Hierarchical Classification of Diatom Images using 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*

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 with diatom images. We apply several feature extraction techniques that can be used in the context of diatom images. Moreover, we investigate whether combination of these techniques increase the predictive performance. The experiments show that our system outperforms the most widely used approaches for image annotation.

Keywords: Diatoms, Automatic Image Annotation, Hierarchical Classification, Predictive Clustering Trees, Feature Extraction from Images

[*] Corresponding author (telephone: +389 2 3099156)

Email addresses: ivicad@feit.ukim.edu.mk (Ivica Dimitrovski), Dragi.Kocev@ijs.si (Dragi Kocev), suze@feit.ukim.edu.mk (Suzana Loskovska), Saso.Dzeroski@ijs.si (Sašo Džeroski)

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 and then 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 [1]. This is not a trivial task, taking into consideration that taxonomists estimate that there may be 200000 different diatom species, half of them still undiscovered, and many of these extremely hard to distinguish on the basis of morphology [2]. Furthermore, this is very tedious and repetitive work, thus any degree of automation can greatly help.

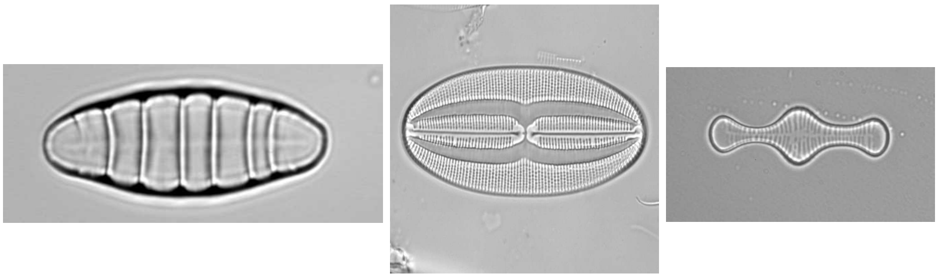


Figure 1: Example images, from left to right: *Diatoma mesodon*, *Fallacia sp.5* and *Tabellaria flocculosa*.

Having this in mind, we propose a system for automatic diatom classification. The system contains two parts: 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). The goal of the complete system is to assist a taxonomist in identifying a wide range of different diatoms.

The remainder of the paper is organized as follows. In Section 2, we present the related work. Section 3 describes the techniques we use for feature extraction from images. Section 4 introduces predictive clustering trees and their use for HMC. In Section 5, we explain the experimental setup. The obtained results and a discussion thereof are given in Section 6. Section 7 concludes the paper and points out some directions for further work.

2. Background and related work

The process of automatic diatom classification can be divided in three main stages [2]: image segmentation (i.e., contour extraction), feature extraction and image classification. The goal of the image segmentation is to locate and obtain the contour of the diatom image. 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. Here, we shortly describe each one of these stages.

2.1. Contour extraction

The problem of contour extraction of gray-scale diatom images can be solved by four approaches: threshold techniques, boundary-based methods, region-based methods, and hybrid techniques, which combine both boundary and region criteria [3]. *Threshold techniques* 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. Because these methods cannot ensure continuous edge-detection, an edge-linking step must be used to produce closed contours. Active contour methods automatically produce closed contours and usually provide better edge localization, but they 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. Representative methods from this technique are region growing, split-and-merge techniques and clustering methods. The main advantage of region-based methods 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. Two important representatives of this class are morphological watershed segmentation and seeded region growing All in all, there is a variety of approaches that one can choose for the problem at hand.

2.2. Feature extraction

After the contour extraction and segmentation step, different features extraction techniques can be employed on the diatom images [4]. These feature extraction techniques include various measures of contour, area, shape, intensity, texture etc. [5].

The diatoms can be primarily distinguished by evaluating properties of the valve's outline. Moreover, these features can be easily interpreted by the taxonomists. Hence, contour features are of high importance in automatic diatom classification. The contour features measure the symmetry, global and local shape characteristics, as well as geometric properties, such as length and width of the diatoms [6], [7].

An important characteristic of diatoms is also the ornamentation of the valve face, which is a specific type of texture [8]. There are several known feature sets able to measure these texture properties: features derived from gray level co-occurrence matrices, Gabor wavelets [9], scale invariant feature transform (SIFT) [10] and local binary patterns (LBP) [11]. 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 part of an automatic classification system is the classification phase. In this phase, a machine learning algorithm is first used to learn a classifier using the extracted features from the previous two stages and the annotations/labels of the images. Then, the obtained classifier maps the images of unidentified specimens to labels of trained taxa, i.e., provides annotations for previously unseen images. In the context of diatom image classification, most typically used machine learning algorithms are neural networks, naïve Bayes, Support vector machines (SVMs) and decision trees.

Santos and Du Buf [9] 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 the same number of units as the input layer, and the output layer has as many units as there are classes (taxa). The neural network is trained until the error rate on a validation set reaches a local minimum.

