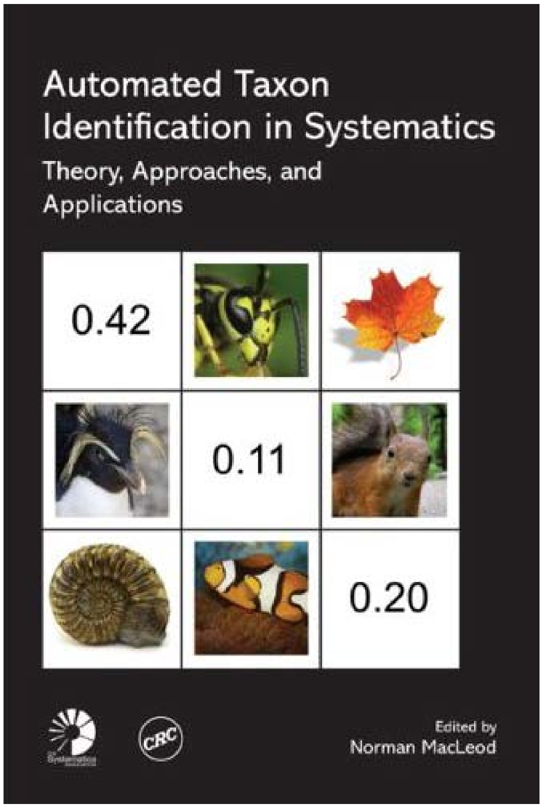

Automated Taxon Identification in Systematics
Theory, Approaches, and Applications
0.42
0.11
0.20
CRC
Edited by
Norman MacLeod

Chapter 9

Series: Systematics Association Special Volumes, 74
Publisher: CRC Press
ISBN: 9780849382055
ISBN 10: 084938205X
Publication Date: 7/24/2007
Number of Pages: 368

9 Introducing SPIDA-Web: Wavelets, Neural Networks and Internet Accessibility in an Image-Based Automated Identification System

Kimberly N. Russell, Martin T. Do, Jeremy C. Huff and Norman I. Platnick

CONTENTS

INTRODUCTION

Attempts to understand how biodiversity originates and is maintained, and how it contributes to ecosystem functioning and human services, are hindered by lack of complete information. To understand the complexity of ecosystem function, and the likely impacts of human activities on these functions, ecologists and conservation scientists need to understand species interactions across multiple scales. Most studies to date have attempted to gain this understanding by looking at a very small subset of species, focusing primarily on vertebrates and other well-known or 'charismatic' groups. Unfortunately, recent syntheses (e.g. Goldwasser and Roughgarden, 1997; Platnick, 1999) suggest that such studies are not adequate in terms of predicting biodiversity patterns or signatures of disturbance. Additional information on lesser known groups is required to complete the picture. Yet, most studies avoid collecting data on diverse groups such as insects and arachnids precisely because they are less well known!

The inclusion of these groups in biodiversity studies has traditionally required both trained personnel who are able to identify known species correctly, and a systematist who can recognize and describe specimens new to science. Even when knowledgeable personnel can be found, the process of identification and description of new species takes time and money – assets in short supply for most ecologists, conservation biologists and wildlife managers. Non-specialists do not have the training or access to the materials necessary to produce accurate and consistent identifications on their own. The combined effect of this has heretofore led to the use and interpretation of questionable data or, more commonly, the complete abandonment of data from those taxonomic groups that comprise the bulk of biodiversity. We cannot hope to understand the complexity of ecosystem function and the relationship of human activities with ecosystem function without knowing how many, and what kinds of organisms are present.

Faced with these problems, and the increasing demand internationally for biodiversity research, some partial solutions have been pursued that attempt to delay or circumvent altogether the need for identifications. The use of technicians, or parataxonomists, to collect, sort and catalogue specimens prior to the input of a specialist has met with some success in Costa Rica (Instituto Nacional de Biodiversidad [INBio], 2001). The designation of RTUs (recognizable taxonomic units), or morphospecies, by non-specialists in order to obtain rapid richness estimates without requiring species-level identifications has proved reasonably accurate and useful in some cases (Oliver and Beattie, 1993, 1996). Certainly, the creation of biodiversity data-bases that catalogue collected specimens – particularly those that incorporate digital images of whole specimens and search procedures (similar to interactive keys) to help with identification (e.g. VirBas in Australia; Oliver et al., 2000) – will facilitate rapid, albeit cursory, biodiversity assessments. Although these methods provide a way to obtain quick species counts for initial richness comparisons, they do not provide enough information for in-depth biological or ecological studies. For serious analyses, identity is important. Therefore, tools must be developed to make routine identifications of specimens by non-experts both accurate and efficient.

An ideal identification system is one that encapsulates the knowledge of a systematist, requires little user input, and yields quick and accurate identifications. Some computer-aided identification systems such as interactive keys, multi-access keys, hypertext keys and expert systems are a significant improvement over the traditional, printed dichotomous key, but still require significant input from the user (and therefore require basic knowledge of the morphology and terminology of the target group; see Edwards and Morse, 1995; Dodd and Rosendahl, 1996; Rambold and Agerer, 1997). Methods that exhibit some level of automation are likely to be more accessible to non-specialists.

Many partly automated identification systems for multicellular organisms make use of digital imaging (e.g. Gerhards et al., 1993; Dietrich and Pooley, 1994; Chtioui et al., 1996; Weeks et al., 1997; Kwon and Cho, 1998; Do et al., 1999; Mancuso and Nicese, 1999; Weeks et al., 1999; Theodoropoulos et al., 2000). In very general terms, information is extracted from images in the form of specific measurements (taken manually or with the help of image tool programs), or the image itself is processed into a form that can be expressed numerically. The extracted observations are then subjected to statistical analysis (e.g. PCA, discriminant analysis), or submitted to some form of artificial neural network (ANN) in order to characterize and subsequently classify the species. Artificial neural networks are programming algorithms that simulate the structure of the brain and its processing of information (see Boddy et al., 1990, for an introduction). Species identification using ANNs, although similar in principle to statistical classification, relies on the ANN itself to create the group 'classifiers' by selectively weighting the input characters and adjusting its own internal configuration to maximize identification accuracy.

In the development of our identification system, we chose to focus on the ANN approach. This decision was based on a number of factors, including previous studies showing that in situations where both statistical and ANN-based approaches were tried using the same data as inputs, the ANNs almost always achieved equivalent or superior levels of accuracy (Chtioui et al., 1996; Goodacre et al., 1996; Wilkins et al., 1996; Parsons and Jones, 2000). The advantage of using ANNs is greatest when traditional identification procedures rely on somewhat subjective, qualitative characters that cannot be simply quantified (or even necessarily described). Qualitative features are subject to inter-and intra-observer variability arising from the user's level of knowledge, experience and frequency of use (Theodoropoulos et al., 2000).

There have already been many promising studies evaluating the potential of neural networks for the identification of cell types and organisms. ANNs have been used successfully in medical research to identify and classify cancer cells (Maollemi, 1991; Jiang et al., 1996; Hurst et al., 1997); to identify microorganisms of various kinds, including bacteria, yeasts and phytoplankton (Rataj and Schindler, 1991; Kennedy and Thakur, 1993; Goodacre et al., 1996; Wilkins et al., 1996; Goodacre et al., 1998; Wit and Busscher, 1998); and to identify macro-organisms, including plants of agricultural interest (Chtioui et al., 1996; Kwon and Cho, 1998; Mancuso and Nicese, 1999), parasitic larvae (Theodoropoulos et al., 2000), spiders (Do et al., 1999) and bats (from their echolocation signals – Parsons and Jones, 2000).

Of course, there are many different kinds of neural networks, ways of structuring an identification system and approaches to making such a system available to the public and there are many challenges to be faced when working with real data. Our system, SPIDA (*sp*ecies *id*entification, *a*utomated), or the web-accessible version, SPIDA-web, was created as a generalized identification system that can be tailored for virtually any group of organisms that can be distinguished visually (i.e. prior testing had demonstrated early versions' ability to distinguish five species of Ichneumonid wasp [unpublished data], six species of Lycosid spiders [Do et al., 1999] and twelve species of North American bees [Russell et al., in prep]). That said, by choosing to develop and refine our system using real data with which we have succeeded in creating a working prototype, we have of necessity had to face a number of challenges that will be common to most if not all automated identification systems.

