Machine Vision and Applications (2006) 17:297–307
DOI 10.1007/s00138-006-0035-1

ORIGINAL PAPER

A model of diatom shape and texture for analysis, synthesis and identification

Y. A. Hicks · D. Marshall · P. L. Rosin · R. R. Martin · D. G. Mann · S. J. M. Droop

Received: 16 March 2005 / Accepted: 23 May 2006 / Published online: 11 July 2006

Abstract We describe tools for automatic identification and classification of diatoms that compare photographs with other photographs and drawings, via a model. Identification of diatoms, i.e. assigning a new specimen to one of the known species, has applications in many disciplines, including ecology, palaeoecology and forensic science. The model we build represents life cycle and natural variation of both shape and texture over multiple diatom species, derived automatically from photographs and/or drawings. The model can be used to automatically produce drawings of diatoms at any stage of their life cycle development. Similar drawings are traditionally used for diatom identification, and encapsulate visually salient diatom features. In this article, we describe the methods used for analysis of photographs and drawings, present our model of diatom shape and texture variation, and finish with results of identification experiments using photographs and drawings as well as a detailed evaluation.

Keywords Classification · Automatic drawing synthesis · Principal curves · Texture analysis · Diatoms

1 Introduction

Diatoms are unicellular microscopic algae found in practically any moist environment. The study of diatoms is of importance for a variety of reasons, one being that they perform 20% of all the photosynthetic fixation of carbon dioxide and hence generate 20% of the oxygen produced each year [8]. Identification of diatoms, i.e. assigning a new specimen to one of the previously described species, finds applications in ecology, palaeoecology and forensic science.

Each diatom has an outside silica shell; the shell contains two larger elements called valves, one on either side of the cell, which bear species-specific patterns. Many diatom valves are sufficiently flat to give a repeatable two-dimensional (2D) view in all photographs. In the rest of the article, we refer to the external 2D contours of such valves as the shape of the diatoms, and to the valve patterns as the texture of the diatoms. There is a great variety in the size, shape and texture of such silica shells, and these characteristics are traditionally used by experts to identify new diatom specimens by comparing them to photographs and drawings of previously described species. The identification task is very difficult due to the huge number of species estimated to exist (approximately 2×10^5 [7]), close similarity in appearance of some species on one hand, and individual and life cycle related variation in diatom size, shape and pattern within a species on the other hand.

Over the years, a wealth of diatom drawings and photographs has been accumulated in the biological literature (Fig. 1). Recently, work has been undertaken to digitise this material and place it in searchable databases (e.g. [15]). A system for automatic identification of diatoms using digitised photographs was a natural progression (e.g. the ADIAC project described in [1]). However, to the best of our knowledge, there has not been any research done on automatic identification of diatoms in drawings.

Inclusion of digitised drawings in the system and providing the ability to compare photographs and drawings can produce several benefits. First of all, drawings contain mainly

Y. A. Hicks (✉)
Cardiff School of Engineering, Cardiff University,
Queen's Buildings, P.O. Box 925, Cardiff, CF24 0YF Wales, UK
e-mail: hicksya@cf.ac.uk

D. Marshall · P. L. Rosin · R. R. Martin
School of Computer Science, Cardiff University, Cardiff, UK

D. G. Mann · S. J. M. Droop
Royal Botanic Garden Edinburgh, Edinburgh, UK

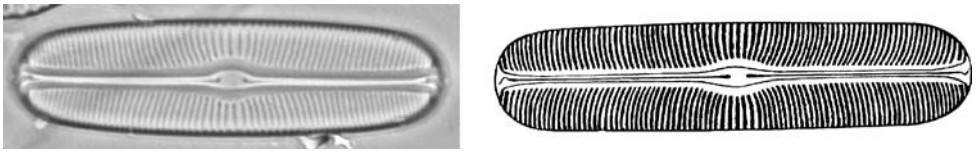

Fig. 1 A photograph of a diatom valve and a drawing of a similar valve by a biologist

the salient information required for identification and thus serve as models of each species. Secondly, they represent a known example of a species. Thirdly, there is a vast archive of drawings of diatoms from 200 years of diatom research, which cannot currently be searched other than via the names (which are often wrong) that have been attached to them by previous generations of researchers. Finally, the type specimens used to define species have generally been illustrated using drawings rather than photographs and the specimens used to generate the drawings are rarely specified, because they cannot easily be re-located among many others in a given sample. (For other classes of organisms, type specimens are usually single labelled specimens.)

In this article, we present a novel system for automatic identification of diatoms in photographs and drawings, which it does with the help of a model. The model is derived from the specimens of diatoms represented in photographic and/or drawing form. Each diatom specimen is represented in the model as a point in a multidimensional space; only the salient features used for identification are represented. In particular, we represent each diatom with a vector of size, shape and texture descriptors, as we explain in Sects. 3 and 4. A drawing and a photograph of the same specimen are ideally represented by the same point in this space, which thus serves as a mapping between photographs and drawings. Each species is modelled by a principal curve, which goes through the middle of a point cloud representing the specimens of this species. The life cycle related variation of diatom shape, texture and size in a species is modelled along this curve, while individual variations are modelled in the directions orthogonal to the curve. It is possible to synthesise a drawing of a diatom from the feature parameters stored at any point on a principal curve, as we show in this article. There are several advantages to being able to synthesise drawings of diatoms from points on a principal curve. Firstly, such a facility provides a visual check on how closely the shape, size and pattern of a species are being modelled. Secondly, it supplies us with new images for taxonomy.

In the rest of the article, we concentrate on the analysis and modelling of the appearance of pennate diatom species with striae patterns on their shells; most diatom species are of this kind. The variety of patterns occurring in diatoms is very great. A complete system would need to perform a series of tests to detect the type of pattern and then choose a suitable set of analytical tools to measure the values of appropriate pattern parameters.

