Weighted Morphometric Shape Analysis of Diatoms

H. Shahbazkia, T. Candeias, R. Oliveira, L. Santos, F. Tomaz

Universidade do Algarve – UCEH

BIF Laboratory, Campus de Gambelas

8000-810 Faro, Portugal

hshah@ualg.pt

Abstract -**In this paper we present an approach for diatom classification on the basis of contour information. First an initial feature set is analysed. Then we present a method to reduce this initial feature space while preserving its discriminative power. Both results based on the initial feature space and the reduced one are compared and the classification method is discussed.**

1. INTRODUCTION

This paper focuses on the diatom contour. One should keep in mind that from the diatomist's point of view there are no sharp boundaries to separate the different contour classes. A descriptive definition of the contour is generally used, e.g. slightly asymmetrical about the apical axis, almost elliptical, quite elongated, etc. Therefore the shape-analysis approach, which is compatible with this descriptive definition, must be based on continuous shape descriptors. Such descriptors should satisfy the following conditions: (1) simple and computationally fast, (2) semantically coherent with the diatomists' description, and (3) continuous and capable of representing the different properties of a contour.

The only descriptors that satisfy these conditions are morphometric ones. Some of these descriptors were modified to fit the particular diatom problem. For example, we had to define symmetry as a continuous measurement rather than a presence/absence feature. This paper presents a first attempt to identify the minimum feature set that is required to classify diatom outlines.

2. SIMPLE MORPHOMETRIC MEASUREMENTS

In this section, 10 shape descriptors are revised. Some of them have been widely used, whereas some others are new or adapted. Diatom outlines are usually characterised by the following keys: symmetric, convex, circular, elliptical, rectangular, elongated, etc. Hence, the choice of the initial shape descriptors was straightforward. They are described below.

Symmetry (SX, SY): The detection of symmetry has often been studied [4,5,6]. Nevertheless, most of the studies consider symmetry as a binary feature: either it exists or it does not exist. In addition, the exact mathematical definition of symmetry is inadequate to describe and quantify it in real images. Even perfectly symmetrical objects loose their exact symmetry when projected onto the image plane due to occlusions, noise, perspective transformation, digitisation, etc. The diatomists label diatoms as being symmetric, slightly asymmetric or asymmetric [6], but they do not have formal definitions of these classes.

In order to compute a symmetry value we use the correlation between the two half-contours obtained by folding the part of a contour lying on one side of an assumed axis onto the other side. However, in the case of concave and non-bijective shapes the correlation algorithms don't work. For this reason we choose to define symmetry as the difference between the two half-curves as half of the area that they enclose. This measure allows us to process concave or non-bijective shapes and it takes into account all the points of the contour; see Fig . 1.
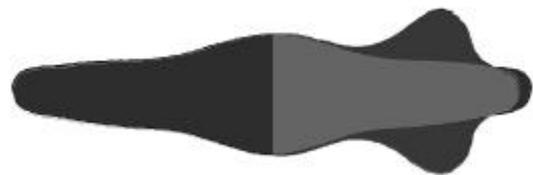

Figure 1. Folding and image symmetry
based on measuring areas

Computing mathematically this area is very complicated, so we prefer to compute it by using an image-based (discrete space) process. The method is as follows: (1) find one point inside the contour on one side of the axis, (2) use a region growing method to change the intensity of this side to C1, (3) find one point inside the contour on the other side of the axis, (4) use a region growing method to change the intensity of this side to C2, (5) fold one side onto the other one, adding the pixel intensities, (6) count all points on the folded side with an intensity different from C1 + C2. In this way, we obtain two descriptors: apical symmetry percentage and transapical symmetry percentage.

Convexity (CP, CA): Many applications use the ratio of the perimeters of the convex hull (minimum convex covering) and the original contour as a measure of convexity. For diatoms this definition of convexity is obviously insufficient, see Fig. 2. A better measure is the ratio of the *areas* of the convex hull and the original contour. In the following section we will see that these two measures are independent and can therefore be used together.

