THE ROYAL SOCIETY

Published online 22 March 2004

Automated species identification: why not?

Kevin J. Gaston[1*] **and Mark A. O'Neill**[2†]

[1]*Biodiversity and Macroecology Group, Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK*

[2]*Bee Systematics and Biology Unit, Hope Entomological Collections, Oxford University Museum of Natural History, Parks Road, Oxford OX1 3PW, UK* (*mao@herald.ox.ac.uk*)

Where possible, automation has been a common response of humankind to many activities that have to be repeated numerous times. The routine identification of specimens of previously described species has many of the characteristics of other activities that have been automated, and poses a major constraint on studies in many areas of both pure and applied biology. In this paper, we consider some of the reasons why automated species identification has not become widely employed, and whether it is a realistic option, addressing the notions that it is too difficult, too threatening, too different or too costly. Although recognizing that there are some very real technical obstacles yet to be overcome, we argue that progress in the development of automated species identification is extremely encouraging that such an approach has the potential to make a valuable contribution to reducing the burden of routine identifications. Vision and enterprise are perhaps more limiting at present than practical constraints on what might possibly be achieved.

Keywords: automated identification; computer-assisted taxonomy; routine identification; taxonomic impediment

1. INTRODUCTION

The 'taxonomic impediment' to biodiversity studies is multi-faceted. Alongside the oft-cited difficulties created by the rather low proportion of species that have been formally taxonomically described, there are several others. These include:

(i) the lack of an agreed list of described species (Hawksworth & Kalin-Arroyo 1995; May 1990, 2000);
(ii) the highly biased nature of the set of species that have been formally described (May 1988; Hammond 1992; Gaston 1993, 1994);
(iii) the many existing species names that are yet to be recognized as synonyms of other species names, and the revisionary efforts required to do so (Gaston & Mound 1993; Mound & Gaston 1993; Solow *et al.* 1995; Bouchet 1997);
(iv) the scattered, fragmented and taxonomically biased nature of the taxonomic workforce and its resources (Gaston & May 1992; Simonetti 1997);
(v) the general decline in the taxonomic workforce (Gaston & May 1992; Hopkins & Freckleton 2002);
(vi) the difficulties in becoming proficient in the identification of many taxa (Culverhouse *et al.* 1996; Do *et al.* 1999);
(vii) the difficulties of using traditional taxonomic products without adequate reference collections and extensive knowledge of arcane specialist terminology (Weeks *et al.* 1999*b*); and
(viii) the vast numbers of specimens (often of common species) for which routine identifications are required.

The last of these components of the taxonomic impediment has been little commented on, but is nonetheless highly significant, and unlikely to be resolved simply by initiatives that have been proposed or are being undertaken to address the other problems (although in some cases these will help significantly (Godfray 2002; Patterson 2003; Wilson 2003)). Although acknowledging that at some level repeated experience with specimens of different (often common) known species helps build the expertise of members of the taxonomic community, all else being equal such 'routine identifications' may distract taxonomic effort away from more fundamental revisionary and descriptive activities (of course, all else may not be equal, and such efforts may not easily be redeployed in this way). But, more importantly, much biodiversity work simply cannot be conducted, not because the species concerned are unknown to science, but because the taxonomic resources required to identify them are not available (Edwards & Morse 1995). Demand for routine identifications far outstrips the capabilities of the taxonomic community. To give an extremely parochial but not atypical example, recent studies by K.J.G. and colleagues in the UK (which has one of the best documented floras and

[*] Author for correspondence (k.j.gaston@sheffield.ac.uk).

† Present address: The Henry Wellcome Building for Neuroecology, University of Newcastle School of Biology, Framlington Place, Newcastle upon Tyne NE2 4HH, UK.

One contribution of 19 to a Theme Issue 'Taxonomy for the twenty-first century'.

DOI 10.1098/rstb.2003.1442

faunas) on the biodiversity associated with domestic gardens (increasingly significant in this region given the negative impacts of agricultural intensification and urbanization) were severely constrained by the impossibility of obtaining accurate identifications of very many of the invertebrate species encountered, although these were almost certainly in most cases formally taxonomically described. Such problems are even more severe in other circumstances, in other kinds of environment and in other regions of the world, where often their practical consequences may be much more acute.

Of course, the problem posed by the demand for routine identifications extends much beyond biodiversity studies *per se*. In many spheres the volumes of plant or animal specimens that can usefully be obtained, particularly using modern sampling methods, vastly outstrip any capacity to identify this material. This constraint forcefully limits progress in areas such as, for example, the monitoring of resource, pest, toxic, vector, disease and alien species populations and the generation of palynological and other data for long-term studies of environmental change (Culverhouse *et al.* 1996; McCall *et al.* 1996; France *et al.* 2000). Moreover, the demands for routine identifications are likely steadily to increase, as the proportion of previously undescribed species in local, national or regional floras and faunas declines, and as the requirement or desirability of biodiversity inventories and other such surveys grows.

Several solutions have been proffered to reduce the burden of routine identifications. These include

(i) improving the ease of use of existing tools for identification of known species, through for example, more accessible paper-based taxonomic keys, multi-access taxonomic keys, hypertext taxonomic keys, expert systems and providing material in other languages (White & Scott 1994; Edwards & Morse 1995; Dodd & Rosendahl 1996; Rambold & Agerer 1997; Dallwitz *et al.* 1998; Jarvie & Stevens 1998);

(ii) training parataxonomists and other such individuals to increase the workforce available to conduct such identifications (or, where appropriate, sufficient approximations thereof) (Gamez 1991; Cranston & Hillman 1992; Basset *et al.* 2000; Hyde *et al.* 2000); and

(iii) automating the identification process in some way (Culverhouse *et al.* 1996; Weeks *et al.* 1997, 1999*a*,*b*; Gauld *et al.* 2000).

Much of (i) and all of (iii) have generically been referred to as Computer Assisted Taxonomy or CAT (Chesmore 1999, 2000).

More generally, where possible, automation has been a common response when humankind has been faced with some activity that has to be repeated numerous times, when the labour costs of so doing have been deemed too high, when the required level of labour could not be obtained (a skills shortage), or when automation offers a faster, more replicable or more accurate outcome. The burden of routine taxonomic identifications meets at least some of these same criteria, and under some circumstances all of them. So far, however, the development and application of an automated approach to taxonomic identifications has remained a minority interest, with a small associated literature and little discussion in the wider arena. In this paper, we consider why this is so, and whether such an approach might offer a serious solution particularly to the burden of routine species identifications.

To address these issues, we take in turn each of several simple arguments as to why automated species identifications are not currently the norm, addressing the notions that they are too difficult, too threatening, too different or too costly. In so doing, emphasis will be laid principally on the automation of species identification based on morphological characters, acknowledging that the field is much broader and might include identifications made through molecular methods (Jonker *et al.* 2000; Garland & Zimmer 2002; Tautz *et al.* 2002, 2003; Hebert *et al.* 2003; Blaxter 2004), mass spectrometry (Jarman *et al.* 2000), sound (Vaughan *et al.* 1997; Chesmore *et al.* 1998; Chesmore 1999, 2000, 2001; Parsons & Jones 2000; Parsons 2001; Schwenker *et al.* 2003), movements (e.g. wingbeat patterns; Moore & Miller 2002), and radar and sonar (Simmonds *et al.* 1996). Many of the points raised apply equally to these other approaches. However, we contend that, despite widely acknowledged and sometimes severe limitations, just as morpho-taxonomy has thus far remained the backbone of taxonomic work (if for no other reason than that most of the material that is available or suitable to work with comprises dead specimens that have variously been pressed, dried and/or pickled), so identifications of individuals of described species will continue to depend heavily on such an approach.

2. IT IS TOO DIFFICULT

Perhaps the simplest explanation for why automated identifications have not become the norm for routine identifications is that such an approach is too difficult. In the limit this argument is undoubtedly wrong. There is no question that automated species identifications are possible, and have been so for a considerable period. Given appropriate images of two species that differ drastically in morphology, it is a relatively trivial task for an automated system to distinguish between them. But then, such identifications do not typically require the services of an expert taxonomist either. The issue is thus not whether automated identifications *per se* are too difficult, but whether morphologically similar species can be distinguished in an automated fashion with sufficient accuracy.