Our test case, the Australasian ground spiders of the family Trochanteriidae, provided good examples of these challenges, including, among others, intraspecific variability (which itself varies in degree across species), variability in sample quality (due to debris or imaging techniques) and small sample sizes. In addition, we decided to tackle the problem of identifying all the closely related species included in a major taxon instead of the much simpler problem of distinguishing the species that happen to co-occur in a single area, most of which are only distantly related to each other and hence relatively easy to separate. Finally, spiders are considered by some to be one of the more difficult groups in terms of assigning species-level identifications, even compared with other arthropods. In the USA, only a tiny fraction of the roughly 3500 species are identifiable without the use of a microscope and the appropriate technical keys. Traditionally, one needs first to determine family membership with one key, genus membership with a different key (focusing on entirely different structures) and then, finally, species membership focusing on the complex structures of the genitalia, described in dizzying technical detail in published monographs. In sum, we have given ourselves a difficult task. But by doing so, we can more realistically assess the

TABLE 9.1A
Total Images Available for Training

	Pro		Anti	
Species	Unique	Total	Unique	Total
Desognaphosa kuranda	34	272	1171	13,393
Desognaphosa massey	20	160	1185	13,505
Desognaphosa millaa	26	208	1179	13,457
Desognaphosa yabbra	58	463	1147	13,202
Hemicloena julatten	20	80	1185	13,585
Longrita insidiosa	28	244	1177	13,441
Morebilus diversus	26	208	1179	13,457
Morebilus fumosus	25	100	1180	13,565
Morebilus plagusius	50	392	1155	13,273
Rebilus bulburin	24	192	1181	13,473
Trachytrema garnet	27	107	1178	13,558
Trachycosmus allyn	32	194	1173	13,471
Trachycosmus sculptilis	136	1055	1069	12,610

Notes: Statistics for the 13 species with 20 or more unique specimens. Total number of images available for training the species ANNs. 'Pro' images are from the species the ANN is being trained to recognize. 'Anti' images refer to all the images from the other species in the group (i.e. all the images that are not in the 'pro' set). Unique images are those taken from unique specimens, so this number reflects the number of individuals we had available for each species. Multiple images were taken of each specimen, so the 'total' number includes every image we have for a given species (pro) or all the remaining species (anti).

challenges of developing automated identification systems and the utility of our unique approach in meeting these challenges.

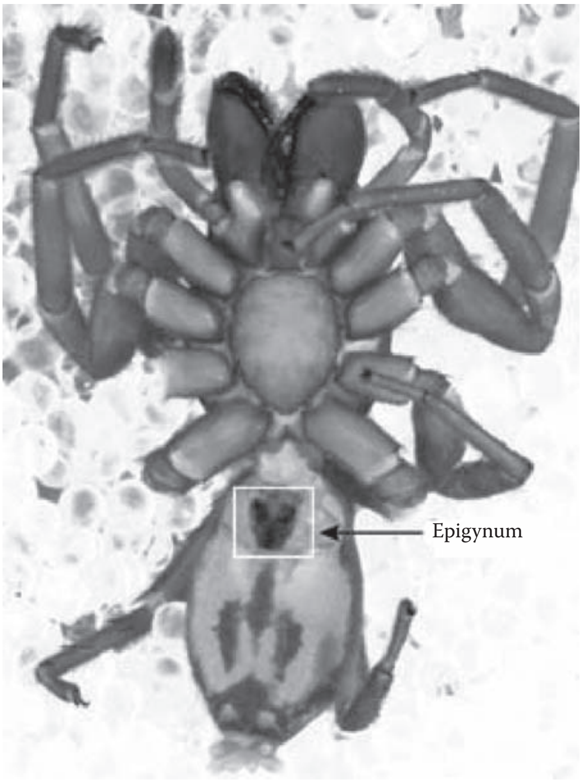
METHODS

The Data

We selected the recently revised (Platnick, 2002) Australasian ground spiders of the family Trochanteriidae as our prototype group. This decision was made primarily because of familiarity with the taxonomy of the group, and because specimens of all species were readily available and the size of the family – 121 species in 14 genera – seemed a reasonable and practical starting point for a practical identification system. Although some species in this family are relatively common, almost 80 per cent were represented by less than 10 individuals (of either sex); more than 50 per cent had fewer than 5. Thirteen species had 20 or more individuals (see Table 9.1A).

Species-level discrimination in spiders is based primarily on shape of the male and female genitalia. Anyone attempting identification to species, or a systematist describing new species, would need to examine these structures. Therefore, these are the structures we use for submission to SPIDA-web. This chapter will focus exclusively on the discrimination of the female specimens, as this work is entirely completed.

Female spider reproductive structures, known as epigyna, can be very complex or quite simple and so present a range of detail that will be useful in assessing the applicability of our system to other groups of organisms. The epigynum is found on the ventral side of an adult female and is



Ventral view

FIGURE 9.1 Adult female of *Desognaphosa bartle* Platnick. The female reproductive structure, the epigynum, is considered species diagnostic for most spiders. We used a single ventral view of the external structure visible here for input into SPIDA-web.

noticeable without dissection (see Figure 9.1). Although specialists often use structures of the internal epigynum (dorsal view) for species description and in constructing species-level keys, we chose to focus exclusively on the external features for the sake of simplicity and ease of use.

Once all the images were in place, it became apparent just how difficult a task we were about to put to a computer algorithm. Figure 9.2 provides illustrative examples of problems an automated ID system will need to overcome.

- Species similarity. Although the distinction between some species is easy to recognize, even between congeners (see Figure 9.2A), other groups appear to lack clear diagnostic structures in the ventral view (Figure 9.2B). Figure 9.2C illustrates the minor differences separating five species in the genus *Desognaphosa,* though in this case there is plenty of visible detail.
- Limited data. The third image in Figure 9.2C is from the single representative of the species *Desognaphosa karnak*. First, the structure is damaged. The human eye is able to compensate adequately, but getting the ANN to ignore this flaw is not so easy without any replicate specimens. But it is likely that even a human worker would have trouble separating this image from the previous image (of *Desognaphosa finnigan*) without additional specimens.
- Intraspecific variation. The relative degree of inter-and intraspecific variation in this group is not always predictable, as illustrated in Figures 9.2D, E and F. Figure 9.2D shows five

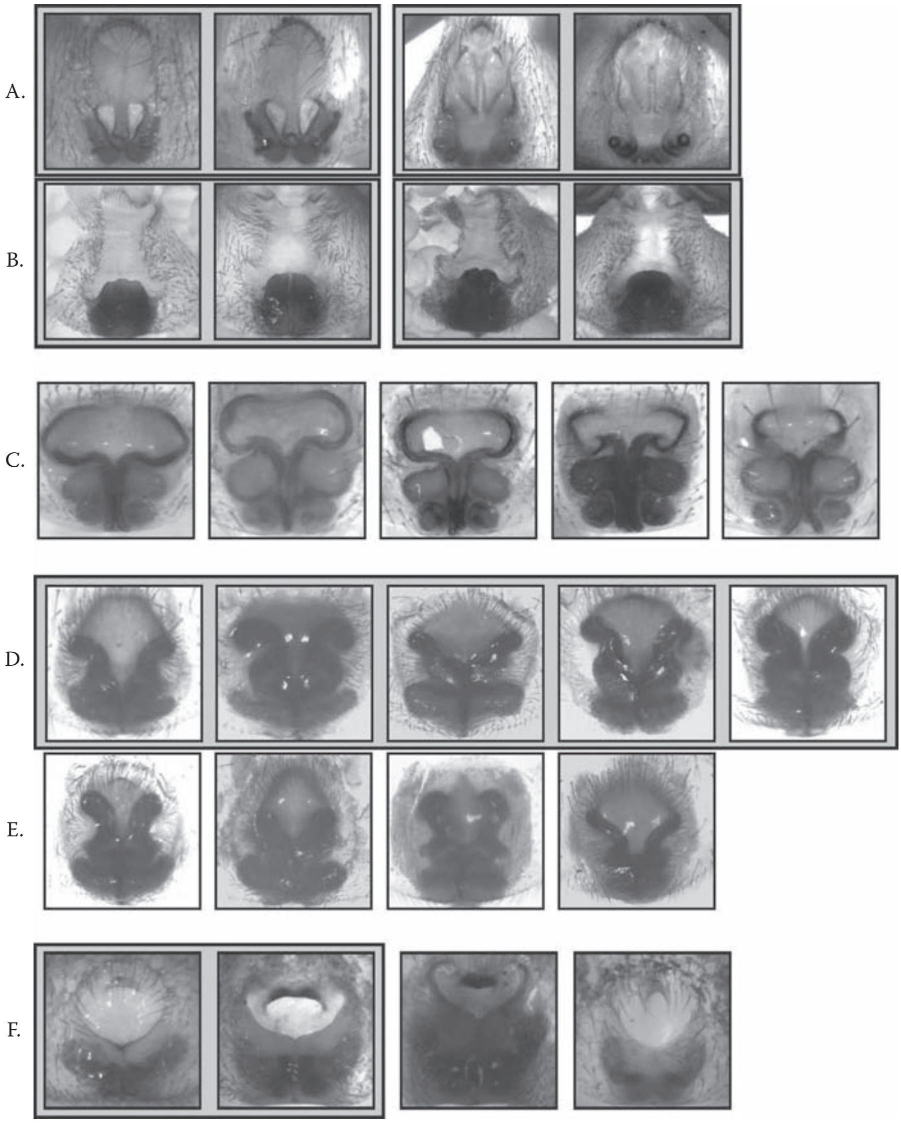


FIGURE 9.2 Illustrative examples of inter-and intraspecific variation in epigynal images among species in the spider family Trochanteriidae. Images enclosed in the same box are from the same species. (A) Two images each of two easy to distinguish congeners, *Trachyrema castaneum* and *T. garnet.* (B) Two images each of two congeners, *Longrita millewa* and *L. yuinmery,* with little information present in the epigynal images. (C) Five similar congeners in the genus *Desognaphosa*: *D. halcyon, D. finnigan, D. karnak, D. bartle* and *D. windsor.* (D) Five disparate individuals of the species *Rebilus bulburin.* (E) Representative images from four species in the same genus as (D), *R. lugubris, R. credition, R. brooklana* and *R. bilpin.* (F) The only two specimens from the species *Platorish nebo* and two related species, *P. churchillae* and *P. flavitarsus.*

images taken from individuals of the same species, *Rebilus bulburin.* This species exhibits one of the highest levels of intraspecific variation in this group. Other species in this genus, however, show much less intraspecific variation and the interspecific differences among some of these congeners (Figure 9.2E) is arguably similar in degree to the variation seen in *R. bulburin.* Finally, Figure 9.2F shows first epigynal pictures from the two individuals known of the species *Platorish nebo* and individual pictures of two other species in the genus, *P. churchillae* and *P. flavitarsus.* At first glance, it is difficult to see

> what the first two images have in common relative to the group. Without more unique specimens (or more information from other structures), it would be difficult to form a useful species image that would enable consistent species determinations for this group.