Figure 2. The contour perimeter/convex hull perimeter ratios of these two contours are equal but the area ratio is different.

Circularity (Ec.): A well-known measure of circularity is the eccentricity as given by:

$$e = (m_{20} - m_{02})^2 + m_{11}^2 / (m_{20} + m_{02})^2,$$

where m_{ij} is the ij central moment of the outline.

Ellipticity (Ell.): This descriptor is often used by astronomers to describe the shape of galaxies and is given by:

$$\left(\sqrt{(m_{20} + m_{02}) + \sqrt{(m_{20} + m_{02})^2 + 4m_{11}^2}} \right) / \left(m_{20} + m_{02} - \sqrt{(m_{20} - m_{02})^2 + 4m_{11}^2} \right)$$
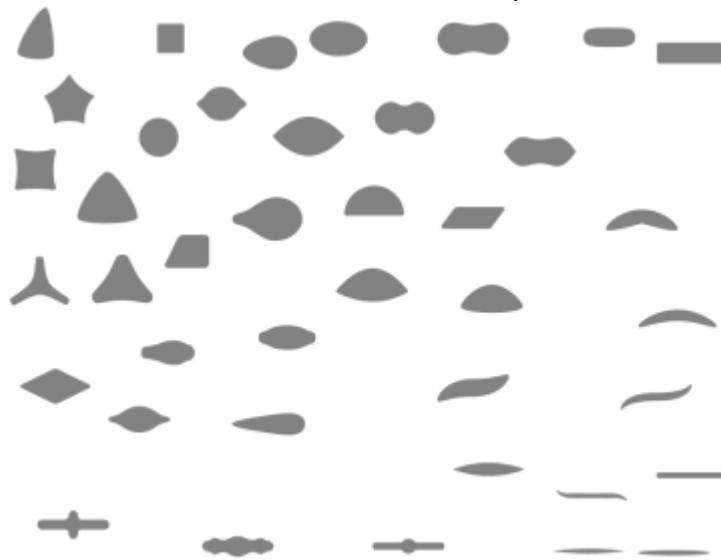
Rectangularity (Rec.): The rectangularity is often used by diatomists, and is computed by

$$rec = \text{area of the outline} / (length \cdot width)$$

Elongatedness (El.):

$$E = area / \text{max thickness of min bounding box}$$

Size ratio (Rs.): $R = width / length$

Compactness (Com.): This is usually given by

$$area / perimeter^2$$

We normalise this using the circle compactness; therefore we use:

$$4p \cdot area / perimeter^2$$

3. APPLICATION

The 10 shape descriptors listed in the previous section were applied to a set of 40 characteristic diatom outlines. This test set contains one synthesised sample for each of the 40 principal diatom contour classes. A 10-component feature vector is extracted for every sample. Then these 10-dimensional vectors are mapped to a 2dimensional space using a Kohonen self-organising map. The output is shown in the Fig. 3.

The correlation index matrix based on the covariance matrix of the feature set is computed in order to assess their discriminative power. As expected, Table 1 shows that the eccentricity, compactness, ellipticity, size ratio and elongatedness are highly correlated, whereas the convexity perimeter-ratio and the convexity area-ratio are rather independent.

