NWACC Final Report
Submitted April 10, 2009

LepID: An Online Image Analysis and Pattern Recognition Tool For Lepidoptera Identification Using A "Signature" Image

Project Director: Dr. Jeffrey C. Miller
Department: Department of Rangeland Ecology and Management
Institution: Oregon State University
Eligibility Code: G804-17
Email: jeffrey.miller@oregonstate.edu
Telephone: 541-737-5508
Other Participants: Dr. Hans Luh

PROJECT DESCRIPTION

In the biology curriculum, species classification and identification are two important topics to help students understand disciplines as varied as physiology, behavior, ecology, evolution, and biodiversity. Classification is the process of defining and naming groupings of organisms based upon the similarity of their attributes, while identification is the process of assigning a specimen to a discretely defined entity. Both classification and identification are based on comparative descriptions which are framed in terms of a list of diagnostic characters. Each character contributes to a set of features that can be used to distinguish one organism from another. The measure of features may be discrete (e.g., number of antennae) or continuous (e.g., width of head). A series of characters with specified values are organized as identification rules that can be organized into diagnostic dichotomous keys, which have been in use for more than 200 years. The use of diagnostic keys often requires that the users fully understand the species characteristics and the measurement of conditional values for each diagnostic feature. It is usually a difficult and frustrating task for non-experts to use diagnostic keys. Traditional diagnostic keys are designed to mimic the thinking of professional taxonomists who can easily carry out the identifications because they are already familiar with the characteristics used in the keys. However, biology students and non-academic nature-lovers are often frustrated by the jargon and structural make-up of dichotomous keys. This type of experience is counterproductive to the appreciation and learning of biological concepts. We propose to take the cumbersome and tedious process of keying-out specimens and replace it with a positive feedback mechanism based on digital imaging, patterns, and colors. We fully recognize certain groups of species, such as most insects that happen to be very difficult to identify, are more amenable to this approach than others.

With an increasing use of computer technology, many traditional diagnostic keys have been digitized and are available online. Two examples of online programs are the Lichen Synoptic Key - http://ocid.nacse.org/lichenland/synopticKey/index.php - and BugBytes - http://ipmnet.org/ent3/bugbytes/. Although a computer-based–tool provides an effective

means for acquiring an identification when designed for interactive use, the information embedded into the system is still a text-based description. Some identification web sites include an image hyperlink to each attribute so users can compare the description to a real image, but the core search engine is still based on text matching to a diagnostic key. The technology of image analysis and image pattern recognition has not been applied to any computer-based web-interactive program for obtaining a name for a species. It is our goal to create a web-interactive program that is based on digital image analysis and recognition through matching patterns and colors between an authoritatively identified reference library and a submitted image of an unknown taxon. The initial proof-of-concept system will provide rapid and accurate identification of moths and butterflies.

OBJECTIVES

1. Develop an online image processing and pattern recognition tool for helping biology students, professionals, and non-academic nature-lovers identify species of Lepidoptera.

2. Evaluate the speed and accuracy of template-matching and signature-mark directed searches while comparing an uploaded image with a resident reference library.

HOW GOALS WERE MET - WORK ACCOMPLISHED

1. Creation and development of the project web site, including, integration of a web-based 'image upload' function; 'automated framing' function in Java script to resize uploaded images to a common set of dimensions without warping the image; a Java script user controlled 'framing function' for delimiting, with built-in restricted sizing, the signature mark in the uploaded image; a Microsoft SQL database component, creation of an original 'advanced search' criterion menu; integration of the "template matching" function in Victor Image Analysis Library software that was modified in the source code by us for more accurate and faster searching capabilities; the design for the results presentation of the template matching search with scores representing the quantitative fit of the match; along with written instructions and a DEMO module regarding step by step guidelines on how to use the site. We also provide the original grant proposal and an introduction; and, lastly, availability of a select set of 'test' images for users to give the site a trial run.