2 Previous research

Diatoms have been studied for many years and their classification, i.e. the systematic division of the taxa into different genera, species, etc., is based on the sizes, shapes and patterns of their silica shells. Classification of diatoms is still developing, e.g. certain diatoms that were considered to be the same species for decades are being split into several species and new species are being found and described. Taking account of the above facts, it can be difficult to identify a diatom, even for a trained biological expert. There is clearly a need for a system that can assist in identification and classification of diatoms, especially for use by non-experts in application areas such as those mentioned above.

In recent years, there have been various efforts aimed at quantitative analysis of shape variation within individual diatom species, as well as the development of automatic systems for diatom identification. Stoermer and Ladewski [13] modelled *Gomphoneis herculeana* shapes by representing specimen outlines using Legendre polynomials and performing principal component analysis (PCA) on them. They were able to reconstruct outlines from such models, which showed that variation in shape corresponding to the first principal component was highly correlated with diatom length and thus with the stage in the life cycle. Goldman et al. [3] later used the same method to describe two populations of *Surirella fastuosa*, which showed that the populations were partially overlapping when projected in the space of the two largest eigenvectors. Mou et al. [9] used PCA on Fourier descriptors of diatom contours belonging to several subgroups of *Tabellaria flocculosa*.

In the ADIAC project [1], a diatom identification system was developed, based on shape, size, pattern features, and decision trees. The system was able to identify specimens from 37 species with around 97% accuracy. However, ADIAC is not capable of reconstructing diatom shape, nor appearance in drawings, nor is it capable of modelling the life cycle.

To date, there has been no attempt to develop a single system for modelling shape and texture variation both within and between a relatively large number of diatom species. One of the reasons for this could be the difficulty of obtaining images of a sufficient number of specimens for each modelled species at differing stages of its life cycle. Neither has there been an attempt to include the information available from type drawings into such a system. Finally, no one has

attempted to produce drawings automatically from photographs of diatoms.

In the following sections, we introduce a system that models changes in size, shape and texture over the life cycles of pennate diatom species, and allows the production of simple diatom drawings. Our system does not require a large number of specimens for training, and allows identification of novel specimens with a success rate comparable to other diatom identification systems such as ADIAC [1].

3 Contour shape analysis and synthesis

As we mentioned earlier, many diatom valves are sufficiently flat to give a repeatable 2D view in all photographs. Traditionally, when analysing diatom shape, researchers performed analysis of diatom valve contours in such view. However, for various reasons it is not an easy task to extract the 2D contours from photographs automatically. Overlapping debris, diffraction effects at the small scale of diatoms, and the compromises made on focus sharpness to capture a diatom in a single view, all make it hard to locate the contour. In the course of ADIAC, several sophisticated methods for contour extraction were developed. In our work, we used the extracted contours provided to us by the ADIAC partners, where each contour is represented as a connected set of points.

We need a general way to describe diatom shapes belonging to a large number (possibly hundreds) of different species. Fourier descriptors have been successfully used to provide a compact and informative description of diatom shapes [9] in previous research. There are several types of Fourier descriptors (FD) for plane closed curves which could be used for diatom contours. We adopted the FD developed by Zahn and Roskies [14], defines as follows:

Let γ be a clockwise-oriented simple closed curve with parametric representation $(x(l), y(l)) = Z(l)$, where l is the arc length and $0 \leq l \leq L$. Let us define the cumulative angular function $\phi(l)$ as the net amount of angular turn between the tangent at starting point $l = 0$ and the tangent at point l. The domain of $\phi(l)$ is normalised to the interval $[0, 2\pi]$. The formal definition of a normalised variant ϕ^* whose domain is $[0, 2\pi]$ is $\phi^*(t) = \phi(Lt/2\pi) + t$; note that ϕ^* is invariant under translations, rotations and scaling, which is of significant benefit in this application. We now expand ϕ^* as a Fourier series $\phi^* = \mu_0 + \sum_{k=1}^{\infty}(A_k \sin(\varphi_k + kt))$.

We describe each diatom contour using a 200 element vector consisting of 100 amplitude values and 100 corresponding phase angles obtained from Fourier descriptors. We found that this number of descriptors allows us to achieve accurate contour reconstructions. Reconstructing the original shape to produce a drawing of the contour, using the above description, is straightforward and is described in detail in [6].

4 Texture analysis and synthesis

Our goal here is to analyse the diatom silica shell patterns and represent them in a way suitable for synthesis. As we mentioned earlier, we restricted our approach to analysis of pennate diatom species with striae patterns on their shells. The striae are transverse lines of pores between the silica ribs coming out from the diatom's long axis (raphe-sternum or sternum). The patterns formed by the striae are characterised by frequency and orientation, as well as by the sternum shape. For simplicity, we model striae as straight, which is a good approximation in the majority of cases considered.

4.1 Texture analysis

In ADIAC [1], Gabor wavelets were used to detect the frequency and orientation of the striae and to segment the internal pattern. However, unless the pattern orientations and frequencies are known beforehand, or of a very limited range, a large bank of filters needs to be applied, which is a time consuming operation. In ADIAC, only 28 filters were used, covering a range of four different orientations and seven different frequencies.

Fourier analysis [4] provides a more general approach to detecting the frequency and orientation of patterns, and is more suitable for the purpose given the type of patterns and range of possible frequencies and orientations. We perform fast Fourier transform (FFT) [4] within a sliding window of size 48×48 at each pixel inside the diatom contour. This size ensures that at least three striae fit inside the window (at our image resolution) for robust detection of pattern orientation and frequency. A typical diatom photograph we worked with is approximately of a size 800×400 pixels.

