Modelling life cycle related and individual shape variation in biological specimens

Y.A.Hicks[1], A.D.Marshall[1]
P.L.Rosin[1], R.R.Martin[1]
M.M.Bayer[2], D.G.Mann[2]
[1] Department of Computer Science
University of Cardiff
Cardiff, CF24 3XF, UK
{Y.A.Hicks, Dave, Paul.Rosin, Ralph}@cs.cf.ac.uk
[2] Royal Botanic Garden Edinburgh
Edinburgh, EH3 5LR, UK
{M.Bayer, D.Mann}@rbge.org.uk

Abstract

The main purpose of this research is to develop methods for automatic identification of biological specimens in digital photographs and drawings held in a database. Incorporation of taxonomic drawings into a visual indexing system has not been attempted to date. Diatoms are a single cell microscopic algae that provide a particularly suitable case study. Identification of diatoms is a challenging task due to the huge number of the species, blurred boundaries between species, and life cycle related shape changes. A novel model based on principal curves representing the life cycle related shape variation of a number of diatom species has been developed. Our model is suitable for reconstruction purposes, allowing us to produce drawings of a variety of diatom shapes, thus providing a link between the photographs and drawings. We present the classification results of photographed and drawn specimens based on the model and compare our results to another recent system for diatom identification. Finally, given a diatom specimen, we are able not only to identify the species it belongs to but also to pinpoint the stage in the life cycle it represents.

1 Introduction

The ultimate aim of our research is to develop methods for automatic identification of biological specimens in photographs and drawings held in a database. Automatic identification of objects represented in visual form and held in a database is also referred to as visual indexing. Biological specimens are frequently represented in visual form for taxonomic and other purposes. Vast catalogues of specimen material have been accumulated over many years in the form of microscope slides, drawings and photographs. Recently, efforts have been made to digitize such data for electronic storage and transmission.
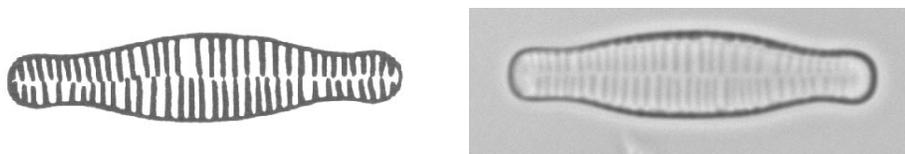
323

Figure 1: A drawing and a photograph of *Fragilariforma bicapitata* diatom specimens.

There has been some recent work in indexing between images in biological databases [2, 14]. However, there is a clear need to extend such indexing capabilities; the inclusion of biological drawings is a natural extension. The approach taken in this research is novel in that it seeks to incorporate taxonomic drawings as a prime source of taxonomic data and to develop methods to enable indexing between digital photographic images and drawings stored in a biological database (Figure 1).

There are several ways in which inclusion of drawings can benefit a visual indexing system. First of all, they represent an example of the species; in some cases it could be the only example. Secondly, they contain only salient for classification information, thus serving as a model of the species. Finally, type specimens are often defined in taxonomy literature using drawings. Production of biological drawings by hand is a time-consuming and a difficult task, which makes it a perfect candidate for automation.

We propose to transform the high-dimensional image space of both photographs and drawings into a lower-dimensional space where only relevant features are represented and use this space for visual indexing and automatic production of drawings. To date we have focused our attention on diatoms. However, we are planning to make our system more general by extending its capabilities to work with other species, such as desmids (another group of microscopic algae) and acari (water mites).

In this paper we investigate different data representation methods involving dimensionality reduction and present our model of diatom shapes based on principal curves. Our model is suitable for reconstruction purposes, allowing us to produce the drawings of diatom life cycle related shape changes, thus providing a link between the photographs and drawings. We apply our model to classification of photographed and drawn specimens obtaining the results comparable to other diatom identification systems. Finally, given a diatom specimen, we are able not only to identify the species it belongs to but also to pinpoint the stage in the life cycle it represents. Throughout this paper, when we mention photographs of diatoms, we are referring to digital photographs obtained using a digital camera and an optical microscope. In case of diatom drawings, we are referring to high-resolution digital scans of the drawings.