The point of this discussion is to emphasize the reality of the data used for our prototype in terms of the kinds of features available for species classification in spiders and to bring attention to the kinds of issues that are certainly *not* unique to spiders, such as intraspecific variation and specimen damage, with which any automated identification system will have to contend.

IMAGING PROTOCOL

All specimens were imaged using a Leica MZ 12.5 microscope fitted with a Q-imaging MicroPublisher CCD camera. Illumination was provided by an EIS fiber optic light source with a dual chrome gooseneck. The camera was connected to either a Dell Dimension 8200 Series or an Apple Titanium Powerbook G4 laptop. Images were converted to greyscale, cropped square, enhanced and resized as necessary in Adobe Photoshop.

Much like human students, ANNs distinguish objects by learning to focus their attention on particular aspects of an image, giving more weight to features that vary reliably between groups. If trained on only one image from each species, it is quite probable that a feature could be found to distinguish the two images that has nothing to do with the two species, but in reality is an artefact of the images themselves (e.g. presence of glare spots or background debris). Without multiple examples, the ANN can also form a much too specific 'vision' of a species, which could result in high numbers of false negatives when intraspecific variation is high. This is why the construction of an adequate training set is so important: the goal is to force the ANN to focus on the structures that are critical for distinguishing species, but also to encapsulate the likely variation both in the structures themselves and in the imaging of these structures (e.g. rotation, lighting, background, etc.).

Ideally, one would use many unique examples from every species to train the ANN, thereby encapsulating both types of likely variation. As previously stated, we did not have an adequate number of replicate specimens for most species. Hoping to compensate partially for this lack of unique samples, we collected either 4 or 12 images of every specimen, depending on how many specimens were available (greater than or less than 15, respectively). An attempt was made to introduce variation in the process by altering the lighting, repositioning the specimen and/or changing the rotation slightly between each picture. These replicate images were kept distinct from images taken of unique individuals. In a further attempt to generate more data for training, we created 'flipped' versions of each image using Adobe Photoshop. The female genitalia for this group are known to be bilaterally symmetrical. Flipping the images horizontally introduces some variation useful in training, particularly for species represented by fewer than three individuals.

IMAGE ENCODING

All automated identification systems face the task of reducing the feature space of the input data (i.e. reducing the total amount of information presented to the system), in order to minimize noise and facilitate more efficient classification criteria. Our approach to this is to use an encoding technique called wavelet transformation (Graps, 1995). Wavelet transforms are similar to the more commonly encountered Fourier transform. These are based on an iterative procedure in which an image is successively reduced to a coarser version of itself, through the removal of high-frequency information contained in wavelet coefficients (sometimes referred to as detail coefficients). These coefficients are parameters that modify the shape of a predetermined function, called a wavelet. Once the information in an image is parsed out into low- and high-frequency elements, the user can selectively eliminate the high-frequency information (usually noise, e.g. spines, hair, debris), keeping the more important shape information. Our previous work made use of the Daubechies 4 wavelet function, described in detail in Do et al. (1999).

The Daubechies 4 wavelet function requires that the input image be a square with a dimension of $2^j \times 2^j$, where j is an integer. We determined that an ANN with 4096 neurons in the input layer is the largest ANN that can be trained in a reasonable amount of time on a Sun Blade 100 or a Pentium 4 computer, which were the computers used in this part of the project. This means that wavelet coefficients in vector spaces V_0, V_1, V_2, V_3, V_4 and V_5 were used, producing an input matrix with a dimension of $2^6 \times 2^6$. This size input matrix can only be generated from an image scaled to 256 × 256 pixels ($2^8 \times 2^8$) prior to Daubechies 4 encoding.

For work with the Trochanteriidae, it was decided to investigate the Gabor wavelet function as well, as this type of filter had recently been applied to the problem of face recognition (e.g. Howell and Buxton, 1995; Krüger et al., 2000; Zhu et al., 2004; Bazanov et al., 2005) and had certain advantages, including being more robust to minor differences in lighting, orientation and scale.

The Gabor filter decomposes the image into data of varying resolutions by using banks of Gabor masks of different sizes to sample the image. The process is actually modeled after the receptive fields of the simple cells in the primary visual cortex of the mammalian eye (Pollen and Ronner, 1981). A set of image masks for different resolutions is available for this filter. At the first (coarsest) resolution, six masks are used. Each mask covers the entire image and represents real and imaginary components of the image with the orientations of 0, 120 and 240°. At the next level of resolution, 24 masks are used (four sets of six masks) each set of six masks covering one quadrant of the image and representing real and imaginary components of the image with the orientations of 0, 120 and 240°. The next level of resolution includes 16 sets of six masks, the next 64 sets of six masks, and the last (finest) level of resolution employs 256 sets of six masks (see Figure 9.3). Each mask yields a Gabor coefficient that was used as an input into the ANN, resulting in a total of 2046 inputs. Experiments indicated it was most efficient to use images scaled to the dimension of 51 × 51 pixels, as this limited computing time while providing acceptable accuracy (see following).

CONSTRUCTION OF THE IDENTIFICATION ENGINE

Although our proof-of-principle study (Do et al., 1999) indicated that back-propagation ANNs were an appropriate computing algorithm to use as an identification engine, we decided to investigate three other commonly used techniques: radial basis function ANNs (RBF), support vector machine (SVM) and the continuous *n*-tuple classifier (with log polar encoding). We wanted to be certain that we had chosen the identification engine that was most likely to succeed considering the organisms and data we had to work with. Do wrote the necessary software to test these other algorithms on a subset of the spider data. None of them performed as well as the back-propagation (specifically, cascade correlation) ANNs in our preliminary tests.

ARTIFICIAL NEURAL NETWORK ARCHITECTURE

Despite the decision to continue working with cascade correlation ANNs, there was a recognized need to change the way our ANNs were structured in order to address two separate issues: one a common criticism of automated identification systems in general and the other specific to back-propagation approaches:

- Classification of unknowns. One problem many automated identification systems face is the proper classification of unknowns (i.e. images from species the system was not trained to recognize; see Edwards and Morse, 1995; Morris and Boddy, 1995). Often these objects are forced into an erroneous classification. This was the case with our pilot study. The system was structured such that there was one ANN for the set of six species in the

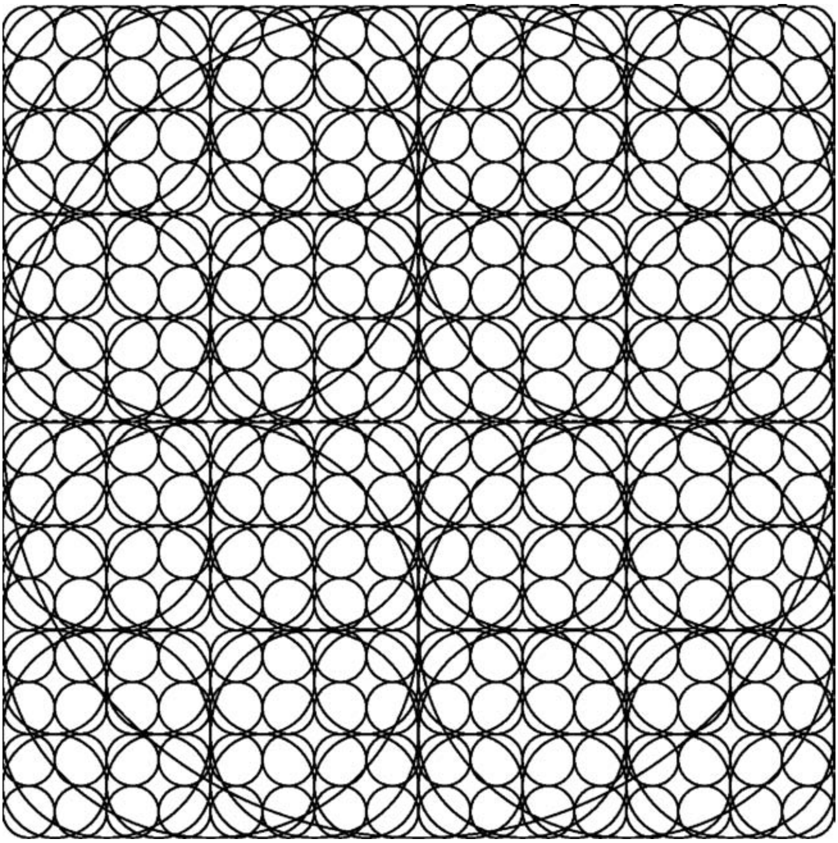


FIGURE 9.3 Sampling positions for the Gabor sampling scheme (modified from Howell and Buxton, 1995). Gabor filters decompose the image into data of varying resolutions by using banks of Gabor masks of different sizes to sample the image. Each circle on this figure represents a set of six Gabor masks representing real and imaginary components of the images with orientations of 0, 120 and 240°. Each mask yields a Gabor coefficient for use as an input into the ANN, resulting in a total of 2046 inputs.

trial, with an output node corresponding to each species. In this situation, an image submitted to the system would be forced through to one of the output nodes.

