Marine Micropaleontology 72 (2009) 66–75

Contents lists available at ScienceDirect

Marine Micropaleontology

journal homepage: www.elsevier.com/locate/marmicro

# Feasibility of computer-aided identification of foraminiferal tests

Kamal Ranaweera [a], Adam P. Harrison [a], Santo Bains [b], Dileepan Joseph [a,*]

[a] Electrical and Computer Engineering, ECE Research Facility, University of Alberta, Edmonton, AB, Canada T6G 2V4
[b] Research and Technology Group, BP Plc, UK

ARTICLE INFO

*Article history:*
Received 19 March 2009
Accepted 23 March 2009

*Keywords:*
microfossils
optical microscopy
foraminifera
automatic classification

ABSTRACT

Over the past two decades, several investigators have worked on computerized systems to accelerate the identification of foraminiferal tests (forams). Leading examples have focused on fully-automatic identification using neural networks and supervised learning. This paper introduces an alternative semi-automatic or *computer-aided* approach. Such an approach reduces the workload associated with foram identification without the challenges of training set collection and fully-automatic recognition. The proposed method begins by photographing a collection of specimens sprinkled on a microscopy slide. Segmented images are then mapped into a canonical space where position, rotation, and scale are normalized. Specimens are clustered based on image similarity in the canonical space. A specialist then identifies the clusters by inspecting representative templates. Experimental results show that the identification effort can be reduced by 35%, yet the accuracy remains comparable to when every specimen is individually identified. Further reduction of effort was prevented by the significant variability of illumination direction in the canonical images. These results encourage further work on a computer-aided approach to foram identification.



## 1. Introduction

The taxonomic study of foraminiferal tests (forams) is a driving force behind biostratigraphy (Simmons et al., 1997). Forams also play a great role in paleoclimatology and paleooceanography as the analysis of their morphological variance, e.g. Kelly et al. (1996), and chemical composition, e.g. Rathburn et al. (1997) or Kennett and Stott (1991), yields important information from the geological record. In addition, foram study is a significant source of data in developing geological models used to locate hydrocarbon accumulations (Breard et al., 1993). These applications rely on having high resolution knowledge of foram taxonomy.

However, foram taxonomy is hampered mainly in two ways. First, traditional taxonomy requires manual examination of particles under a microscope. This is time-consuming work requiring expert labour. Secondly, building up a good taxonomic base necessitates the analysis of very large volumes of particles. When coupled together, these two issues present serious challenges for any application that depends on detailed foram taxonomy. One way to manage these challenges is to inject a degree of autonomy into the identification process. This has been a recognized goal for over two decades (Thierstein et al., 1987).

Early systems attempted to incorporate the same rules that micropaleontologists use into a computer-aided identification process (Brough and Alexander, 1986; Riedel, 1989). A defining characteristic of these systems was their focus not on providing automation, but on lessening human error and training requirements. For instance, the Visual Identification Expert System (VIDES) functions by presenting the user with a set of identifying attributes (Swaby, 1992). For each specimen in question, the user specified as many of the attributes as possible until the system was able to infer the identity of the specimen, on the basis of pre-defined rules, with good confidence. However, as with all of these early systems, VIDES still required the manual study of each specimen in question.

In contrast to the rule-based approach, later systems use artificial neural networks (ANNs) to classify microfossils. A leading example is COGNIS, developed by the Micropaleontology/Geophysiology group at ETH Zurich (Bollmann et al., 2004). The authors report on two main experiments. In the first one, scanning electron microscopy was used for multi-species classification, producing good results. The other experiment involved optical microscopy, and it tested whether COGNIS could identify a single species (*F. profunda*) from a collection of images. While the system had a high sensitivity to the species, it suffered from a high false positive rate. Further investigation is required to determine whether COGNIS can produce acceptable results with optical microscopy, the simplest and cheapest modality. The authors concluded that it may be possible to extend their single species system to one that identifies multiple species through the use of parallel networks.

Along this vein, another leading system named SYRACO 2 used the concept of parallel neural networks but applied it in a different manner than what was proposed by the developers of COGNIS. Originally, SYRACO 2 used a single fat neural network that operated on

* Corresponding author. Tel.: +1 780 492 9164; fax: +1 780 492 1811.
*E-mail address:* dil.joseph@ualberta.ca (D. Joseph).

0377-8398/$ – see front matter. 
doi:10.1016/j.marmicro.2009.03.005

images pre-processed to be rotation and translation invariant (Dollfus and Beaufort, 1999). Only brief details of the pre-processing were given, and the authors were not satisfied with it. As a result, the latest version of SYRACO 2 used six parallel neural networks to further transform the images in five possible ways (Beaufort and Dollfus, 2004). However, even in its original form, the SYRACO 2 network used at least 800,000 free parameters to process low-resolution $64 \times 64$ pixel images. According to ANN theory, to successfully train such a network would require at least 800,000 images (Tarassenko, 1998). As their number of training images was much lower than this figure, the authors admitted that they were unable to fully explain how their network performed so well as it did in coccolith identification. The difficulty in understanding SYRACO 2's performance remains an obstacle to widespread adoption of this ANN-based approach.

