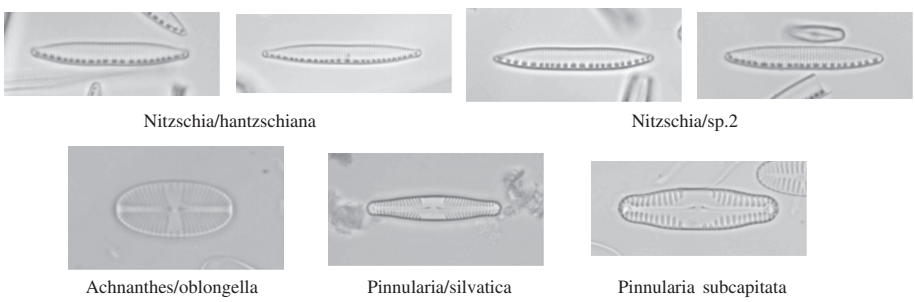

Fig. 5. Example images that were difficult to annotate correctly by our system.

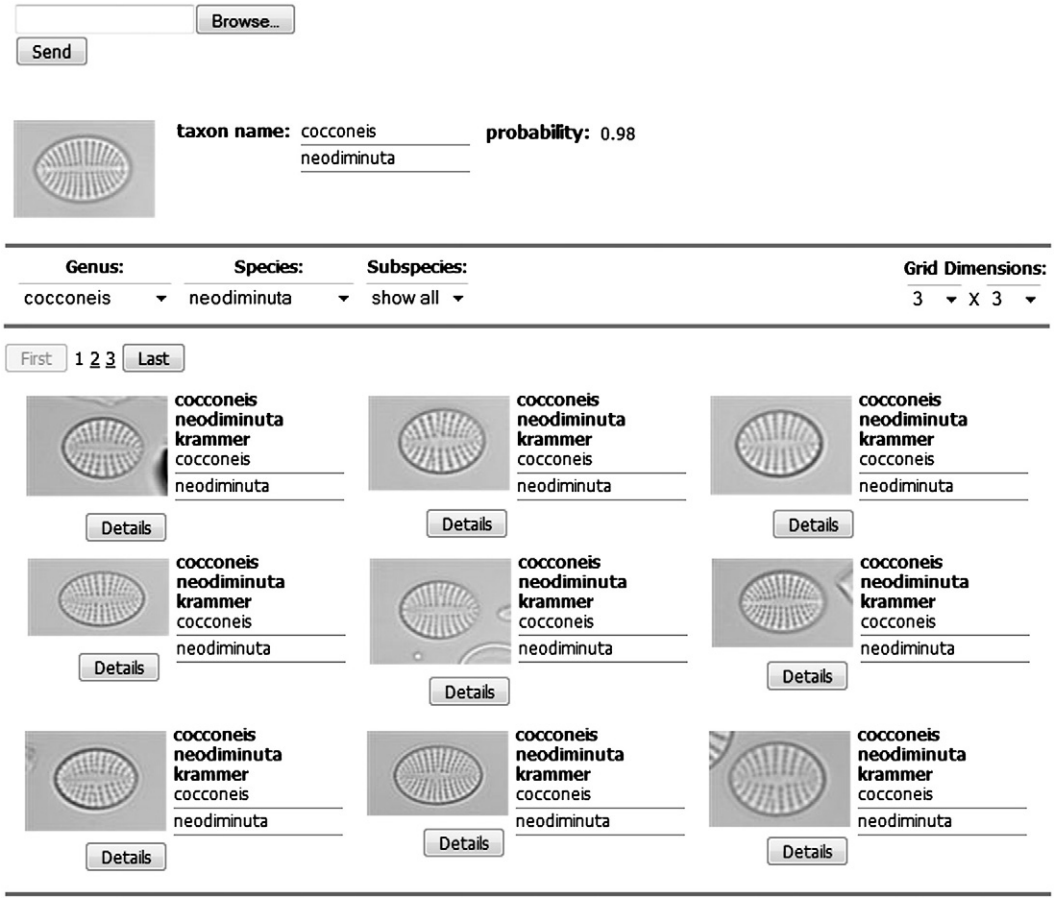

Fig. 6. Web-based user interface for our system. The query image and the reported identification results (taxon name and probability) are shown in the upper part. The bottom part shows the reference images for the identified taxon name that are present in the database.

5.3. Web-based diatom classification/identification interface

The goal of the complete system is to assist a taxonomist in identifying a wide range of different diatoms. We have developed a web-based user interface (Fig. 6) to demonstrate the usefulness of the proposed system. The user (a taxonomist) can upload a diatom query image and submit a request to identify the taxon for the given image.

After the request is submitted to the classification system, automatic image segmentation is performed first. Then, in the feature extraction module, the visual descriptors are generated. At the end, the classifier returns the full taxon name of the query image including the probability that the given example is annotated with the returned taxon name. The user can then click on the taxon name to submit a query to the system, which will return a collection of reference images for that taxon. Fig. 6 shows the user-interface of our system with a selected query diatom image, the returned taxon name, and some reference images returned for that taxon name.

6. Conclusions

We propose a novel approach to taxonomic identification from microscopic images. We combine different feature extraction approaches and hierarchical multi-label classification. We learn ensembles of predictive clustering trees that predict the taxonomic rank of the diatom in the image by using the features of the image and taking into account the hierarchical structure of the taxonomy.

We evaluate the proposed approach on the ADIAC diatom image database. We compare the different feature extraction techniques and suggest that the combination of contour-based and texture-based features is most suitable for automatic classification of diatom images. We directly compare the predictive performance and efficiency of ensembles of PCTs and SVMs. The results show that random forests of PCTs have better predictive performance and are more efficient than the trained SVMs.