- Scalability. The issue of scalability is a common criticism of back-propagation ANNs. The traditional way of using such networks to classify species is to structure the ANN with an output node for each group the system attempts to distinguish. In this case, adding another species to the identification system would require the retraining of the entire ANN after adding another output node to represent the new group. This could be a very lengthy process, depending on the size of the group and the number of training images.

The solution to these problems eventually adopted for SPIDA was to create an ANN for each species in the group (Figure 9.4). Each of these ANNs has two output nodes, one positive and one negative, and is trained on images from the target species (the 'pro' training set) and a selection of images from other species in the group (the 'anti' training set). An image submitted for identification is presented to each ANN in the group. In this system, a true unknown can cycle through the group of ANNs without eliciting a positive response in any. If it is determined the unknown truly belongs in no group, adding the new species is as simple as training another relatively small ANN for the new species alone. One potential disadvantage to this approach is that, at some point, if the number of secondarily added ANNs exceeds some threshold, it is possible that accuracy may decrease, as

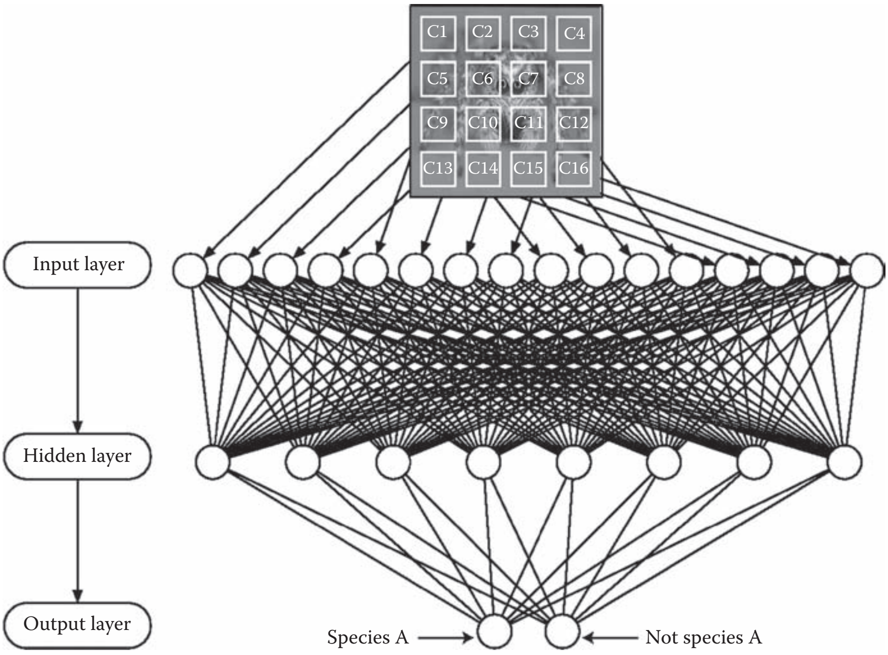


FIGURE 9.4 Simplified diagram of a back-propagation ANN, illustrating the basic structure of a typical species ANN in SPIDA-web. The output layer has only two nodes: species A and not species A (i.e. 'yes' or 'no'). Each species ANN is trained on information from members of that species ('pro' images) and from other species ('anti' images). Wavelet coefficients generated from the images of spider epigyna comprise the input data.

the newly added species will not be in the 'anti' training sets of the original set. In this case, more extensive retraining may be required, though it should be limited to closely related species only.

In order to test SPIDA's ability to classify previously unseen species correctly as unknowns, we randomly selected 20 images from species in related families and submitted them for identification.

SYSTEM STRUCTURE

As mentioned previously, SPIDA is structured as a collection of individually trained species-level ANNs. Images submitted to the system are cycled through all the ANNs in a predetermined grouping (e.g. family or genus). The positive output values for each ANN are saved to a file and ranked. The top three are selected and information on these species is retrieved from a data-base for presentation to the user. If the highest output value is above 0.59, then it is considered a positive ID and presented to the user as such. Although it is theoretically possible to have more than one ANN in the group return a value above 0.59, we did not experience this in the course of our study and in subsequent testing efforts. In fact, the difference between the first and second highest output values was usually very pronounced. However, we chose always to include information on the top three species in the event that a near tie were to occur, thus alerting the user to the fact that further scrutiny is required before a definitive ID can be assigned.

TRAINING

Generalized Training

Back-propagation ANNs consist of multiple layers of simple computing elements with many interconnections among the layers. The initial architecture of the ANN is established according to the amount of data making up each image (which determines the number of input neurons) and the number of groups that the system is designed to distinguish. The initial ANN consists of a layer of input neurons and a layer of output neurons fully interconnected between layers by random initial weights. Through a process of supervised learning, the network essentially enhances some features highly while diminishing the influence of others, using a complex method of averaging input parameters. The training process establishes additional neurons in a hidden layer between the input and output layers. In some cases, the number of hidden layers is fixed prior to training.

Other training algorithms, such as cascade correlation, allow neurons to be added one at a time as necessary, thereby minimizing the size of the ANN. These hidden neurons act as feature detectors that respond to specific patterns (e.g. a pattern unique to a given genus or species). The idea is to 'teach' the ANN to set the output neuron assigned to a given genus or species to its maximum value of 1.0 whenever a pattern indicative of that genus or species is presented, and set all other output neurons to their minimum value of 0.0. In practice, the ANN sets the output neurons to an intermediate value depending on the certainty of its identification (e.g. an output of 0.9999 indicates virtual certainty, whereas 0.6000 indicates lower confidence). The resulting output vector is then evaluated against the target function to compute an error. This error is then used to modify the weights in the connections. Training continues until the desired level of accuracy is attained. Once trained, the network is tested with previously unseen individuals to assess its ability to classify them into the correct groups (i.e. the network's ability to generalize from the training set to unknowns).

SPIDA Training

We used cascade correlation in conjunction with quick propagation (Fahlman, 1988; Fahlman and Lebiere, 1991) to train SPIDA ANNs using the Stuttgart Neural Network Simulator version 4.2 (SNNSv4.2). The training procedure was optimized for development of the final prototype system, which made use of all available data. In order to maximize the accuracy and future ability of the system to generalize, training of each ANN was a highly iterative and closely supervised procedure. For each species, a random set of images was selected for the training set, usually from a single individual. All other images were used for testing the network after it had been trained on this limited data-set. Then, the image that gave the most incorrect identification was added into the training set and the ANN was retrained and tested. This continued until all images were identified correctly. Finally, to ensure the system would likely be able to generalize appropriately (i.e. give accurate identifications to newly submitted images), images were sequentially removed from the training set to determine whether the ANN could still accurately identify all other images in the testing set. This process continued until the smallest possible training set that could accurately identify all the remaining images in the testing set had been defined.

In addition, for species that seemed to require the largest training sets, it was important to review the log files of this process to pick out any potentially contaminating images and remove them from the training sets. Contaminating images were defined as those with large amounts of debris, damage or occasionally questionable species designation (i.e. due to human error). These were noticed only when the identification logs indicated a persistent misclassification. Sometimes, when such an image was then added to the training set, the accuracy of subsequent identifications actually decreased after retraining. In other situations, it was merely a matter of examining the images in the 'pro' training set and picking out the oddball image if the ANN appeared to have trouble generalizing appropriately.

ACCESSIBILITY AND SPIDA-WEB

As important as creating a system with the ability to discriminate species using minimal data and computing power is making such a system adequately accessible to those most able to benefit from the technology. The overall goal was to create a user-friendly system, requiring a minimum of taxonomic knowledge and specialized equipment, that could be accessed from anywhere in the world. SPIDA-web is essentially SPIDA with an Internet interface, allowing users to have access to previously trained ANNs housed on a server, designed to distinguish specific organism groups.