Another popular approach for automatic diatom classification is the naïve Bayes classifier, also called maximum-likelihood classifier [12]. This classifier estimates the probability density function of the features for each class from the training set. 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 machine learning techniques for image annotation in general. There are several studies concerning automated taxonomic classification that use SVMs as classifiers [13], [14]. The evaluation of these approaches was done for a variety of organisms [15]. However, for diatom classification, the neural networks and decision trees are typically used [2].

Current state-of-the-art results in automatic diatom classification are achieved using decision trees and bagging as ensemble learning technique [5]. The decision trees have several advantages: no prior assumptions for the probability distribution of the dependent and the independent variables, discrete and/or continuous independent variables, elegant handling of 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 [16].

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. The predictive clustering trees (PCTs), on the other hand, exploit the hierarchical taxonomy and simultaneously predict all taxonomic ranks [17]. This approach yields a very efficient classifier (a PCT) that offers high predictive performance.

3. Contour and feature extraction from images

The first two steps from the procedure of annotation of diatom images are contour extraction and feature extraction. With the first steps, an image is segmented: relevant objects/regions are detected and located. On these regions are then applied feature extraction techniques to obtain the descriptors of the image. These descriptors, together with the annotations of the images (for the ones that these annotations are provided beforehand, i.e., the images from the training set), are used to learn a classifier. In the following, we describe the contour and feature extraction techniques that we use in this study.

3.1. Contour extraction

Automatic extraction of diatom contours is the first phase in diatom classification. This is a difficult and non-trivial task, because diatoms may lie on top of each other, or be surrounded by debris. For automatic diatom segmentation and contour extraction, we used the procedure described in [18]. This procedure can be summarized in two major steps: 1) pre-segmentation, where multiple objects in an image are detected by suppressing the background (Section 3.1.1), and 2) detailed analysis of the remaining regions to find the exact location of the object contours (Sections 3.1.2 and 3.1.3). After these two steps are completed, we can proceed with extraction of the image descriptors.

3.1.1. Pre-segmentation of an image

The microscopic images of diatoms consist of two parts: the diatom itself and a background. In contrast to the areas occupied by the diatom objects, the background of the microscopic images is mostly unstructured with very smooth gray level transitions and a small variance. Using this property, the images are segmented into regions of possible diatom objects (structured regions) and background (unstructured regions).

The pre-segmentation of an image involves threshold selection, identification of regions with structured objects and merging nested regions. The selection of the two thresholds that separate the diatom objects from dark or light debris is done by analyzing the histogram of the entire image. So, the parts of the image that have gray level outside these two thresholds are considered debris and are not further analyzed. The remaining parts of the image are then analyzed using local variance of the gray level in a 3x3 neighborhood. The local variance makes it possible to distinguish between the contour and debris: if the local variance is low then it is debris, otherwise it is part of a contour. However, some very dark or light pixels, that were eliminated with the first two thresholds can cause gaps in the result from the analysis of the local variance. To alleviate this, the regions are merged together using neighborhood graph of connected regions. At the end, the smallest rectangular bounding box is determined for each object region and the corresponding part of the original input image is used for further processing. Because an object image can still include two (or more) connected or overlapping objects, further post-processing is necessary to find the exact location of the diatoms outlines.

3.1.2. Edge-based thresholding for contour extraction

The pre-segmentation detects initial regions/objects in the images. Using these, the next step is to determine the location of the boundary between the objects and the background and produce binary (black and white) images with the diatom contours. To obtain these closed diatom contours, we employ edge-based thresholding [19]: Specifically, we use the Canny edge detector [20], which uses hysteresis-based thresholding.

The edge-based thresholding uses 'edge detection operators' that locate the gray level differences across edges. Edges correspond to image locations with strong transitions. The edge detection algorithms output a map of the most significant edges. The non-significant/weak edges usually correspond to noise and are removed by applying some threshold.

The idea behind the application of a hysteresis in the Canny edge detector is that the weak edges can belong to real edges if they are connected to any of the pixels with strong response. It applies two thresholds: high and low. So, the Canny edge detector first selects edge pixels that have strength (the local variance from the previous Section) above a given (high) threshold. Then, it applies an additional (low) threshold to the connected pixels that can also be considered as edge pixels. The detected edges are marked in black and all other pixels are white.

3.1.3. Contour following

The last step in the process of contour extraction is the contour following. The contour following traces the region borders in the binary image. 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 3x3 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’.

The contours that are obtained with the procedure are then post-processed by evaluating their warping. If a contour is significantly warped, implying contour deformations by noise or debris, then it is checks whether there is a better contour candidate with a smaller curvature. Also, this procedure rejects the contours that have less pixels than the minimum diatom size.

3.2. Feature extraction

Using the regions/objects identified by the contour extraction procedure described above, we proceed with feature extraction techniques. Given the specific problem of annotating microscopic diatom images, we apply three techniques that can and have been used in this context: shape, Fourier and SIFT descriptors. In the following, we shortly describe each of these.