At the heart of automated species identification based on morphological characteristics lies the need for computerized pattern recognition systems. These have found a wide range of applications, including the recognition of human faces, fingerprints, palmprints and handwriting (Turk & Pentland 1991; Banarse *et al.* 2000; Wu & Zhou 2002; Guo *et al.* 2003; He *et al.* 2003; Lu *et al.* 2003; Tsalakanidou *et al.* 2003). Nonetheless, reliable automated species identification that would make a valuable contribution to addressing the burden of routine identifications constitutes a particularly hard problem. The primary difficulties are threefold. First, individuals of a given species may vary hugely in their morphology. Some of this variation is systematic, particularly the allometric scaling

Table 1. Some automated taxonomic identification systems based on morphological characteristics.

name	method	reference
Automated Leafhopper Identification System (ALIS)	discriminant function	Dietrich & Pooley (1994)
Digital Automated Identification (SYstem) (DAISY)	Lucas continuous *n*-tuple classifier/PSOM network	Gauld *et al.* (2000), O'Neill *et al.* (2000); http://chasseur.usc.edu/pups/projects/daisy.html
Automatic Identification and characterization of Microbial PopulationS (AIMS)	artificial neural networks (ANNs)	Jonker *et al.* (2000); http://www.flowcytometry.org/default.htm
Automatic Bee Identification System (ABIS)	support vector machines, kernel discriminant analysis	Arbuckle *et al.* (2001); Arbuckle (2002); http://www.informatik.uni-bonn.de/~arbuckle/abis/

of many features. But, much of the variation is idiosyncratic, reflecting the expression of individual genotypic variation, and phenotypic variation related to such things as age, environmental conditions experienced and accidents; taxonomists regularly distinguish specimens that exhibit key diagnostic features particularly well, and often place them in museum collections for this reason. In consequence, automated species identification is a matter of a one-to-many matching, in which the identification of a single specimen requires its matching to the morphological pattern of the species to which it belongs as characterized by several other individuals of that species. By contrast, human face matching, for example, is one-to-one matching, in which the identification of an individual face requires its matching to just the one face, albeit images of that face may vary in expression, pose, lighting, etc. and may be corrupted to varying extents during image capture.

The second problem for automated species identification is that closely related species may be extremely similar to one another. Indeed, the fine resolution of the morphological differences that discriminate between them comes as a surprise to many biologists, let alone people working in other fields. Detailed patterns in the form of particular morphological structures may be crucial, and may not always be readily captured in, for example, digital images of specimens.

The fact that the number of possible species to which a specimen may belong is effectively unbounded or at least the bounds are extremely broad, constitutes a third problem for automated species identification. In the extreme, globally there are at least a few million extant species of organisms, and perhaps many more (Hawksworth & Kalin-Arroyo 1995; May 2000), and many taxonomic groups may comprise thousands or tens of thousands of species exhibiting basically the same body plan, a proportion of which may be entirely unknown to science. Even for more narrowly constrained taxonomic groups and regions, many hundreds of species may share the same basic characteristics. Thus, ideally, an automated species identification system needs not only to be able to match an individual specimen with one of a set of known species, but if necessary also to be able to reject it as belonging to a species that is not part of this set. This particular kind of challenge is one that is not shared with many other problems that computerized pattern recognition systems have been employed to solve.

(a) *Approaches*

Accepting these difficulties, a small but growing number of studies have sought to develop and test systems for automated species identification based on morphological characters (tables 1 and 2). Typical steps in automated identification are outlined in figure 1, and comprise two processes, which are usually conducted sequentially, but may to some extent be performed in parallel. The first process comprises the analysis of specimens that have been independently and accurately identified to species, and are used to generate a training set on which the differences between species are determined. The second process comprises the analysis of specimens that are to be automatically identified (the 'unknowns').

The first step in automated identification based on morphological features, that of capturing digital imagery of the specimens (figure 1; a pattern vector), is common to both the specimens for the training set and the unknowns to be identified by the system. This can be accomplished in several ways, including

(i) offline capture using a digital camera, with captured imagery subsequently being uploaded to the identification system;
(ii) online capture using a flatbed scanner (this option is particularly good for two-dimensional objects such as slide-mounted insect wings); and
(iii) online capture using a charge couple device camera attached to an imaging card in a computer (this option is good for data capture from microscopes with a camera attachment).

Pattern vectors are then pre-processed to convert them into the correct format for subsequent analysis (figure 2). Typical operations applied to morphological image data at this stage may include cropping, edge extraction, histogram equalization and rectification to transform the images to a standard pose. In addition, many identification systems also try to reduce the dimensionality of the data, which may increase both throughput and/or accuracy, depending on the precise nature of the approach to identification that is used.

Pre-processed pattern vectors for specimens for a training set are analysed by using an identifier of some form (see below) to discriminate among those belonging to different species. Unknowns are effectively compared with

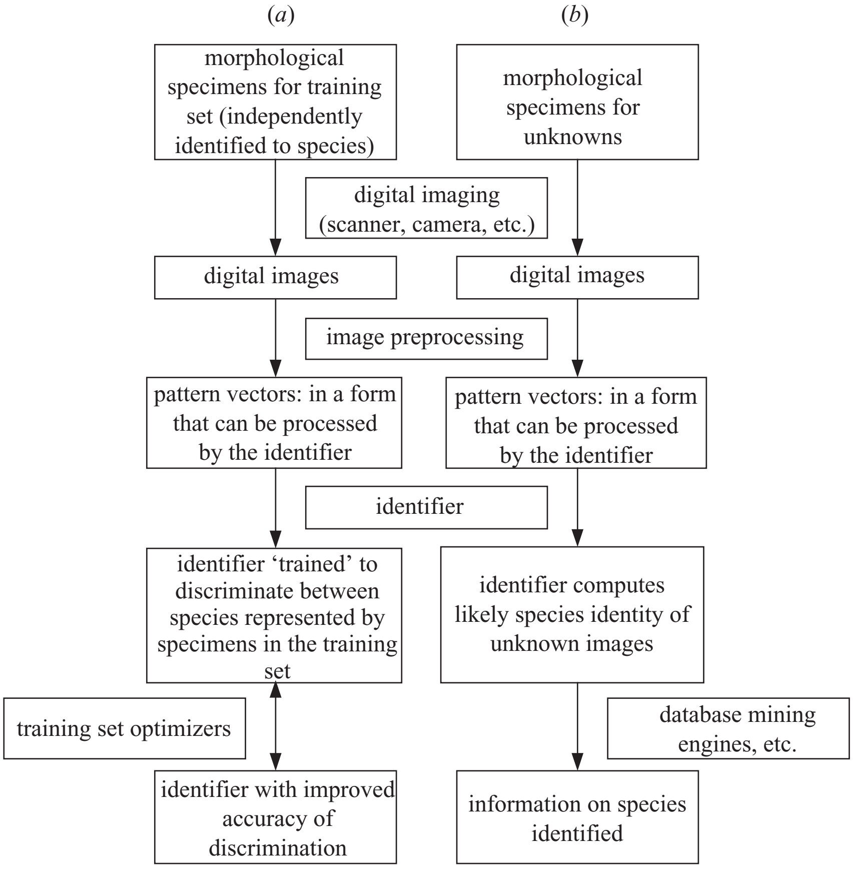
(*a*)
(*b*)
morphological specimens for training set (independently identified to species)
morphological specimens for unknowns
digital imaging (scanner, camera, etc.)
digital images
digital images
image preprocessing
pattern vectors: in a form that can be processed by the identifier
pattern vectors: in a form that can be processed by the identifier
identifier
identifier 'trained' to discriminate between species represented by specimens in the training set
identifier computes likely species identity of unknown images
training set optimizers
database mining engines, etc.
identifier with improved accuracy of discrimination
information on species identified

Figure 1. Schema showing basic processing pathways in a typical automated species identification system. (*a*) Procedure for specimens in the training set and (*b*) procedure for unknowns. See § 2a for details.

the training patterns, with the detail of the comparison being algorithm dependent. The identifier returns the pattern (or perhaps a short list of patterns) present in the training set that most closely resembles the unknown. In addition, if information about the organism identified is available, automated species identification systems may also present this to the user (e.g. as pages of HTML).

More complex automated species identification systems are also able to run concurrent background tasks to dynamically optimize system performance and/or throughput. For example, a stochastic training set optimizer may be used to optimize the composition of the training set for throughput, accuracy or both. Similar functionality can be added to affect other aspects of dynamic learning (e.g. plastic self-organizing map (PSOM) dynamic learning, manifold reduction).