The website consists partly of static html pages designed to give users some basic information about the project, such as how the system works and what its limitations are, as well as information on how to prepare and submit images. The remainder of the website is constructed with Java server pages, which essentially allow the software to display dynamic data in response to the user's input. These Java server pages are supported by a number of Java servlets served up by the open source Tomcat server software. Servlets are essentially computer codes written in the Java computer language designed to respond to a specific query submitted by the users via the Web. The servlets are capable of obtaining data from a data-base in order to process the user's request and send the results back to the Java server pages.

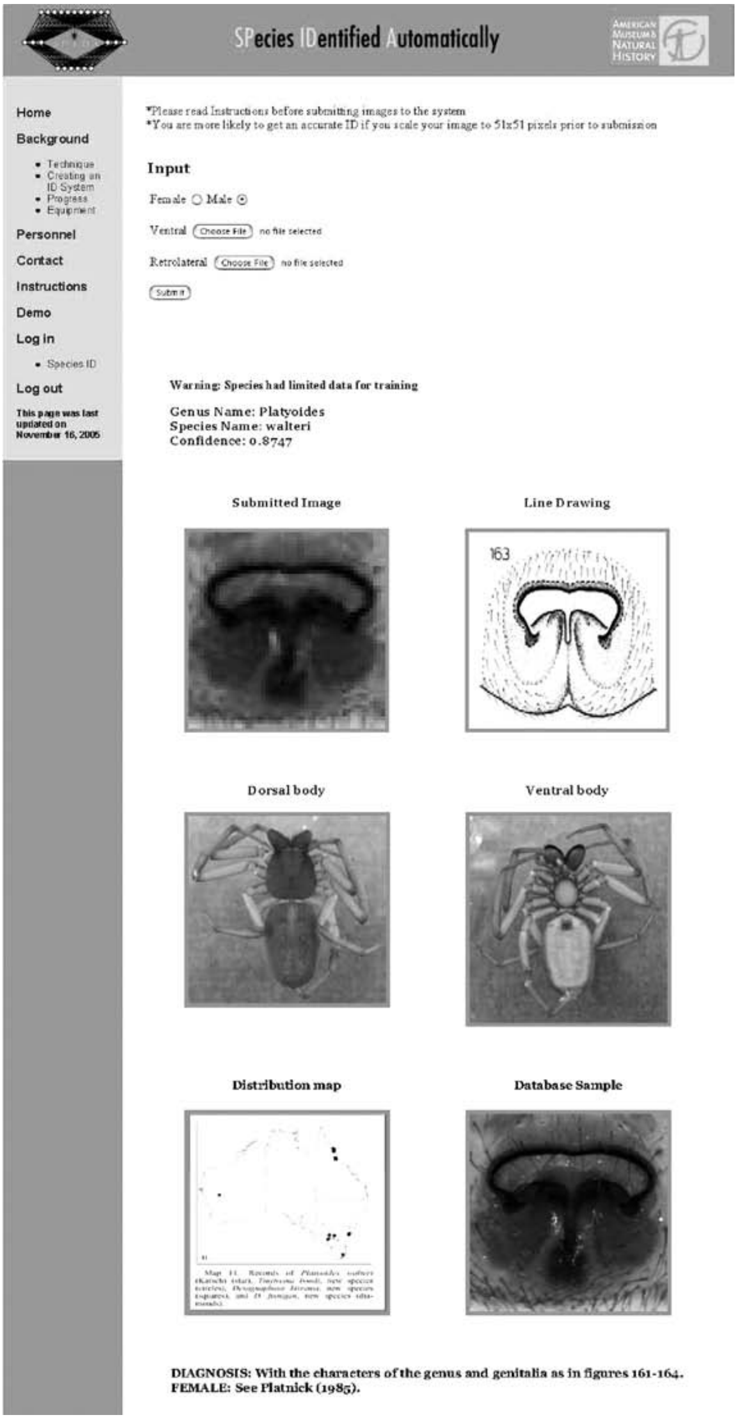
SPIDA-web supports a number of functions. First, all users are required to log in before accessing the site. This allows SPIDA-web managers to communicate with users to obtain more inputs for the data-base as well as more data to improve its ANNs. Once logged in, users can then access the Java server page that will allow them to select images for submission by browsing a local hard drive. Once selected, the images are submitted to a servlet that uses Sun's Java Advance Imaging (JAI) library to convert them to a usable format and scale them to a proper size. These images then undergo wavelet transformation and the resulting information is submitted to the trained ANNs obtained from the data-base via another servlet. The system then forwards the resulting identification to a Java server page that displays the identification (along with a second and third choice based on the ANN confidence values) and information from the data-base to the user (Figure 9.5). This information includes distribution maps, line drawings of genitalia, whole-body images, technical species descriptions and training images.

The entire process requires only a few seconds to complete on a local machine, but the ultimate speed of SPIDA-web will be dependent on the server and connection speed. SPIDA-web also has a number of administrative functions, such as adding and deleting user information; adding, deleting and editing genus and species information; and viewing user activity on the site. At this point, SPIDA-web is not set up to make automatic use of new data, such as images from new or under-represented species. These data will, however, be saved to a data-base, and reviewed and integrated by the SPIDA-web managers as needed.

RESULTS

SPIDA-web ANNs were trained successfully (i.e. convergence was attained) for each of the 121 species in the Australasian ground spider family Trochanteriidae, including those represented by only one or two individuals. Because so many species were data poor, the training method was optimized (as described previously) to produce the best result possible with the available information. In many cases, this meant that SPIDA-web had to use at least one of all the unique images available for a species in the training process. Also, the iterative training method, though leading to the best use of the data available, left any calculations of accuracy suspect as each ANN was trained intentionally until it was able to identify accurately all the images in the testing set. Until new users begin submitting images to SPIDA-web, there will be no truly new data from the Trochanteriidae on which to test the system.

One way to evaluate performance, however, is to look at the number of images required to train each ANN. Table 9.1B gives this information for the 13 species that had at least 20 unique individuals for the Gabor and Daubechies 4 wavelet encoding scheme. Networks trained with the Gabor-encoded images required a smaller percentage of images to achieve 100 per cent accuracy than the networks trained with the Daubechies 4 encoded images, though the difference was not great (49 vs. 53% unique for the 'pro' set and 3 vs. 6% for the 'anti' set, respectively). This result indicates that Gabor encoding may lead to ANNs better able to generalize, so all of the ANNs



This page was last updated on November 16, 2005

*Please read Instructions before submitting images to the system
*You are more likely to get an accurate ID if you scale your image to 51x51 pixels prior to submission

Input

Female ○ Male ⦿

Ventral no file selected

Retrolateral no file selected

Warning: Species had limited data for training

Genus Name: Platyoides
Species Name: walteri
Confidence: 0.8747

DIAGNOSIS: With the characters of the genus and genitalia as in figures 161-164.
FEMALE: See Platnick (1985).

FIGURE 9.5 Screen shot of SPIDA-web. This is an example of a positive identification output after a user has logged in and submitted an image from his or her hard drive. In addition to the top ID, SPIDA-web is designed to show the second and third highest matches along with accompanying data-base information, which appears directly below the positive match, requiring the user only to scroll down.

TABLE 9.1B
Percent of Images Used for Training

	Gabor training				Daubechies training			
	Pro		Anti		Pro		Anti	
Species	% of unique	% of total	% of unique	% of total	% of unique	% of total	% of unique	% of total
Desognaphosa kuranda	62	14	5	0.53	41	7	3	0.36
Desognaphosa massey	75	21	4	0.39	75	14	4	0.44
Desognaphosa millaa	62	13	4	0.46	65	15	6	0.66
Desognaphosa yabbra	31	6	3	0.33	24	5	5	0.48
Hemicloena julatten	55	16	3	0.26	N/A	N/A	N/A	N/A
Longrita insidiosa	54	9	3	0.30	64	20	6	0.70
Morebilus diversus	65	15	6	0.56	69	30	14	1.54
Morebilus fumosus	52	22	4	0.34	68	37	6	0.67
Morebilus plagusius	20	3	1	0.12	30	7	5	0.47
Rebilus bulburin	63	10	4	0.32	71	20	7	0.83
Trachytrema garnet	44	15	2	0.22	N/A	N/A	N/A	N/A
Trachycosmus allyn	50	12	3	0.31	53	12	5	0.59
Trachycosmus sculptilis	10	2	3	0.21	18	4	6	0.67
Average	49	12	3	0.33	53	16	6	0.67

Notes: Statistics for the 13 species with 20 or more unique specimens. Percentage of unique and total images used in training each species ANN. Images were added to the training set iteratively until all the remaining images in the testing set were accurately identified. Gabor and Daubechies are the two different wavelet encoding techniques used in this project.

currently in place for SPIDA-web were produced using this technique. However, these same data also hinted that Daubechies 4 may be useful for some groups. Therefore, both algorithms are still being tested. On average, the number of unique images required to train an ANN to distinguish a given Trochanteriidae species adequately from all the rest was 16.

The SPIDA-web site is located at http://research.amnh.org/invertzoo/spida. Users have successfully logged on and submitted images (though disappointingly few from the Trochanteriidae).