3.2.1. Simple geometric properties and simple shape descriptors

Several features, such as length, width, size and the length-width ratio, can be easily computed from the extracted contour of the diatom image [5]. The length and width are calculated using the contours principal axes. The direction of the minor axis is selected perpendicular to the major axis. The length L is defined as the maximum distance between the intersections of the contour and the major axis. The width W is calculated in the same way, but using the minor axis. The length-width ratio is defined as $R = L/W$. The size S of a contour is equal to the number of pixels it encloses. At the end, using the pixel size of the images, we convert the length and width to μm and the size to μm^2.

The shape descriptors usually rely on some simple heuristics and yield acceptable results in the case of simple shapes. These descriptors cannot be used for a reconstruction, and they do not work well for complex shapes, but experts can easily interpret the values. Heuristic descriptors that we use in our experiments are: rectangularity, triangularity, compactness, ellipticity and circularity [21]. The rectangularity of an object is defined as the ratio of the object area to the area of its minimum enclosing rectangle. The triangularity is defined as state or quality of having the shape of a triangle. The compactness, ellipticity and circularity are measures that are connected with the circular shape of the contour. They measure how close is a given shape to a circle, ellipse ot elongated polygon. These descriptors can provide satisfactory information for shape classification and discrimination.

3.2.2. Fourier descriptors

The Fourier descriptors see a closed curve (diatom contour) as a periodic function and represent it by a set of Fourier coefficients. The Fourier descriptors are obtained through Fourier transform on a complex vector derived from the coordinates of the shape boundary. The complex vector is given as the difference of the boundary points to the centroid of the shape. This representation is invariant to translation because the centroid substraction represents the position of the shape from boundary coordinates. Fourier transformation is then applied to the complex vector to obtain the Fourier coefficients.

The magnitudes of the obtained Fourier coefficients are normalized by the magnitude of the first coefficient. The Fourier descriptors are invariant to translation, rotation and scaling. The high frequency noise can be significantly reduced by limiting the number of coefficients (the effect of lowpass filtering). At the same time, this will preserve the main details of the patterns.

By limiting the number of coefficients k, we are able to reduce the high frequency noise to a great extent, leaving at the same time the main details of the patterns. Thus, the application of limited number of Fourier descriptors has the effect of lowpass filtering. On the other hand, this reduction can lead to loss of spatial information in terms of fine detail. Following the reccomendations from [5], we consider 30 coefficients as sufficient to distinguish between most shapes.

3.2.3. 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 some local, smaller regions of an image can provide significant information for distinguishing and discrimination of the images. To this end, many different techniques for detecting and describing local image regions have been developed. The most widely used techniques is the Scale Invariant Feature Transform (SIFT), which was proposed as a method of extracting and describing key-points which are reasonably invariant to changes in illumination, image noise, rotation, scaling, and small changes in viewpoint [10].

The descriptors using local features can be quite big because an image may contain many key-points and each key-point is described by a 128 dimensional vector with numerical values. To reduce the descriptor's size, we use histograms of local features. With this approach, 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, the key-points are extracted from all database images. For the key-point extraction and descriptor calculation, we use the default parameters proposed by Lowe [10]. The key-points are clustered into 200 clusters using k-means. Please note that we consider only the key-points that are detected inside the shape of the contour, while the key-points detected outside of the shape contour are discarded. Afterwards, for each key-point, we discard all information 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 to 1. This results in a 200 dimensional histogram for each image.

4. Ensembles of PCTs for HMC

The descriptors that were obtained using the procedures from the previous Section, combined together with the annotations of the images, are used to train a classifer. The annotations/labels 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) [22] for hierarchical multi-label classification (HMC) [17]. 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 multi-label hierarchical classification. We then present the PCTs for HMC and the ensembles of PCTs for HMC.

4.1. 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

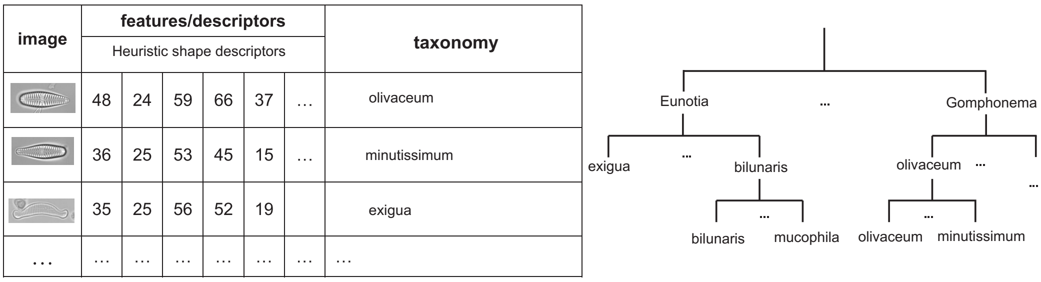


image	features/descriptors						taxonomy
	Heuristic shape descriptors						
	48	24	59	66	37	…	olivaceum
	36	25	53	45	15	…	minutissimum
	35	25	56	52	19		exigua
…	…	…	…	…	…	…	…

Figure 2: 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).

