Features for the Classification of Marine Microfossils

Siegfried Brechner and Frank Ade

ETH Zürich, Communication Technology Lab, Image Science Group, ETZ F 85,
Gloriastrasse 35, CH-8092 Zürich, Switzerland

Abstract. Features for a new application in pattern recognition are presented. The objects to be classified are coccoliths (marine microfossils). Since the objects differ both in their outline and in their internal structure the features developed take into consideration both kinds of variability.

The features concerning the internal structure are computed on the grey-value images produced by a scanning electron microscope. These features include measures of both of the angular variation, using the autocorrelation function, and of the radial variation within the objects by an approximation with Legendre polynomials.

The features concerning solely the shape of the objects include the content of long and short straight segments of the outline, and measures based upon ellipse fits and the smallest bounding rectangles to the outlines. In addition, moments are calculated on the objects' binary filled interior.

Keywords: statistical pattern recognition, feature extraction, marine microfossils, coccoliths

1 Introduction

The analysis of coccoliths (marine microfossils) has been a very successful but time-consuming technique applied in the earth sciences in the past. In order to classify several thousands of samples per day we are developing a fully-automated pattern recognition system in collaboration with the geological institute of the ETH-Zürich. Large data sets are required because they enhance the statistical reliability of the determination of the relative abundances of coccolith-species, offering the possibility of geological high-resolution stratigraphy and paleoceanography, as well as providing a better basis for studies of plankton diversity.

The ultimate goal is to recognize and classify about 40 morphologically distinct coccolith species, which are spread among other sediment particles on a specimen stub, with a success rate of better than 90 %.

This paper reports on the features developed so far and an error rate is presented for a set of 103 samples belonging to eleven different coccolith species. Figure 1 gives an overview over the species considered so far, which are the following: Calcidiscus leptoporus (CL), Coccolithus pelagicus (CP), Discosphaera

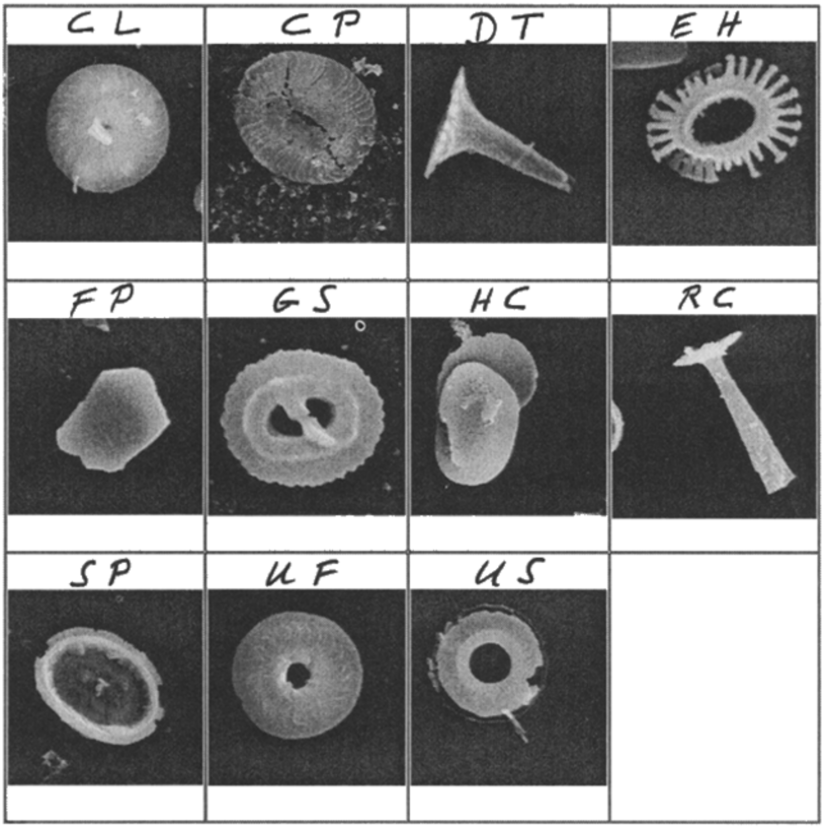

Fig. 1. overview over the species considered

tubifera (DT), Emiliana huxleyi var. huxleyi (EH), Florisphaera profunda (FP), Gephyrocapsa (GS), Helicosphaera carteri (HC), Rhabdosphaera clavigera (RC), Syracosphaera pulchra (SP), Umbilicosphaera sibogae var. foliosa (UF), Umbilicosphaera sibogae var. sibogae (US).

Since the size of the coccoliths is of the order of one micrometer a scanning electron microscope (SEM) is used for image aquisition. The image segmentation is based on local edge detection but since the goal of this report is to show the discriminatory ability of the features, the segmentation results of all of the images available were inspected visually and only those images, for which the segmentation algorithm performed reasonably well, were selected. For these 'good' images the features were computed and the subsequent statistical analysis was done. The set of 'good' images consists of 103 samples.

