

# BUILDING SHAPE AND TEXTURE MODELS OF DIATOMS FOR ANALYSIS AND SYNTHESIS OF DRAWINGS AND IDENTIFICATION

Y.Hicks, D.Marshall, R.R.Martin, P.L.Rosin  
Cardiff School of Computer Science  
UK  
email: y.a.hicks, dave,  
ralph, p.l.rosin @cs.cf.ac.uk

S.Droop, D.G.Mann  
Royal Botanic Garden Edinburgh  
UK  
email: D.Mann@rbge.org.uk

## Abstract

We describe tools for automatic identification of diatoms by comparing their 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, paleoecology and forensic science. The model we build represents life cycle and natural variation of both external shape and internal texture over multiple species and is based on *principal curves*. The model is also suitable for automatically producing 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 to analyse photographs and drawings, present our model of diatom shape and texture variation, and illustrate our approach with a collection of drawings synthesised from our model and derived from example photographs. Finally, we present the results of identification experiments using photographs and drawings.

**Keywords:** *Classification, automatic drawing synthesis, principal curves, diatoms.*

## 1 Introduction

Diatoms are unicellular algae with a highly ornate silica shell around each specimen. The shell contains two larger elements called valves, one on either side of the cell, which bear species-specific patterns. Identification of diatoms, *i.e.* assigning a new specimen to one of the known species, has applications in many disciplines, including ecology, paleoecology and forensic science. Specimens are usually identified by highly trained specialists by considering diatom morphological characteristics, including shape and texture, and comparing them to photographs and drawings of previously identified specimens. This task is challenging due to a huge number of diatom species, similarities between species and life cycle related changes in shape and texture.

Recently there have been various efforts in quantitative analysis of diatom shape variation [2, 6, 7]. A system for automatic identification of diatom specimens in photographs, based on the silica shell shape, size and pattern characteristics, was developed in the ADIAC project [1]. We seek to extend such capabilities through the inclusion of biological drawings. There is a wealth of diatom specimen drawings in the biological literature accumulated over many years. The drawings contain mainly the salient information required for identification and thus may serve as models of each species. Hence, including digitised drawings in the system and providing the ability to compare photographs and drawings has significant benefits for the biological community.

A different issue is automatic production of diatom drawings. Diatom type specimens are traditionally defined in the taxonomic literature using drawings and, although photographs have been used much more

often in the last 20 years, there remain a significant number of genera for which drawings are more appropriate. Automating the production of drawings would be especially useful as it is a time consuming and difficult task (Figure 1).

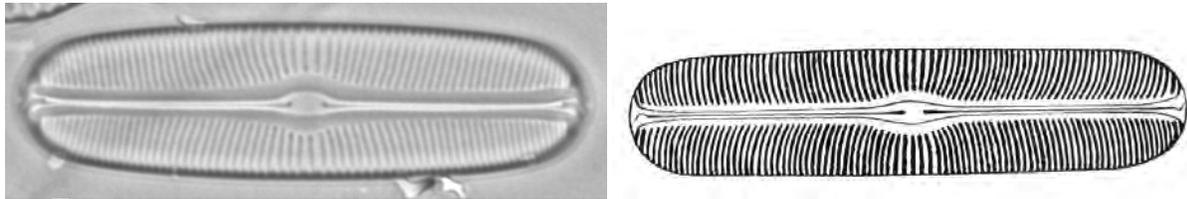


Figure 1: A photograph of a diatom valve and a drawing of a similar valve by a biologist.

In recent years, the problem of finding a mapping between photographs and drawings came to the attention of computer vision and computer graphics communities. For example, A.Hertzmann *et al.* [4] learn the mapping through correspondence of low-level pixel statistics in a drawing and a photograph. However, such approaches are unsuitable for the task at hand due to their requirement for an exact match between the drawings and the photographs, which is usually not available in biological materials.

Our approach is to transform the high-dimensional image space of both photographs and drawings into a lower-dimensional space where only relevant features are represented. We then use this space for the comparison of different specimens as well as for automatic production of drawings.