Though extensive testing of SPIDA-web with new data from the Trochanteriidae was not possible at present, its ability to correctly classify species from other, related families as unknowns could be examined. Of the 20 randomly selected outgroup images, SPIDA-web correctly classified 19, or 95 per cent, as unknowns. The one image it missed was matched with a species that had limited data available for training. The lack of data may have created an ANN focused on inappropriate features that happened to be shared with the image from the outgroup species.

In lieu of new unique data, a handful of quick tests were performed to determine how SPIDA-web would react to reprocessed images. The system should be able to handle variation in user preprocessing techniques such as cropping and rotation. In the website documentation, users were instructed (with illustrations) how to crop and orient images, but minor variation is inevitable. A sample of 15 images that SPIDA-web had previously identified correctly was selected. Ten of these images we recropped, either zooming in or zooming out prior to resizing. Five images were deliberately cropped poorly, with the epigyna clearly not centred. Results are shown in Table 9.2A. All of the zoomed images and three of the five off-centre images were identified correctly. The two errors, though producing a 'no match found' output, had the correct species as the closest match found. A different selection of 10 images was also used to test the effect of rotation. Images were first rotated 2°, then 4° clockwise before being resized and submitted to SPIDA-web. The 2° rotation

TABLE 9.2A
Effects of Cropping

Image	Confidence		
	Original	Recropped	Type
mbpf2	0.8919	0.885	Zoom out
mbpf8	0.6554	0.7756	Zoom out
dgyff8a	0.9903	0.9934	Zoom out
dgyff11c	0.976	0.9743	Zoom out
dgyff55a	0.7972	0.741	Zoom out
dgyff6b	0.9812	0.9378	Zoom in
mbpff9	0.9561	0.9733	Zoom in
mbpf6	0.7807	0.8157	Zoom in
falf1	0.8793	0.8076	Zoom in
pdw1a	0.8183	0.9312	Zoom in
pdw1a	0.8183	0.3887	Off centre
mbpf8	0.6554	0.5439	Off centre
dgyff13a	0.9925	0.9876	Off centre
dgyff57c	0.8193	0.4503	Off centre
mbpff2	0.9239	0.9137	Off centre

TABLE 9.2B
Effects of Rotation

Image	Confidence		
	Original	Rotate 2°	Rotate 4°
mbpf1	0.7526	0.8648	0.3746
mbpf2	0.8919	0.8679	0.6472
mbpf3	0.427	0.3813	nmf
mbpf4	0.6887	0.6071	nmf
mbpf5	0.7355	0.8136	0.6159
mbpf6	0.7598	0.7845	Wrong ID
mbpf7	0.9694	0.9662	0.9113
mbpf8	0.7473	0.6019	nmf
mbpf9	0.874	0.8917	0.5371
mbpf10	0.6712	0.7065	0.2905

Notes: Effects of user variability in image cropping and rotation. The identification confidence from the species ANNs is reported before and after recropping and rotation. 'Off centre' was a deliberate attempt to badly crop an image. Numbers greater than 0.5 indicated a correct identification. Numbers less than 0.5 are reported when the system returned 'no match found', but the first choice was the correct species. 'nmf' refers to situations when there was no positive identification and the first choice was not the correct species.

produced no changes in the outcome of the identifications (Table 9.22B). A 4° rotation, however, reduced identification accuracy to 40 per cent.

DISCUSSION

Taking steps toward practical implementation is absolutely critical to the advancement of the field of automated object recognition, as it must be proven to be more than just a 'pie-in-the-sky' idea that works only in the abstract. This investigation was given the task of putting the reasonably established reality of automated species identification into practice in the form of a usable, accessible system. The goal, therefore, was not necessarily to create the *most* accurate or *most* flexible or *most* easily used system, but rather to design and implement an ID system from beginning ('John Smith', a collector who finds himself with a spider needing an ID) to end (technician recording the scientific name and confidence on a data-sheet). That goal was met. A prototype was successfully developed to identify the 121 Australasian species in the spider family Trochanteriidae.

The base system, SPIDA-web, receives digital images of unidentified specimens via the Internet, encodes the information in these images using wavelet transformation, circulates this information through a set of ANNs trained on sets of identified images, and returns identifications to the user – all in a matter of seconds. Output is structured to give users basic information on each species, including distribution maps, drawings, pictures and technical descriptions as well as an indication of confidence and alternative choices.

But how do we evaluate the true success of an automated object recognition system like this? Four criteria to determine the utility of such a system might be: (1) accuracy, (2) accessibility, (3) scalability and (4) flexibility. How does our prototype of SPIDA-web stack up?

Accuracy

Many potential users would rank this at the top of the list in terms of importance. That said, results of informal surveys of arachnologists suggest that acceptable cutoffs for accuracy vary widely and often depend on the background of the respondents. Systematists or taxonomic specialists demand the highest accuracy levels – 95 per cent minimum for such a system to be useful for them. Ecologists and conservationists would be happy with 85–90 per cent if it meant they could have a species list to work with. Certainly, automated ID systems should do the same or better than untrained or quickly trained novices (e.g. technicians, students). Ideally, they would do the same or better than those workers with moderate amounts of training (e.g. entomologists or PhD students after taking one or more courses in spider taxonomy).

As stated previously, we did not structure our training protocol in such a way that would allow us to measure accuracy in the true sense. We trained the ANNs to be able to identify all the images in the testing set accurately, but only after first expanding and then reducing the size of the training set. We can only offer that the accuracy of identification for well represented species (15–20 unique samples available for training) was consistently high – in the range of 90–96 per cent – in the present study when tested on subsets of data prior to final training. We suspect that accuracy levels for under-represented groups (<10 unique specimens) are much lower, perhaps below 75 per cent based on a very small test set. Unfortunately, we do not currently have access to more data on which to further test the prototype, as all known specimens from the Trochanteriidae are already in our possession.

Use of replicate, processed and flipped images certainly helps the system be more robust against variability due to lighting, image preprocessing and minor damage or debris. However, for species with very few individuals, these do not adequately replicate the information necessary to force ANNs to converge on the most useful and appropriate features in the images. This is absolutely necessary if an ANN system is going to be able to generalize and recognize what constitutes intraspecific variation versus interspecific variation.

The problem of small training set size will certainly not be limited to spiders. Most invertebrate communities consist of a few common species and many rare ones. One reason these organisms are often so difficult to identify is the fact that so many were described based on just a handful of individuals. Lack of data is likely going to be a common problem for all automated identification systems. One way to ameliorate this problem partially is to design ID systems to be evolving such that, as they are used, they improve. We have designed SPIDA-web to store all submitted images so that new images can be incorporated into the appropriate training sets as needed to improve accuracy. We would suggest that all systems be designed to have this capability.

Of course, we can measure the prototype's accuracy in terms of its ability to identify unknown, or 'new' species by testing it with out-group images. Having a system that is perhaps overly sensitive to the detection of unknowns is more useful than a system that errs on the side of misclassification. A non-identification forces the user to re-evaluate the specimen, perhaps setting it aside for a specialist to review. An incorrect, but positive identification is much more likely to be overlooked, as a technician will be less likely to question it. Though not perfect (only 95% accuracy from our small test set), SPIDA-web is much better at detecting unknowns than our previous system, which employed a more traditional ANN architecture (Do, 1996). For the data-set described earlier, most misclassifications (of true unknowns) were limited to species that were trained on inadequate data.

In most respects ANNs are unpredictable, as there is no way of knowing what information from the training images they weight highly as features. For example, if an ANN is trained on images from a single individual, and that individual had a tear or large piece of debris attached, it is possible that the ANN would consider those anomalies to be features. If an unknown happened to have a similar 'feature', real or not, it could result in a false positive identification. That is why, once again, it is always better to have a multitude (>15) of training specimens.

Accessibility

Accessibility refers to both the ease with which non-specialists can navigate through the identification process as well as the ability of users to gain access to the ID system. SPIDA-web ranks high in both categories. Because the input to SPIDA-web is the whole image of a structure (in the case of the prototype, a picture of the external genitalia of spiders), there is no need for users to measure or dissect or even know the name of what they are taking a picture of. It is as simple as finding the structure, centering it and snapping an image. Instructions on how to find the structure are included in the introductory pages of the website. Aside from rudimentary image processing, such as converting the image to greyscale and cropping it square, users can submit an image without having any technical software or technical knowledge.

What *do* they need? They need to have a digital camera (or access to a scanner) and any computer with Internet access. In the case of our prototype, they need to know they have a spider and what family that spider is in. One could imagine an auto-ID system that does not even require the user to know they have a spider (vs. a mite vs. a harvestman), but it seems unlikely that a single image could be used to identify everything from order to species (as unlikely, perhaps as a single portion of DNA that could be use to distinguish all species in the animal kingdom). Therefore, it is likely that all practical auto-ID systems will need to be hierarchical to some degree (see following discussion).

Scalability

Scalability is certainly an important issue, as any relevant auto-ID system will need to distinguish large numbers of species. As the taxonomy of most difficult to recognize groups (e.g. insects, arachnids, etc.) is relatively fluid, auto-ID systems must have the capacity to expand and/or be modified without requiring an excessive amount of computing time. As stated previously, traditionally structured back-propagation ANNs require full retraining each time a new species is either added or removed. Some forms of ANNs (e.g. plastic self-organizing maps, or PSOMs) can accept new species almost indefinitely without major adjustments.