2 Previous research in diatom shape analysis

Diatoms are unicellular algae that live practically in any moist or aquatic environment. Each specimen is enclosed in a silica shell with a regular geometric pattern. Diatoms can undergo dramatic changes in size and shape during their life cycle (Figure 2). Diatoms have been studied for many years and their classification, i.e. the systematic division of the taxa into different genera, species, etc., is traditionally based on the size, shape and pattern of their silica shells. Identification of diatoms, i.e. assigning a specimen to one of
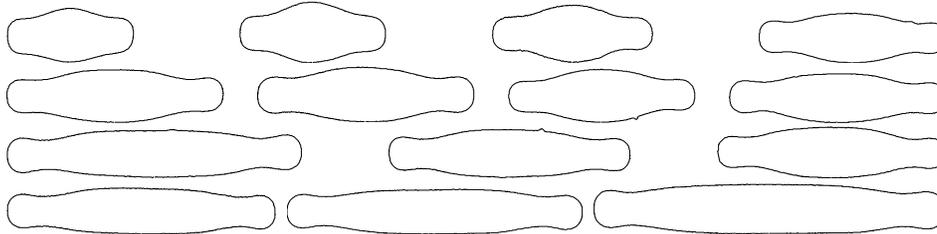
Figure 2: *Fragilariforma bicapitata* specimen contours in the order of projection onto the principal curve.

the known species on the basis of the above features has applications in different areas, including environmental monitoring and forensic science. The difficulties in identification of diatoms are introduced by the large number of the species, sometimes poorly defined borders between species, the changes in shape diatoms undergo during their life cycle, and natural variation related to genetic or environmental factors. There are several thousand of described diatom species and the estimates of undescribed species are in the region of one hundred thousand. The classification of diatoms is still developing, i.e. diatoms that have been 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 could assist in identification and classification of diatoms.

In recent years there have been some efforts in the development of quantitative analysis of shape variation in separate diatom species, as well as in the development of automatic systems for diatom identification. Stoermer and Ladewski [11] modelled *Gomphoneis herculeana* shapes by representing the specimen outlines using Legendre polynomial descriptors and performing principal component analysis on them. They were able to reconstruct the shapes from the model, which showed that the 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.* [6] 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. Finally, Mou *et al.* [9] used principal component analysis on Fourier descriptors of diatom contours belonging to several subgroups of *Tabellaria flocculosa*. They showed that the distribution of the whole taxon formed a nonlinear curve in the space of the first three principal components. The groups within *Tabellaria flocculosa* formed smaller curves, representing their growth trajectories in the shape space.

In the recent ADIAC project [2, 5], a diatom identification system was developed, based on shape, size, internal pattern features, and decision trees. The system can identify specimens from 37 different species with around 97% accuracy. However, the ADIAC is not capable of reconstructing diatom shape or appearance in drawings, nor it is capable of modelling the life cycle.

To date, there has been no attempt to develop a single system for modelling shape variation of a relatively large number of diatom species. One of the reasons for this could be the difficulty of obtaining a sufficient number of specimens for each modelled species. Neither has there been an attempt to include the information from available specimen

325

drawings into such a system. Finally, no one has attempted to produce the specimen drawings automatically from the photographs.

In the next sections we introduce a system that models the diatom shape development through the life cycle for a large number of species, and allows the production of shape-drawings. Our system does not require a large number of specimens for training, and allows identification of the new specimens with a success rate comparable to other diatom identification systems.

3 Modelling diatom contours

We need to find a general way to describe diatom shapes belonging to a large number (possibly hundreds) of different species. We require this method to be general enough to describe not only previously encountered shapes, but also previously unseen specimens and even the specimens of previously unseen species. Finally, we desire the description to be suitable for reconstruction of the original shapes.

We considered describing variation in diatoms by modelling the distribution of a number of automatically chosen landmarks on a diatom's contour [8]. While this method works well for modelling a single diatom species, it runs into difficulties when trying to match the landmarks on the contours of different species, which is not surprising considering the variety of diatom shapes. Even manual landmarking wouldn't work here owing to the different number of landmarks on contours of different species, as well as the high degree of ambiguity concerning which landmarks should be matched to which. Alternatively, we could choose a uniform landmarking approach, such as equal distance sampling of points over the arc length of the shape. However, unless we use a very high number of landmarks, this could lead to missing important detail in the shape.