For each window, we find the FFT amplitude (energy) [4] values corresponding to the Fourier coefficients. Then, we set to zero the components corresponding to the frequencies of 0, 1, 1/2, as we expect at least three striae in each window. We also set to zero the values corresponding to almost horizontal orientations, as we do not expect to find striae with such orientations. Finally, we find the maximum among the remaining FFT energy values to give the orientation and frequency. Thus we obtain three maps from FFT processing. The first one contains the stria orientation values for each pixel inside the diatom contour, the second contains the stria frequency for each pixel inside the diatom contour, and the third map contains energy values for each pixel inside the diatom contour (Fig. 2). We use the maps at a later stage to find the average stria orientation and frequency values in different areas of a diatom.

Apart from mapping the stria orientation and frequency, we also need to detect the borders of the central area of the diatom where there are no striae (the sternum or raphe-sternum). The energy map gives us some idea of where there are

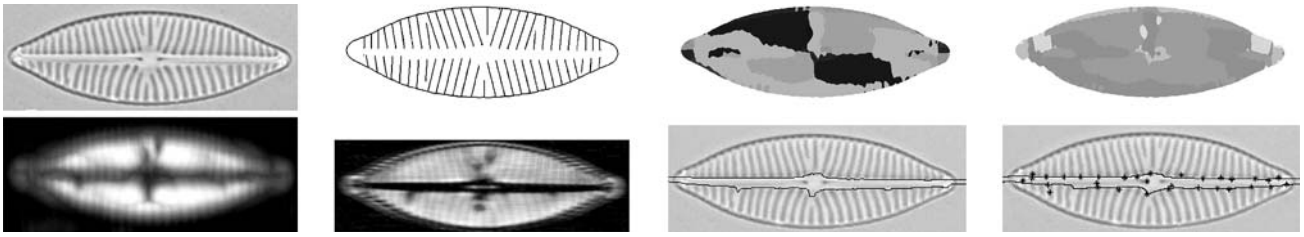

Fig. 2 From left to right top down: a photograph of a diatom, synthesised drawing, orientation map, frequency map, energy map (using 48 × 48 window), energy map (using 2 × 48 window), central part borders, fitted splines

striae. However, its borders are hard to pinpoint due to the large size of the sliding FFT window. We perform a second windowed FFT on the whole image, this time using a window of size 2 × 48, finding the largest peaks in the Fourier domain. We find the longitudinal borders of the sternum by traversing the energy values in each column of the map up and down from the centre, looking for the first value above the threshold, which we set experimentally at three quarters of the average energy value over the whole energy map. Finally, we fit a set of six cubic splines to both the top and bottom borders, thus describing each border with 19 spline control points, since the last control point of each spline serves as the first control point of the next spline (Fig. 2). We found that this representation gives us a good approximation to the original shape of the sternum. We use splines to approximate the shape of the sternum rather than the FD, which we used to represent the diatom shapes, due to the fact that the sternum does not have a clear border, as the diatoms do, and hence we need more smoothing when fitting curves to the extracted data.

To obtain parameter values characterising the texture, we split the inside of the diatom contour into a number of parts. In our experiments, we used six areas above the sternum and the same number below the sternum, which reflects on the variety of patterns typical pennate diatoms have. We find the average orientation and frequency inside each of these parts as the average of all orientation and frequency values weighted by the corresponding energy values.

The internal pattern of each diatom is described using a 100 element vector, where 76 elements are the coordinates of the 38 control points and the other 24 values are orientation and frequency values.

4.2 Texture synthesis

To draw the internal structure of the diatom, we draw lines representing the striae between the diatom contour and the edge of the sternum. This is done using the average orientation and frequency values in several areas inside the diatom contour.

To model or mimic actual valves satisfactorily, the requirements for the generated striae are that they should have appropriate orientation and frequency values, and they should be continuous across each area of different orientation and frequency. For example, if two striae diverge far from each other, another stria should appear in between, or if they converge, they should either merge or one of them should disappear eventually.

In our synthesis algorithm, we attempt to follow the way it is believed a diatom shell is formed naturally [12]. The striae are formed gradually: the ones near the centre of the diatom start growing first and may be partially completed by the time the striae further away from the centre start forming. We attempt to model this process in the algorithm outlined below.

1. Starting at the centre of the top sternum border, going out towards the right end of the diatom add one more pixel to the length of all existing striae, keeping all striae of orientations appropriate to the areas of the diatom they are located in, checking that they have not reached the diatom contour yet and that they are not too close (less than half of the striae spacing appropriate to the corresponding area of diatom) or too far (more than twice the striae spacing appropriate to the corresponding area of diatom) from the nearest stria on the left. The threshold values for the striae spacing were derived experimentally to imitate the underlying natural processes.
2. If the stria on the left is too close to the current stria, or the current stria has reached the diatom contour, then the current stria becomes "completed", and in that case no more pixels are added to it in the future.
3. If the stria on the left is too far away, then another stria is inserted between the two that have diverged too far.
4. After we have considered all existing striae on the right from the centre, and if we have not reached the right hand end of a diatom yet, we add one more stria to the right of the rightmost stria.
5. Repeat all the above steps until all the striae are "completed".
6. Repeat all the above steps for the other three quarters of the diatom starting at the centre and going out towards the ends of the diatom along the top or bottom of the sternum.

5 A model of life cycle variation of shape and texture

Having described how we extract and represent the shape and texture of individual specimens, we now discuss the problem of modelling the life cycle related variation within and between species.

The dimensionality of the extracted data describing the diatom outlines, texture and size is very high (200 Fourier descriptors, 100 parameters describing internal texture and one length parameter). The distribution of parameter vectors representing variation during the life-cycle of any one species is non-linear. We need to find a method to reduce the dimensionality of the space and to model the distributions. After careful consideration of a number of methods [6], we decided to model the distribution using a collection of principal curves (PC) [5].

