

Alternative 2D and 3D Form Characterization Approaches to the Automated Identification of Biological Species

Norman MacLeod

Abstract — Few have sought to compare the performance of alternative types of morphological data for biological species identification. This investigation contrasts results of form characterization via form factors, superposed landmark coordinates, landmark-registered semilandmark outlines, 3D semilandmark networks, and raw digital images for a test set of seven Recent planktonic foraminifer species. While all data types performed better than the qualitative assessment of morphological variation by human taxonomists, landmark-registered semilandmark outlines and raw digital images delivered the best performance in the context of approaches that could reasonably serve as the basis for fully automated species identification systems.

Index Terms — automated identification, landmark coordinates, Foraminifers, taxonomy.



The automated identification of biological objects (individuals) and/or groups (e.g., species, guilds, characters) has been a dream of systematists' for centuries. The goal of some of the first multivariate biometric methods was to address the perennial problem of group identification and inter-group characterization [1], [2]. Despite much preliminary work in the 1950s and 60s, progress in designing and implementing practical systems for fully automated specimen identification has proven frustratingly slow. However, as recently as 2004 Dan Janzen updated the dream for a new audience [3].

“The spaceship lands. He steps out. He points it around. It says ‘friendly–unfriendly–edible–poisonous–safe–dangerous–living–inanimate’. On the next sweep it says ‘*Quercus oleoides–Homo sapiens–Spondias mombin–Solanum nigrum–Crotalus durissus–Morpho peleides–* serpentine’. This has been in my head since reading science fiction in ninth grade half a century ago.” (p. 731)

Janzen's solution to this classic problem involved building machines to identify species from their DNA. His predicted budget and proposed research team are "US\$1 million and five bright people." (p. 731). However, recent developments in computer architectures, as well as innovations in software design, have placed the tools needed to realize Janzen's vision in the hands of the scientific community not in several years hence, but now; and not just for DNA barcodes, but for digital images of organisms.

A recent survey of small-scale automated species identification system trials (<50 taxa), shows an average reproducible accuracy of over 85 percent with no significant correlation between accuracy and the number of included taxa or the type of group being assessed (e.g., butterflies, moths, bees, pollen, spores, foraminifera, dinoflagellates, vertebrates) [4]. These figures should be compared with the disturbingly few blind test studies of accuracy and consistency of human taxonomist identifications that have been published to date [5], [10]. Human cognition studies [5] suggest that human experts who are routinely engaged in particular discriminations can return accuracies in the range of 84 to 95 percent. But in the (far more common) cases in which trained personnel must deliver identifications for species they are not dealing with on a day-to-day basis self-consistencies drop to 67-83 percent and consensus consistencies between identifiers to 43 percent. Moreover, semi-automated and automated identifications—often involving thousands of individual specimens—can be made in a fraction of the time required by human experts and can be done on site, on demand, anywhere in the world.

Is there a need for such systems? After all, biology has been getting by without them for millennia. What makes anyone think computers can – much less should – replace human taxonomists or that the taxonomic communities efforts would not be better spent lobbying for increased government funding for tried and true traditional α -taxonomy?

If evidence existed to reassure the scientific community that most taxonomic identifications are accurate and consistent current identification practices situation might be tolerable. There is little such evidence. For example, 1997 a group of geologists organized a blind test to try to resolve a controversy over whether marine animals went extinct before or after the meteorite impact that marked the end of the Cretaceous Period.⁹ Four taxonomic experts were asked to identify species of microscopic foraminifera in a set of rock samples without being told the age of the samples. No consensus on when the animals died out was established – not because of any flaw in the test's design, but because the species lists produced were so different as to be incomparable, in some cases with just 25 percent of species names in common.

Contrary to some voices within the systematics community, these developments could not have come at a better time. As all scientists already know, the world is running out of specialists who can identify the very biodiversity whose preservation has become a global concern. In commenting on this problem in palaeontology as long ago as 1993, Roger Kaesler recognized¹² ...

“... we are running out of systematic paleontologists who have anything approaching synoptic knowledge of a major group of

organisms” (p. 329). “Paleontologists of the next century are unlikely to have the luxury of dealing at length with taxonomic problems ... [Paleontology] will have to sustain its level of excitement without the aid of systematists, who have contributed so much to its success.” (p. 330).