Fourier descriptors have been successfully used to provide a compact and informative description of diatom shapes [5, 9] in a number of cases and do not suffer from above problems. Hence, we chose to work with these.

3.1 Extracting diatom shape contours

We will start this section with a description of the diatom data set we are trying to model. It is the same data set that was used in the ADIAC project [2]. It contains photographs of diatom specimens belonging to 37 species, each represented by approximately 20 photographs. For each species there is also a digitised drawing of a single specimen. In the natural habitat there are usually considerably larger numbers of diatoms at the later stages of their life cycle rather than at the earlier stage, which is reflected in the data, and it is often difficult to find "clean" diatom specimens, i.e. not broken and without any overlapping debris in the photograph. There is also a problem of diffraction effects which can sometimes distort the image. At present we can successfully extract the features from approximately 50% of these specimens, depending on the species.

We start processing the digital photographs of diatoms by extracting the contour of the specimen, which is a difficult problem in its own right because of the diffraction effect around the diatom specimens (Figure 1). We extract the contours by subsequently applying automatic thresholding, area closing and area filling operations. After that, we find all the contour chains in the image and assume that the chain with the largest enclosed area is the contour of the diatom. Extracting the contour from a drawing of a diatom

is relatively straightforward using a similar approach; here there are fewer image distortions (for more details see [8]). Before representing the extracted contours using Fourier descriptors we resample them to the same length in order to obtain the same number of Fourier descriptors for each contour.

We successfully extracted diatom contours from approximately half of the photographs from the original data set. The failures were due mainly to the presence of debris in the photographs and strong diffraction effect in some of the photographs which left the diatom boundaries extremely blurred. We are planning to look into solving the problems caused by diffraction effect in the future. As a result we obtained a total of 268 specimens representing 22 species, which is still an adequate training set.

3.2 Fourier descriptors

In this section we discuss the methods by which we describe the shape contours extracted above. There are several types of Fourier descriptors (FD) for plane closed curves such as diatom contours. We adopt Fourier descriptors developed by Zahn and Roskies [15], which have been successfully used before for describing diatom contours [9]. They define FD 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 define the cumulative angular function $\phi(l)$ as the net amount of angular bend between the starting point $l = 0$ and point l. The domain of $\phi(l)$ can be normalised to the interval $[0, 2\pi]$. The formal definition of a normalised variant ϕ^* whose domain is $[0, 2\pi]$ is

$$\phi^*(t) = \phi(\frac{Lt}{2\pi}) + t$$

and ϕ^* is invariant under translations, rotations and scaling.

We now expand ϕ^* as a Fourier series

$$\phi^* = \mu_0 + \sum_{k=1}^{\infty}(a_k \cos kt + b_k \sin kt).$$

Mou and Stourmer [9] have shown that it is enough to keep the first 30 Fourier descriptors (harmonic amplitude and phase angle values) for a good reconstruction of the original diatom contour from them. We describe each diatom contour using a 60 element vector consisting of 30 amplitude values and 30 corresponding phase angles obtained from Fourier descriptors.

4 Modelling diatom shape variation

Having described how we extract and represent the shape of individual specimens we turn our attention to the problem of modelling the life cycle related shape variation in a single species. This is then extended in Section 5 to a model of life cycle shape variation across several species.

A comparative investigation of potential data representations suitable for our modelling requirements has been undertaken as a part of this project. The dimensionality of the data describing the diatom outlines is very high (60 Fourier descriptors). In addition,

the distributions of vectors representing different species are non-linear. We need to find a method to reduce the dimensionality of the space and to model the distributions. One of the best known methods of dimensionality reduction is principal component analysis (PCA), however it does not model well the non-linearities or discontinuities in the data. There are non-linear dimensionality reduction methods, such as, for example, the constraint shape space point distribution model [1]. This method automatically clusters the data into a combination of Gaussian distributions. Yet the data representing a diatom species often does not form a Gaussian in the distribution space [9]. In addition, there are fewer larger specimens in our data set, as mentioned in Section 3.1, which we do not want to be reflected in our model. Finally, the distributions of different species are very close to each other or even intersecting [6]. In such circumstances any automatic clustering method is likely to combine the samples from different species into one cluster, which we do not want to happen.