In our research we go further by not only developing a system capable of identifying new diatom specimens, but also producing a model describing life cycle related variation in the shape and pattern of multiple diatom species and suitable for synthesising example drawings of the species.

In this article we present methods for analysing diatom shape and texture, produce a model representing variation of shape and texture in multiple diatom species, and illustrate our approach with a number of drawings generated automatically from the model and original photographs. We finish with presenting the results of identification experiments.

## 2 External contour analysis and synthesis

Many diatom valves are sufficiently flat to give a repeatable view in all photographs. Traditionally, when analysing diatom shape, diatomists performed 2D contour analysis in this view. However, due to various reasons it is not an easy task to extract the contours from photographs automatically. Overlapping debris and diffraction effects may make it hard to locate the contour. In the course of ADIAC [1], several sophisticated methods for contour extraction have been developed. In this article we use the extracted contours provided to us from the ADIAC project.

To represent diatom contours in a compact way we use Fourier descriptors as we explain in [5]. Thus each diatom contour is represented with a 200 element vector consisting of 100 amplitude values and 100 corresponding phase angles obtained from Fourier descriptors. It is possible to reconstruct the shape of the diatom from these values, as we do in [5].

## 3 Texture analysis

Our goal here is to analyse the diatom silica shell patterns and represent them in a way suitable for synthesis. 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. In the initial system reported in this article we restricted our approach to the analysis of pennate diatom species with striae patterns on their shells; most diatoms are of this kind. The striae are transverse lines of pores between the silica ribs coming out from

the diatom's long axes (raphe-sternum or sternum). The patterns formed by the striae are characterised by frequency and orientation. For simplicity, we model striae as straight, which is a good approximation in the majority of cases considered.

In ADIAC [1], Gabor wavelets were used to detect the frequency and orientation of the striae and to segment the diatom shells. However, unless the pattern orientation and frequency are known beforehand, or their range is very limited, a large bank of filters needs to be applied. In ADIAC, 28 filters were used, covering a range of 4 different orientations and 7 different frequencies.

Fourier analysis provides a more general approach to detecting the frequency and orientation of the striae patterns, and is more suitable for the purpose given the range of possible frequencies and orientations, thus it is our chosen tool. We perform an FFT within a sliding window of size 48 x 48 at each pixel inside the diatom contour. This size ensures that at least 3 striae fit inside the window (at our image resolution) for robust detection of pattern orientation and frequency.