The set of features used can be divided into three subsets: features derived from the 'raw' outline, features from the binary filled outline and into features obtained from the grey-valued interior of the objects.

2 Features

2.1 Features related to the raw outline

The features derived from the raw outline can be divided into two subsets: The first subset is based upon an approximation of the outline by an ellipse including the following features: the *quality of the ellipse-fit* and the *elongation of the ellipse.*

The second subset is based upon an approximation of the outline by linear segments and includes: the *percentage of long, straight segments* and the *percentage of short, straight segments*, where a long line segment corresponds to more than 10 % and a short one to less than 5% of the entire outline.

2.2 Features related to the binary filled outline

The features extracted from the binary filled outline are standard shape features and are, except for the computation of the moments, supplied within the software package KBVision [2].

The smallest bounding rectangle (SBR) is used after the object has been turned to corrected for rotational variations. Then *pixel–count* is the number of ON-pixels (belonging to the object) within the SBR, *height* and *width* are the (vertical) height and the (horizontal) width of the SBR.

Another notion used is the *perimeter,* where each side of an ON-pixel, which is four-connected to an OFF-pixel (a pixel outside of the object), adds one to the perimeter. Therefore, one single ON-pixel has a perimeter of four.

The features provided within KBVision are the following: $\frac{pixel-count}{height*width}$, $\log_{10} \frac{height}{width}$, $\frac{perimeter}{2*height+2*width}$, $16 * \frac{pixel-count}{perimeter^2}$, and the *elongation of the fitted ellipse* (fitted to the area of the object). In addition, *invariant moments* are computed on the binary filled outlines [3].

2.3 Features derived from the objects' interior

The features describing the grey-valued interior of the objects can be divided into two subsets: *angular features* describing the angular variation and *radial features* describing the radial variation within the objects.

The angular and radial features are tailored to ellipse-like objects (Fig. 1) and take as a starting point an ellipse fit to the outline. Ten ellipses with the same elongation as the fitted ellipse but with smaller major axes are drawn, where the major axes of the 'internal' ellipses are evenly distributed between zero and the major axis of the ellipse fit to the outline, see Fig. 2.