A recently developed method, Isomap [12], finds meaningful low-dimensional structures hidden in the high-dimensional observations. This method only provides a way of reducing dimensionality of the original space. After projecting all the data into a lower-dimensional space we still need to separate the data into the groups representing the species. In addition, this method requires a large amount of data, especially if we want to project any more data into the reduced-dimensionality space once it has been built. Alternatively, we could use the Isomap method to find the low-dimensional representation of each species separately. In this case, there is still a problem of insufficient quantity of data. Apart from that, we would have to find a mechanism of assigning a new sample to one of the Isomap representations in the space of all species.

Kernel PCA [10] models non-linearities in the data by using kernels other than the inner product, but to model the data successfully one needs to know the underlying relationship in order to choose the correct kernel, which is not the case in our application.

Principal curves [7] were introduced about 20 years ago and since then have found applications in different areas including computer vision [3] . They are non-linear generalisations of principal components that model well non-gaussian distributed data, and are especially good for modelling trajectories. Other advantages in modelling data with principal curves is that they do not require a large amount of data and possibly can be trained incrementally.

4.1   Principal curves

The principal curve was first defined by Hastie and Stuetzle [7] as a smooth (C^{∞}) unit-speed 1D curve in $\mathbf{R}^p$ satisfying the self-consistency condition

$$f(x) = E_{\vec{Y}|g(\vec{Y})}\{\vec{Y} \mid g(\vec{Y}) = x\}, \forall x \in \Lambda \subseteq \mathbf{R}$$

where E is the conditional average operator and $g(y)$ is the projection operator given by

$$g(\vec{Y}) = \sup_{\lambda \in \Lambda}\{\lambda : \| \vec{Y} - f(\lambda) \| = \inf_{\mu \in \Lambda} \| \vec{Y} - f(\mu) \|\}.$$

Intuitively, a principal curve is a smooth curve passing through the "middle" of a data distribution. The above definition allows recursive estimation of a principal curve for a given data set. In practice, the conditional expectation is replaced by a smoother or

nonparametric regression estimate, and the curve is approximated with a number of knots and linear segments connecting them.

There are several problems with the original definition of the principal curve which gave a rise to several alternative definitions offering improved estimation algorithms. Tibshirani [13] considered the principal curve as a smoothed mixture of Gaussians along a 1-D manifold in $\mathbf{R}^d$ space. Under this framework, each component density is defined by the conditional Gaussian probability of $\mathbf{y}$ given x, with mean $\mathbf{f}(x)$, covariance $\mathbf{\Sigma}(x)$, and mixing probabilities $P(x)$. Then the standard closed-form Expectation-Maximisation solution can be derived for the mixture of Gaussians.

Chang *et al.* successfully applied principal curves to the classification problem [3] and subsequently improved on the Tibshirani's [13] definition by making use of data projections onto the curve through orienting the covariances to lie orthogonal to the manifold gradient at each node, with variance in the manifold gradient direction attenuated [4]. In the next subsection we adopt Chang's method to model the life-cycle shape trajectory of a single diatom species.

4.2   Applying principal curves to modelling shape variation

Prior to modelling the diatom shape data we find the main modes of variation in the data set of all species through PCA. We propose to model the life-cycle shape variation in a single species using a principle curve going through the middle of the corresponding data set created by the Fourier descriptor vectors projected into the eigenspace. To model several species we follow the procedure described below for each species separately; see Section 5 for more details. This approach allows us to extend the model to include a new species easily, unlike in the case of a decision-tree diatom identification method [2].

We adopt Chang's [4] principal curve estimation method to model the life cycle shape variation of a single species. For illustrative purposes we model the shape variation of *Fragilariforma bicapitata* in this section. In Figure 3 you can see the original data set projected into the space of the first and third eigenvectors (for illustrative purposes) with overlaid corresponding diatom contours, as well as the principal curve fitted into the data. The fitted principal curve follows the growth trajectory of *Fragilariforma bicapitata*, as that provides the main source of shape variation. As a result of fitting a principal curve into the data set we reduced its dimensionality from 60 to 1, and modelled the life cycle related variation simultaneously.

Individual shape variations lie in the dimensions orthogonal to the principal curve and can be modelled, for example, by a "generalised cone" of varying thickness fitted around it. The thickness of the cone can be determined according to the deviation of the specimens from the curve in the different sections of the curve. A priori knowledge can be taken into account as well. For example, it has been noticed that there is more diversity in the shape of *Tabellaria flocculosa* in the life-cycle at a later stage [9].

