Unlike pathway and multiaccess keys, which use diagnostic morphological characters, NemaScope uses point-and-click visual matching that allows users to navigate through a collection of images until images similar to specimens under investigation are found. High-definition multifocal images of genera provide the basis for the initial photos and can be viewed when additional morphological information is needed.

Acknowledgments

I thank Jill Brady, Eamonn Keogh, and Michalous Falutoso for their contribution to the creation and maintenance of NemaScope, and Paul De Ley for his advice during formation of this communication.

References Cited

Blaxter, M., B. Elsworth, and J. Daub. 2004. DNA taxonomy of a neglected animal phylum: an unexpected diversity of tardigrades. Proc. R. Soc. Lond. B 271: S189–S192.

De Ley, P. and W. Bert. 2002. Video capture and editing as a tool for the storage, distribution, and illustration of morphological characters of nematodes. J. Nematol. 34: 296–302.

De Ley, P., I. Tandingan De Ley, K. Morris, E. Abebe, M. Mundo-Ocampo, M. Yoder, J. Heras, D. Waumann, A. Rocha-Olivares, A. H. J. Burr, J. G. Baldwin, and W. K. Thomas. 2005. An integrated approach to fast and informative morphological vouchering of nematodes for applications in molecular barcoding. Phil. Trans. R. Soc. B 360: 1945–1958.

Eyualem, A., R. E. Grizzle, D. Hope, and W. K. Thomas. 2004. Nematode diversity in the Gulf of Maine, USA, and a Web-accessible, relational database. J. Mar. Biol. Assoc. UK 84: 1159–1167.

Yoder, M., I. Tandingan De Ley, I. Wm. King, M. Mundo-Ocampo, J. Mann, M. Blaxter, L. Poiras, and P. De Ley. 2006. DESS: a versatile solution for preserving morphology and extractable DNA of nematodes. Nematology 8: 367–376.

Melissa Yoder is a graduate student working towards a Ph.D. in nematode systematics at the University of California–Riverside. Her research interests include creating and testing user-friendly biological identification keys, and studying the biodiversity of freshwater nematode communities (yoderm01@student.ucr.edu).

Automated Classification of Skippers based on Parts Representation

Bir Bhanu, Rui Li, John Heraty, and Elizabeth Murray

Rapid advances in digital imaging technology, the low cost of cameras, scanners, and storage devices, and the accessibility of the Web make it possible to collect, store, and access huge numbers of images. Advances in areas of digital instrumentation, bioinformatics, and cyberinfrastructure are impressive, but much more is yet to come (MacLeod 2007). The morphological data inherent in these images is enormous but largely unexplored by novel techniques. Overall, morphological data can be tedious to collect and is often qualitative rather than quantitative. Beyond traditional data collection, geometric morphometrics, based on Cartesian coordinates of anatomical landmarks, is widely used in morphometric analyses, but these data are laborious to collect and analyze. For human-directed measurements, statistics-based toolkits have functioned well, but have drawbacks (Zelditch et al. 2004).

What if unidentified specimen images could be amassed in a database and then correctly classified through an automated system? This technique could help biologists identify specimens and lead to searches with images instead of using key words as query items. These kinds of searches are part of available Automated Taxon Identification (ATI) systems such as SPecies IDentification, Automated (SPIDA), Automated Bee Identification System (ABIS), and Digital Automated Identification SYstem (DAISY), which are being improved in accuracy, accessibility, scalability, and flexibility for image-based classification (reviewed in MacLeod 2007). All systems have identification accuracy levels of 95% or higher for certain data sets.

These ATI systems rely upon classical human recognition of taxonomic landmarks or features for accurate classification. Conversely, the automated classification system we describe relies on abstract

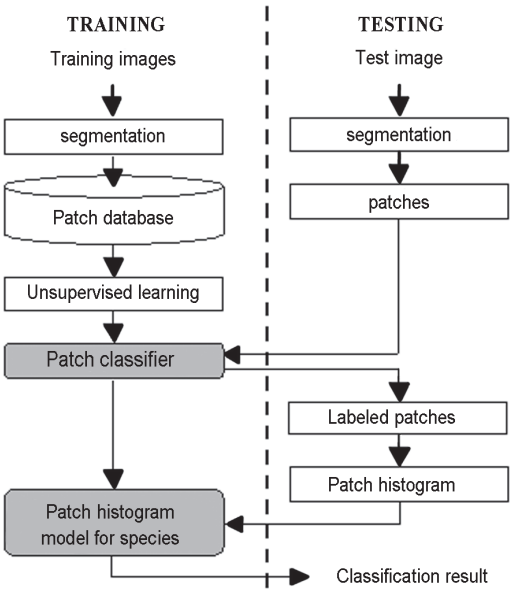

Fig. 1. System diagram.

recognition to differentiate and classify images. Our goal is to develop a system that can recognize features important for grouping (recognition) and separating (phylogenetics and evolution) groups of organisms on the basis of a variety of images.