Different studies of automated identification systems vary substantially in the degree of automation that they practically manage to achieve, but the key issue to be resolved is automation of the step that involves the conversion of one or more pre-processed images of an unknown into a species-level identification. If this can be achieved then the automation of other steps in the process should be relatively straightforward. Two broad approaches have been used in the automated taxonomic identification step itself, although these are increasingly convergent. They are rooted respectively in traditional statistical methods and artificial neural networks (ANNs).

(i) *Traditional statistical methods*

Some of the earlier attempts at automated taxonomic identification sought to discriminate between species (or higher taxa) based on using quite traditional statistical tools, such as PCA and linear discriminant analysis, applied to measures of several (sometimes many) of the sizes and shapes of, and the distances and angles between, features of specimens (Daly *et al.* 1982; Jeffries *et al.* 1984; Longford *et al.* 1990; van de Vooren *et al.* 1992; Yu *et al.* 1992). Although sometimes quite effective, such approaches typically ignore much of the available information about the morphological structure of specimens (concentrating on just a few features) and make overly restrictive assumptions about the statistical nature of the data (Boddy & Morris 1993; Weeks & Gaston 1997), and have now largely been abandoned.

These approaches have, however, in some senses been built upon to develop much more sophisticated techniques, which (i) operate more directly on the patterns of intensities and hues (or grey scales) of a digital image, automatically deriving information from the statistical structure of the imagery, and (ii) employ more appropriate and powerful statistical methods. One of the best developed of these approaches is probably that embodied in the Digital Automated Identification System known as DAISY (Weeks *et al.* 1997, 1999*a*,*b*; Gauld *et al.* 2000; O'Neill *et al.* 2000; see http://chasseur.usc.edu/pups/projects/daisy.html). This was initially motivated by the

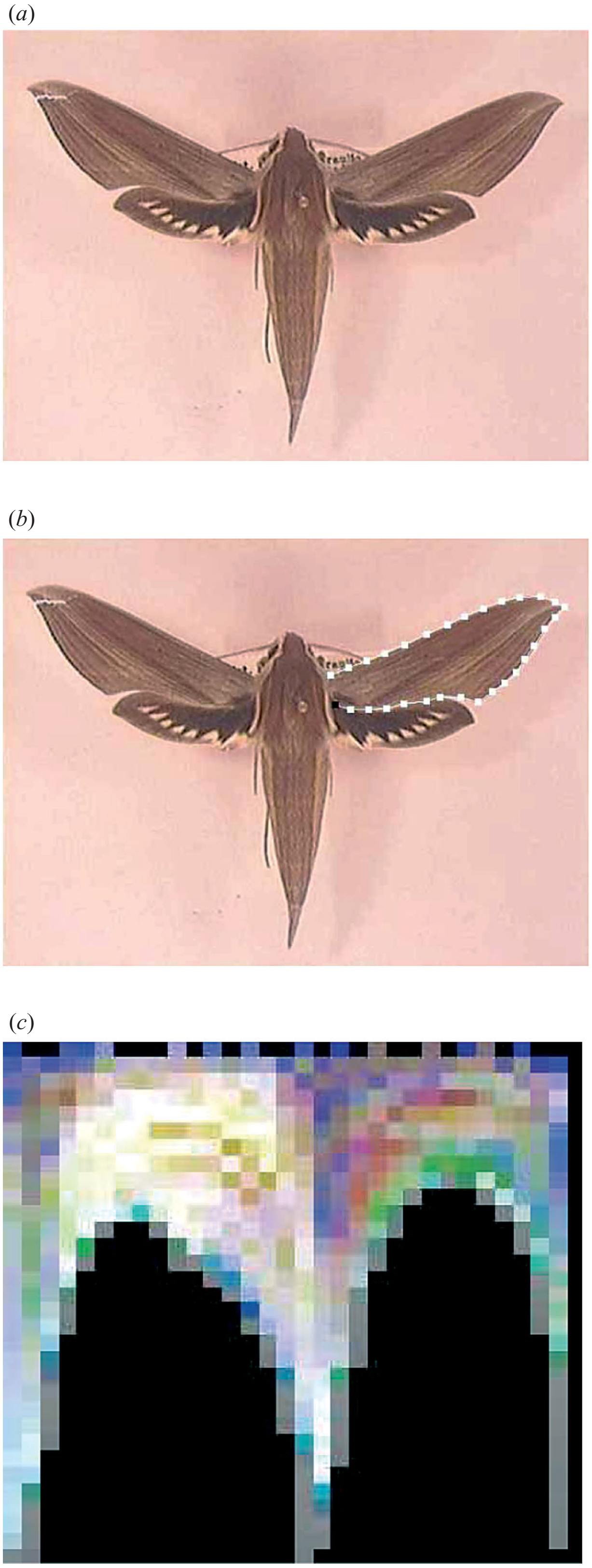

Figure 2. Successive stages of image capture in DAISY for a specimen of the sphingid moth *Xylophanes tersa*: (*a*) input image, (*b*) application of an overlay and (*c*) the resultant pattern vector.

progress that has been made in human face detection and recognition using fuzzy template matching techniques based on decomposing a training set into a linear set of orthogonal eigenimages (principal components; Turk & Pentland 1991). Unknown objects were identified by determining how well they correlated with an optimal linear combination of the principal component eigenfunctions. However, in practice, this approach proved to be computationally slow and error prone. It is slow because every time a new specimen is added to a training set the eigenfunctions (which are a low-dimensionality representation of morphological space) have to be recomputed. The method is prone to error because linear PCA is not good at modelling morphological spaces that may contain nonlinear regions (see Bichsel & Pentland 1994). The current variants of DAISY use a hybrid identification scheme based on the Lucas continuous *n*-tuple classifier (see Lucas 1997; a variant of nearest-neighbour classification; NNC) and the PSOM (Lang & Warwick 2002), which is itself a variant of Kohonen-based classification algorithms (which are essentially unsupervised ANNs, see below) capable of dynamic learning.

An *n*-tuple (NNC) classifier simply compares an unknown image with a set of images (the training set) which have been assigned to pre-defined classes by an expert. The *tuples* are the pixel pairs [$\text{Unknown}_{i,j}$, $\text{Tset}_{i,j,k}$] which are compared by using some variant of cross correlation. Thus

$$\alpha_k = \sum \chi(\text{Unknown}, \text{Tset}_k), \qquad (2.1)$$

where α_k is the *affinity* between the unknown image Unknown and the training set image Tset_k and χ is a cross-correlation function. The continuous *n*-tuple classifier, in its simplest form sets the class of the unknown to that training set image $T_{k,\text{max},}$ which has the greatest affinity $\alpha_{k,\text{max}}$ for the unknown image given cross-correlation function χ.

PSOM is a variant of the continuous *n*-tuple classifier that can learn dynamically. In the case of PSOM, the training set image $T_{k,\text{max}}$, which best correlates with the unknown is moved (in morphological) space in the direction of the unknown by mixing in a small amount of the unknown image that it has just identified

$$T_{k,\text{maxPSOM}} = |\lambda \text{Unknown} + T_{k,\text{max}}|_{\text{norm}}, \qquad (2.2)$$

where λ is the learning ratio (typically much less than 1.0), and norm implies that the vector is to be renormalized so that $\Sigma T_{i,j,k} = 1.0$. The effect of applying dynamic learning in this fashion is to adapt the training set distribution to the actual unknowns presented to the system. Of course, in a realistic implementation, learning algorithms such as PSOM are applied only if the system is certain of the identity of the unknown to a high degree of confidence. Note that in addition to reinforcement learning (λ positive), if an expert user is interacting with the system we can also have inhibition (λ negative). In this case the system is punished by the user for misidentification with the result that $T_{k,\text{max}}$ is moved away from Unknown. The *n*-tuple and PSOM approaches are both capable of dealing with nonlinear morphological spaces, and in addition they are modular: extra pattern vectors can be added to the training set with minimum computational overhead compared to linear PCA.

Unfortunately, so far, the general capabilities of DAISY remain rather poorly explored, with most empirical studies having been conducted using images of the wings of specimens from insect groups for which constituent species are known to be difficult to distinguish from these morphological features alone (Weeks *et al.* 1997, 1999*a*,*b*; Gauld *et al.* 2000). For example, the project has used groups of biting midges, bees, ichneumonid wasps and, most extremely, aphidiine wasps some species of which are impossible for even expert taxonomists to distinguish using morphological features alone. The levels of successful identification achieved for these examples (table 2) suggest, however, that this is likely to be quite a powerful approach for more amenable groups. For example, using only 5–10 training images per species, DAISY has been found to be able to identify more than 80% of unknowns for the British butterfly fauna (*ca.* 60 species) (M. A. O'Neill, unpublished data).