The LepWingID website can be found at http://ipmnet.org/lepid/

2. Data entry for each image placed into the search reference library. These data are used by the advanced search function that allows the user to limit the number of reference library images included in any given search. The information fields placed into the database are: image filename, Family, Genus, species, type (moth or butterfly), body (dorsal or ventral view), and color (one of nine predefined colors). The database currently contains 405 records.

3. Acquisition of digital images showing dorsal and ventral views of fully spread museum specimens of Lepidoptera, moths and butterflies. We currently have a reserve library of more than 1,000 images for selection of additional species to be added to the program in the future. Presently, among the 405 images in the library: 272 show butterflies and 133 show moths; 311 show a dorsal (upperside) view and 94 show a ventral (underside) view. The species depicted in each image are assigned to a color category based on any of nine colors. A total of 18 Families of Lepidoptera are contained in the reference library. The regions represented by the species are: North America (primarily Oregon), Mexico, Central America, South America, South Korea, Taiwan, and Thailand.

4. Graphic editing. On average 12 minutes were required to edit each image selected for placement into the reference library. The specimen depicted in the image must be near perfect regarding tears, scratches, rubbed areas, and orientation of the wings and abdomen. Also, the background in the reference library images must a solid color (we chose either black or white) because the signature image is programmed to search a slightly larger area than the actual signature mark and the background must be 'neutralized'. However, the background in an uploaded image may exhibit any variety of colors and patterns.

5. Testing the search function. The testing of the search speed and accuracy was conducted in multiple phases. The first phase was to see if a signature mark placed over an uploaded image could find that very same image which had been previously placed into the reference library, we termed this 'self-recognition of the twin image'. Failure to find the identical image would mean additional work was needed in reprogramming the Victor Image search engine. The second phase involved using an uploaded image that was totally independent of the library image (not the very same image). The library image could have been acquired from the web, scanned out of a book, or from the portfolio of JCM. Similarly, the uploaded image would originate from a separate source and depict a specimen different than the specimen illustrated in the reference library, we termed this 'independent image matching'. Failure to achieve an accurate result would suggest that the library requires additional images to represent natural variation within each species and/or the 'forgiveness' in the search of the signature mark was either too strict or too sloppy. A third phase of our testing involved changing the size of the signature cropping tool and changing the criteria included in the advanced search. These tests gave valuable information regarding the speed of the search based on the area (pixel width x pixel height) searched for a match in the signature mark in each qualified library image and on the number of images remaining that qualified for the search based on the selected advanced criteria. The fourth phase involved creating a cartoon of four typical wing patterns and assessing the 'strength' of template matching based on artistic re-creation of pattern. This test served many purposes, one, was to determine if we could provide a means for a user to recall a color and pattern seen in the field but for which no digital image is available; two, was to allow a user to test for a range of virtual colors and patterns and to see if a species in nature (as represented in our reference library) had evolved to show the artistic rendition. The fifth testing phase was to ask independent users to log onto the site and conduct a search with an image of their own creation. Of course they needed to use an image of a species listed in our reference library. The sixth and final testing phase was to either use an image of a species not represented in the reference

library or to answer the advanced search criteria questions incorrectly. The results from these intentionally misdirected searches provided interesting insights into the accuracy of identifications (based on the scores provided along with the results image) and the need for a user to invoke some degree of 'sensible judgment' on accepting the results of a search.

Phase I: Self-recognition of twin image tests. A total of 200 tests were conducted. The test achieved a rate of 100% accuracy based on achieving a result where the "unknown" twin was uploaded and the search resulted in the "known" twin being located in the library as the top-ranked match among the library images. These results gave us confidence in the utility of designating a signature reference mark and the capability of the software in finding a suitable match for the submitted pattern. Thus, we were confident in advancing to Phase II and conducting tests involving independent images was warranted.