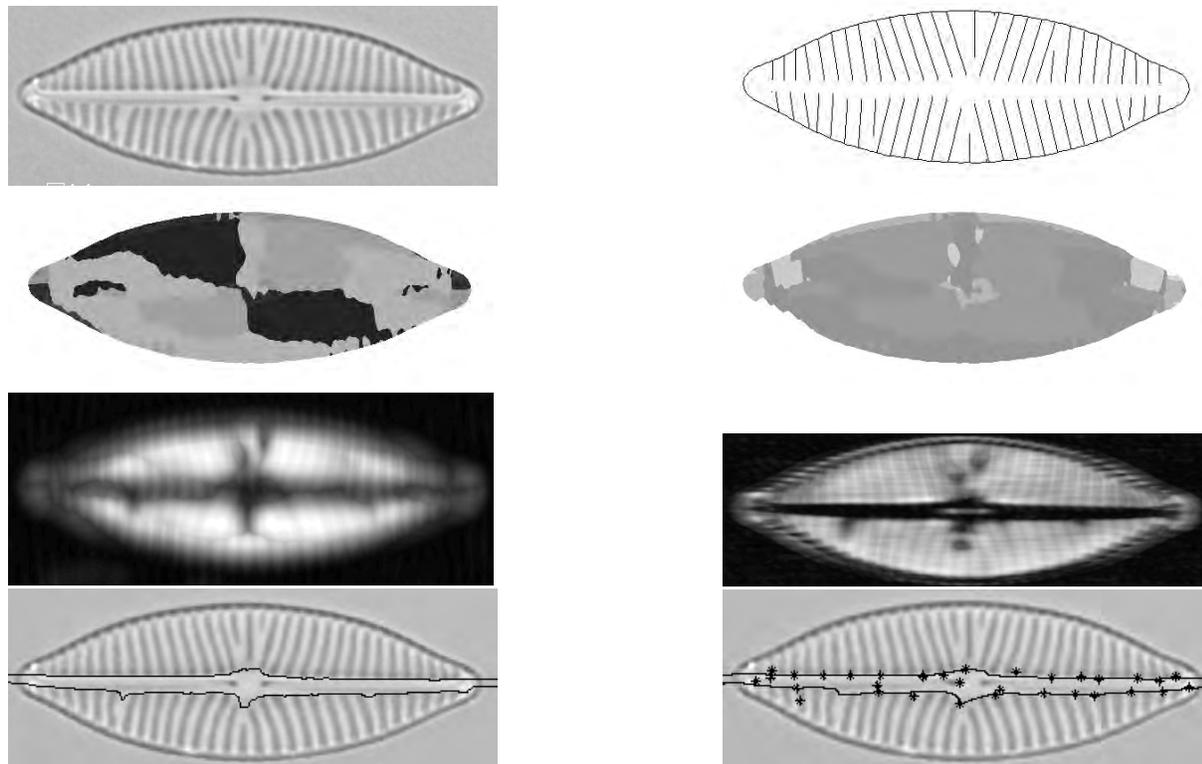


Figure 2: From left to right top down: a photograph of a diatom, synthesised drawing, orientation map, frequency map, energy map (using 48x48 window), energy map (using 2x48 window), central part borders, fitted splines together with control points.

For each window we find the energy values corresponding to the Fourier coefficients. Then we set to zero the DC Fourier component as well as the values corresponding to the frequencies of 1 and 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 in 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 the run of the FFT. The first one contains the striae orientation values for each pixel inside the diatom contour, the second contains the striae frequency for each pixel inside the diatom contour, and the third map contains energy values (FFT amplitude) for each pixel inside the diatom contour (Figure 2). We use these maps at a later stage to find the average striae orientation and frequency values in different areas of a diatom.

Apart from knowing the striae orientation and frequency, we also need to detect the borders of the

central area of the diatom with no striae (the sternum or raphe-sternum). The energy map gives us some idea of where there are striae. However, its borders are hard to pinpoint due to the size of the sliding FFT window. We perform a second windowed FFT on the whole image, this time using a window of size  $2 \times 48$ , finding the largest peaks in the Fourier domain in the same way as before. However, this time we are only interested in the energy map. We find the vertical borders of the central area 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 at three quarters of the average energy value over the whole energy map. Finally, we fit a set of cubic splines into the top and bottom borders, thus describing each border with 19 spline control points.

To obtain parameter values characterising the texture, we split the inside of the diatom contour into 12 parts, 6 above the sternum and 6 below. The borders of the parts are determined by splitting the curves approximating the top and the bottom borders of the central diatom area into equal lengths. We find the average orientation and frequency inside each of these parts as the weighted average of all orientation and frequency values, where the weights are 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 another 24 values are orientation and frequency values.

In conclusion, we would like to point out that the method presented above is suitable for the analysis of diatoms represented in both photographic and drawing form.

## 4 Texture synthesis

To draw the internal structure of the diatom, we draw lines representing the striae between the external contour and the sternum borders. 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 the appropriate orientation and frequency values, and should be continuous across each area of different orientation and frequency. For example, if two striae diverge too far from each other, another stria should appear in between, or if they converge, eventually they should either merge or one of them should disappear.

In our synthesis algorithm we attempt to follow the way it is believed the diatom shell is formed naturally [9]. 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 our iterative synthesis 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 external 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 contour of the diatom, we add one more stria to the right of the rightmost stria at the distance appropriate for the area.