The use of modular identifiers within DAISY means that adaptive (self-learning) identifiers who are capable of modifying their own training data may be implemented relatively easily. Indeed, the implementation of meta-identifiers, which use statistical clustering techniques to build species identifiers in an emergent fashion, have proven effective (Gauld *et al.* 2000). These have been employed reliably to distinguish among morphologically extremely similar groups of taxa, including Africanized and non-Africanized honeybees *Apis mellifera*, and the two mosquito subspecies *Culex pipiens pipiens* and *C. pipiens molestus*. The use of modular identifiers also allows the implementation of manifold reduction: effectively a stochastic optimization algorithm is run in the background, which seeks an optimal subset of pixels for each image in the training set, which maximizes the affinity of that image with other images of the same species while minimizing the affinity of the image to images of other species. This can improve discrimination between very similar species, where the level of overall morphological similarity can mask important localized morphological differences. For example, *Aricia artaxerxes* and *Aricia agestis*, two species of lycaenid butterflies, are readily separable by the human expert as the former has prominent white spots on the forewing whereas the latter does not. Holistic systems may not separate these taxa well, if the signal from the (small) spot is drowned out by the signal from the rest of the forewing (which is very similar in both species).

(ii) *Artificial neural networks*

ANNs are the most commonly employed computerized pattern recognition tool, and have been used widely in automated taxonomic identification based on morphological features (tables 1 and 2). ANNs are information-processing structures modelled after the massively parallel structure of the brain. They comprise nodes interconnected in layers to form a network, and take many forms. ANNs are not rule based, but are trained on examples of the taxa to be identified, an iterative process that can be time consuming, with the internal organization of the network being altered until it can successfully distinguish between these taxa.

The best developed of the applications of ANNs to automated taxonomic identification is the work of Boddy, Morris and colleagues on the identification of phytoplankton species, which has explored many issues in this field (Boddy *et al.* 1994, 1998, 2000, 2001; Wilkins *et al.* 1999; Boddy & Morris 2000; Morris *et al.* 2001). Data acquisition here is rather different from that of most of the other studies cited in this paper (but see also Balfoort *et al.* (1992)), in that it employs flow cytometry, and does not just incorporate information on specimen morphology. A stream of cells is illuminated by a laser, and a set of measurements for each cell is made in relation to the beam, including the time of passage through the beam, various fluorescences, scatter and polarization. These give information on the size of the cell, indications of its structure and chemical content. Processing of these data by ANNs has given reasonably good levels of successful identification (Boddy *et al.* 1994, 2000), with, for example, ANNs trained on 54, 62 and 72 phytoplankton taxa identifying them with overall success rates of, respectively, 73, 77 and 70% (Boddy *et al.* 2000).

Two main classes of ANNs have been used in automated taxonomic identification systems. The first is supervised ANNs, which are used in the flow cytometry system developed by Boddy and colleagues. The second is unsupervised ANNs, which are (effectively) used in the latest variants of the DAISY system. The major disadvantage of supervised ANNs is the time that is required to train them offline. In addition, some types of supervised ANN also suffer from the same type of non-modularity as PCA: if a new pattern vector is added to the training set the entire network needs to be retrained as a consequence. This would clearly be problematic for any production automated taxonomic identification system that must deal with hundreds to thousands of taxa while new material is continuously added to its training sets. The advantage of the supervised approach is that it may be possible both robustly to reject as unidentifiable those specimens which are not in any class (species) known to the system, and to distinguish more effectively between species whose morphology is very similar. The advantages and disadvantages of unsupervised approaches (Kohonen, PSOM) are essentially the reverse, in that the system can be trained while running so that there is no need for offline training, and that they tend to be modular: new material causes only local changes to the network which means that training is fast. Perhaps the best way forward is a hybrid approach with unsupervised foreground training combined with supervised background training. For example, DAISY uses a PSOM network or NNC in the foreground, while running a set of background supervised stochastic optimization processes which adjust training set composition and pixel masking parameters for the training set images to maximize the similarities between members of the same class while minimizing similarities between members of different classes. This process is similar in both concept and execution to the offline training of supervised ANNs. Effectively, this means that DAISY trains in a supervised manner when it has nothing else to do. When an identification is required, the system switches into an online mode, and uses the training set it has generated to date to do the identification. An interesting effect of this is that like a human expert, the capabilities of DAISY will become enhanced over a period of time as it becomes more experienced.

Approaches have also been developed using ANNs for detecting the presence of specimens of particular individual species against a background of many others (Morris *et al.* 2001), a problem that is reasonably commonly encountered in many applied situations (e.g. screening for pest species). Essentially this reduces to the problem of distinguishing species A from not species A, and may be significantly simpler to achieve than a full identification system. Essentially one needs to establish a threshold distance from the centre of the morphological space of the species of concern (e.g. one or more pests), with specimens that fall within this perimeter being deemed to belong to these species, and those falling outside not so. The system may fail if there are species that are not of concern that are morphologically very similar to the species that are.

(b) *Problems*

Almost whatever the approach employed, there are some common significant issues associated with reliable automated species identification. Here, we highlight four: the quality of training sets, the nature of errors in identification, how to scale the process up to differentiate among larger numbers of species, and how to deal with species that a system has not been trained to identify.

(i) *Training sets*

The quality of training sets is important in obtaining reliable identifications. Such sets need to comprise high-quality images of specimens that have been accurately identified, to be reasonably large, and, while avoiding particularly aberrant specimens, to capture sufficient of the breadth of morphological variation exhibited by individuals of each species.

Accurate identification of specimens in training sets is plainly an essential prerequisite to the functioning of any automated identification system. The use of high-quality images in these sets is generally advantageous, in that it is clear that discrimination between species is then based on their morphological characteristics rather than other factors. The degree of standardization of the form of such images that is required is more problematic. Increased standardization of matters such as pose, background and lighting, for example, will tend to narrow the apparent morphological space occupied by specimens of different species, increasing the likelihood of being able to distinguish between them. However, too great a standardization of this form can mean that if images of unknowns that are subsequently presented to the system for identification vary in some of these considerations then they may stand an enhanced likelihood of being misidentified. Clearly, some compromise is required.

Most studies of automated identification systems (with the principal exception of some based on flow cytometry, for which lots of data can be generated quickly) have employed training sets with relatively small numbers of specimens (5–10) per species, while commonly observing that ideally they should be larger and that this would tend to improve the accuracy of identifications. It would seem likely that, in general, performance will increase asymptotically with the size of a training set, with larger sets being required to distinguish effectively between species that are narrowly separated in morphological space. Establishing larger training sets can be difficult where some of the species to be included are rare, and specimens may be problematic to obtain; this problem is exacerbated in situations where any marked variation in the size of training sets for different species may reduce the accuracy of identifications (Boddy *et al.* 2001; Morris *et al.* 2001). As previously mentioned, analytical techniques exist to improve the quality of training sets once they have been created.

(ii) *Errors in identification*

Although the overall frequency of correct identifications of specimens achieved by some automated identification systems can be quite impressive, particularly given that the identification problems being posed may be difficult, for many purposes the error rates are not always trivial (table 2). Whereas studies have tended to concentrate on their successes, to potential users of such systems or of their outputs the failures may be just as, or more, important.

Most studies of automated species identification systems have paid insufficient attention to errors in identification, but these do seem to exhibit some general features. Most significant is that, not infrequently, errors are highly clumped, with specimens of a small proportion of species contributing to a high proportion of errors, and those errors tending to be systematic. This means that more certainty can be attached to the correct identifications of specimens of some species than of others, and for species where that certainty is lower there is a high likelihood of being able to say what any misidentification will be. Attaching levels of certainty to identifications markedly improves the value of automated identification, although absolute probabilities can only be given if the unknowns are drawn from the same statistical set of objects as the training set.