A PC was first defined by Hastie and Stuetzel [5] as a smooth (C^∞) unit-speed 1D curve in $\mathbf{R}^p$ parameterised over $\Lambda \in \mathbf{R}^1$, satisfying the self-consistency condition $\boldsymbol{f}(x) = E_{\mathbf{Y}|g(\mathbf{Y})}\{\mathbf{Y} \mid g(\mathbf{Y}) = x\}, \forall x \in \Lambda \subseteq \mathbf{R}$, where E is the conditional average operator, $\mathbf{Y} \in \mathbf{R}^p$ and $g(\mathbf{Y})$ is the projection operator given by $g(\mathbf{Y}) = \sup_{\lambda \in \Lambda}\{\lambda : \| \mathbf{Y} - \boldsymbol{f}(\lambda) \| = \inf_{\mu \in \Lambda} \| \mathbf{Y} - \boldsymbol{f}(\mu) \|\}$.

Intuitively, a PC is a smooth curve passing through the "middle" of a data distribution. The above definition allows recursive estimation of a PC for a given data set. In practice, the curve is approximated using a number of knots and linear segments connecting them. There are several problems with the original definition of PC, which have given rise to several alternative definitions offering improved estimation algorithms. We adopt Chang's method [2] to model the life cycle shape trajectory of a single diatom species.

5.1 Applying principal curves to modelling diatom shape and texture variation

Prior to modelling the diatom shape and texture data we normalise the data (the set of parameter values described above for all specimens from all species) to have zero mean and standard deviation of one. We find the main modes of variation in the data set of all species through PCA. Then, we model the life cycle size, shape and texture variation in a single species using a principal curve going through the middle of the corresponding data subset created by the Fourier descriptors, texture parameters and size vectors projected into the eigenspace. To build a model incorporating several species, we simply compute a PC for each species using an appropriate training data subset.

We illustrate the modelled variation of *Gomphonema minutum* in Fig. 3, which shows the original data set and the PC fitted to the data. Note that the PC was fitted to the data in the fully dimensional space and projection into the space of the three largest eigenvectors serves an illustrative purpose only. The fitted PC approximately follows the growth trajectory of *Gomphonema minutum*, as that provides the main source of shape variation.

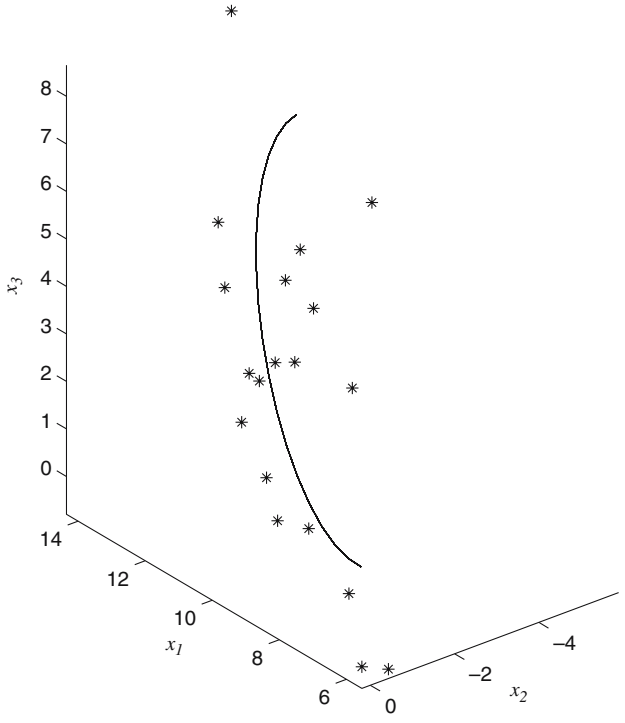

Fig. 3 A principal curve modeling *Gomphonema minutum* and the data used for its training, projected into the space of three largest eigenvectors

In Fig. 4 we show how we can synthesise drawings of diatoms from any point on a PC, allowing us to depict the diatom species at different stages in its life cycle. To produce a drawing of a diatom at any stage of its life cycle, we choose an appropriate point on the corresponding PC, transform it from the PCA space into the original space of Fourier descriptors and texture descriptors, and use them as described in Sects. 3 and 4.2.

5.2 Applying the model of shape and texture variation

We may now apply the PC model to automatic drawing production and identification of new diatom specimens in the form of photographs or drawings. To identify an unknown diatom, firstly, we obtain shape and internal texture parameter measurements as described earlier, giving us a vector of parameter values. After obtaining the parameter vector, we assign it to one of the species included in the model on the basis of the shortest Euclidean distance between the point representing the specimen and any PC. Other distance measures could be used to assign a specimen to one of the PCs, based, e.g. on the Mahalanobis distance. However, given the small number of training samples, such distances may be unreliable.