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 [23].

In diatom classification, the application domain on which we focus, an important problem is the development of an automatic image classification system, which can identify the diatom species. The diatom species are separated/classified in several categories (taxonomic rank). These categories are grouped in a logical system of categories with hierarchical structure based on their characteristics. The most global subdivision is the genus. Within each genus there are many species which can be further divided into subspecies, varieties, forms, morphotypes, etc., such as the one shown in Fig. 2. Each image is represented with: (1) a set of descriptors (in this example, the descriptors are heuristic shape descriptors: rectangularity, compactness, ellipticity, triangularity, and circularity) and (2) a set of labels/annotations. A single image can be annotated with multiple labels at different levels of the predefined hierarchy of taxonomic ranks.

For example, the image in the second row of the table in Fig. 2 has one label: minutissimum which is listed explicitly. Note that this image is also implicitly labeled with the labels: olivaceum and gomphonema. These labels are all ancestors of the explicitly listed label in the given hierarchy.

The data, as presented in the table in the left-hand side of Fig. 2, constitute a data set for HMC. This set can be used by the machine learning algorithm to train a classifier for HMC. The testing set of images contains only the set of descriptors and has no *a priori* annotations.

4.2. PCTs for hierarchical-multi label classification

In the PCT framework [22], 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. 2). 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, while the variance $Var(S)$ is defined by (1). Maximizing the variance reduction maximizes cluster homogeneity and improves predictive performance.

A leaf of a PCT is labeled with/predicts the prototype of the set of examples belonging to it. With appropriate variance and prototype functions, PCTs can handle different types of data, e.g., multiple targets [24] or time series [25]. A detailed description of the PCT framework can be found in [22]. The PCT framework is implemented in the CLUS system, which is available at `http://www.cs.kuleuven.be/~dtai/clus`.

To apply PCTs to the task of HMC, the example labels are represented as vectors with Boolean components. Components in the vector correspond to labels in the hierarchy traversed in a depth-first manner. The i-th component of the vector is 1 if the example belongs to class c_i and 0 otherwise. If $v_i = 1$, then $v_j = 1$ for all v_j's on the path from the root to v_i.

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, i.e.,

$$Var(S) = \frac{\sum\limits_{i} d(v_i, \bar{v})^2}{|S|} \tag{1}$$

The higher levels of the hierarchy are more important: an error at the upper levels costs more than an error at the lower levels. Considering this, a weighted Euclidean distance is used:

$$d(v_1, v_2) = \sqrt{\sum_{i} w(c_i)(v_{1,i} - v_{2,i})^2} \tag{2}$$

where $v_{k,i}$ is the i'th 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 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. [17]. Next, we explain how PCTs are used in the context of an ensemble classifier, in order to further improve the performance of PCTs.

4.3. 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) [16],[26], or more complex combinations.

We use PCTs for HMC as base classifiers. Previously, in the domain of functional genomics [27] and annotation of medical X-ray images [28], it is shown that, both random forests and bagging of PCTs, outperform a single PCT. 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.

We consider two ensemble learning techniques that have primarily been used in the context of decision trees: bagging and random forests. Bagging [16] 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 is obtained equal to the size of the training set. Bagging is applicable to any type of learning algorithm.

A random forest [26] 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.

5. Experimental setup

We evaluate the presented techniques for feature extraction from diatom images and hierarchical multi-label classification on the ADIAC diatom image database [29]. In our experiments we used subset of 1099 images that are classified using the taxonomic rank mentioned above. These images belong to 55 different taxa. For each taxa there are at least 10 images available, up to a maximum of 29 images. The diatoms in this set vary in shape but also in ornamentation (three examples are shown in Fig. 1).

A major set of analyses of this database was performed in [2], where two more versions of this dataset were used. The first one consists of 37 taxa, for which at least 20 images per taxa are available (819 images in total). The second one consists of 48 taxa, for which at least 15 images per taxa are available (1020 images in total). For comparability issues, we conduct experiments on all three variants of the database. Note that additional pre-processing was done on the database in [2], such as removal of some images with low quality. Here, we use the complete database without any additional pre-processing.

The algorithm for learning PCTs requires as input the weight of the depth in the hierarchy. We set w_0 to 0.75 to force the algorithm to make better predictions on the upper levels of the hierarchy. We trained ensembles of 100 un-pruned trees (PCTs) [30]. The size of the feature subset that is retained at each node, when training a random forest, 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.

The evaluation was done using 10-fold cross validation on the train set by reporting the overall recognition rate of the entire taxonomic rank: genus, species, variety, and form. The overall recognition rate is a very common and widely used evaluation measure. In our case it is the fraction of the validation images whose complete taxonomy was predicted correctly.

6. Results and discussion

This section presents the results for the identification of diatom taxa from images. We look at the results from three angles: performance of bagging and random forests of PCTs for HMC on the three variants of the

Table 1: Predictive performance of the feature extraction algorithms and their combination.