Phase II: Independent image tests. A total of 100 tests were conducted. Based on the criterion of the top-ranked result being the correct match these test were 98% accurate. However, using a less stringent criterion that the search was 'successful' if the correct identification was among the top 5 ranked results, then the tests were 100% accurate. One of the two tests that failed to give the top-ranked correct identification involved a green moth where the reference image was of a dark individual. The uploaded image in this test was of a specimen with a lighter green tone. Therefore, we created a new reference image using a lighter exposure (of the dark specimen) and the rerun of the search produced a result with the top-ranked match being the correct identification. The second failed test involved a group of species with very similar markings, silver stripes and patches on a blue-black background. The original search resulted in the correct match being the #3 ranked image. Although the selection of the correct identification was possible simply by visually comparing the photos we reran the search using a different signature mark. The rerun test resulted in the top-ranked result being correct. Firstly, the tests using independent images showed the search software to be highly accurate. Secondly, we had anticipated (and these tests support the concern) the need for a reference library stocked with multiple images of each species showing variation not only in the natural pattern but also the natural tones and hues exhibited by the species. Thirdly, if the user does not acquire a result that is satisfactory (an obvious mismatch) then a second search should be conducted using a different signature mark. Overall, we expect the library should possess at least 5-6 images of each species to have natural variation represented and a user should conduct two searches to confirm the identification if numerous similar looking species are involved in the search results.

Phase III: Area of signature mark and the number of searched reference images. A total of 100 tests were conducted. This set of tests involved running searches using a selected computer (Mac G-5) with a private home broadband connection (Earthlink) to the internet based on the smallest signature mark possible (20 x 20 pixels), a medium sized mark (40-60 x 40-60 pixels), and a large-sized mark 60-80 x 60-80), then timing the search period, and scoring the accuracy of the top-ranked results. In general, the smallest signature mark resulted in the fastest search times with an accuracy of 99.9%. The largest signature mark resulted in the slowest search times with an accuracy of 100%. The single failed test using the small mark was mitigated by using a slightly larger mark that provided an accurate

identification for the top-ranked result. Overall, the searches based on the smallest signature box proceeded at a scanning rate of 16-20 images per minute. Thus, a search of 40 images (obtained by incorporating advanced criteria) took 2.5 seconds while a search of the entire library (n=328 at the time of the test and with no advanced criteria selected) took an average of 16-20 seconds. In comparison, cropping the signature mark to the largest allowed size resulted in a search rate of 5-6 images per minute, equivalent to a scanning period of one hour to evaluate the entire (n=328) library. The longer search time using the largest signature mark possible becomes an issue if advanced search criteria are not included and as the reference library adds more images. However, the smaller signature mark appears to be highly accurate and much faster.

Phase IV: Cartoon tests. A total of ten tests were conducted. These tests were based on the uploading of an artistic version of a moth or butterfly image. No correct identifications were possible but similarities in matching the top-5 result images for color and pattern could be evaluated. Four examples of results demonstrate the utility of these tests and the matching 'power' of the program. Firstly, based on the signature mark of the white dot on a red background a search provided the single species in the library that was the real life model for this artistic pattern. Secondly, based on the eyespot as the signature mark the top-ranked results showed species with eyespots on the hindwing. Thirdly, based on grey and black vertical wavy lines the search found 5 species all with grey and black lines. Fourthly, a search for the well marked wing veins resulted in species exhibiting well marked veins in the hindwing. These tests demonstrated the ability of the program to match species to a pattern when the uploaded pattern was virtually an artistic creation rather than photographic reality.

Phase V: Independent user tests. A total of five tests were conducted. Four of the tests resulted in correct identifications based on the top-ranked result. The single incorrect identification was based on a result where the #3 ranked image was correct. In this test the #1 and #2 ranked species were very similar. A rerun of the search using a larger signature mark in a different location provided the correct identification matched with the top-ranked result.