Different approaches to automated species identification may tend to lead to different error patterns. For example, different pre-processing methodologies (histogram stretch, wavelet transforms, edge extraction etc.) lead to different error patterns in the DAISY identifier (M. A. O'Neill, unpublished data). Empirically these error patterns have been found to be essentially orthogonal. Identification approaches with different pre-processing methodologies may therefore potentially be tied into voting ensembles in which the identity of a specimen is deemed to be the species that receives the most votes. These ensembles may be significantly more reliable at identification than single approaches. It is, of course, possible to take this idea further. For example, one could envisage automated identification being achieved through a voting ensemble of completely different automated methods of species recognition. This sort of approach has parallels with biological vision systems (see, for example, Dietterich 2002).

In particular, because rather little effort has thus far been directed towards so doing, there is every reason to believe that present levels of errors in identification can be significantly reduced, and even some relatively simple steps would pave the way. For example, most studies so far have employed only single morphological structures in generating pattern vectors (e.g. wings), and often ones that tend to be quite information poor, and yet taxonomists typically examine several such structures before making a species identification.

Table 2. Examples of studies of semi-automated and automated species identification based on morphological characteristics.

taxon	structure	number of species	method(s)	success (%)[a]	reference
Africanized and non-Africanized honeybees	various	1 (2 subspecific variants)	discriminant analysis	96	Daly *et al.* (1982)
plants	pollen	6	discriminant analysis	> 87	Longford *et al.* (1990)
algae	cell	8	ANN	> 90	Balfoort *et al.* (1992)
mushrooms	sporophore	1 (8 subspecific variants)	linear statistical methods	80	van de Vooren *et al.* (1992)
ichneumonid wasps	wing	5	univariate methods, discriminant analysis	100	Yu *et al.* (1992)
phytoplankton	cell	40	ANN	> 70	Boddy *et al.* (1994)
dinoflagellates	cell	23	ANN, discriminant analysis, k-nearest neighbours	56–83	Culverhouse *et al.* (1996)
phytoplankton	cell	9/14	discriminant analysis	> 83	McCall *et al.* (1996)
plant fungal pathogens	spore	16/19	ANN, discriminant analysis, k-nearest neighbours	63–79	Morgan *et al.* (1998)
phytoplankton	cell	5	optical correlation	> 90	Pech-Pacheco & Alvarez-Borrego (1998)
wolf spiders	genitalia	6	ANN	81	Do *et al.* (1999)
ichneumonid wasps	wing	5	principal component associative memories	94	Weeks *et al.* (1997, 1999*a*)
biting midges	wing	49	principal component associative memories	86	Weeks *et al.* (1999*b*)
phytoplankton	cell	34	ANN	92	Wilkins *et al.* (1999)
phytoplankton	cell	62–72	ANN	70–77	Boddy *et al.* (2000)
plants	pollen	3	ANN	83	France *et al.* (2000)
aphidiine braconids	wing	15	Lucas continuous *n*-tuple classifier	40–60	Gauld *et al.* (2000)
Africanized and non-Africanized honeybees	wing	1 (2 subspecific variants)	Lucas continuous *n*-tuple classifier	100	Gauld *et al.* (2000)
mosquitoes (subspecies)	wing	2	Lucas continuous *n*-tuple classifier	> 98	Gauld *et al.* (2000)
bumble-bees	wing	4	support vector machines, kernel discriminant analysis	> 95	Arbuckle *et al.* (2001)
macrolepidoptera	wing	35	Lucas continuous *n*-tuple classifier	83	Watson *et al.* (2003)
bumble-bees	wing	4	Lucas continuous *n*-tuple classifier	> 94	M. A. O'Neill *et al.* (unpublished data)
wolf spiders	genitalia	6	Lucas continuous *n*-tuple classifier	84	M. A. O'Neill (unpublished data)

[a] Some studies conducted a variety of tests of the level of successful identification; the figures given here are overall levels of success or typical values, or the lower bounds thereof.

(iii) *Scaling*

If sufficiently reliable automated identifications can be made among a few species, the issue remains as to how to scale the process up to differentiate among large numbers (the largest implementation of which we are aware was for *ca*. 200 species using DAISY, which achieved *ca*. 90% accurate identification; M. A. O'Neill, unpublished data). This is a non-trivial problem. First, in general, increasing the number of species amounts to increasing the breadth of taxa under consideration. Different taxa often require different sets of features to be examined to generate species-level identifications, which requires different kinds of species identifier. This raises the possibilities either of having a semi-automated system in which identification to higher taxa is done manually, or of employing some form of hierarchical automated identification, in which individuals are first identified to higher taxonomic groups and then identified to species. The latter is generally regarded as providing a more workable approach, at least for the foreseeable future, although comparisons of hierarchical and non-hierarchical automated identification approaches have reached variable conclusions (Boddy *et al.* 1994, 2000).

In general, it is often easier correctly to identify a specimen to a higher taxon than it is to identify it to species, especially in the case of taxa containing species swarms and sibling species that are morphologically very similar. However, it is also computationally more efficient to split the identification process into multiple steps. For example, when identifying bees, ABIS (see http://www.informatik.uni-bonn.de/~arbuckle/abis/) first identifies a list of possible genera. These are passed to species identifiers for final identification. The gain here is that the genus identifier effectively removes the need to process many species training sets. Other systems can also be run in a similar configuration. In the case of DAISY, algorithms exist to generate optimal genera training sets which are populated by the best species images for each genus (e.g. those that are most typical of the genus and which therefore lie towards the centre of the species clusters in morphological space) and are therefore in some way typical of these species. Such approaches may, in principle, be able to identify the higher taxon to which specimens belong even if no appropriate species identifiers are included in the system.

Second, using at least some approaches, the proportion of accurate identifications declines with the number of species the system can potentially identify (Gauld *et al.* 2000). Among closely related species the problem seems to be an increasing overlap in the morphological space occupied by different species. How severe this problem could become is unknown, because there has been too little work in building systems to identify significant (or for most purposes even realistic) numbers of species. However, it may be lessened by the use of manifold reduction (see § 2a) to highlight areas of morphological difference between species, or of methods to amplify the weightings of particular morphological features, using for example contrast stretching or feature extraction algorithms, such as local feature analysis (see Penev & Atick 1996) or compact wavelet transforms (see Press *et al.* 1992). Contrast stretching or feature extraction algorithms are passive, and are applied in the hope that they amplify interspecific differences while minimizing intraspecific ones without actively optimizing for this state of affairs. Essentially, the use of compact wavelets to amplify interspecific differences while minimizing intraspecific ones is another example of an active manifold reduction optimization. Like the simple manifold reduction mentioned already, it is driven by a stochastic optimization process. In this case, the optimization is in a space which is the wavelet transformation of morphological space, and the stochastic optimizer seeks a set of wavelets φ which accomplish this. In the straightforward manifold reduction, a set of pixels θ is stochastically sought which fulfil this criterion.

Third, because automated identification is computationally intensive, as the number of possible species to be identified increases, the acquisition of sufficient computing power rapidly becomes limiting. The provision of a computational environment that scales seamlessly is a difficult computational problem. Traditionally parallel systems have been based on message passing (e.g. Parallel Virtual Machine (PVM); see, for example, Geist *et al.* 2003). Whereas these sorts of system work well for large-scale scientific data processing (weather forecasting, computation of protein tertiary structure, molecular engineering, etc.), they are not the most appropriate methods for automated species identification systems. In parallel computational terms, these systems are coarse grained compared with physical computations. Consequently, the kinds of technology developed for the implementation of distributed databases may be more appropriate to parallel automated species identification system implementation. DAISY, which has been designed to run in parallel from the outset, uses an environment called P3M (O'Neill *et al.* 2002). This agent-based system allows DAISY to make optimal use of both computing clusters and parallel machines in a dynamic fashion. It also provides novel methodologies for building networks of cooperating identifiers in an emergent fashion using ideas derived from biological systems. As far as we know, DAISY is the only automated species identification system that can use cluster and parallel computational environments at the present time.

(iv) *Novel species*

The development and testing of systems for automated species identification has almost exclusively been concerned with closed assemblages, in which a training set is established for a set of species, and the system is tested using other specimens of these same species. In most cases, specimens of other species will tend to be identified as belonging to one or other of this assemblage, although it may be possible to screen these out on the basis of low measures of their likelihood of being a species in the training set (Wilkins *et al.* 1999). Systems that discriminate specimens of single species against a background of many others (the species A, not species A problem) can potentially be employed in multiple forms (one for each species in the training set) to address this issue (Jonker *et al.* 2000; Morris *et al.* 2001).