For biological images, local patterns/features are critical for defining different species. We propose a patch-based system of analysis for the groups of interest (Fig. 1). Compared with classic pattern recognition approaches based on global features, we use information-rich local patches. This patch-based representation could be used to explore significant differences between images and may help to exploit more information on species classification and evolution.

Related Work and Contributions

In different applications of pattern recognition, researchers use various features, which include raw pixel intensities, features obtained via global image transformations, and local features such as edge fragments, rectangle features, Gabor filter-based representations, and wavelet features. Object parts are extracted that are rich in information content and use part-based representation. As an example, Agarwal et al. (2004) extracted square patches around interest points. Intensity pixel values are used to represent the patches, and the sparse representation is used for describing the image. The method is shown to have a good performance in detecting motor vehicles (cars) in side view. Compared with this algorithm, our approach extracts patches of various shapes that contain more precise local information. It also uses a more compact description for the whole class. This representation can help exploit the significant visual difference between classes.

Technical Approach

Patch Extraction – Segmentation with Normalized Cut. Different from classic interest-point search algorithms, our system uses segmentation to extract information-rich patches. A normalized-cut algorithm, related to the graph theory of grouping, is used. The sets of points in an arbitrary feature space are represented as a weighted undirected graph $G = (V, E)$, where the nodes of the graph (V) are the points in feature space and an edge (E) is formed between each pair of nodes. The weight on each edge, $w(i, j)$ is a function of the similarity between nodes i and j. G can be partitioned into two disjoint sets, $A, B, A \cup B=V, A \cap B=\phi$, by simply removing the edges connecting A and B. The degree of dissimilarity between these two pieces can be computed as a total weight of the edges that have been removed. In graph theory language, this is called the *cut*, so that $\text{cut}(A,B)=\Sigma_{u\in A, v\in B} w(u,v)$.

The optimal bipartitioning of a graph minimizes this *cut* value. However, this minimum cut criteria favors cutting small sets of isolated nodes in the graph. So Shi and Malik (2000) defined a new criterion called *normalized cut* (*Ncut)*. Solving the minimized *Ncut* problem reduces to solving the generalized eigenvalue problem. The eigenvector corresponding to the second smallest eigenvalue is used to bipartition the graph. If the current partition should be subdivided, the algorithm is recursively run to make more segments.

Patch Classification—Unsupervised Learning of a Gaussian Mixture Model

After segmentation, the training set is partitioned into patches and allocated to a patch database. A feature vector of each patch is extracted, and the patches classified on the basis of these features. We assume that feature vectors $X=\{\mathbf{x}^1,\ldots, \mathbf{x}^N\}$ are samples of a Gaussian mixture model. $\mathbf{x}^1,\ldots, \mathbf{x}^N$ represent the outcome of a random variable $\mathbf{X}$. $\mathbf{X}$ follows a C-component mixture model, as shown in equation given below, where θ_i is the set of parameters for the ith mixture component, and α_i is the component weight. All α_i must be positive and sum to 1. We assume that all components follow a *Gaussian Mixture Model* (GMM), where θ_i is the mean vector $\mathbf{u}_i$ and covariance matrix Σ_i.

$$p_i(\mathbf{X}|\theta_i) = \sum_{i}^{C} \alpha_i N(\mathbf{u}_i, \Sigma_i),\ \alpha_i > 0,\ i = 1,\cdots, C, \text{ and } \sum_{i=1}^{C} \alpha_i = 1$$

When C is known, the classic *Expectation Maximization* (EM) algorithm could be used to estimate the parameters and classify the feature vectors. However, in most cases, C is unknown. Figueriedo and Jain (2002) proposed a variant of EM. This algorithm seamlessly integrates model selection (finding the number of clusters) and model estimation (Gaussian component parameter estimation) in the iterative process. It incorporates a *Minimum Description Length* (MDL) criterion for model selection and achieves the best estimation of the mixture parameters.

This algorithm classifies the patches into several classes and learns a Bayesian patch classifier.

Training Model—Patch Histogram Model

Each patch in the patch database is labeled after the Bayesian patch classifier is found. We assume

Table 1. Skipper dataset.