Classifier	Descriptor	# features	Overall recognition rate [%]		
			55 diatom taxa	48 diatom taxa	37 diatom taxa
Bagging	Geometric and shape descriptors	9	76.3	76.7	77.2
	Fourier descriptors	30	86.7	88.1	88.6
	SIFT histograms	200	88.4	89.2	91.3
	Geometric and shape desc.+Fourier desc.+SIFT hist.	239	96.2	98.1	98.8
Random Forests	Geometric and shape descriptors	9	76.3	76.7	77.2
	Fourier descriptors	30	86.6	88.1	88.7
	SIFT histograms	200	88.2	87.9	91.1
	Geometric and shape desc.+Fourier desc.+SIFT hist.	239	96.2	98.1	98.7

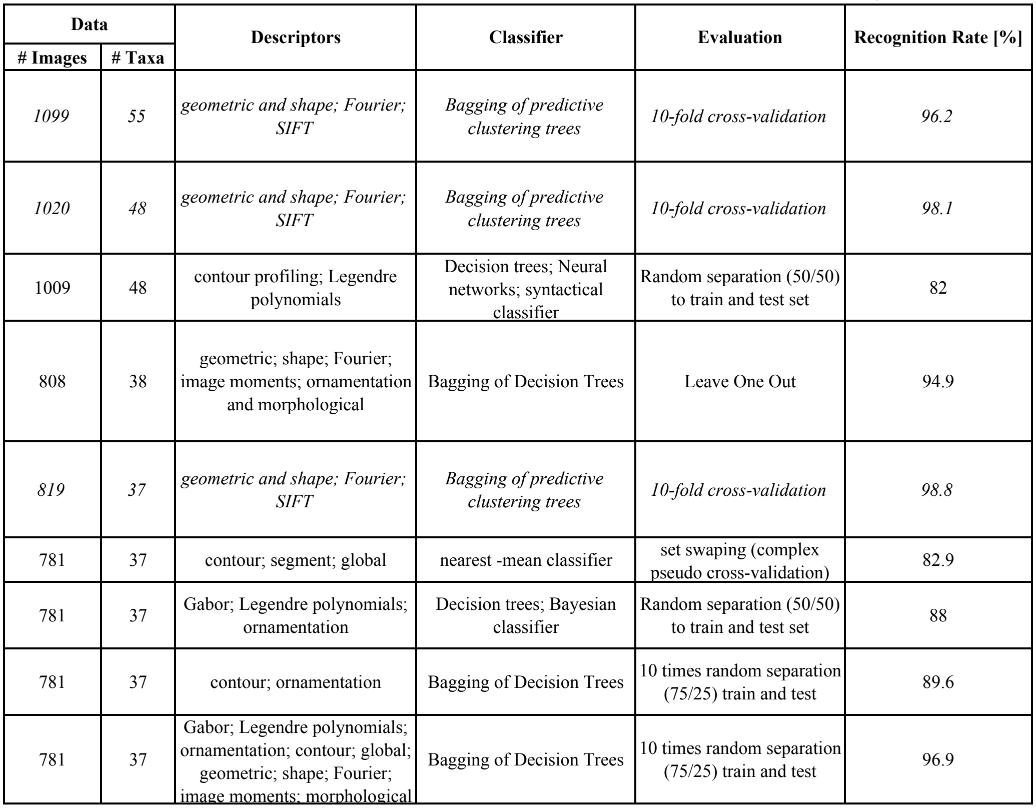
database (Table 1), performance comparison with other approaches (Table 2) and identification results per taxon (Table 3).

Table 1 summarizes the recognition results for the diatom image dataset with 55 different taxa, using the different feature extraction algorithms. We can note the high predictive performance of the SIFT histogram: it is most capable of capturing the hierarchical structure of the taxonomic ranks of the diatom images. The Fourier descriptors give the second best recognition rate. The simplest descriptors (combination of geometric properties with the heuristic shape descriptors) performed quite well on the selected database with overall recognition rate of 76.3%.

Inclusion of more than one type of features in the classification process contributes to better representation of the hierarchical nature of the images and offers orthogonal information to the classifier. This helps to further improve the predictive performance. The best recognition rate is obtained by concatenation of the individual descriptors and then learning a classifier using the larger feature set. The predictive performance in this case is 96.2%. This implies that no single set of features allows to discriminate all different taxa, furthermore the shape descriptors and the texture descriptors are very important in the classfication process. Most of the diatoms can be distinguished by evaluating properties of the valve outline, hence contour features are of high importance in automatic diatom classification. However, the ornamentation of the valve face, which is a specific type of texture, is also an important characteristic of diatoms.

On the smaller ('cleaner') variants of the dataset, the performance is higher. The best overall performance (98.7%) is on the smallest dataset, again achieved with the combination of all features. We can also note that there is no difference in the performance of the bagging and random forests ensemble learning methods.

Table 2: Comparison of the performance of the ensembles of PCTs (given in italic typeface) to the performance of the approaches from Du Buf and Meyer [2]. For each approach, we present the number of images, number of different taxa, used feature extraction techniques and classifiers, the evaluation of the performance and reported overall recognition rate.