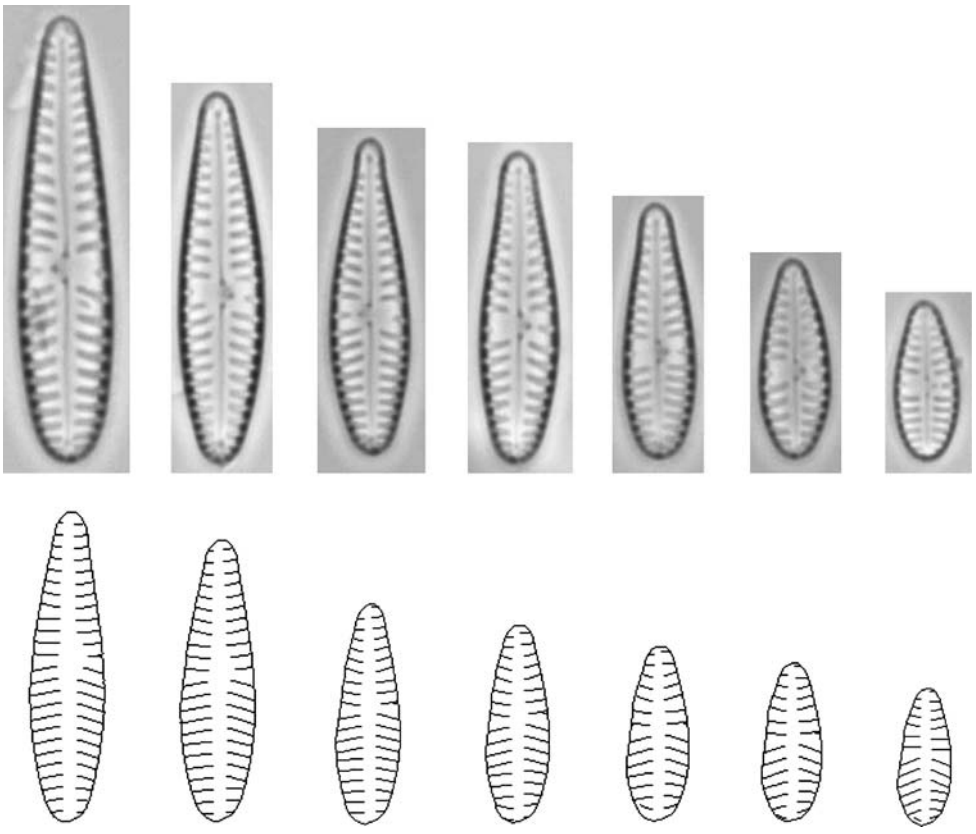

Fig. 4 *Top row* representative subset of photographs used to train a model of *Gomphonema minutum*. *Bottom row* drawings of *Gomphonema minutum* generated automatically from the principal curve modelling shape, texture and size of the diatom through its life cycle

6 Experiments

In the following sections, we present the results of an independent expert assessment of the quality of the automatically produced drawings, followed by results of extensive experiments testing automatic diatom identification in photographs and drawings using our model. The photographs used for this research were originally obtained for the ADIAC project and have a resolution of 8 pixels/μm or more; for more detailed information, please refer to [1].

6.1 Diatom analysis and automatic drawing generation

Our test data included 303 photographs of 13 different species, such as *Placoneis constans, Sellaphora bacillum, Navicula rhynchocephala, Gomphonema augur, Cymbella hybrida, Cymbella subaequalis, Navicula capitata, Caloneis amphisbaena, Navicula menisculus, Gomphonema minutum, Gomphonema* sp., *Navicula radiosa, Navicula viridula* (examples are shown in Fig. 5). We used these to produce drawings directly from each photograph. The quality of the drawings produced degraded gracefully with decreasing quality of the original photographs. Please note, that due to the reduced size of the photographs, it may be difficult to see the stria orientation and frequency of *Caloneis amphisbaena* in Fig. 5. A better representation of the stria orientation and frequency for this species can be seen in a drawing made by a biologist in Fig. 6.

To assess the quality of the produced drawings in a non-biased experiment, we supplied an external expert in diatoms with the collection of 303 produced drawings in a random order. The expert had not been informed of the names or the number of the diatom species in the selection and was asked to identify the species in each drawing. In total, 9 out of 13 present species were identified correctly, with 140 out of 303 diatoms in drawings identified to the species, and 225 out of 303 diatoms identified to the genus. The expert commented that in some misidentified cases, he had not previously encountered the species; while in other cases, a detailed representation of the raphe slits was required in the drawings to make a correct identification down to the species. On the whole, the expert was positive about the result of the experiment, pointing out that he correctly identified 9 out of 13 present species, which is comparable to the results of identification experiments in diatom photographs achieved by human experts in the ADIAC project [1].

6.2 Identifying diatoms from photographs

In the following experiments, we built a model of 13 diatom species, and tested its accuracy in identification experiments.