SPIDA, with its collection of individually trained species-level ANNs, falls somewhere in between. Species can be added without affecting the rest of the species' ANNs in the group. The relatively small number of 'anti' images (images not belonging to the species an ANN is being trained on) required in the training process (Table 9.1B) supports the ease with which new species could be added to established systems. That said, it might be necessary to retrain some of the other ANNs if the addition of the new species caused a decrease in accuracy. There is a limit beyond which it would be advisable to retrain the whole set, if too many species are added. However, we do not see this as major limitation since any sensible identification system must be structured hierarchically.

It is unrealistic to think that any one morphological structure would be universally applicable to all groups of organisms. Even among spiders, certain families will likely require either different or supplemental information, not just genitalic structures, for accurate species designations. In fact, SPIDA was originally set up hierarchically, with a set of genus ANNs trained first to classify an image to genus, then to circulate the same image through only the set of species ANNs within that genus (instead of all 121). Though initial results gave very high accuracy for genus-level identifications (99%), for some genera, the genus ANNs failed to converge (i.e. during training, a consistent solution was never found), making them essentially useless as a discriminatory tool.

This happened with very large, very diverse genera. This was not surprising since, in order for a hierarchical system to work, the same characters used for species determination must also be used

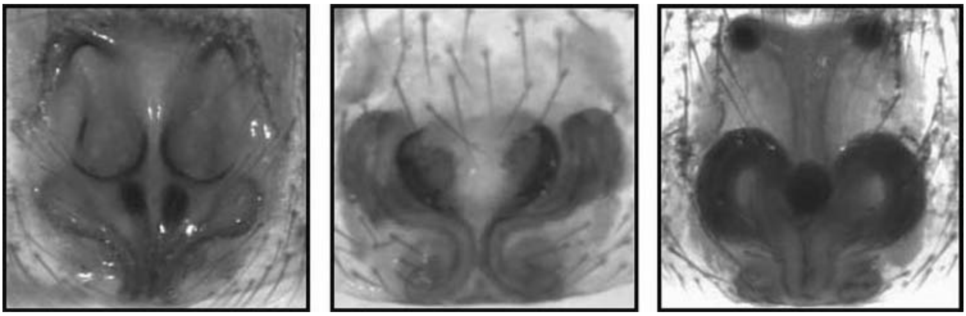


FIGURE 9.6 Epigyna of three species in the genus *Desognaphosa*: *D. massey, D. kroombit* and *D. bulburin.* These structures appear to share very few discernable features, if any. This illustrates the intrageneric diversity in many of the larger genera in this group, prohibiting the convergence of genus-level ANNs.

at higher levels of classification. For spiders, we know this is only sometimes the case. Typically, specialists do not use genitalic characters for genus classification. Despite this, for most genera, variation of epigynal structures within the genus was significantly less than variation between genera. In some cases, particularly in the larger genera (e.g. Desognaphosa with 26 species), the variation within a genus was extreme (Figure 9.6). Time was spent investigating potential solutions to this problem, but results proved inconclusive. Therefore, the final system was structured non-hierarchically, with each submitted image being circulated through all 121 species ANNs.

This simply illustrates the point that different information will be needed to distinguish different groups of organisms, so it makes more sense to design ID systems that operate within information type. Otherwise, one runs the risk of accidental similarities between processed images of very different structures. Perhaps the most efficient approach is to create systems tailored to manageable groups, adding a semi- or fully automated top-level system designed to shuttle the user to the appropriate subgroup. This could be accomplished with multiple-access keys or a separate ANN system based on different data. For example, family classification in spiders might be possible using images of eye-pattern or carapace shape (Roberts et al., in review). Finally, perhaps the most immediately useful application of this technology will be the identification of collections of species from a particular region or habitat, of which there will be a limited number of species in the pool. In either case, SPIDA is adequately scalable for most reasonable applications.

Flexibility

Flexibility is a measure of how easily an ID system can be applied to different groups of organisms. Some systems are amazingly accurate at distinguishing certain sets of organisms based on characters specific to that group – for example, ABIS (see Steinhag et al., this volume) using wing cell shape characters to identify bee species. Others have been shown to be amazingly flexible when tested on many different types of organisms and objects – for example, DAISY (see O'Neill, this volume). SPIDA, as defined by the combination of a wavelet encoding scheme and sets of individually trained ANNs, is not tied to any specific organism group, as its input is simply an image. This *is* a mild constraint, as at the moment, SPIDA can only be trained to identify organisms where the relevant characters can be imaged rather easily. Still, it has been tested with single images of wasp wings and bee wings and multiple images of male spider genitalia, all with successful results.

In summary, SPIDA-web is accurate (with adequate training data), highly accessible, reasonably scalable and quite flexible. There is plenty of room for improvement; as previously stated, our goal was to go from theory to practice, beginning to end. As that is now accomplished, the focus will shift to finding ways to improve (and better measure) accuracy, streamline the training process, explore the limits of SPIDA's scalability and further test its flexibility with new organism groups.

CONCLUSIONS

Automating the identification of specimens to species is a difficult task. There is no reason to believe that teaching a computer to identify species will be any easier than teaching a person to do so. In fact, it is likely a trickier process altogether, considering the amazing ability of the human mind to compensate for missing information and recognize the similarity in objects. The advantage, however, is that computers are fast, consistent and, once taught, do not forget. We chose to test our system on the most difficult of tasks: distinguishing individuals from closely related species. We also chose a challenging group, with species diagnostic characters that are difficult to quantify or even describe, making the use of traditional taxonomic keys problematic for the (relative) novice. Though sometimes strikingly similar within genera, these structures often vary widely in shape, size and dimensionality across genera, making even the basic description of differences complicated (if not impossible) for anyone but specialists in the group. Despite this, our results have been promising – even more so when you consider the growing evidence that specialists may not be as accurate or consistent as they think they are when assigning species names (e.g. Culverhouse, this volume). More to the point, we suspect that higher accuracy (~95–100%) will be attainable for ecological samples, as species will be more disparate.

As mentioned previously, we envision the most useful application of this technology to be in ecological and/or conservation studies. The majority of studies to date looking for ecological patterns in diversity, distribution, response to disturbance, etc. have relied on only a handful of surveys, often only one (Spellerberg, 1991; Green et al., 2005). This 'snapshot' approach limits our understanding of the processes governing the dynamic nature of species and communities and can often yield misleading data. There has been much discussion in the literature of promoting more multisurvey, long-term studies of biological communities to ameliorate this problem; there is widespread agreement about the need for repeated surveys to help expand our understanding of ecological phenomena, especially in the face of increasing human impacts (landscape alteration, global climate change, etc.; see Balmford et al., 2005; Green et al., 2005).

That said, most funding agencies require results in 2–3 years and often conservation efforts face even more pressing deadlines in terms of averting ecological disasters. When working with arthropods in particular, though they are relatively easy to sample, the identification process can often take years. This time delay is often used as an argument to not include arthropods in conservation studies and/or biological monitoring efforts. Our development of SPIDA-web targeted this type of need. We foresee identification modules being developed on data from the first set of surveys conducted at a site. Once trained, these modules could then be used to identify all subsequent surveys, leading to quick analysis of community dynamics. Technicians could be easily taught to image the specimens and submit them for identification via SPIDA-web; then, only the few individuals not recognized by SPIDA-web need be examined by a specialist, thus saving vast amounts of time.

In addition, there is no need to wait until newly collected species are given a proper scientific name – they could be added to the module based on a morphospecies designation in the short term. This has the added advantage of guaranteeing consistency in morphospecies classification throughout the monitoring period. This is relevant, as we anticipate SPIDA-web being most useful when trained on collections of species from a particular region, thereby being applied to pressing problems of biodiversity and conservation.

But what of the future beyond SPIDA-web? At a minimum, there needs to be cooperation and data sharing between groups working in the field of automated object recognition. We are at the stage where alternate approaches can be tested and evaluated based on the criteria outlined previously using real data. There will certainly be no single solution and no single approach that can be labeled as 'the best' for all tasks and/or organism groups. Yet seeing where each succeeds (or fails) will yield practical data and can only propel the field forward. To make automated object identification practically useful, enough infrastructure must be built up to make the creation and maintenance of identification modules suitably efficient and relatively independent of the whims of short-term funding agencies. Perhaps the creation of a research centre for automated identification or, on a smaller scale, the establishment of permanent research positions in this field at major research museums would provide adequate stability and resources. Of course, this will only happen if the powers that be are convinced that automated object recognition is an integral part of the future of taxonomy.

ACKNOWLEDGEMENTS

Funding for this project was provided by the National Science Foundation, grant #0119578. We thank the editor of this volume, Professor Norman MacLeod, and two anonymous reviewers for their helpful comments on the manuscript.