Figure 3. Self-organising map of the 40 synthesised samples for a 10-component feature vector

Ec.	El.	Rs.	Rec.	Com.	Ell.	CP	CA	SX	SY
100	86	81	3	79	87	11	5	29	23
86	100	98	1	93	80	12	7	28	18
81	98	100	11	95	75	7	3	32	9
3	1	11	100	14	25	28	57	58	68
79	93	95	14	100	74	18	15	33	15
87	80	75	25	74	100	22	36	0	52
11	12	7	28	18	22	100	30	32	27
5	7	3	57	15	36	30	100	25	68
29	28	32	58	33	0	32	25	100	38
23	18	9	68	15	52	27	68	38	100

Table. 1 The scaled correlation matrix

By excluding the features with a high correlation, we obtained a set of only 6 descriptors. Using this feature vector as it is gives an equal importance to each feature, but in the case of diatoms the *a priori* knowledge can be used to give more or less importance to each feature. This weights all features by multiplying them with a constant before the mapping process. The weighting process is not done in a formal manner as the knowledge to be incorporated is in form of expertise. The size-ratio and the symmetry play a very important role in the case of diatoms whereas convexity is some-how not very important. A second mapping on the basis of the weighted six features is shown in Fig. 4. As we can see, all symmetric and circular shapes have the tendency to occupy the upper-left side of the map, whereas elongated and asymmetric ones go to the lower-right side.

4. CLASSIFICATION

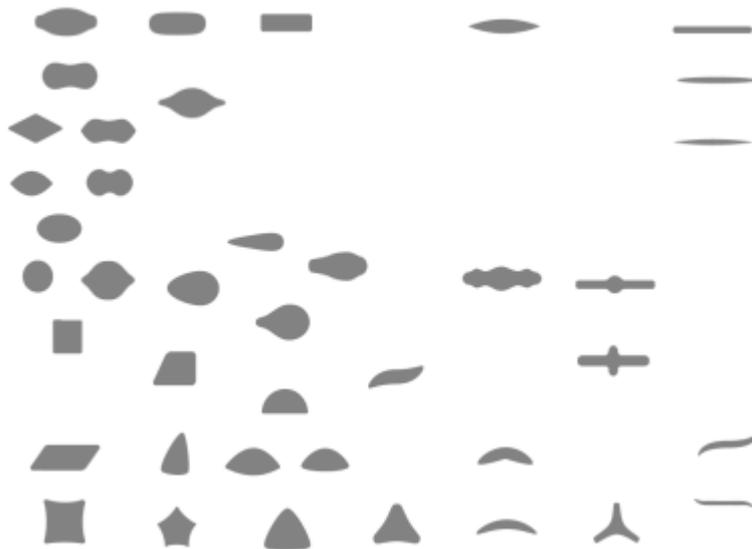Once the discriminative power of the feature vector has been assessed, a classification can be applied in order to produce ergonomic results for the diatom experts. The classification can be done simply by computing the distance to the nearest-neighbour classes. The labelling of the contour is then done by using the same vocabulary as the diatomists use but with a quantification of the attributes (Ex: distance of 1.6 from ideal circle and distance of 1.4 from ideal ellipse, which means an elliptic-circular outline). A selected set of 50 real outlines (black shapes) was tested and the results are shown in Fig. 5. The correct hit rate is 95%.

5. DISCUSSION

When two different scientific communities work together, much effort goes into understanding each other, because the knowledge and vocabulary can be quite different. In our case, a simple description in terms of symmetric, slightly asymmetric and asymmetric can become a real problem. Since diatomists do not wish to use very rigid classes for the contour labelling nor for other diatom features, a continuous labelling should be used. With respect to a contour classification, many methods do not allow the user to have a verbal description of the shape. One example of this are Hu-moments.

The *a priori* knowledge about the shapes to be classified and the descriptions obtained from domain experts have enabled us to describe the diatom outline by using a reduced set of perfectly adapted features. The CPU time is kept low and the exchange of information with the experts is simplified.