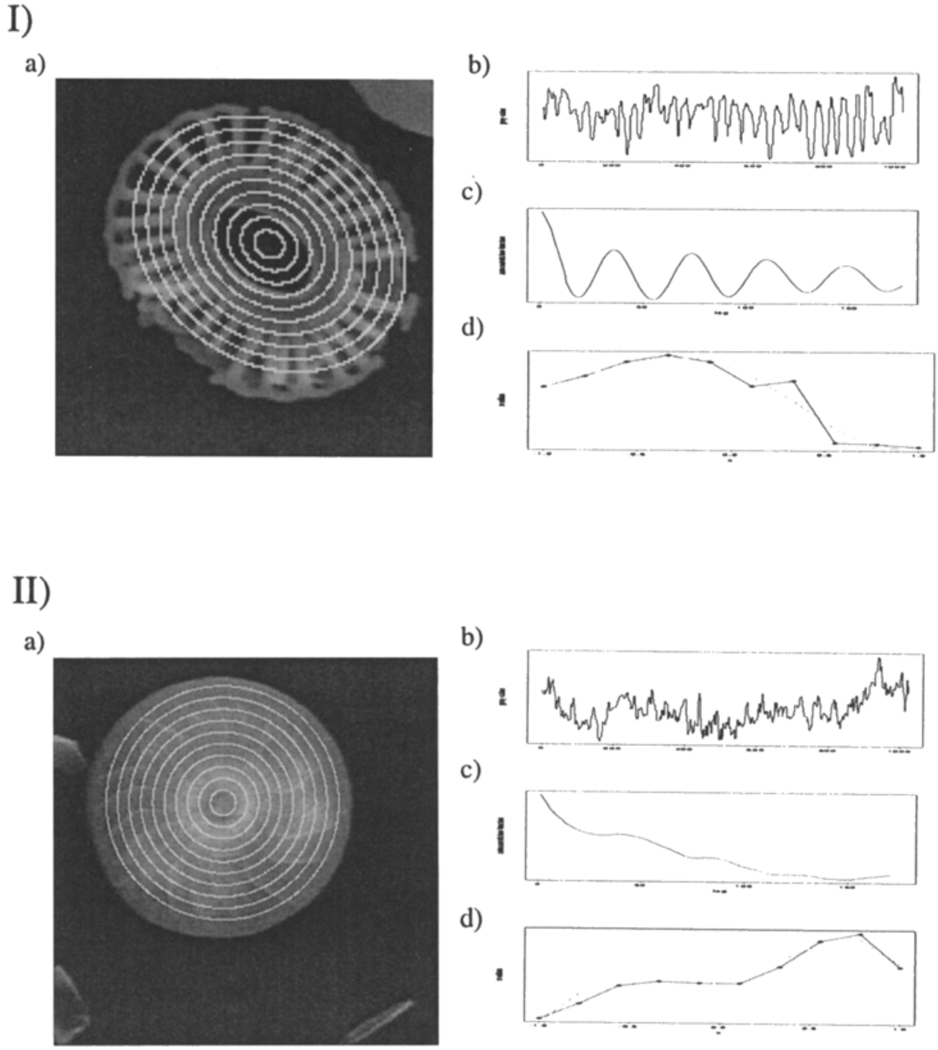

Fig. 2. Grey-value features for a sample of EH (I) and of CL (II): (a) internal ellipses plotted on the grey-value image (b) grey-values on the internal ellipse, which contains the 'best' peak (2nd outmost for EH and 8th outmost for CL); the grey values are taken at equal intervals in angle, one interval corresponding to $\frac{2\pi}{1024}$ (c) autocorrelation on the function of grey-values of b (d) the median values (solid, thick lines) for each ellipse are plotted as a function of radial distance from the center of the ellipse: $x = -1.0$ corresponds to the outermost and $x = 1.0$ to the innermost ellipse; the dashed line shows the approximation by Legendre polynomials

For the angular feature the grey-values on each ellipse are taken as a function of the the azimuth ϕ with $0 \leq \phi \leq 2\pi$, and the autocorrelation function is computed for each grey-value function for each ellipse separately. Now some simple criterion (for example the greatest height of all of the peaks of the autocorrelation function relative to its neighbouring inflection points) is used to distinguish between smoothly decaying autocorrelaton functions and those, which show distinct peaks. The highest value of this criterion for all of the autocorrelation functions on the internal ellipses is taken as the feature for this particular sample. Thus any information on the radial distribution is lost, but the algorithm is reasonably robust against small displacements of the 'true' center of the ellipse and against small occlusions of the object, which occur at certain radial distances only.

In order to measure the radial variation within the objects the median values of the internal ellipses are taken as a function of radial distance to the center of the objects and approximated by Legendre polynomials of fourth order. The Legendre coefficients are the features describing the radial variation within the objects. Since Legendre polynomials form an orthonormal basis for L_2 functions within $-1 \leq x \leq 1$, the functions are approximated within the interval $-1 \leq x \leq 1$. Let $x = -1.0$ correspond to the outermost and $x = 1.0$ to the innermost ellipse.

Figure 2 gives examples for the grey-value and autocorrelation functions of the internal ellipse, which yields the best peak in the autocorrelation spectrum. The radial variation of the meadian values together with its approximation by Legendre polynomials, where fourth order is sufficient for the overall characterization of the curves, is plotted as well. It can be seen, that both the radial and the angular variations are quite different for the samples of the two species, implying the usefulness of the corresponding features.

3 Statistical Analysis

Since the angular and radial features of Sect. 2.3 are tailored towards ellipse-like shapes and because the invariant moments perfectly distinguish between (more or less) ellipse-like objects and the hammer-like objects of the classes DT and RC (Fig. 1), a hierarchical classificator with two stages is employed: the first stage uses the invariant moments to divide the set of samples into two subsets, one comprising the two species DT and RC (Fig. 1) with 31 samples and the other comprising all of the other species with 72 samples. In the second stage the samples of these two subsets are classified separately and independently, first using linear discriminant analysis and then applying the Bayesian approach, where the a-priori-probabilities are determined by the proportions of the samples belonging to the various classes. Since rather few samples were available for both subsets the assumption of a multivariate distribution with equal covariance matrices for the classes of each subset was made. For the determination of the error rate the leaving-one-out method was used [1].

For the 31 samples of the classes DT and RC only the first two invariant moments of second order of [3] were used for the classification. Three samples were misclassified.

For the 72 samples of the remaining species all of the features introduced in Sect. 2 were used, except for the invariant moments, since these were not useful in distinguishing between ellipse-like shapes. Here only one sample was misclassified.

The overall error was therefore at four misclassifications for 103 samples, corresponding to an error rate of four percent.

4 Conclusions and outlook

Features for the classification of well preserved coccoliths were presented which yield an experimental error rate of four percent for eleven distinct coccolith species. No non-coccolith objects were allowed. For a future automatic pattern recognition system a foreground-background-separation mechanism must be elaborated. For the elliptic and hammer-like objects presented Hough transforms for straight lines and for ellipses appear to be appropriate. Furthermore, the feature extraction needs to be extended to discriminate between more coccolith species and to cope with occlusions and missing parts in the outline.

References

1. R.O. Duda and P.E. Hart: Pattern Classification and Scene Analysis, John Wiley & Sons, 1973
2. The KBVision System: Task Reference Manual, Amerinex Applied Imaging, Inc., 409 Main Street, Amherst, MA 01002, edition 3.2, December 1996
3. M. R. Teague, Image Analysis via the general theory of moments, J. Opt. Soc. Am., August 1979, vol. 70, no. 8, 920–930