Phase VI: Intentionally misdirected tests. A total of ten tests were conducted. These ten tests involved intentionally conducting a misguided search to observe how the search program responded with the top five results, including the matching score. The reality of this type of test is based on two points: 1) a user is likely to submit an image of a species not in the reference library but may obtain a search result that provides a lead to the correct genus, and 2) the advanced search criteria may be selected incorrectly. One example of the results is illustrated by selecting advanced criteria that were incorrect, such as the wrong family. In these cases the top five search results were obviously not correct and the matching scores were relatively low, typically below 475. In the tests where an image of a species not in the reference library was submitted the "near accuracy' of the results depended on the pattern exhibited by the uploaded image. In the case of submitting a species with a hindwing eyespot the search provided a top five 'best' results that also showed hindwing eyespots with scores in the high 400's and low 500's. However, depending on the particular search

the correct genus was identified. These tests reinforced the need to provide images of at least the top five 'best' matches and the scores of all the species in the reference library that fit the search criteria.

HOW UTILIZED IN TEACHING, RESEARCH, and LEARNING

The development of our website did not mature until the winter term at OSU, thus, to date, it has not been used by any courses. Miller has already incorporated the program into his research on international insect biodiversity studies in Taiwan, Thailand, and South Korea. Also, the program has been introduced to students through classroom lectures and workshops in Thailand. We created a six-step instructional guideline for website use. These steps are also presented in an animated DEMO.

STEP 1: Submit An Image For Identification Assistance.
A) Prior to uploading your digital photograph of an unknown species edit the image to be: jpg; RGB or sRGB; not to exceed 400 pixels in width (this allows our program to automatically proportionately resize your image to fit our requirements for matching images with one another). *HINT: The image should show a flat wing surface that is more or less perpendicular to the plane of the camera lens. Also, during editing you should crop your image so the subject fills most of the frame.* If you are interested in using one of our "test" images left click on TEST IMAGES in the sidebar menu.
B) Please enter your name and email address. We ask for this information because every upload is tagged with a time stamp and we are conducting a study of user interests.
C) Left click on BROWSE to review your folders/files and select the image to be uploaded by highlighting the selected file and then left click on UPLOAD. Wait a few seconds for your image to appear.

STEP 2: Define The Image Boundary.
A) Left click and hold and drag the cropping tool to frame the subject in as tight a box as possible. *HINT: start the cropping drag at the upper left corner to get the box edges to just miss touching the front edge and outer edge of the wing; drag to the bottom right corner to enclose the specimen with little or no extraneous background showing.*
B) Click on SUBMIT to enter your defined image boundary.
C) You are now viewing your bounded image, if OK then click on SAVE or to go back and redo the boundary click on REDEFINE.

STEP 3: Define The Reference-Signature Mark.
A) Left click and hold to drag the cropping tool to enclose the area of the wing you wish to select as the reference mark that our program will use for image matching. *HINT: a smaller reference area will result in a faster search. Avoid crossing the line between the forewing and the hindwing and avoid referencing pixels outside the wing margins.*
B) Left click on SAVE REFERENCE MARK. You are now looking at the area of the wing to be submitted for matching your image with an image in our reference library.

C) To reject (repeat) your selection of the reference mark left click on REDO CROP; to accept your selection of reference mark left click on CONTINUE.

STEP 4: Choose Search Mode.
Either, left click on SIMPLE to search only our initial 34 images of Oregon butterflies. The program will automatically initiate the search. *HINT: this is slow and with advancements made later in the development of the website relatively incomplete.* The simple search will be eliminated after further development of the reference library for Oregon butterfly species. Or, left click on ADVANCED to use certain criteria (color, wing surface, group, Family, Genus) for a faster and more selective search.