Solving the problem of what to do with novel species is probably the most serious challenge remaining to the development of automated species identification systems. It is, however, also a problem that plagues other approaches to solving the burden of routine species identifications. For example, trying to identify specimens of species that are not included in a taxonomic key, but which belong to the same group, may often lead to false identifications, with no necessary indication that this is the case.

(c) *Synthesis*

It is far too simplistic to argue that automated species identification should not be pursued as a solution to the burden of routine identifications because it is too difficult. With many studies having focused on groups of species that are known to be hard to distinguish (see Weeks *et al.* 1997, 1999*a*,*b*; Pech-Pacheco & Alvarez-Borrego 1998; Gauld *et al.* 2000; Arbuckle *et al.* 2001), the high success rates often achieved (in some cases greater than those achievable by experts working with the same material) suggest that technically reliable systems of wide applicability are well within our grasp, and the low success rates sometimes encountered should not be too discouraging (table 2). Indeed, one might argue that extraordinary progress has been achieved for, viewed in the wider context, surprisingly little effort and resources. Further investments should readily serve to ensure developments that increase levels of successful identifications, and systems capable of handling many species and novel species.

3. IT IS TOO LABOUR INTENSIVE

Automating species identifications does not remove the need for human involvement in the process. First, material for identification needs to be prepared. Doing this correctly and in a standardized fashion may be critical. Second, images of the material need to be obtained, and again this commonly has to be done in a standardized way, minimizing differences in such things as pose and illumination, for example. In some instances, such as working with pollen grains, even locating on a slide, for example, the material to be imaged (differentiating it from other material) may be a non-trivial task to automate (France *et al.* 2000). Third, some element of preprocessing of images is usually required, often to standardize and enhance these images to enable extraction of the important features. In principle, it should be possible to automate this step, but this is not always straightforward, or computationally efficient. For example, it is entirely possible to extract regions of interest (e.g. insect wings) from a cluttered image using statistical edge detectors; however, this may take a long time (of the order of minutes to hours). It may also be possible to circumvent this problem by using a brute force approach, for example employing very large training sets that span typical variation in pose, illumination and scale (in addition to the phenotypical variation exhibited by the species). This is attractive, but would only work if appropriate parallel computing architectures (e.g. the Computational MicroBot (CMB) machine proposed by O'Neill & Curtis-Rouse (2002)) or quantum computing architectures (see, for example, Grover 1997) were available as realizable hardware.

In establishing identifications of many specimens, performing these activities by hand may constitute substantial amounts of work (as well as comprising repetitive operations, with high risks of operator fatigue and boredom). This could perhaps be argued to some degree to undermine the utility of automated identification. However, the skills required to conduct these steps are fundamentally different from those involved in the identification step itself, and do not require the expert taxonomic identification skills that are the constraint on many routine identifications. Automating these other steps would, nonetheless, obviously be attractive.

4. IT IS TOO THREATENING

Another possible explanation for why automated taxonomic identification has not become the norm for routine identifications is that there has been resistance from the taxonomic community to the development of such an approach. Such a response, conscious or otherwise, might be expected on the grounds first that the fundamentals of how routine species identifications are conducted have remained essentially the same since Linnaeus, and the suggestion that hard-won human expertise can in any sense be replaced by machines has historically often proven threatening. Against this background the vision necessary to fund, explore, develop and embrace an entirely different approach will always be difficult to sustain, particularly in the face of competing demands for resources and given the lack of a proven prototype system.

To the extent that there is any truth in this explanation, it does not undermine the practical possibilities that automated species identification may provide for routine identifications. It should, however, encourage sensitivity in the development and promulgation of such techniques, and the clear recognition and communication of what they can and cannot achieve.

5. IT IS TOO DIFFERENT

In a related vein, the development of the tools for automated species identification undoubtedly requires access to sets of skills that are not typically encountered among systematists or within the departments and institutions in which the bulk of formal taxonomic identifications are conducted. This may well have hindered the development of automated identification and will probably continue to do so. Developing such approaches requires novel collaborations between biologists and computer scientists, and personnel who have significant knowledge of both biology and computing science. The present climate of encouragement for interdisciplinary research could do much to fulfil these needs.

6. IT IS TOO COSTLY

A final explanation for why automated species identification has not become the norm could be that it is too costly to implement. The difficulty here, as often encountered in other fields, is in resourcing product development. Almost all studies so far have been proof of principle/concept, apparently funded through scientific research grants and the like. Turning this work into reliable tools that can be deployed widely requires funding from other sources. Because fully functional automated identification systems are complex combinations of hardware and software, the implementation of a production system is an expensive undertaking. At present, this does not seem to be regarded as commercially viable. We contend, however, that this results largely from a significant underestimation of the possible scale on which such a system might be employed, with its possible applications in museums, universities, research institutes, pest/health control laboratories and programmes, customs points, schools, and so forth.

7. IN CONCLUSION

Although the enormous potential that faster and more sophisticated computing offers to the field of pattern recognition has been exploited in many fields, in that of species identification this remains at a comparatively early stage. Indeed, the principal obstacle to the widespread application of such an approach to resolving the burden of routine identifications at present is arguably more one of a lack of vision and enterprise than any real practical considerations. The technical challenges remain considerable; however, the huge progress that has been made by a few largely exploratory projects is impressive. It suggests that the bounds on just what it is possible to achieve remain to be established. Cultural issues and availability of adequate and appropriate resources have undoubtedly severely constrained progress so far. However, given the magnitude of the possible prize—a generic automated species identification system that could open up vistas of new opportunities for pure and applied work in biological and related fields—it would seem foolish not to find ways of overcoming these obstacles in the future.

The authors thank the editors for the invitation to write this piece. Discussions with several people have helped shape their views on the role and practice of automated identification; in particular they are grateful to I. D. Gauld and P. J. D. Weeks. A. Cannon, M. Denos, S. Gaston, I. D. Gauld and O. L. Petchey commented on the manuscript.

REFERENCES

Arbuckle, T. 2002 Automatic identification of bees' species from images of their wings. In *Proc. 9th Int. Workshop on Systems, Signals and Image Processing*, pp. 509–511. Manchester, UMIST.

Arbuckle, T., Schröder, S., Steinhage, V. & Wittmann, D. 2001 Biodiversity informatics in action: identification and monitoring of bee species using ABIS. In *Proc. 15th Int. Symp. Informatics for Environmental Protection, ETH Zurich, 10–12 October 2001*, vol. 1, pp. 425–430. Zurich: Metropolis.

Balfoort, H. W., Snoek, J., Smits, J. R. M., Breedveld, L. W., Hofstraat, J. W. & Ringelberg, J. 1992 Automatic identification of algae: neural network analysis of flow cytometric data. *J. Plankton Res.* **14**, 575–589.

Banarse, D. S., France, I. & Duller, A. W. G. 2000 Analysis and application of a self-organising image recognition neural network. *Adv. Engng Software* **31**, 937–944.

Basset, Y., Novotny, V., Miller, S. E. & Pyle, R. 2000 Quantifying biodiversity: experiences with parataxonomists and digital photography in New Guinea and Guyana. *BioScience* **50**, 899–908.

Bichsel, M. & Pentland, A. P. 1994 Human face recognition and the face image set's topology. *CVGIP: Image Understanding* **59**, 254–261.

Blaxter, M. L. 2004 The promise of a DNA taxonomy. *Phil. Trans. R. Soc. Lond.* B **359**, 669–679. (DOI 10.1098/rstb.2003.1447.)

Boddy, L. & Morris, C. W. 1993 Analysis of flow cytometry data—a neural network approach. *Binary* **5**, 17–22.

Boddy, L. & Morris, C. W. 2000 Artificial neural networks for identification. In *Proc. Inaugural Meeting of the BioNET-INTERNATIONAL Group for Computer-Aided Taxonomy (BIGCAT)* (ed. D. Chesmore, L. Yorke, P. Bridge & S. Gallagher), pp. 29–37. Egham: BioNET-INTERNATIONAL Technical Secretariat.

Boddy, L., Morris, C. W., Wilkins, M. F., Tarran, G. A. & Burkill, P. H. 1994 Neural network analysis of flow cytometric data for 40 marine phytoplankton species. *Cytometry* **15**, 283–293.

Boddy, L., Morris, C. W. & Morgan, A. 1998 Development of artificial neural networks for identification. In *Information technology, plant pathology and biodiversity* (ed. P. Bridge, P. Jeffries, D. R. Morse & P. R. Scott), pp. 221–231. Wallingford: CAB International.