We also contrast our results with earlier results on this dataset, which used specialized features developed for diatom images. Previous work also used a smaller portion of this dataset, with fewer species, and focused on images of high quality. We show that our approach outperforms the current state-of-the-art in the field and offers very high predictive performance. Furthermore, we demonstrate the possible usage of the developed system by taxonomists.

Several directions for further work call for attention. First, we can consider using other diatom image databases, as quite a few have become available recently. Second, we can consider identifying multiple species in the same sample at the same time: this would truly exploit the multi-label aspect of hierarchical multi-label classification. Finally, we can consider using the same approach to address taxon identification problems for other types of organisms.

To summarize, we propose a system for automatic diatom classification that consists of two parts: image processing (feature extraction from images) and image classification. It offers very high predictive performance — the best reported performance on the considered dataset. The proposed approach can be extended with new feature extraction techniques. It can thus be applied to other similar tasks, such as the taxonomic classification of other groups of organisms.

References

ADIAC, 2011. Diatom Image Database from ADIAC Projectaccessed on 01.07.2011 http://rbg-web2.rbge.org.uk/ADIAC/pubdat/pubdat.html.

Bauer, E., Kohavi, R., 1999. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Machine Learning 36 (1), 105–139.
Blockeel, H., Raedt, L.D., Ramong, J., 1998. Top-down induction of clustering trees. Proc. of 15th International Conference on Machine Learning. Morgan Kaufmann, pp. 55–63.
Breiman, L., 1996. Bagging predictors. Machine Learning 24 (2), 123–140.
Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.
Chang, C.-C., Lin, C.-J., 2001. LIBSVM: a Library for Support Vector Machinesaccessed on 01.07.2011 http://www.csie.ntu.edu.tw/cjlin/libsvm.
Ciobanu, A., du Buf, H., 2002. Identification by contour profiling and legendre polynomials. In: du Buf, H., Bayer, M.M. (Eds.), Automatic Diatom Identification. World Scientific Publishing, pp. 167–185.
Dimitrovski, I., Kocev, D., Loskovska, S., Džeroski, S., 2011. Hierarchical annotation of medical images. Pattern Recognition 44 (10–11), 2436–2449.
du Buf, H., Bayer, M.M., 2002. Automatic Diatom Identification. World Scientific Publishing.
Fischer, S., Bunke, H., 2002. Identification using classical and new features in combination with decision tree ensembles. In: du Buf, H., Bayer, M.M. (Eds.), Automatic Diatom Identification. World Scientific Publishing, pp. 109–140.
Jalba, A.C., Wilkinson, M.H., Roerdink, J.B., 2004. Automatic segmentation of diatom images for classification. Microscopy Research and Technique 65, 72–85.
Kocev, D., 2011. Ensembles for predicting structured outputs, Ph.D. thesis, IPS Jožef Stefan, Ljubljana, Slovenia.
Kocev, D., Vens, C., Struyf, J., Džeroski, S., 2007. Ensembles of multi–objective decision trees. Proc. of the 18th European Conference on Machine Learning – LNCS 4701. Springer, pp. 624–631.
Loke, R.E., du Buf, H., 2002. Identification by curvature of convex and concave segments. In: du Buf, H., Bayer, M.M. (Eds.), Automatic Diatom Identification. World Scientific Publishing, pp. 141–165.
Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60 (2), 91–110.
MacLeod, N., 2008. Automated Taxon Identification in Systematics: Theory, Approaches and Applications. CRC Press, Taylor & Francis Group.
Mensink, T., Csurka, G., Perronnin, F., Sanchez, J., Verbeek, J.J., 2010. LEAR and XRCe's participation to visual concept detection task - imageclef 2010. CLEF (Notebook Papers/LABs/Workshops).
Morris, C.W., Autret, A., Boddy, L., 2001. Support vector machines for identifying organisms — a comparison with strongly partitioned radial basis function networks. Ecological Modelling 146 (1–3), 57–67.
Ojala, T., Pietikainen, M., Maenpaa, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (7), 971–987.
Santos, L.M., du Buf, H., 2002. Identification by gabor features. In: du Buf, H., Bayer, M.M. (Eds.), Automatic Diatom Identification. World Scientific Publishing, pp. 187–220.
Schietgat, L., Vens, C., Struyf, J., Blockeel, H., Kocev, D., Dzeroski, S., 2010. Predicting gene function using hierarchical multi-label decision tree ensembles. BMC Bioinformatics 11, 1–14.
Silla, C., Freitas, A., 2011. A survey of hierarchical classification across different application domains. Data Mining and Knowledge Discovery 22 (1–2), 31–72.
Sosik, H.M., Olson, R.J., 2007. Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. Limnology and Oceanography: Methods 5, 204–216.
Stoermer, E., Smol, J., 2004. The Diatoms: Applications for the Environmental and Earth Sciences. Cambridge University Press.
Vens, C., Struyf, J., Schietgat, L., Džeroski, S., Blockeel, H., 2008. Decision trees for hierarchical multi-label classification. Machine Learning 73 (2), 185–214.
Westenberg, M.A., Roerdink, J.B.T.M., 2002. Mixed-method identifications. In: du Buf, H., Bayer, M.M. (Eds.), Automatic Diatom Identification. World Scientific Publishing, pp. 245–257.
Wilkinson, M.H.F., Jalba, A.C., Urbach, E.R., Roerdink, J.B.T.M., 2002. Identification by mathematical morphology. In: du Buf, H., Bayer, M.M. (Eds.), Automatic Diatom Identification. World Scientific Publishing, pp. 221–244.
Zhang, H., 2004. The optimality of naive bayes. Proc. of the 17th International Florida Artificial Intelligence Research Society (FLAIRS) Conference.