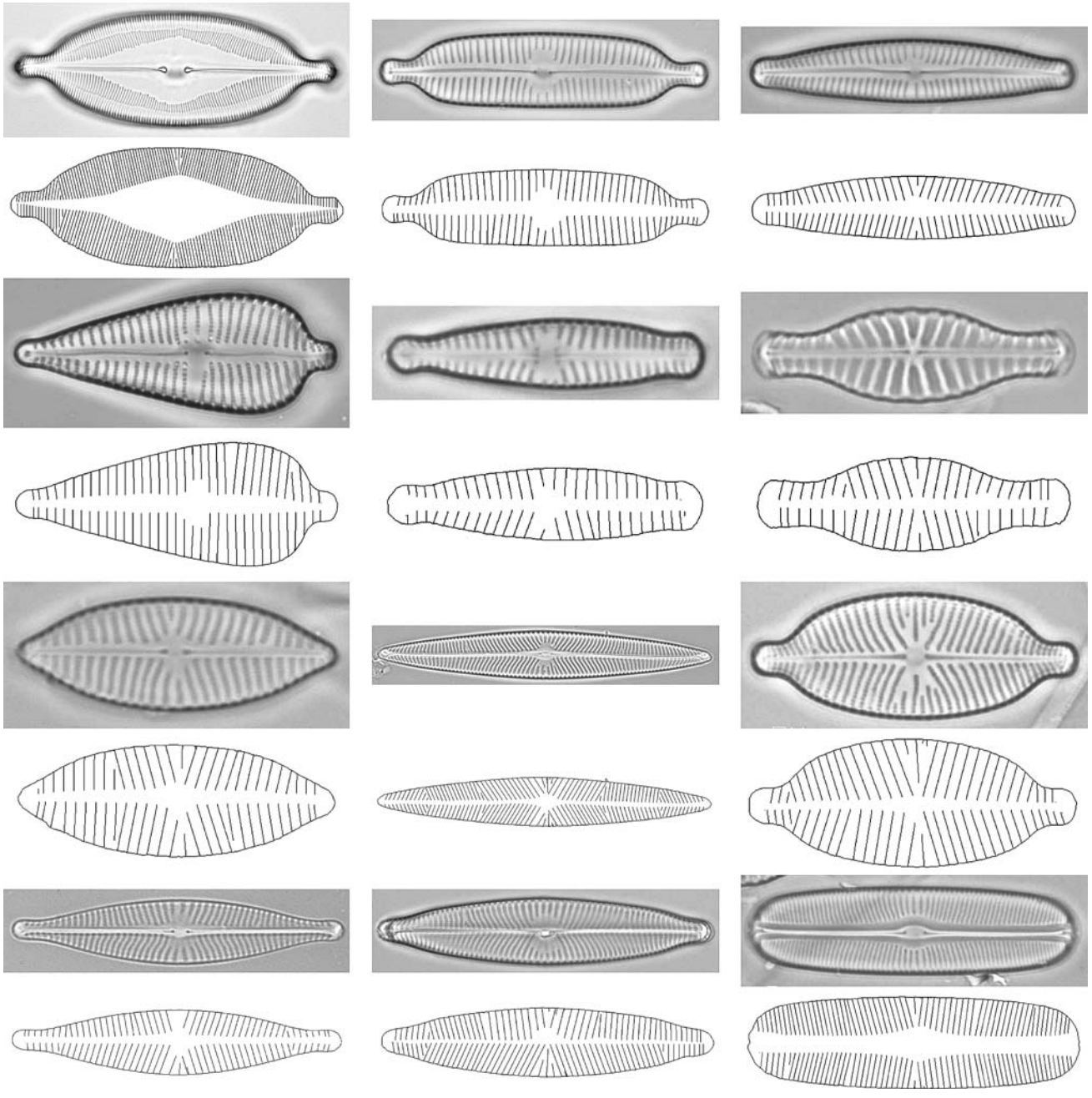

Fig. 5 Photographs and drawings generated automatically from the photographs of 12 species. The species are in the following order (*top-down*, *left-to-right*): *Caloneis amphisbaena, Cymbella hybrida, Cymbella subaequalis, Gomphonema augur, Gomphonema sp., Navicula capitata, Navicula menisculus, Navicula radiosa, Placoneis constans, Navicula rhynchocephala, Navicula viridula, Sellaphora bacillum*

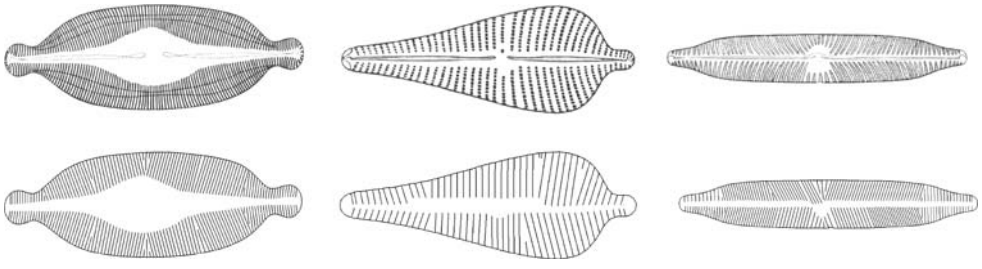

Fig. 6 Drawings of *Caloneis amphisbaena*, *Gomphonema augur* and *Navicula viridula* specimens made by biologists and the drawings automatically generated by the system from the original drawings

Table 1 Confusion matrix for contour shape and diatom length

	C. am	*C. hy*	*C. su*	*G. au*	*G. mi*	*G. sp*	*N. ca*	*N. me*	*N. ra*	*P. co*	*N. rh*	*N. vi*	*S. ba*
C.am	10	0	0	0	0	0	0	0	0	0	0	0	0
C.hy	3	14	4	0	0	0	0	0	0	0	0	0	0
C.su	0	0	12	0	0	0	0	1	0	0	0	0	0
G.au	0	0	0	4	0	0	0	0	0	0	0	0	0
G.mi	0	0	0	0	16	0	0	0	0	0	0	0	0
G.sp	0	0	0	0	2	19	0	0	0	0	0	0	0
N.ca	0	0	0	0	0	0	13	0	0	0	0	0	0
N.me	0	0	1	0	0	0	0	14	0	0	1	0	0
N.ra	0	0	0	0	0	0	0	0	5	0	1	0	0
N.co	1	0	0	0	0	0	2	2	0	17	0	0	7
N.rh	0	0	0	0	1	0	0	2	0	0	13	2	0
N.vi	0	1	0	1	0	0	0	0	0	0	1	6	0
C.ba	0	0	0	0	0	0	2	0	0	0	0	0	0

The columns represent the real species and the rows represent identified species

Table 2 Confusion matrix for texture data only

	C. am	*C. hy*	*C. su*	*G. au*	*G. mi*	*G. sp*	*N. ca*	*N. me*	*N. ra*	*P. co*	*N. rh*	*N. vi*	*S. ba*
C.am	14	0	0	0	0	0	0	0	0	0	0	0	0
C.hy	0	15	0	0	0	0	0	0	0	0	0	0	0
C.su	0	0	16	0	0	0	0	0	0	0	0	0	0
G.au	0	0	0	4	0	0	0	0	0	0	0	0	0
G.mi	0	0	0	0	16	3	0	0	0	0	0	0	0
G.sp	0	0	1	0	3	16	0	0	0	0	0	0	0
N.ca	0	0	0	0	0	0	16	0	0	0	0	0	0
N.me	0	0	0	0	0	0	1	19	0	0	0	0	0
N.ra	0	0	0	0	0	0	0	0	5	0	0	0	0
N.co	0	0	0	1	0	0	0	0	0	17	1	0	0
N.rh	0	0	0	0	0	0	0	0	0	0	15	1	0
N.vi	0	0	0	0	0	0	0	0	0	0	0	7	0
C.ba	0	0	0	0	0	0	0	0	0	0	0	0	7

We selected the 178 best photographs out of those described above, to make sure that the models produced were reliable and did not contain any errors from the analysis stage, and built a model of diatom shape, size and internal texture variation over the life cycles of the above 13 species.