Boddy, L., Morris, C. W., Wilkins, M. F., Al-Haddad, L., Tarran, G. A., Jonker, R. R. & Burkill, P. H. 2000 Identification of 72 phytoplankton species by radial basis function neural network analysis of flow cytometric data. *Mar. Ecol. Prog. Ser.* **195**, 47–59.

Boddy, L., Wilkins, M. F. & Morris, C. W. 2001 Pattern recognition in flow cytometry. *Cytometry* **44**, 195–209.

Bouchet, P. 1997 Inventorying the molluscan diversity of the world: what is our rate of progress? *Veliger* **40**, 1–11.

Chesmore, D. 1999 Technology transfer: applications of electronic technology in ecology and entomology for species identification. *Nat. Hist. Res.* **5**, 111–126.

Chesmore, D. 2000 Methodologies for automating the identification of species. In *Proc. Inaugural Meeting of the BioNET-INTERNATIONAL Group for Computer-Aided Taxonomy (BIGCAT)* (ed. D. Chesmore, L. Yorke, P. Bridge & S. Gallagher), pp. 3–12. Egham: BioNET-INTERNATIONAL Technical Secretariat.

Chesmore, E. D. 2001 Application of time domain signal coding and artificial neural networks to passive acoustical identification of animals. *Appl. Acoustics* **62**, 1359–1374.

Chesmore, E. D., Femminella, O. P. & Swarbrick, M. D. 1998 Automated analysis of insect sounds using time-encoded signals and expert systems—a new method for species identification. In *Information technology, plant pathology and biodiversity* (ed. P. Bridge, P. Jeffries, D. R. Morse & P. R. Scott), pp. 273–287. Wallingford: CAB International.

Cranston, P. & Hillman, T. 1992 Rapid assessment of biodiversity using biological diversity technicians. *Aust. Biol.* **5**, 144–154.

Culverhouse, P. F. (and 11 others) 1996 Automatic classification of field-collected dinoflagellates by artificial neural network. *Mar. Ecol. Prog. Ser.* **139**, 281–287.

Dallwitz, M. J., Paine, T. A. & Zurcher, E. J. 1998 Interactive keys. In *Information technology, plant pathology and biodiversity* (ed. P. Bridge, P. Jeffries, D. R. Morse & P. R. Scott), pp. 201–212. Wallingford: CAB International.

Daly, H. V., Hoelmer, K., Norman, P. & Allen, T. 1982 Computer-assisted measurement and identification of honey bees (Hymenoptera. Apidae). *Ann. Entomol. Soc. Am.* **75**, 591–594.

Dietrich, C. H. & Pooley, C. D. 1994 Automated identification of leafhoppers (Homoptera: Cicadellidae: *Draeculacephala* Ball). *Ann. Entomol. Soc. Am.* **87**, 412–423.

Dietterich, T. G. 2002 Ensemble learning. In *The handbook of brain theory and neural networks*, 2nd edn (ed. M. A. Arbib), pp. 405–408. Cambridge, MA: MIT Press.

Do, M. T., Harp, J. M. & Norris, K. C. 1999 A test of a pattern recognition system for identification of spiders. *Bull. Entomol. Res.* **89**, 217–224.

Dodd, J. C. & Rosendahl, S. 1996 The BEG Expert System—a multimedia identification system for arbuscular mycorrhizal fungi. *Mycorrhiza* **6**, 275–278.

Edwards, M. & Morse, D. R. 1995 The potential for computer-aided identification in biodiversity research. *Trends Ecol. Evol.* **10**, 153–158.

France, I., Duller, A. W. G., Duller, G. A. T. & Lamb, H. F. 2000 A new approach to automated pollen analysis. *Quatern. Sci. Rev.* **19**, 537–546.

Gamez, R. 1991 Biodiversity conservation through facilitation of its sustainable use: Costa Rica's national biodiversity institute. *Trends Ecol. Evol.* **6**, 377–378.

Garland, E. D. & Zimmer, C. A. 2002 Techniques for the identification of bivalve larvae. *Mar. Ecol. Prog. Ser.* **225**, 299–310.

Gaston, K. J. 1993 Spatial patterns in the description and richness of the Hymenoptera. In *Hymenoptera and biodiversity* (ed. J. LaSalle & I. D. Gauld), pp. 277–293. Wallingford: CAB International Press.

Gaston, K. J. 1994 Spatial patterns of species description: how is our knowledge of the global insect fauna growing? *Biol. Conserv.* **67**, 37–40.

Gaston, K. J. & May, R. M. 1992 Taxonomy of taxonomists. *Nature* **356**, 281–281.

Gaston, K. J. & Mound, L. A. 1993 Taxonomy, hypothesis testing and the biodiversity crisis. *Proc. R. Soc. Lond.* B **251**, 139–142.

Gauld, I. D., O'Neill, M. A. & Gaston, K. J. 2000 Driving Miss Daisy: the performance of an automated insect identification system. In *Hymenoptera: evolution, biodiversity and biological control* (ed. A. D. Austin & M. Dowton), pp. 303–312. Collingwood, VIC: CSIRO.

Geist, A., Berguelin, A., Dongarra, J., Weicheng, J., Manchek, R. & Sunderam, V. 2003 *PVM: Parallel Virtual Machine. A users guide and tutorial for network parallel computing*. Cambridge, MA: MIT Press.

Godfray, H. C. J. 2002 Challenges for taxonomy—the discipline will have to reinvent itself if it is to survive and flourish. *Nature* **417**, 17–19.

Grover, L. K. 1997 Quantum computers can search arbitrarily large databases by a single query. *Phys. Rev. Lett.* **79**, 4709–4712.

Guo, B., Lam, K-M., Lin, K-H. & Siu, W.-C. 2003 Human face recognition based on spatially weighted Hausdorff distance. *Pattern Recognition Lett.* **24**, 499–507.

Hammond, P. M. 1992 Species inventory. In *Global biodiversity: status of the Earth's living resources* (ed. B. Groombridge), pp. 17–39. London: Chapman & Hall.

Hawksworth, D. L. & Kalin-Arroyo, M. T. 1995 Magnitude and distribution of biodiversity. In *Global biodiversity assessment* (ed. V. H. Heywood), pp. 107–199. Cambridge University Press.

He, Y., Tian, J., Luo, X. & Zhang, T. 2003 Image enhancement and minutiae matching in fingerprint verification. *Pattern Recognition Lett.* **24**, 1349–1360.

Hebert, P. D. N., Cywinska, A., Ball, S. L. & deWaard, J. R. 2003 Biological identifications through DNA barcodes. *Proc. R. Soc. Lond.* B **270**, 313–321. (DOI 10.1098/rspb.2002.2218.)

Hopkins, G. W. & Freckleton, R. P. 2002 Decline in the numbers of amateur and professional taxonomists: implications for conservation. *Anim. Conserv.* **5**, 245–249.

Hyde, K. D., Ho, W. H., Taylor, J. E. & Hawksworth, D. L. 2000 Estimating the extent of fungal diversity in the tropics. In *Nature and human society: the quest for a sustainable world* (ed. P. H. Raven & T. Williams), pp. 156–175. Washington, DC: National Academy Press.

Jarman, K. H., Cebula, S. T., Saenz, A. J., Petersen, C. E., Valentine, N. B., Kingsley, M. T. & Wahl, K. L. 2000 An algorithm for automated bacterial identification using matrix-assisted laser desorption/ionization mass spectrometry. *Analyt. Chem.* **72**, 1217–1223.

Jarvie, J. K. & Stevens, P. F. 1998 Interactive keys, inventory, and conservation. *Conserv. Biol.* **12**, 222–224.

Jeffries, H. P., Berman, M. S., Poularikas, A. D., Katsinis, C., Melas, I., Sherman, K. & Bivins, L. 1984 Automated sizing, counting and identification of zooplankton by pattern recognition. *Mar. Biol.* **78**, 329–334.

Jonker, R., Groben, R., Tarran, G., Medlin, L., Wilkins, M., Garcia, L., Zabala, L. & Boddy, L. 2000 Automated identification and characterisation of microbial populations using flow cytometry: the AIMS project. *Scientia Mar.* **64**, 225–234.

Lang, R. & Warwick, K. 2002 The plastic self organising map. See http://www.cyber.rdg.ac.uk/research/publications/01183.pdf.