Data		**Descriptors**	**Classifier**	**Evaluation**	**Recognition Rate [%]**
# Images	**# Taxa**				
1099	*55*	*geometric and shape; Fourier; SIFT*	*Bagging of predictive clustering trees*	*10-fold cross-validation*	*96.2*
1020	*48*	*geometric and shape; Fourier; SIFT*	*Bagging of predictive clustering trees*	*10-fold cross-validation*	*98.1*
1009	48	contour profiling; Legendre polynomials	Decision trees; Neural networks; syntactical classifier	Random separation (50/50) to train and test set	82
808	38	geometric; shape; Fourier; image moments; ornamentation and morphological	Bagging of Decision Trees	Leave One Out	94.9
819	*37*	*geometric and shape; Fourier; SIFT*	*Bagging of predictive clustering trees*	*10-fold cross-validation*	*98.8*
781	37	contour; segment; global	nearest -mean classifier	set swaping (complex pseudo cross-validation)	82.9
781	37	Gabor; Legendre polynomials; ornamentation	Decision trees; Bayesian classifier	Random separation (50/50) to train and test set	88
781	37	contour; ornamentation	Bagging of Decision Trees	10 times random separation (75/25) train and test	89.6
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	96.9

The most elaborate work so far on the problem of diatom identification is presented by Du Buf and Bayer [2] and is summarized in Table 2. We can note that the best performing approach is the one that uses the combination of the various feature sets and applies bagging of decision trees. Its recognition rate is 96.9%. For the 37 taxa variant, our approach has almost 2% better recognition rate than the best performing method. For the 48 taxa variant, where the best reported recognition rate so far is 82%, we achieve a 98.1% recognition rate – an absolute improvement of 16%. Moreover, our approach addresses a more difficult problem (55 taxa) and uses a less pre-processed database of images.

At the end, we would like to summarize the recognition rates per taxon

Table 3: The recognition rate per taxa obtained with combination of the feature extraction techniques and random forests of PCTs for HMC.

		Overall recognition rate [%]		
taxon	**#images**	**55 diatom taxa**	**48 diatom taxa**	**37 diatom taxa**
Navicula reinhardtii var. reinhardtii Grunow in Van Heurck	29	100.00	100.00	100.00
Surirella brebissonii Krammer & Lange-Bertalot	28	100.00	100.00	100.00
Navicula lanceolata (Agardh) Ehrenberg	27	100.00	100.00	100.00
Nitzschia sp.2	27	92.59	92.59	92.59
Diatoma mesodon (Ehrenberg) Kutzing	26	100.00	100.00	100.00
Cymbella helvetica Kutzing	26	96.15	96.15	96.15
Fallacia forcipata (Greville) Stickle & Mann	26	92.30	92.30	92.30
Encyonema silesiacum (Bleisch in Rabenhorst) Mann	25	100.00	100.00	100.00
Gomphonema minutum (Agardh) Agardh	24	100.00	100.00	100.00
Navicula sp.	24	100.00	100.00	100.00
Cocconeis stauroneiformis (W. Smith) Okuno	23	100.00	100.00	100.00
Tabellaria quadriseptata Knudson	23	100.00	100.00	100.00
Eunotia denticulata (Brebisson) Rabenhorst	22	100.00	100.00	100.00
Navicula constans var. symmetrica Hustedt	22	100.00	100.00	100.00
Denticula tenuis Kutzing	22	95.45	100.00	100.00
Cymbella subequalis Grunow in Van Heurck	21	100.00	100.00	100.00
Navicula radiosa Kutzing	21	100.00	100.00	100.00
Pinnularia kuetzingii Krammer	21	100.00	100.00	100.00
Eunotia tenella (Grunow) Hustedt in Schmidt	21	95.23	95.23	90.00
Tabularia investiens (W. Smith) Williams & Round	21	95.23	95.23	95.23
Gomphonema sp.1	20	100.00	100.00	100.00
Gyrosigma acuminatum (Kutzing) Rabenhorst	20	100.00	100.00	100.00
Navicula capitata Ehrenberg var. capitata	20	100.00	100.00	100.00
Nitzschia dissipata (Kutzing) Grunow	20	100.00	100.00	100.00
Parlibellus delognei (Van Heurck) Cox	20	100.00	100.00	100.00
Petroneis humerosa (Br?bisson ex Smith)Stickle & Mann	20	100.00	100.00	100.00
Tabellaria flocculosa (Roth) Kutzing	20	100.00	100.00	100.00
Tabularia sp.1	20	100.00	100.00	100.00
Cymbella hybrida var. hybrida Grunow in Cleve & Moller	20	95.00	100.00	100.00
Diatoma moniliformis Kutzing	20	95.00	100.00	100.00
Eunotia incisa Gregory	20	95.00	95.00	95.00
Gomphonema augur var. augur Ehrenberg	20	95.00	95.00	95.00
Nitzschia sinuata (Thwaites) Grunow var. sinuata	20	95.00	95.00	95.00
Opephora olsenii Moller	20	95.00	95.00	100.00
Fragilariforma bicapitata Williams & Round	20	90.00	100.00	100.00
Meridion circulare (Greville) Agardh	20	90.00	100.00	100.00
Nitzschia hantzschiana Rabenhorst	20	85.00	85.00	100.00
Cocconeis neodiminuta Krammer	19	100.00	100.00	N/A
Cocconeis placentula var. placentula Ehrenberg	19	100.00	100.00	N/A
Epithemia sorex var. sorex Kutzing	19	100.00	100.00	N/A
Navicula viridula var. linearis Hustedt	19	100.00	100.00	N/A
Stauroneis smithii Grunow	19	100.00	100.00	N/A
Navicula rhynchocephala Kutzing	19	94.73	94.73	N/A
Caloneis amphisbaena (Bory) Cleve	18	100.00	100.00	N/A
Navicula menisculus Schumann	18	100.00	100.00	N/A
Sellaphora bacillum (Ehrenberg) D.G. Mann	18	100.00	100.00	N/A
Fallacia sp.5	17	88.24	88.24	N/A
Staurosirella pinnata (Ehrenberg) Williams & Round	16	87.50	87.50	N/A
Pinnularia subcapitata var. hilseana (Janisch ex Rabenhorst) O. Moller	14	92.85	N/A	N/A
Achnanthes oblongella Oestrup	12	91.66	N/A	N/A
Eunotia bilunaris var. bilunaris (Ehrenberg) Mills	12	83.33	N/A	N/A
Navicula gregaria Donkin	11	90.90	N/A	N/A
Encyonema neogracile Krammer	10	90.00	N/A	N/A
Pinnularia silvatica Petersen	10	90.00	N/A	N/A
Achnanthes minutissima var. minutissima Kutzing	10	80.00	N/A	N/A