In the first experiment, we measured the accuracy of our model in identification experiments with photographs. We used the standard "leave one out" approach, where the model was trained on all the specimens apart from one and the remaining specimen was identified using the trained model; this was repeated for each specimen out of the total 178. We compared the identification accuracy between a model trained on the shape and contour length data, a model trained on the texture data only, and a model trained on shape, texture and contour length data.

The error rate when using the external contour and length data was 19.66%. Table 1 displays the confusion matrix in this experiment, where the columns represent the real species and the rows represent identified species. For the texture data only, the error rate was 6.18% (Table 2). Using shape, texture and contour length data the error rate decreased to 3.37% (Table 3), which is a significant improvement over using either contour or texture data alone, and is similar to the error rate achieved in the ADIAC project in similar experiments. Note, however, the data set used in the ADIAC included a larger number of species, some of which had non-stria patterns.

We used several other standard classification methods on the same data set in leave-one-out experiments for comparison with our model. Using Rifkin's implementation of a support vector machine [11] with a linear kernel gave us a classification error rate of 6.18% on the normalised data, and a 19.1% error rate was achieved using an OC1 decision tree approach [10] on the raw data without prior normalisation.

The identification experiments presented in this section showed that our system performed better than the general classification methods we have tested on our data, and achieved similar identification rates to the system developed specifically for diatom identification in the ADIAC project.

6.3 Identifying diatoms from drawings

In this section, we tested the accuracy of our system when identifying diatoms in drawing representation. We tested our

Table 3 Confusion matrix for contour shape, texture and diatom length

	C. am	*C. hy*	*C. su*	*G. au*	*G. mi*	*G. sp*	*N. ca*	*N. me*	*N. ra*	*P. co*	*N. rh*	*N. vi*	*S. ba*
C.am	14	0	0	0	0	0	0	0	0	0	0	0	0
C.hy	0	15	0	0	0	0	0	0	0	0	0	0	0
C.su	0	0	16	0	0	0	0	0	0	0	0	0	0
G.au	0	0	0	4	0	0	0	0	0	0	0	0	0
G.mi	0	0	0	0	17	0	0	0	0	0	0	0	0
G.sp	0	0	0	0	2	19	0	0	0	0	0	0	0
N.ca	0	0	0	0	0	0	17	0	0	0	0	0	0
N.me	0	0	1	0	0	0	0	19	0	0	0	0	0
N.ra	0	0	0	0	0	0	0	0	5	0	0	0	0
N.co	0	0	0	1	0	0	0	0	0	17	1	0	0
N.rh	0	0	0	0	0	0	0	0	0	0	15	1	0
N.vi	0	0	0	0	0	0	0	0	0	0	0	7	0
C.ba	0	0	0	0	0	0	0	0	0	0	0	0	7

Table 4 Confusion matrix for drawings

	C. am	*C. hy*	*C. su*	*G. au*	*G. mi*	*G. sp*	*N. ca*	*N. me*	*N. ra*	*P. co*	*N. rh*	*N. vi*	*S. ba*
C.am	1	0	0	0	0	0	0	0	0	0	0	0	0
C.hy	0	0	0	0	0	0	0	0	0	0	0	0	0
C.su	0	0	0	0	0	0	0	0	0	0	0	0	0
G.au	0	0	0	1	0	0	0	0	0	0	0	0	0
G.mi	0	0	0	0	0	0	0	0	0	0	1	0	0
G.sp	0	0	0	0	0	0	0	0	0	0	0	0	0
N.ca	0	0	0	0	0	0	1	0	0	0	0	0	0
N.me	0	0	0	0	0	0	0	0	0	0	0	0	0
N.ra	0	0	0	0	0	0	0	0	1	0	0	0	0
N.co	0	1	0	0	0	0	0	0	0	0	0	0	0
N.rh	0	0	0	0	0	0	0	0	0	0	0	1	0
N.vi	0	0	0	0	0	0	0	0	0	0	0	0	0
C.ba	0	0	0	0	0	0	0	0	0	0	0	0	0

system on a selection of only seven drawings representing seven different species, as the number and variety of digitised drawings of diatoms available to us was quite limited.

To identify a diatom in a drawing we used the same procedure as for the photographs. Four drawings were identified correctly (Table 4). In the two out of three misidentified drawings, the stria frequency extracted from the drawing was double the real value due to the artistic technique used in the drawings. After we manually corrected the frequency values for these two drawings, one of them was identified correctly, but another one was still misidentified, with a different species this time.

To assess the success of our system in identifying diatoms from drawings made by biologists, we compared the drawings generated automatically by our system to the original drawings made by biologists (Fig. 6). In the first drawing made by a biologist, where the striae are represented by thick lines, the stria frequency and orientation were detected correctly, which you can see in the drawing produced automatically. In the second drawing, where the striae are represented by intermittent thick lines, the stria frequency and orientation were detected closely to the real values as well. However, in the third drawing, where each stria is represented by two vertical thin lines, while the stria orientation was detected correctly, the frequency was estimated as double the real value. In the future, this problem could be corrected through additional tests.

While the results of the identification experiments in this section were not as good as the results of automatic identification experiments in photographs, the accuracy was still similar to that of human experts [1].