No.	Subfamily	Species name	Training sample no.	Testing sample no.
1	Hesperiinae	Perichares philetes	15	5
2	Hesperiinae	Vettius aurelius	18	14
3	Pyrginae	Astraptes SENNOV	47	11
4	Pyrginae	Entheus matho	13	5
5	Pyrginae	Phocides pigmalion	13	5
6	Pyrginae	Urbanus belli	32	26

K patch classes, and therefore a *K*-bin histogram can be built for each image. Patch histograms for all images in one class are averaged to form a *patch histogram model.*

Testing—Patch Histogram Matching

A test image is segmented into patches; each patch is labeled by the patch classifier, and then a patch histogram is built. The χ^2 distance between this patch histogram and the patch histogram model is calculated for every class. Classification of the test image corresponds to the one with the shortest distance.

Experimental Results Dataset

We built an image dataset for six species (see Table 1 and Fig. 2) of Hesperiidae containing 138 training images and 66 test images. Some images in the dataset have shadows and missing body parts.

Patch Extraction

An *Ncut* algorithm is applied to each image, and 40 segments are found for each image. An example is shown in Fig. 3: 3a is the original color image, and 3b is the segmentation result superimposed on the intensity image. In 3c, background segments are removed. Using 3c, the skipper is scaled to fill the image frame and resized to 150×150 pixels, as shown in 3d. We call this *normalization*. Every training and testing image is segmented and normalized. All patches from the training dataset are gathered to form a patch database of 2,709 patches.

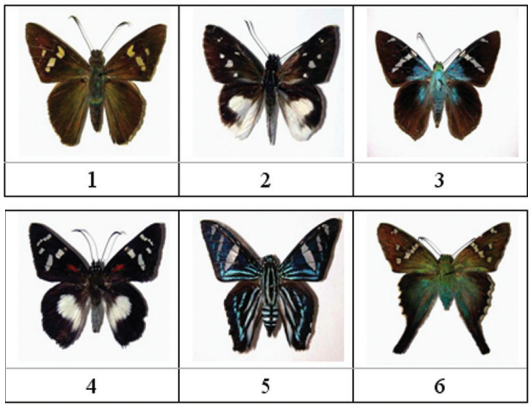

Fig. 2. Sample image of each species.

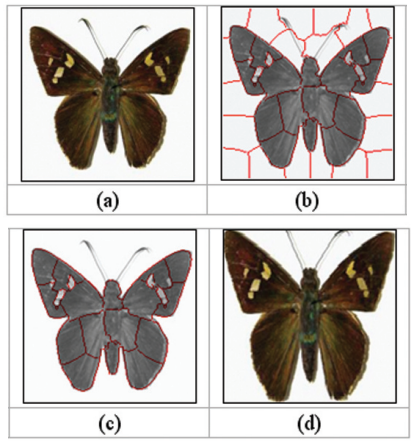

Fig. 3. Image segmentation and normalization.

Patch Classification

We extract six color features (means and variances for hue, saturation, value) for each patch. Thus, the patch database contains 2,709 6D feature vectors. The classification results for the 18 patch classes are shown in Fig. 4.

Build Training Model

We build a patch histogram for each image and average all the histograms in one class to form a patch histogram model. The six patch histogram models are shown in Fig. 5. Each has 18 bins, corresponding to 18 patch classes. From

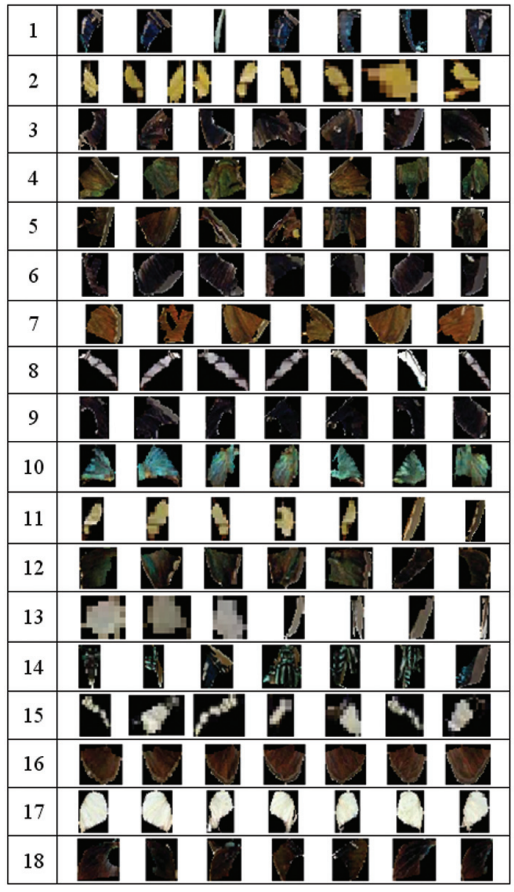

Fig. 4. Classified patches.