REFERENCES

Balmford, A., Crane, P., Dobson, A., Green, R.E. and Mace, G.M. (2005) The 2010 challenge: data availability, information needs and extraterrestrial insights. *Philosophical Transactions of the Royal Society, Series B,* 360: 221–228.

Bazanov, P.V., Buryak, D.Y., Mun, W.J., Murynin, A.B. and Yang, H.K. (2005) Comparison of Gabor wavelet and neural network-based face detection algorithms. In *Signal and Image Processing* (ed M.W. Marcellin), ACTA Press, Honolulu, Hawaii, pp. 200–208.

Boddy, L., Morris, C.W. and Wimpenny, J.W.T. (1990) Introduction to neural networks. *Binary,* 2: 179–185.

Chtioui, Y., Bertrand, D., Dattee, Y. and Devaux, M.-F. (1996) Identification of seeds by colour imaging: comparison of discriminant analysis and artificial neural network. *Journal of the Science of Food and Agriculture,* 71: 433–441.

Dietrich, C.H. and Pooley, C.D. (1994) Automated identification of leafhoppers (Homoptera: Cicadellidae: Draeculacephala Ball). *Annals of the Entomological Society of America,* 87: 412–423.

Do, M.T. (1996) *Pattern Recognition System Using Artificial Neural Networks and Wavelets for Taxonomic Identifications.* Master's thesis. University of Tennessee, Knoxville.

Do, M.T., Harp, J.M. and Norris, K.C. (1999) A test of a pattern recognition system for identification of spiders. *Bulletin of Entomological Research,* 89: 217–224.

Dodd, J.C. and Rosendahl, S. (1996) The BEG expert system – A multimedia identification system for arbuscular mycorrhizal fungi. *Mycorrhiza,* 6: 275–278.

Edwards, M. and Morse, D.R. (1995) The potential for computer-aided identification in biodiversity research. *Trends in Ecology and Evolution,* 10: 153–158.

Fahlman, S.E. (1988) An empirical study of learning speed in back propagation network. Technical Report (CMU-CS-88-162). Pittsburgh, PA, Carnegie Mellon University.

Fahlman, S.E. and Lebiere, C. (1991) The cascade-correlation learning architecture. Technical Report (CMU-CS-90-106). Pittsburgh, PA, Carnegie Mellon University.

Gerhards, R., Nabout, A., Sokefeld, M., Kuhbauch, W. and Nour Eldin, H.A. (1993) Automatic identification of 10 weed species in digital images using Fourier descriptors and shape parameters. *Journal of Agronomy and Crop Science,* 171: 321–328.

Graps, A. (1995) An introduction to wavelets. *IEEE Computational Sciences and Engineering,* 2: 50–61.

Green, R.E., Balmford, A., Crane, P.R., Mace, G.M., Reynolds, J.D. and Turner, R.K. (2005) A framework for improved monitoring of biodiversity: responses to the World Summit on Sustainable Development. *Conservation Biology,* 19(1): 56–65.

Goldwasser, L. and Roughgarden, J. (1997) Sampling effects and the estimation of food-web properties. *Ecology,* 78: 41–54.

Goodacre, R., Timmins, E.M., Burton, R., Kaderbhai, N., Woodward, A.M., Kell, D.B. and Rooney, P.J. (1998) Rapid identification of urinary tract infection bacteria using hyperspectral whole-organism fingerprinting and artificial neural networks. *Microbiology-UK,* 144: 1157–1170.

Goodacre, R., Timmins, E.M., Rooney, P.J., Rowland, J.J. and Kell, D.B. (1996) Rapid identification of *Streptococcus* and *Enterococcus* species using diffuse reflectance-absorbance Fourier transform infrared spectroscopy and artificial neural networks. *FEMS Microbiology Letters,* 140: 233–239.

Howell, A.J. and Buxton, H. (1995) Receptive field functions for face recognition. In *Proceedings of the 2nd International Workshop on Parallel Modelling of Neural Operators for Pattern Recognition (PAMONOP)*, Faro, Portugal, pp. 83–92.

Hurst, R.E., Bonner, R.B., Ashenayi, K., Veltri, R.W. and Hemstreet, G.P.I. (1997) Neural net-based identification of cells expressing the p300 tumor-related antigen using fluorescence image analysis. *Cytometry,* 27: 36–42.

Instituto Nacional de Biodiversidad. (2001) National Biodiversity Inventory. <www.inbio.ac.cr/en/invn/Invent.html> (March 11).

Jiang, Y., Nishikawa, R.M., Schmidt, R.A., Wolverton, D.E. and Comstock, C.E. (1996) Evaluation of a computerized classification scheme for clustered microcalcifications for computer-aided diagnosis. *Radiology,* 201: 1304–1304.

Kennedy, M.J. and Thakur, M.S. (1993) The use of neural networks to aid in microorganism identification: a case study of *Haemophilus* species identification. *Antonie van Leeuwenhoek,* 63: 35–38.

Krüger, V., Bruns, S. and Sommer, G. (2000) Efficient head pose estimation with Gabor wavelet networks. In *Proceedings of the 11th British Machine Vision Conference,* University of Bristol. UK, pp. 72–81.

Kwon, Y.K. and Cho, R.K. (1998) Development of identification method of rice varieties using image processing technique. *Agricultural Chemistry and Biotechnology,* 41: 160–165.

Mancuso, S. and Nicese, F.P. (1999) Identifying olive (*Olea europaea*) cultivars using artificial neural networks. *Journal of the American Society for Horticultural Science,* 124: 527–531.

Maollemi, C. (1991) Classifying cells for cancer diagnosis using neural networks. *IEEE Expert,* 6(6): 8–12.

Morris, C.W. and Boddy. L. (1995) Artificial neural network identification and systematics of eukaryotic microorganisms. *Binary,* 7: 70–76.

Oliver, I. and Beattie, A.J. (1993) A possible method for the rapid assessment of biodiversity. *Conservation Biology,* 7: 562–568.

Oliver, I. and Beattie, A.J. (1996) Invertebrate morphospecies as surrogates for species: a case study. *Conservation Biology,* 10: 99–109.

Oliver, I., Pik, S., Britton, D., Dangerfield, J.M., Colwell, R.K. and Beattie, A.J. (2000) Virtual biodiversity assessment systems. *Bioscience,* 50: 441–450.

Parsons, S. and Jones G. (2000) Acoustic identification of twelve species of echolocating bat by discriminant function analysis and artificial neural networks. *Journal of Experimental Biology,* 203: 2641–2656.

Platnick, N.I. (1999) Dimensions of biodiversity: Targeting megadiverse groups. In *The Living Planet: Biodiversity Science and Policy* (eds J. Cracraft and F.T. Grifo), Columbia University Press, New York, pp. 33–52.

Platnick, N.I. (2002) A revision of the Australasian ground spiders of the families Ammoxenidae, Cithaeronidae, Gallieniellidae, and Trochanteriidae (Araneae: Gnaphosoidea). *Bulletin of the American Museum of Natural History,* 271: 1–243.

Pollen, D.A. and Ronner, S.F. (1981) Phase relationships between adjacent simple cells in the visual cortex. *Science,* 212: 1409–1411.

Rambold, G. and Agerer, R. (1997) DEEMY – The concept of a characterization and determination system for ectomycorrhizae. *Mycorrhiza,* 7: 113–116.

Rataj, T. and Schindler, J. (1991) Identification of bacteria by multilayer neural networks. *Binary,* 3: 159–164.

Roberts, A.K., Smith, D.M., Guralnick, R.P., Cushing, P.E. and Krieger, J. (in review). Quantitative prediction of fossil biodiversity: eocene spiders from Florissant, Colorado. *Palaios*.

Spellerberg, I.F. (1991) *Monitoring Ecological Change*. Cambridge University Press, Cambridge, 350 pp.

Theodoropoulos, G., Loumos, V., Anagnostopoulos, C., Kayafas, E. and Martinez-Gonzales, B. (2000) A digital image analysis and neural network based system for identification of third-stage parasitic strongyle larvae from domestic animals. *Computer Methods and Programs in Biomedicine,* 62: 69–76.

Weeks, P.J.D., Gauld, I.D., Gaston, K.J. and O'Neill, M.A. (1997) Automating the identification of insects: a new solution to an old problem. *Bulletin of Entomological Research,* 87: 203–211.

Weeks, P.J.D., O'Neill, M.A., Gaston, K.J. and Gauld, I.D. (1999) Automating insect identification: exploring the limitations of a prototype system. *Journal of Applied Entomology,* 123: 1–8.

Wilkins, M.F., Boddy, L., Morris, C.W. and Jonker, R. (1996) A comparison of some neural and non-neural methods for identification and phytoplankton from flow cytometry data. *Cabios,* 12: 9–18.

Wit, P. and Busscher, H.J. (1998) Application of an artificial neural network in the enumeration of yeasts and bacteria adhering to solid substrata. *Journal of Microbiological Methods,* 32: 281–290.

Zhu, J., Vai, M.I. and Mak, P.U. (2004) Gabor wavelets transform and extended nearest feature space classifier for face recognition. In *Proceedings of the Third International Conference on Image and Graphics*, Hong Kong, pp. 246–249.