5   Classifying diatom specimens

We have described how we build a model of shape variation across a single species. Now lets turn our attention to classifying specimens over a wide range of species.

We fit an individual principal curve into each of the available 22 species shape data sets as described in Section 4.2. The fitted principal curves can be viewed as a drastically
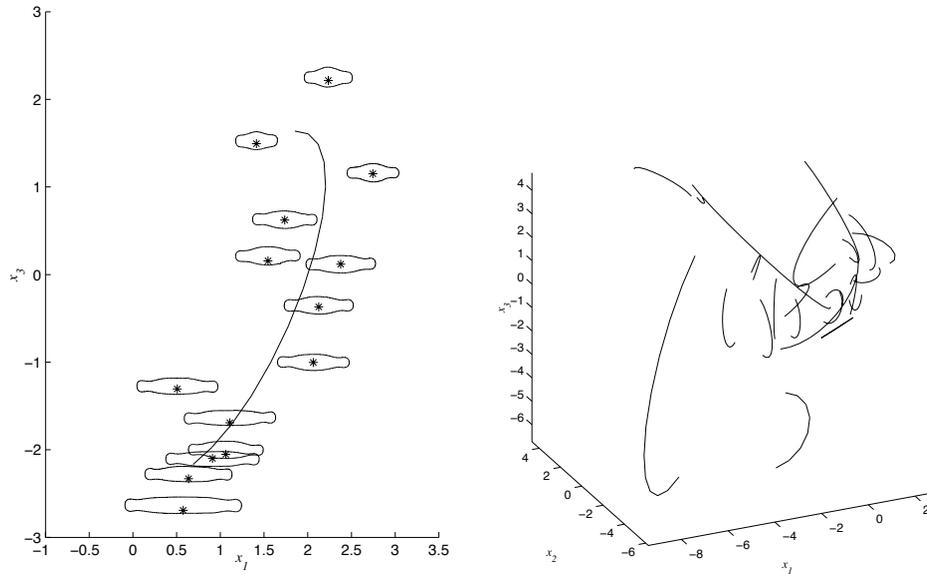
329

Figure 3: **Left**: the distribution of vectors representing shape of *Fragilariforma bicapitata* specimens (represented with *'s) projected onto the first and third eigenvectors together with the principal curve fitted into the distribution. **Right**: 22 principal curves modelling shape variation in 22 diatom species projected into the space of the first three eigenvectors.

reduced dimensional description of the life-cycle variation across the species (Figure 3). Since the curves are in the same shape space of the 60 Fourier descriptors, we can place a new specimen into the same space and then estimate Euclidian distance from the specimen to its projection onto the curves for identification purposes. This is not the only approach to specimen identification in our model. For example, the likelihood of the specimen belonging to each of the curves instead of Euclidian distance can be estimated. Rejection, i.e. not assigning a diatom to any species, also can be easily incorporated. For reasons of simplicity, we base our identification experiments below on Euclidian distance.

We tested the model in identification experiments using the standard "leave one out" approach, by training the model on all specimens apart from one and classifying that specimen using the model. The experiment was repeated to omit each specimen out of the total 268. Approximately 84% were classified correctly compared to 77% achieved in ADIAC project when using only Fourier descriptors for identification. However, we used only a third of the set used in the ADIAC, so firm conclusions cannot be made.

Next, we tested how our model performs when identifying the drawings of diatoms. We tested the model on 11 drawings of the species included in the model. The model correctly identified 7 out of the 11 drawings, which is comparable to identification results on the photographs. We expect the identification accuracy to improve for both photographs and drawings in the future when we include internal features descriptors into the model.

As we mentioned above, given a photograph (or a drawing) of a diatom specimen we can not only identify the species it belongs to but pinpoint the stage of the life cycle it represents. To illustrate this, we projected the vectors representing the shape of *Fragilar-*
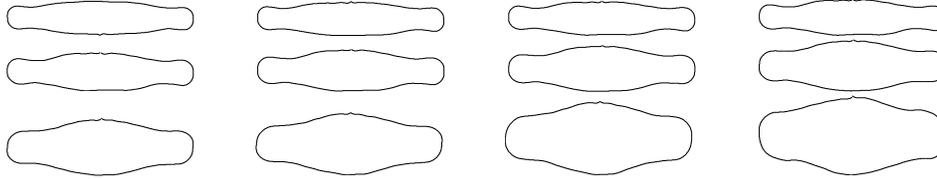
Figure 4: The drawings of *Fragilariforma bicapitata* (scaled to the same length) reconstructed from the knots on the principal curve following its length.

iforma bicapitata specimens onto the corresponding principal curve and then sorted them in the order of their principal curve projection points. The results were very close to the expected order, as you can see in the Figure 1.