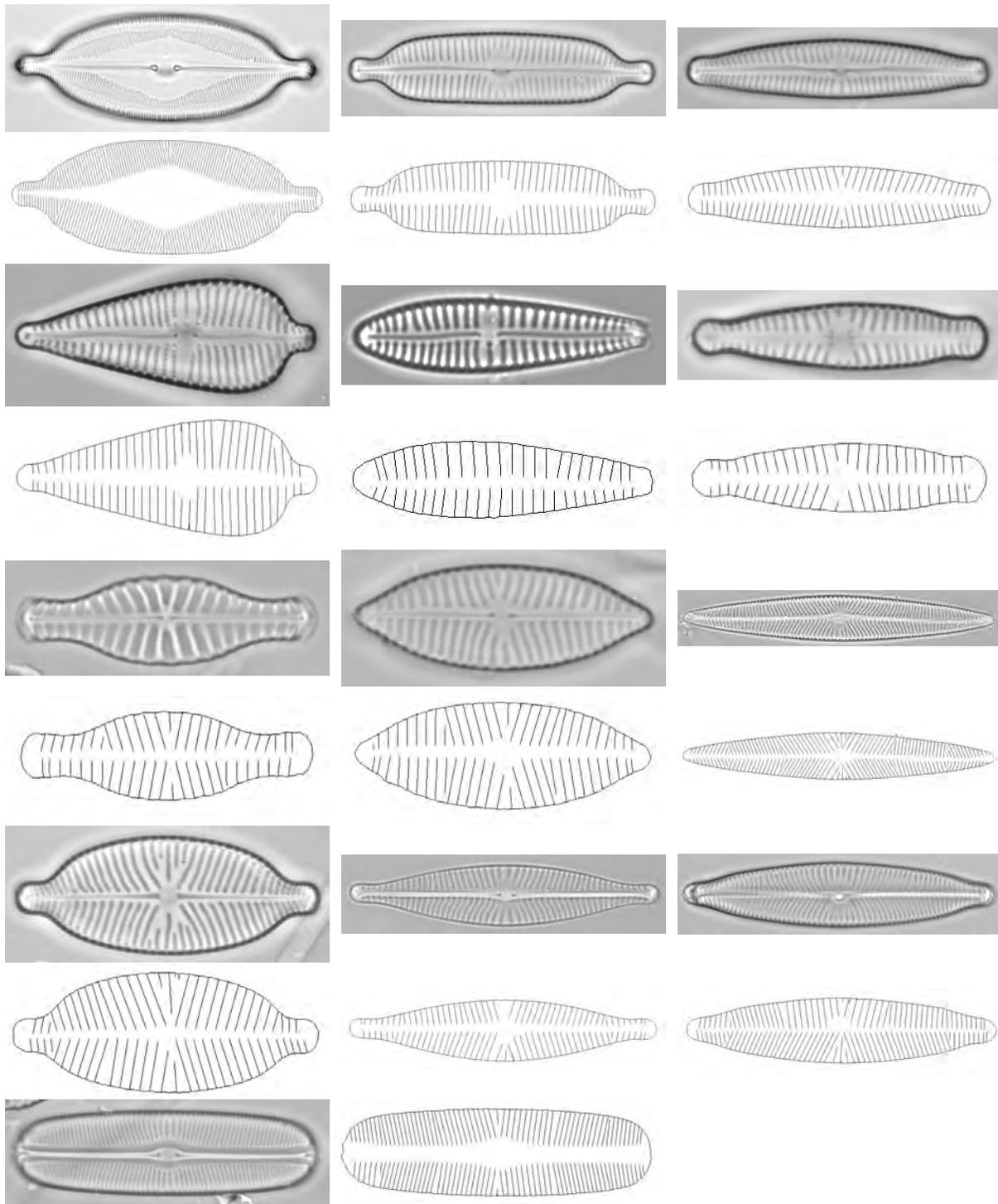


Figure 3: Photographs and drawings generated automatically from the photographs of 13 species. The species are in the following order: *Caloneis amphisbaena*, *Cymbella hybrida*, *Cymbella subaequalis*, *Gomphonema augur*, *Gomphonema minutum*, *Gomphonema species 1*, *Navicula capitata*, *Navicula menisculus*, *Navicula radiosa*, *Navicula constans*, *Navicula rhynchocephala*, *Navicula viridula*, *Sellaphora bacillum*. Please note that the original images are very high resolution and contain high frequency information which may not be adequately printed or displayed on some devices.