The state of the art in calcareous microfossil identification is defined by extremes in terms of autonomy. At one end are the fully-interactive rule-based systems, and at the other are the fully-automatic ANN systems. While the former does not reduce identification workload by much, the latter suffers from the drawbacks of supervised learning. Problems associated with traditional foram identification, which have motivated computerized foram identification in the first place, present serious obstacles to the data collection needed to train an ANN able to classify large numbers of species. Added to this problem is the rigid nature of the training process. ANNs and other supervised learning systems can only handle species included in the training set. The addition of new species requires a re-training of the entire system. This is not to say that the ANN approach does not hold promise. However, these points motivate efforts toward an alternative direction.

An alternative approach is to use a *computer-aided* system to reduce, rather than to eliminate, identification workload. This paper introduces such an approach and reports on preliminary experiments. The proposed computer-aided system automatically clusters foram images based on their visual similarity, without requiring pre-defined knowledge about the genera and species in the specimen set. Each such cluster is represented by a template, which is chosen automatically. Providing that clusters are homogeneous, i.e. they only contain images that are very similar, micropaleontologists need only inspect the templates to identify entire clusters. Since the number of specimens that must be identified is reduced to the number of clusters, this method reduces the workload associated with identification. In terms of autonomy, an unsupervised clustering system can be seen to lie in between the early rule-based systems and the later ANN approaches.

## 2. Materials and methods

The proposed system begins by capturing photographs of sieved particles dropped on an opaque glass slide. By scanning the slide under a microscope, images of all specimens may be captured. These images are then segmented and mapped to a canonical space using an *invariant transform* (Section 2.1). Next, visual similarity is estimated among these canonical images (Section 2.2) and used to automatically cluster the specimens (Section 2.3). For each cluster, a template is chosen automatically (Section 2.4) such that the template has the highest overall similarity to all images in the cluster. A micropaleontologist can use either the template image or the physical particle indicated by the chosen template to identify the entire cluster. It is only for this step that a specialist is needed. All other steps are automatic.

Section 3 describes a test of the proposed system using a data set composed of 244 specimens, made up primarily of three genera: *Morozovella* (72.5%), *Acarinina* (15.2%), and *Subbotina* (9.0%). Each specimen has a classification based on examination of the actual particle under a microscope, and a classification based on a *default view* image only, obtained without particle manipulation. The data set suffices to demonstrate that identification workload can be reduced significantly with the computer-aided approach. More details on the composition and collection of the data set can be found in a related paper (Ranaweera et al., 2009).

### 2.1. Invariant transform

Specimens may be arbitrarily positioned and orientated in images. As well, they may have different scales, for example due to differences in zoom level. Invariance against these attributes may be enforced by transforming images prior to computerized identification. Initially, SYRACO 2 did this using only computations of object centre, major axis, and extremes of image luminosity (Dollfus and Beaufort, 1999). But no details were given and the authors went on to extend their pre-processing with six parallel neural networks (Beaufort and Dollfus, 2004). This section details a robust normalization technique, which uses principal component analysis (PCA) and third central moments, for the proposed computer-aided system.

#### 2.1.1. Segmentation

For each photograph, the non-linear *gamma correction* built into digital cameras (IEC, 1999) is first inverted. Resulting pixel values are directly proportional to light intensities, and are mapped to a 0–1 range. Next, the invariant transform segments images of specimens and generates binary silhouettes using a threshold of 0.02, which represents the maximum background pixel value. Any holes in the foreground are filled in morphologically (Gonzalez and Woods, 2006).

Silhouettes in a segmented image represent the physical shape of specimens in the input image. Since the photographed particles were obtained from sieved size fractions, limits of valid silhouette dimensions are known a priori. Thus, the system automatically identifies specimens that were broken after sieving, as well as those that landed attached to each other, by simple computation of silhouette diameter. These problematic cases are automatically filtered out.

#### 2.1.2. Normalization

Position, rotation, and scale are independent of illumination, so working with a silhouette leads to an invariant transform robust to illumination. The centroid of the silhouette is computed and is taken as the origin for further calculations. The principal components of points that make up the silhouette are computed in the new coordinate space, thereby determining the rotation and scale of principal axes (Gonzalez and Woods, 2006), as shown in Fig. 1(a) and (b).

Once the silhouette attributes have been computed, a transform is defined using basic image processing operations to normalize the silhouette such that:

- The canonical silhouette is centered in an image of $640 \times 640$ pixels.
- The orientation of the major axis, which is the orientation of the first principal component, is horizontal in the canonical silhouette.
- The length of the first principal component, shown in Fig. 1 as half the length of the major axis, is 128 pixels in the canonical silhouette.

Fig. 1(c) and (d) show the result of applying this transform to the silhouette and the original image, respectively. Both of these are required (Section 2.2).

On its own, the PCA-based mapping does not provide a unique transformation as there is a rotational ambiguity that must be resolved. To understand this ambiguity, consider the two silhouettes shown in Fig. 1(c) and (e). The second silhouette is a 180° rotated version of the first. Both of these silhouettes conform to the criteria listed above. Exactly which one is obtained depends on numerical round-off and the internal workings of the numerical method used to implement PCA.