Longford, M., Taylor, G. E. & Flenley, J. R. 1990 Computerised identification of pollen grains by texture analysis. *Rev. Palaeobot. Palynol.* **64**, 197–203.

Lu, G., Zhang, D. & Wang, K. 2003 Palmprint recognition using eigenpalms features. *Pattern Recognition Lett.* **24**, 1463–1467.

Lucas, S.M. 1997 Face recognition with the continuous n-tuple classifier. In *British Machine Vision Conference, 1997*, pp. 222–231. See http://www.bmva.ac.uk/bmvc/1997/papers/113/paper.html.

McCall, H., Bravo, I., Lindley, J. A. & Reguera, B. 1996 Phytoplankton recognition using parametric discriminants. *J. Plankton Res.* **18**, 393–410.

May, R. M. 1988 How many species are there on earth? *Science* **241**, 1441–1449.

May, R. M. 1990 How many species? *Phil. Trans. R. Soc. Lond.* B **330**, 293–304.

May, R. M. 2000 The dimensions of life on earth. In *Nature and human society: the quest for a sustainable world* (ed. P. H. Raven & T. Williams), pp. 30–45. Washington, DC: National Academy Press.

Moore, A. & Miller, R. H. 2002 Automated identification of optically sensed aphid (Homoptera: Aphidae) wingbeat waveforms. *Ann. Entomol. Soc. Am.* **95**, 1–8.

Morgan, A., Boddy, L., Mordue, J. E. M. & Morris, C. W. 1998 Evaluation of artificial neural networks for fungal identification, employing morphometric data from spores of *Pestalotiopis* species. *Mycol. Res.* **102**, 975–984.

Morris, C. W., Autret, A. & Boddy, L. 2001 Support vector machines for identifying organisms—a comparison with strongly partitioned radial basis function networks. *Ecol. Model.* **146**, 57–67.

Mound, L. A. & Gaston, K. J. 1993 Conservation and systematics—the agony and the ecstasy. In *Perspectives on insect conservation* (ed. K. J. Gaston, T. R. New & M. J. Samways), pp. 185–195. Andover: Intercept.

O'Neill, M. A. & Curtis-Rouse, M. 2002 An engineering study for the design of a Computational MicroBot (CMB). Technical Document. Tucson, AZ: Morpho Inc.

O'Neill, M. A., Gauld, I. D., Gaston, K. J. & Weeks, P. J. D. 2000 Daisy: an automated invertebrate identification system using holistic vision techniques. In *Proc. Inaugural Meeting BioNET-INTERNATIONAL Group for Computer-Aided Taxonomy (BIGCAT)* (ed. D. Chesmore, L. Yorke, P. Bridge & S. Gallagher), pp. 13–22. Egham: BioNET-INTERNATIONAL Technical Secretariat.

O'Neill, M. A., Burns, G. A. P. C. & Hilgetag, C. C. 2002 The PUPS-MOSIX environment: a homeostatic environment for neuro- and bio-informatic applications. In *Neuroscience databases: a practical guide* (ed. R. Kötter), pp. 187–202. Boston, MA: Kluwer.

Parsons, S. 2001 Identification of New Zealand bats (*Chalinobus tuberculatus* and *Mystacina tuberculata*) in flight from analysis of echolocation calls by artificial neural networks. *J. Zool. (Lond.)* **253**, 447–456.

Parsons, S. & Jones, G. 2000 Acoustic identification of twelve species of echolocating bat by discriminant function analysis and artificial neural networks. *J. Exp. Biol.* **203**, 2641–2656.

Patterson, D. J. 2003 Progressing towards a biological names register. *Nature* **422**, 661.

Pech-Pacheco, J. L. & Alvarez-Borrego, J. 1998 Optical–digital system applied to the identification of five phytoplankton species. *Mar. Biol.* **132**, 357–365.

Penev, P. S. & Atick, J. J. 1996 Local feature analysis: a general statistical theory for object representation. *Network: Comput. Neural Systems* **7**, 477–500.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. 1992 *Numerical recipes in C: the art of scientific computing*, 2nd edn. Cambridge University Press.

Rambold, G. & Agerer, R. 1997 DEEMY—the concept of a characterization and determination system for ectomycorrhizae. *Mycorrhiza* **7**, 113–116.

Schwenker, F., Dietrich, C., Kestler, H. A., Riede, K. & Palm, G. 2003 Radial basis function neural networks and temporal fusion for the classification of bioacoustic time series. *Neurocomputing* **51**, 265–275.

Simmonds, E. J., Armstrong, F. & Copland, P. J. 1996 Species identification using wideband backscatter with neural network and discriminant analysis. *ICES J. Mar. Sci.* **53**, 189–195.

Simonetti, J. A. 1997 Biodiversity and a taxonomy of Chilean taxonomists. *Biodiv. Conserv.* **6**, 633–637.

Solow, A., Mound, L. A. & Gaston, K. J. 1995 Estimating the rate of synonymy. *Syst. Biol.* **44**, 93–96.

Tautz, D., Arctander, P., Minelli, A., Thomas, R. H. & Vogler, A. P. 2002 DNA points the way ahead in taxonomy. *Nature* **418**, 479.

Tautz, D., Arctander, P., Minelli, A., Thomas, R. H. & Vogler, A. P. 2003 A plea for DNA taxonomy. *Trends Ecol. Evol.* **18**, 70–74.

Tsalakanidou, F., Tzovaras, D. & Strintzis, M. G. 2003 Use of depth and colour eigenfaces for face recognition. *Pattern Recognition Lett.* **24**, 1427–1435.

Turk, M. & Pentland, A. P. 1991 Eigenfaces for recognition. *J. Cogn. Neurosci.* **3**, 71–86.

van de Vooren, J. G., Polder, G. & van de Heijden, G. W. A. M. 1992 Identification of mushroom cultivars using image analysis. *Trans. ASAE* **35**, 347–350.

Vaughan, N., Jones, G. & Harris, S. 1997 Identification of British bat species by multivariate analysis of echolocation call parameters. *Int. J. Anim. Sound Record.* **7**, 189–207.

Watson, A. T., O'Neill, M. A. & Kitching, I. J. 2003 Automated identification of live moths (macrolepidoptera) using Digital Automated Identification SYstem (DAISY). *Syst. Biodiv.* **1**, 287–300.

Weeks, P. J. D. & Gaston, K. J. 1997 Image analysis, neural networks, and the taxonomic impediment to biodiversity studies. *Biodiv. Conserv.* **6**, 263–274.

Weeks, P. J. D., Gauld, I. D., Gaston, K. J. & O'Neill, M. A. 1997 Automating the identification of insects: a new solution to an old problem. *Bull. Entomol. Res.* **87**, 203–211.

Weeks, P. J. D., O'Neill, M. A., Gaston, K. J. & Gauld, I. D. 1999*a* Species-identification of wasps using principal component associative memories. *Image Vis. Comput.* **17**, 861–866.

Weeks, P. J. D., O'Neill, M. A., Gaston, K. J. & Gauld, I. D. 1999*b* Automating insect identification: exploring the limitations of a prototype system. *J. Appl. Entomol.* **123**, 1–8.

White, I. M. & Scott, P. R. 1994 Computer identification resources for pest identification: a review. In *The identification and characterisation of pest organisms* (ed. D. L. Hawksworth), pp. 129–137. Wallingford: CAB International.

Wilkins, M. F., Boddy, L., Morris, C. W. & Jonker, R. R. 1999 Identification of phytoplankton from flow cytometry data by using radial basis function neural networks. *Appl. Environ. Microbiol.* **65**, 4404–4410.

Wilson, E. O. 2003 The encyclopedia of life. *Trends Ecol. Evol.* **18**, 77–80.

Wu, J. & Zhou, Z-H. 2002 Face recognition with one training image per person. *Pattern Recognition Lett.* **23**, 1711–1719.

Yu, D. S., Kokko, E. G., Barron, J. R., Schaalje, G. B. & Gowen, B. E. 1992 Identification of ichneumonid wasps using image analysis of wasps. *Syst. Entomol.* **17**, 389–395.

GLOSSARY

ABIS: Automatic Bee Identification System
ANN: artificial neural network
CAT: Computer Assisted Taxonomy
DAISY: Digital Automated Identification SYstem
NNC: nearest-neighbour classification
PCA: principal component analysis
PSOM: plastic self-organizing map
PVM: Parallel Virtual Machine