The Kohonen network allows an exploration of high dimensional data and creates an easy visualisation of the inputs in two dimensions. It is an ideal tool to assess possible separation limits when using a given set of features.
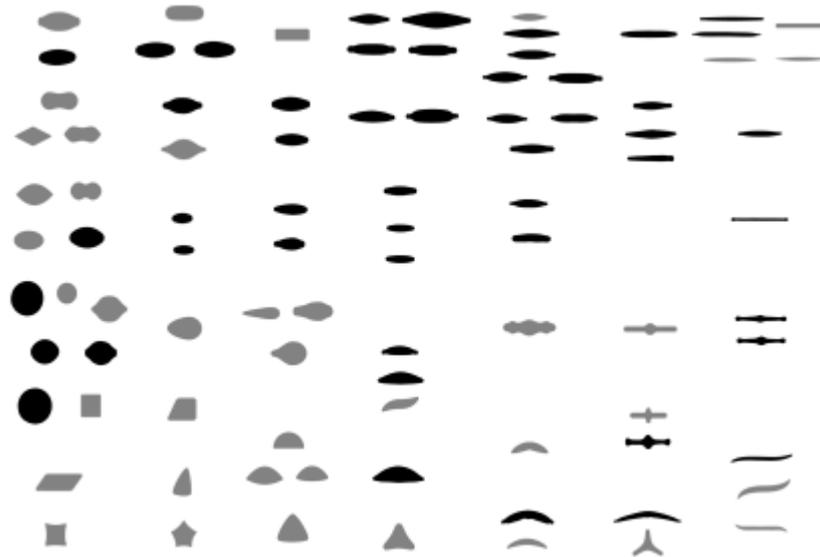

Figure 4. Self-organising map of the 40 synthesised samples for a 6-component feature vector

Figure 5. Self-organising map of the 40 synthesised samples (gray) plus 50 real outlines (black)
for a 6-component feature vector

6. CONCLUSION

We started our work by considering a large set of 10 continuous features, with which we could analyse the diatom shape. The Kohonen mapping proved to be clear and conclusive, and a visual inspection of the results allows to distinguish easily many different kinds of diatoms.

The covariance matrix revealed a high correlation between some of the features and allowed to reduce the set to only 6. An interesting result concerning the convexity was that its calculation based on the areas' quotient proved to be independent from the calculation based on the perimeters' quotient. The reduced feature set applied to a selected set of 50 real (noisy) contours resulted in a 95% correct shape classification.

The classification based on the reduced feature set allows us to describe a diatom's shape in a formal way, such that it is coherent with the diatomists' vocabulary. Ongoing research concerns (1) the assessment and improvement of the feature extraction using thousands of real contours, and (2) the extraction of ornamentation features that complement the contour features in order to develop a complete diatom identification scheme.

7. ACKNOWLEDGEMENTS

REFERENCES

[1] Davies, E.R., Machine vision, Academic Press, 1997.

[2] Klette, R. and P. Zamperoni, Handbook of Image Processing Operators, Wiley, 1996.

[3] Gonzales, R. and R. Woods, Digital Image processing, Addison-Wesley, 1992.

[4] J. Bigun, "Recognition of local symmetries in gray value images by harmonic functions", Proc Int Conf Pattern Recognition, 1988.

[5] G. Marola, "On the detection of the axes of symmetry and almost symmetric planar images", IEEE Trans. Pattern Analysis and Machine Intelligence, Vol 11, nr. 1, 1989.

[6] H.G. Barber and E.Y. Haworth, "A Guide to the Morphology of the Diatom Frustule with a Key to the British Freshwater Genera", Freshwater Biological Association Sci. Publ., 1981.

[7] H. Zabrodsky, "Computational aspects of pattern chracterization. Continuous symmetry", PhD thesis, Hebrew Univ., Jerusalem, 1993.

[8] H. Zabrodsky, "Continuous symmetry measure. Chirality", J. Am. Chemical Soc., Vol 117, 1995.

[9] H. Zabrodsky "Symmetry of fuzzy data", Proc Int Conf. Pattern Recognition, TelAviv, 1994.

[10] Hans du Buf et al., "Diatom identification: a Double Challenge Called ADIAC", Proc 10th Int Conf on Image Analysis and Processing, Venice, Sept. 27-29, 1999, pp.734-739.