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 shape and texture

Previously [5], we presented a model of shape variation during the life cycle of several diatom species. The model was based on a collection of principal curves, where each curve modelled the growth trajectory of a diatom species. Individual shape variations within species are defined in the dimensions orthogonal to the principal curve.

Principal curves were first defined by Hastie and Stuetzle [3]. Intuitively, a principal curve is a smooth curve passing through the “middle” of a data distribution. Principal curves are estimated recursively for a given data set. In practice the curves are approximated with a number of knots and linear segments connecting them.

We have now extended our earlier model based on diatom contours to represent diatom texture as well. Prior to modelling the diatom shape and texture data (the set of parameter values described in Sections 2 and 3, for all specimens from all species) we normalise the data to have zero mean and standard deviation of one. We find main modes of variation in the data of all species through PCA. Then we model the life cycle shape and internal texture variation in each species using a principal curve going through the middle of the corresponding data set. This approach allows us to extend the model to include a new species easily, which is more difficult for a decision-based diatom identification method [1].

## 6 Experiments

### 6.1 Diatom analysis and automatic drawing generation

Our test data includes over 300 photographs of 13 different species, namely, *Navicula constans*, *Sellaphora bacillum*, *Navicula rhynchocephala*, *Gomphonema augur*, *Cymbella hybrida*, *Cymbella subaequalis*, *Navicula capitata*, *Caloneis amphisbaena*, *Navicula menisculus*, *Gomphonema minutum*, *Gomphonema species 1*, *Navicula radiosa*, *Navicula viridula* (examples are shown in Figure 3). We used these to produce drawings directly from each photograph. The quality of the produced drawings 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 striae orientation and frequency of *Caloneis amphisbaena* in Figure 3.

### 6.2 Building a model and reconstructing drawings from the model

For this experiment, we selected the best quality photographs described in the previous section to make sure that the models produced were reliable and did not contain any errors from the analysis stage. The number of the specimens in each species set ranged from 5 for *Gomphonema augur* and *Navicula radiosa* to 20 for *Gomphonema minutum*, giving a total of 178 specimens. Prior to using principal curves to model the diatom shape data, we normalised the data and then found the main modes of variation in the data set of all species through PCA, as described earlier.

We built a model of diatom shape, length and internal texture variation over the life cycles of the above 13 species by fitting an individual principal curve to each of the available 13 data sets (Figure 6.2).

In Figure 5 we synthesise the drawings of diatoms from the principal curve nodes depicting the diatoms at different stages in their life cycle. Note that there may be no corresponding photograph for that stage – here the drawings are generated solely from the model, not directly from a photograph.

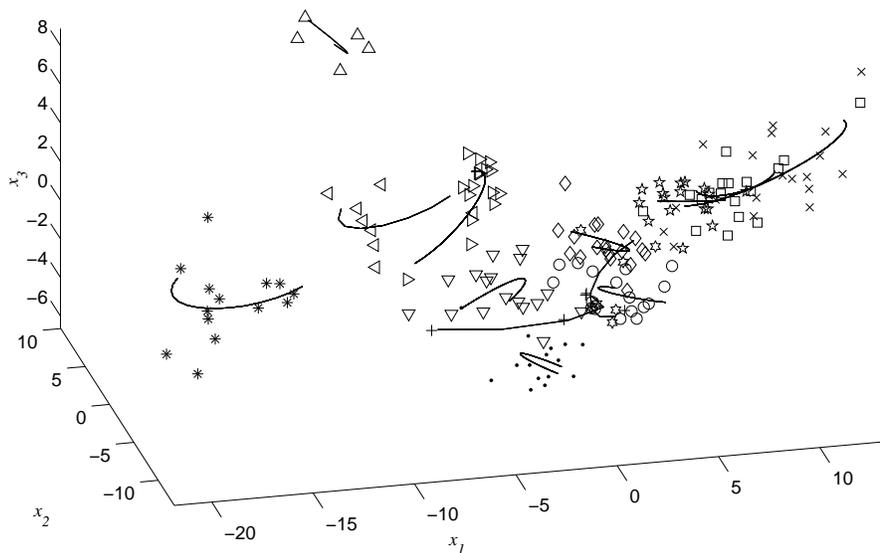


Figure 4: Principal curves and the data used for their training, projected into the space of three largest eigenvectors. Different species are represented with different symbols.