This expertise deficiency cuts as deeply into those commercial industries that rely on accurate identifications (e.g., agriculture, biostratigraphy) as it does into a wide range of pure and applied research programmes (e.g., conservation, biological oceanography, climatology, ecology).

If truth be told, it is commonly, though informally, acknowledged that the technical, taxonomic literature of all organismal groups is littered with examples of inconsistent and incorrect identifications. This is due to a variety of factors, including taxonomists being insufficiently trained and skilled in making identifications (e.g., using different rules-of-thumb in recognizing the boundaries between similar groups), insufficiently detailed original group descriptions and/or illustrations, inadequate access to current monographs and well-curated collections and, of course, taxonomists having different opinions regarding group concepts. Peer review only weeds out the most obvious errors of commission or omission in this area, and then only when an author provides adequate representations (e.g., illustrations, recordings, gene sequences) of the specimens in question.

Few have sought to compare the performance of alternative types of morphological data for biological species identification. This investigation contrasts results of form characterization via form factors, superposed landmark coordinates, landmark-registered semilandmark outlines, 3D semilandmark networks, and raw digital images for a test set of seven Recent planktonic foraminifer species. While all data types performed better than the qualitative assessment of morphological variation by human taxonomists, landmark-registered semilandmark outlines and raw digital images delivered the best performance in the context of approaches that could reasonably serve as the basis for fully automated species identification systems.

Systematics too has much to gain, both practically and theoretically, from the further development and use of automated identification systems. It is now widely recognized that the days of systematics as a field populated by mildly eccentric individuals pursuing knowledge in splendid isolation from funding priorities and economic imperatives are rapidly drawing to a close. In order to attract both personnel and resources, systematics must transform itself into a “large, coordinated, international scientific enterprise” [13] (p. 4). Many have identified use of the internet—especially via the world-wide web—as the medium through which this transformation can be made. While establishment of a virtual, GenBank-like system for accessing morphological data, audio clips, video files and so forth would be a significant step in the right direction, improved access to observational information and/or text-based descriptions alone will not address either the taxonomic impediment or low identification consistency issues successfully. Instead, the inevitable subjectivity associated with making critical decisions on the basis of qualitative criteria must be reduced, or at the

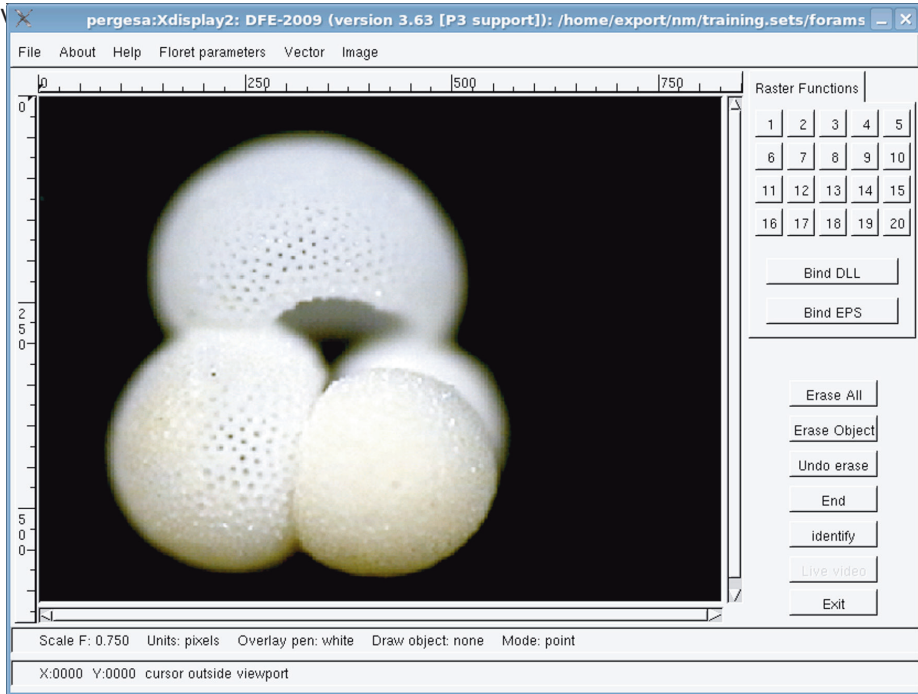


Fig. 1 – Example of the DAISY system interface displaying a planktonic foraminifer specimens from the test dataset. For this group, chamber arrangement, primary aperture position, and wall texture are among the primary taxonomic characteristics used to identify species.