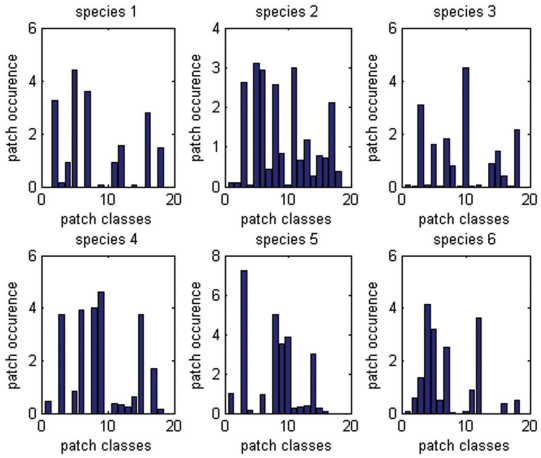

Fig. 5. Patch histogram model for each species.

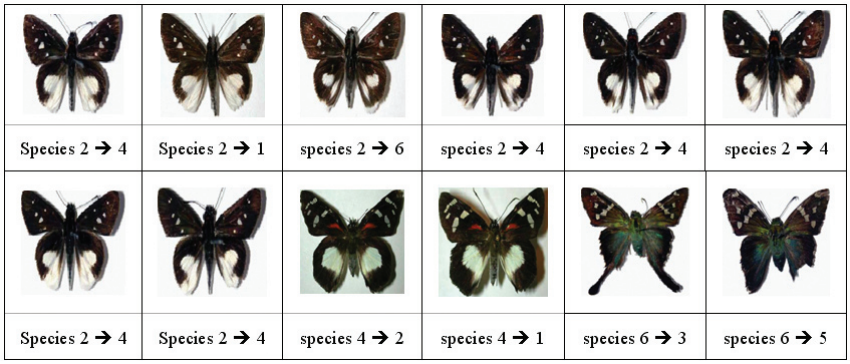

Fig. 7. Misclassified test images with their ground-truth labels (species label before arrow is the ground-truth label and after the arrow is the misclassification label assigned by the system).

5	0	0	0	0	0
1	6	0	6	0	1
0	0	11	0	0	0
1	1	0	3	0	0
0	0	0	0	5	0
0	0	1	0	1	24

Fig. 6. Confusion matrix for the test dataset.

the histogram, we can recognize which features are most significant for each species. For example, for species 1, patches 2, 5, 7 and 16 appear most often, and for species 3, patches 3 and 10 are most common, so we could use these to differentiate the two classes (species).

Testing Results

We calculate the χ^2 distance between the test image and each species histogram model. The classification precision is 81.8% (54 species out of 66 species are correctly classified) (Fig. 6). The taxa across the diagonal are correctly classified and those across the off-diagonal misclassified.

Conclusions

Image data can help to understand species evolution from a new perspective. In this paper, we propose a parts-based (patch-based) representation for biological images. Experimental results show this compact model as efficient and effective for representing and classifying skipper images. The results can be further improved by exploiting symmetry of the shape and increasing the quality of image segmentation.

Acknowledgements

The authors thank Dr. Dan Janzen, University of Pennsylvania, for supplying test images (http://janzen.sas.upenn.edu). The authors also thank Cameron Allen and Yu Sun for their helpful comments. This research was supported in part by NSF grant 0641076.

References Cited

Agarwal, S. et al. 2004. Learning to detect objects in images via a sparse, part-based representation. IEEE Trans. Pattern Anal. Mach. Intell. 26: 1475–1490.

Figueriedo, M. A., and A. Jain. 2002. Unsupervised learning of finite mixture models. IEEE Trans. Pattern Anal. Mach. Intell. 24: 382–396.

MacLeod, N. 2007. Automated taxon identification in systematics: theory, approaches and applications. CRC Press, Boca Raton, FL.

Shi, J., and J. Malik. 2000. Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 22: 888–905.

Zelditch, M. et al. 2004. Geometric morphometrics for biologists. Elsevier Academic Press, Boston.

Bir Bhanu is a professor of Electrical Engineering and Cooperative Professor of Computer Science and Engineering at the University of California, Riverside (bhanu@cris.ucr.edu). **Rui Li** completed her Ph.D. in ELectrical Engineering at the University of California at Riverside. Her interests are in pattern recognition, machine learning and databases. **John Heraty** is a professor of entomology at the University of California, Riverside. He specializes in systematics, phylogeny, and biogeography of parasitic wasps (Chalcidoidea, Hymenoptera). His interests are biological image databases, computer vision, pattern recognition, learning and data mining (john.heraty@ucr.edu). **Elizabeth Murray** is a Ph.D. student in systematics at the University of California at Riverside. She is beginning a phylogenetic revisionary investigation on a genus of parasitic Hymenoptera.