Finally, we reconstructed diatom contours from the knots on the principal curve representing *Fragilariforma bicapitata* species. You can see the reconstructed drawings in the order of the knots on the curve in the Figure 4. All the contours are scaled to the same length, however, if we incorporate the diatom length into the model, they can be scaled to the correct length.

6 Discussion and future work

An investigation of different data representation methods with application to modelling the shape of biological specimens has been carried out. A model of the shape variation in diatom specimens belonging to 22 different species has been presented. Our model not only allows identification of the specimens but also models their shape variation through different life stages. The model provides a link between photographs and drawings of the specimens and allows us to create a drawing of a diatom shape from its photograph. The results of classification of drawn and photographed diatom specimens based on our model are comparable to the results of another recent system for diatom identification [2].

So far we have concentrated on modelling the shape of diatoms. Our next step is to extend the model to represent their internal structure, which, as shown in ADIAC, should improve identification accuracy. The extended model will also support automatic production of drawings of the internal structures.

The presented model extends naturally not only to modelling of other biological objects and organisms, such as desmids and acari, but can be used for modelling of other classes of objects where intraclass variation can exceed that of interclass variation, which is often the case in medical and biological applications.

7 Acknowledgments

References

[1] R. Bowden. Learning non-linear Models of Shape and Motion, PhD Thesis, Dept Systems Engineering, Brunel University, UK, 2000.

[2] 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. In press.

[3] K. Chang and J. Ghosh. Principal curves for nonlinear feature extraction and classification. SPIE Applications of Artificial Neural Networks in Image Processing III, pp. 120-129, vol. 3307, 1998.

[4] K. Chang and J. Ghosh. A unified model for probabilistic principal surfaces. IEEE Trans. Pattern Analysis and Machine Intelligence, pp. 22-41, vol. 23, no. 1, 2001.

[5] S. Fischer *et al.*. Feature based retrieval of diatoms in an image database using decision trees. Institute of Computer Science and Applied Mathematics, University Bern. Technical Report IAM-00-001.

[6] N. Goldman *et al.*. Quantitative analysis of shape variation in populations of *Surirella fastuosa*. Diatom Research, pp. 25-42, vol.5, 1990.

[7] T. Hastie and W. Stuetzle. Principal Curves. Journal of the American Statistical Association, pp. 502-516, vol.84, issue 406, June 1989.

[8] Y. Hicks *et al.*. Automatic Landmarking for Building Biological Shape Models. Accepted to IEEE ICIP 2002 Conf. at Rochester, New York, USA.

[9] D. Mou and E. Stourmer. Separating Tabellaria (Bacillariophyceae) Shape Groups Based on Fourier Descriptors. Journal of Phycology, pp. 386-395, 28, 1992.

[10] B. Schlkopf *et al.*. Kernel principal component analysis. In B.Schlkopf, C.J.C.Burges, A.J.Smola, editors, Advances in Kernel Methods - Support Vector Learning, pp. 327-352. MIT Press, Cambridge, MA, 1999.

[11] E.F. Stourmer and T.B. Ladewski. Quantitative Analysis of Shape Variation in Type and Modern Populations of *Gomphoneis herculeana*. Nova Hedvigia, Beih. pp.347-386, 73, 1982.

[12] J. Tenenbaum *et al.*. A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science magazine, pp. 2319-2323, Vol.290, 22 December 2000.

[13] R. Tibshirani. Principal curves revisited. Statistics and Computation, pp. 183-190, vol. 2, 1992.

[14] P.J. Weeks *et al*. Species identification of wasps using principal component associative memories. Image and Vision Computing, pp.861-866, 17, 1999.

[15] C.T. Zahn and R.Z. Roskies. Fourier descriptors for plane closed curves. IEEE Trans. Computers, pp. 269-281, vol. c-21, no.3, March 1972.