Figure 5: Some of the *Gomphonema minutum* photographs used for training a principal curve and drawings generated automatically from the principal curve at other stages of the life cycle.

### 6.3 Identifying diatoms from photographs and drawings using our model

The first experiment consisted of identifying diatoms whose images were not used for constructing the model. For this experiment 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. We repeated the experiment omitting each specimen out of the total 178 used in Section 6.2. We compared the identification accuracy between a model trained on the diatom shape and length data, a model trained on the texture data only, and a model trained on shape, texture and length data.

The error rate when using the external contour and length data was 19.66%. For the texture data only, the error rate was 6.18%. Using shape, texture and length data the error rate decreased to 3.37%, which is a significant improvement to using either contour or texture data alone, and is similar to the error rate achieved in the ADIAC project in similar experiments. However, the data set used in the ADIAC included a larger number of species, some of which had non-striae patterns.

We used several other standard classification methods on the same data set in leave-one-out experiments for comparison with our model. Using a support vector machine (SVM), developed by Ryan Rifkin at MIT’s Center for Biological and Computational Learning 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 OC1 decision tree approach [8] on the raw data without prior normalisation.

To identify a diatom in a drawing we used the same procedure as for the photographs. We obtained parameter values by image analysis of seven drawings of seven different diatom species also represented in the above photograph set. Four drawings were identified correctly. In the two out of three misidentified

drawings, the striae frequency was found to be double the real value due to the artistic technique used in the drawings. After we manually corrected the frequency values for these drawings, one more was identified correctly.

## 7 Evaluation and future work

We have presented a means of modelling shape, length and texture variation in multiple diatom species. The model is built from data automatically extracted from photographs, 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. It is also suitable for reconstructing drawings of diatoms at any stages of their life cycles, including those not explicitly represented in the original training set.

We have 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, ADIAC data set was larger and included some diatoms with non-striae patterns.

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

## 8 Acknowledgments

This project is funded by the BBSRC/EPSRC under the Bioinformatics Programme, grant 754/BIO14261. In our experiments we used Chang's implementation of Probabilistic Principal Curves as a part of LANS Pattern Recognition Toolbox, <http://www.lans.ece.utexas.edu/~kuiyu/>. The data set of diatom photographs, used in the project, was provided to us by the ADIAC partners.

## References

- [1] H. du Buf and M.M. Bayer (eds.). Automatic Diatom Identification. Vol. 51, Series in Machine Perception and Artificial Intelligence, World Scientific Publishing Co., Singapore, 2002.
- [2] N. Goldman *et al.*. Quantitative analysis of shape variation in populations of *Surirella fastuosa*. Diatom Research, vol.5, pp.25–42, 1990.
- [3] T. Hastie and W. Stuetzle. Principal Curves. Journal of the American Statistical Association, vol.84, issue 406, pp.502–516, June 1989.
- [4] A. Hertzmann *et al.*. Image Analogies. SIGGRAPH'2001 Proceedings, pp.327–340, 2001.
- [5] Y.A. Hicks *et al.*. Modelling life cycle related and individual shape variation in biological specimens. Proc. BMVC'2002, Sept 2-5, Cardiff, Wales, Volume 1, pp.323–332, 2002.
- [6] Y. Hicks *et al.*. Automatic Landmarking for Building Biological Shape Models. Proc. ICIP 2002, Rochester, NY, USA Vol II, pp.801-804, 2002.
- [7] D. Mou and E.F. Stourmer. Separating *Tabellaria* (Bacillariophyceae) Shape Groups Based on Fourier Descriptors. Journal of Phycology, vol.28, pp.386–395, 1992.
- [8] S. Murthy *et al.*. System for Induction of Oblique Decision Trees. Journal of Artificial Intelligence Research, vol.2, pp.1–33, 1994.
- [9] F.E. Round *et al.*. The diatoms. Biology and morphology of the genera. Cambridge University Press, 1990.