7 Discussion

When fitting a PC to the diatom data, we noticed that even though we do not need a large number of specimens for the curve to approximate the diatom growth trajectory, we do need the specimens to sample different stages in diatom development well. For example, if the training data set contains a number of small specimens and only one large specimen, the principal curve may be biased towards the small specimens, thus failing to model the true life cycle trajectory. The latter may manifest itself in the system in unexpected

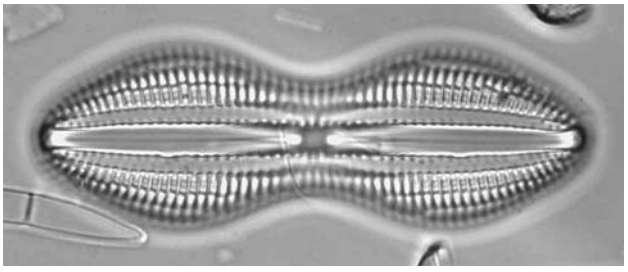

Fig. 7 A diatom with a complex stria pattern

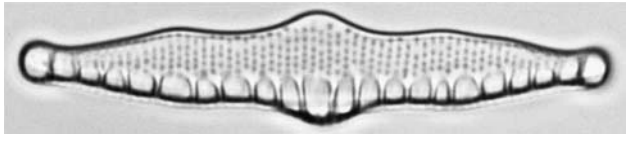

Fig. 8 A diatom with a non-stria pattern

misidentifications. For example, in our experiments, *Sellaphora bacillum* was misidentified as *Placoneis constans* when using only the contour shape and length parameters. The difference in the shape of these two species is apparent, meaning that one of the PCs did not represent the species well enough to enable it to be separated from other species. However, when we added the texture data to the model, the problem was corrected with all *Sellaphora bacillum* specimens identified correctly. Thus the addition of texture parameters to the model provides extra stability, as we hypothesised in [6]. We believe that unexplained misidentifications in the drawings were also caused by this problem. The solution to the problem would be to make sure that the training specimens sample the underlying data distributions well.

We developed our model for the majority of pennate diatoms, which have simple stria patterns, but it could be adapted quite easily to cope with more complex morphologies. For example, Diploneis (Fig. 7) has a complex stria structure, with longitudinal chambers within the cell wall. For this, our system could be extended fairly easily by increasing the number of areas used to record the stria pattern, subdividing each of the six sectors on either side of the raphe-sternum horizontally, into two or more compartments.

For diatom patterns of a completely different nature (Fig. 8), firstly, varying types of pattern would have to be identified and then an appropriate set of tools applied to obtain a set of parameters describing the pattern.

8 Conclusions

We have presented a means of modelling shape, size and texture variation in multiple diatom species. One of the main novelties of our research is in the ability of our model to map between drawings and photographs of diatoms, thus making it possible to automatically identify diatoms in traditional drawings made by biologists and match them with the diatoms represented in photographic form. The model can be built from data automatically extracted from photographs or drawings, and is based on diatom features which are present in both photographs and drawings and used for diatom identification.

The model is suitable for identification of previously unseen diatoms represented in photographic or drawing form. The model is also suitable for reconstructing drawings of diatoms at any stages of their life cycles, including those not explicitly present in the original training set. We presented drawings produced by our methods and the results of identification experiments. Identification experiments achieved a similar accuracy to those resulting from the ADIAC project; however, they were conducted on a smaller data set.

Currently biologists are working on applying the system presented to classification problems in a biological context (taxonomy).

Acknowledgements This project was funded by the BBSRC/EPSRC under the Bioinformatics Programme (grant no. 754/BIO14262). In our experiments, we used Chang's implementation of Probabilistic Principal Curves [2] as a part of LANS Pattern Recognition Toolbox, Murthy's implementation of an OC1 decision tree approach [10] and Rifkin's implementation of SVM [11]. The data set of diatom photographs, used in the project, was provided to us by the ADIAC [1] partners together with the extracted diatom contours.

References

1. du Buf, H., Bayer, M.M. (eds.): Automatic Diatom Identification. vol. 51, Series in Machine Perception and Artificial Intelligence, World Scientific, Singapore (2002)
2. Chang, K., Ghosh, J.: A unified model for probabilistic principal surfaces. IEEE Trans. Pattern Anal. Mach. Intell. **23**(1), 22–41 (2001)
3. Goldman, N., Paddock, T.B.B., Shaw, K.M.: Quantitative analysis of shape variation in populations of *Surirella fastuosa*. Diatom Res. **5**, 25–42 (1990)
4. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Addison-Wesley, Reading (1993)
5. Hastie, T., Stuetzel, W.: Principal curves. J. Am. Stat. Assoc. **84**(406), 502–516 (1989)
6. Hicks, Y.A., Marshall, A.D., Rosin, P.L., Martin, R.R., Bayer, M.M., Mann, D.G.: Modelling life cycle related and individual shape variation in biological specimens. Proc. BMVC **1**, 323–332 (2002)
7. Mann, D.G., Droop, S.J.M.: Biodiversity, biogeography and conservation of diatoms. Hydrobiologia **336**, 19–32 (1996)
8. Mann, D.G.: The species concept in diatoms (Phycological Reviews 18). Phycologia **38**, 437–495 (1999)
9. Mou, D., Stoermer, E.: Separating *Tabellaria (Bacillariophyceae)* shape groups based on fourier descriptors. J. Phycol. **28**, 386–395 (1992)
10. Murthy, S., Kasif, S., Salzberg, S.: System for induction of oblique decision trees. J. Artif. Intell. Res. **2**, 1–32 (1994)
11. Rifkin, R.: SvmFu package. Available from http://five-percent-nation.mit.edu/SvmFu/

12. Round, F.E., Crawford, R.M., Mann, D.G.: The diatoms. Biology and Morphology of the Genera. Cambridge University Press, Cambridge (1990)
13. Stoermer, E.F., Ladewski, T.B.: Quantitative Analysis of Shape Variation in Type and Modern Populations of *Gomphoneis herculeana*. **73**, 347–386, Nova Hedwigia, Beih (1982)
14. Zahn, C.T., Roskies, R.Z.: Fourier descriptors for plane closed curves. IEEE Trans. Comput. **c-21**(3), 269–281 (1972)
15. British Diatomists of the 19th Century database. Available at http://rbg-web2.rbge.org.uk/DIADIST/dia_ intro.htm