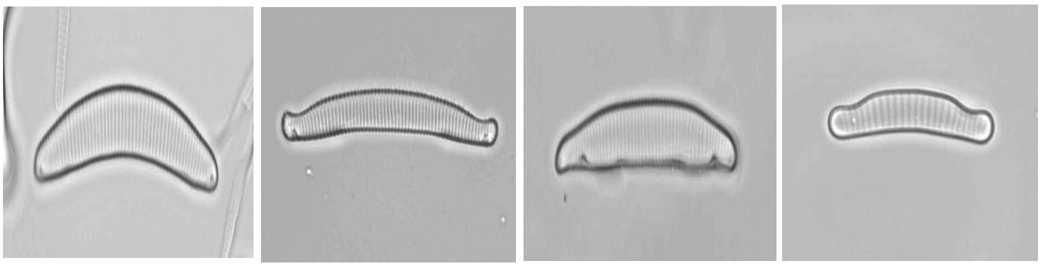


Figure 3: The taxa that are most difficult to identify because of the similarity of the images. From left to right, *Eunotia bilunaris var. bilunaris*, *Eunotia denticulata*, *Eunotia incisa*, *Eunotia tenella*.

(given in Table 3). On all the datasets, our method achieves maximal (100%) recognition rates for the majority of taxa. Namely, for the dataset with 55 taxa, the maximal recognition rate is achieved for 30, for the dataset with 48 taxa, for 35 taxa and for the smallest one with 37 taxa, it is achieved for 29 taxa. Lower recognition rates are achieved for taxa that are very similar to each other and difficult to distinguish. For example, the *Eunotia* diatoms depicted in Figure 3 are different to tell apart. Also, the *Fallacia* diatoms are similar to each other. Furthermore, four images for *Nitzschia* diatoms have a significant amount of debris in the background.

To summarize, we presented an approach for automatic taxa identification in microscopic diatom images. The presented approach has very high predictive performance (ranging from 96.2% to 98.7%) on the three variants of the image database. Also, on the majority of taxa, it achieves 100% recognition rate. These results are significantly better than then the ones reported in [2] for $2-16\%$ in terms of recognition rate. Moreover, the results are better than the ones obtained from human annotators: Du Buf and Bayer [2] report 63.3% recognition rate. All in all, our results are the best reported results on this database so far.

7. Conclusions

We propose a novel approach to taxonomic identification of taxa from microscopic images. We combine different feature extraction approaches and hierarchical multi-label classification: The predictive modeling problem that we consider is to learn predictive clustering trees the taxonomic position of the diatom in the image by using the hierarchical structure and the features of the image. We evaluate the proposed approach on the ADIAC Diatom Image Database.

We compare the different feature extraction techniques and suggest that the combination of all features is most suitable for automatic classification of diatom images. We also contrast our results with earlier results on this dataset, which used specialized features developed for diatom images. Previous work also used only a small portion of this dataset, with just a handful of 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.

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 this dataset. The proposed approach can be easily 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

[1] E. Stoermer, J. Smol, The Diatoms:Applications for the Environmental and Earth Sciences, Cambridge University Press, 2004.