Properly designed, flexible, and robust, automated identification systems, organized around distributed computing architectures and referenced to authoritatively identified collections of training set data (e.g., images, gene sequences) can, in principal, provide all systematists with access to the electronic data archives and the analytic tools necessary to handle routine identifications of common taxa. Properly designed systems can also recognize when their algorithms cannot make a reliable identification and refer that image to a specialist. Such systems will, inevitably, include elements of artificial intelligence that will allow them to improve their performance the more they are used. Most tantalizingly, once morphological (or molecular) models of a species have been developed and demonstrated to be accurate, these models can be queried to determine which aspects of the observed patterns of variation and variation limits are being used to achieve the identification, thus opening the way for the discovery of new and (potentially) more reliable taxonomic characters.

As has been demonstrated repeatedly through human history, scientific progress lies, in part, in constructing machines that do what machines do best and allowing humans to do what humans do best. Far from making taxonomists obsolete, the creation of automated identification systems will free them from the drudgery of delivering routine identifications to focus on the more conceptually

difficult issues of discovering, revising and describing species concepts, understanding how species fit into higher taxonomic and ecological groups and establishing how species function within natural systems. Getting high-throughput machine-learning systems on the agenda of research communities and scientific research funding councils, as well as into the study programmes of all sorts of disciplines, is required if taxonomy is to regain the sense of mission that will allow it to fulfil its potential as a twenty-first century science.

REFERENCES

- [1] R. R. Sokal and P. A. Sneath, *Principles of numerical taxonomy*. W. H. Freeman, San Francisco, 1963.
- [2] P. H. A. Sneath and R. R. Sokal, *Numerical taxonomy: the principles and practice of numerical classification*. W. H. Freeman, San Francisco, 1973.
- [3] Janzen, D. H., "Now is the time". *Philosophical Transactions of the Royal Society of London, Series B*, vol. 359, pp. 731–732, 2004.
- [4] K. J. Gaston and M. A. O'Neill, "Automated species identification—why not?" *Philosophical Transactions of the Royal Society of London, Series B*, vol. 359, pp. 655–667, 2004.
- [5] P. F. Culverhouse, R. Williams, B. Reguera, V. Herry and S. González-Gils, "Do experts make mistakes?" *Marine Ecology Progress Series*, vol. 247, pp. 17–25, 2003.
- [6] W. P. Colquhoun, "The effect of a short rest pause on inspection efficiency". *Ergonomics*, vol. 2, 367–372, 1959.
- [7] W. J. Zachariasse, W. R. Riedel, A. Sanfilippo, R. R. Schmidt, M. J. Brolsma, H. J. Schrader, R. Gersonde, M. M. Drooger and J. A. Brokeman, *Micropaleontological counting methods and techniques – an exercise on an eight meters section of the Lower Pliocene of Capo Rossello, Sicily*. Utrecht Micropaleontological Bulletins, vol. 17, pp. 1–265, 1978.
- [8] R. Simpson, P. F. Culverhouse, R. Ellis and R. Williams. Classification of *Euceratium gran*. pp. 223-230. Neural Networks. IEEE International Conference on Neural Networks in Ocean Engineering. IEEE, Washington, D. C., 1991.
- [9] R. N. Ginsburg, "Perspectives on the blind test". *Marine Micropaleontology*, 29, pp. 101–103.
- [10] Kelly, M. G. 2001. "Use of similarity measures for quality control of benthic diatom samples". *Water Research*, vol. 35, pp. 2784–2788, 1997.
- [11] K. W. Gobalet, "A critique of faunal analysis; inconsistency among experts in blind tests". *Journal of Archaeological Science*, vol. 28(4), pp. 377-386, 2001.
- [12] R. L. Kaesler, "A window of opportunity: peering into a new century of paleontology". *Journal of Paleontology*, vol. 67, pp. 329–333, 1993.
- [13] Q. D. Wheeler, "Transforming taxonomy". *The Systematist*, vol. 22, pp. 3–5, 2003.