[2] H. du Buf, M. M. Bayer, Automatic diatom identification, World Scientific Publishing, 2002.

[3] A. C. Jalba, M. H. Wilkinson, J. B. Roerdink, Automatic segmentation of diatom images for classification, Microscopy research and technique 65 (2004) 72–85.

[4] M. A. Westenberg, J. B. T. M. Roerdink, Mixed-method identifications,

in: H. du Buf, M. M. Bayer (Eds.), Automatic diatom identification, World Scientific Publishing, 2002, pp. 245–257.

[5] S. Fischer, H. Bunke, Identification using classical and new features in combination with decision tree ensembles, in: H. du Buf, M. M. Bayer (Eds.), Automatic diatom identification, World Scientific Publishing, 2002, pp. 109–140.

[6] R. E. Loke, H. du Buf, Identification by curvature of convex and concave segments, in: H. du Buf, M. M. Bayer (Eds.), Automatic diatom identification, World Scientific Publishing, 2002, pp. 141–165.

[7] A. Ciobanu, H. du Buf, Identification by contour profiling and legendre polynomials, in: H. du Buf, M. M. Bayer (Eds.), Automatic diatom identification, World Scientific Publishing, 2002, pp. 167–185.

[8] M. H. F. Wilkinson, A. C. Jalba, E. R. Urbach, J. B. T. M. Roerdink, Identification by mathematical morphology, in: H. du Buf, M. M. Bayer (Eds.), Automatic diatom identification, World Scientific Publishing, 2002, pp. 221–244.

[9] L. M. Santos, H. du Buf, Identification by gabor features, in: H. du Buf, M. M. Bayer (Eds.), Automatic diatom identification, World Scientific Publishing, 2002, pp. 187–220.

[10] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[11] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (7) (2002) 971–987.

[12] H. Zhang, The optimality of naive bayes, in: FLAIRS Conference, 2004.

[13] H. M. Sosik, R. J. Olson, Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry, Limnology and Oceanography: Methods 5 (2007) 204–216.

[14] C. W. Morris, A. Autret, L. Boddy, Support vector machines for identifying organisms – a comparison with strongly partitioned radial basis function networks, Ecological Modelling 146 (1-3) (2001) 57–67.

[15] N. MacLeod, Automated Taxon Identification in Systematics: Theory, Approaches and Applications, CRC Press, Taylor & Francis Group, 2008.

[16] L. Breiman, Bagging predictors, Machine Learning 24 (2) (1996) 123–140.

[17] C. Vens, J. Struyf, L. Schietgat, S. Džeroski, H. Blockeel, Decision trees for hierarchical multi-label classification, Machine Learning 73 (2) (2008) 185–214.

[18] S. Fischer, H. R. Shahbazkia, H. Bunke, Contour extrction, in: H. du Buf, M. M. Bayer (Eds.), Automatic diatom identification, World Scientific Publishing, 2002, pp. 93–107.

[19] D. Ziou, S. Tabbone, Edge detection techniques an overview, Pattern Recognition and Image Analysis 8 (24) (1998) 537–559.

[20] J. Canny, A computational approach to edge detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 8 (6) (1986) 679–698.

[21] P. L. Rosin, Measuring shape: Ellipticity, rectangularity, and triangularity, Machine vision and applications 14 (3) (2003) 172–184.

[22] H. Blockeel, L. D. Raedt, J. Ramong, Top-down induction of clustering trees, in: International Conference on Machine Learning, Morgan Kaufmann, 1998, pp. 55–63.

[23] C. Silla, A. Freitas, A survey of hierarchical classification across different application domains, Data Mining and Knowledge Discoverydoi:10.1007/s10618-010-0175-9.

[24] D. Kocev, C. Vens, J. Struyf, S. Džeroski, Ensembles of multi-objective decision trees, in: European conference on Machine Learning, Lecture Notes In Artificial Intelligence, 2007, pp. 624–631.

[25] I. Slavkov, V. Gjorgjioski, J. Struyf, S. Džeroski, Finding explained groups of time-course gene expression profiles with predictive clustering trees, Molecular BioSystems 6 (4) (2010) 729–740.

[26] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.

[27] L. Schietgat, C. Vens, J. Struyf, H. Blockeel, D. Kocev, S. Dzeroski, Predicting gene function using hierarchical multi-label decision tree ensembles, BMC Bioinformatics 11 (2010) 1–14.

[28] I. Dimitrovski, D. Kocev, S. Loskovska, S. Džeroski, Hierarchical annotation of medical images, in: The 11th International Multiconference Information Society - IS 2008, 2008, pp. 174–181.

[29] P. consortium ADIAC, Diatom image database from ADIAC project (2010).
URL `http://rbg-web2.rbge.org.uk/ADIAC/pubdat/pubdat.html`

[30] E. Bauer, R. Kohavi, An empirical comparison of voting classification algorithms: Bagging, boosting, and variants, Machine Learning 36 (1) (1999) 105–139.