STEP 5: Advanced Search Criteria.
Pick one attribute within each category A-E; 'not sure' is the default answer.
A) Wing surface. Choose among: 'not sure', 'dorsal', or 'ventral').
B) Identify the color. Based on the forewing (even if you selected a reference mark on the hindwing) choose the primary color considering the surface (dorsal or ventral) you selected in the preceding step. If you selected 'not sure' for the wing surface then select the color exhibited on the dorsum. *HINTS: white includes: crème, very light grey, and off-white; grey includes silver and light black; black includes: dark grey and charcoal; brown includes: tan and chocolate; blue includes: navy blue, purple, turquoise, and aqua; green includes: lime/lemon green; yellow includes: mustard and gold; orange includes: butterscotch; red includes: pink and maroon.* Because many species show a wide range in the degree of brown-grey; red-orange, black-blue, and so on, we have scored the species liberally. Many species have been assigned multiple color categories. So, if a wing looks black in the shadows but deep navy blue in the sun then we have scored the color both ways, black and blue. If you have difficulty determining the color to select, and your advanced search does not offer a satisfactory match to your specimen, then repeat the search based on a different color.
C) Group of Lepidoptera. Choose among: 'not sure, 'butterfly', or 'moth'.
D) Family name. Choose either 'not sure', or, if you are sure of (or want to guess) the family name then select that name within the drop-down list. Among the butterflies we have included danaines within the Nymphalidae but list satyrines as Satyridae. This is not a comment on classification but an aid to create a smaller search unit. Similarly, we have retained the traditional Family names for the macromoths.
E) Run search. You have completed the selection of advanced search criteria.
To run the search left click SEARCH LIBRARY.

STEP 6: Interpretation of Results.
A set of five images will appear as results of the search. The program selects and ranks the five best matches to the signature mark. We have designed the program to rank the images (top to bottom) based on a numeric value that represents the degree of matching of our library image to your submitted reference mark. Note that this is not a ranking of best identifications, but instead a value that measures the match in pattern, position, and color of your previously selected reference mark with an image in our library. The top ranked image will not necessarily provide the most likely species identification, although in many cases it

will. Please use this information with caution in your desire to put a name to the subject of your submitted image. Use the images in the search results to assist you in obtaining an identification based on your ability to see similarities in the butterfly or moth when viewed in its entirety. Furthermore, caution in accepting a name presented in the search results is required because: 1) the submitted reference marks are not necessarily diagnostic and unique to any one species, and 2) the resident library of images that are searched by our program is restricted to a small number of species of butterflies and moths. We do not have images of all of the known species in our library. Therefore, we expect the user to determine if the similarity between the submitted image and the search results provides a suitable match. *HINT: one way to test the search results is to conduct a second search using a different reference mark and alter any advanced criteria that were chosen with less than 100% confidence.*

OTHERS USING THIS PROGRAM

A related project, BugBytes (http://ipmnet.org/bugbytesnorthamerica/), is operated by Dr. Andy Moldenke (Oregon State University). Moldenke (partially funded by a prior NWACC grant) created a web-based insect identification system based on the matrix of characters rather than the typically used dichotomous key. Moldenke has been involved with the progress of our Signature Image project and has used the website. He will be involved in the future development of the project.

An OSU Ph.D. graduate student, Michael Liu, from the Department of Science and Mathematics Education plans to include our program in his research. Mr Liu recognizes that our program is a good learning tool for biological identification using morphological traits. When a high school student uses this web site to identify a butterfly or moth image, Michael will observe how the student chooses (i.e., crop) a morphological pattern from the wing. He will also record their responses when the students obtain the searching results and how they improve the following search from their previous experiences.

PAPERS AND PRESENTATIONS

Three invited seminars were presented by Miller where the 'Signature Image' project has been incorporated into the theme of insect biodiversity and taxonomy. Two of the seminars were presented in Thailand at Naresuan University, one in October, 2008, the other in December, 2008. The third seminar was presented to a formal discussion group during the annual meeting of the Pacific Northwest Lepidopterist Society held on the OSU campus in November, 2008. Seven additional seminars to be presented by Miller are currently scheduled for 2009 on the topic of insect biodiversity and taxonomic challenges for projects based on species assemblages exceeding 1,000 species: one seminar in Thailand (Naresuan University) and another in Taiwan (Taiwan Forest research Institute), with five presentations in South Korea (Incheon University; National Seoul University; South Korean Forestry Service; Mokpo University; Women's Normal University, Seoul). No manuscripts have

been published yet but we have a draft plan for at least one paper on the project to date for submission to the Lepidopterists Society of America.

ADDITIONAL FUNDING

One grant proposal is pending. A full-version proposal to the OSU-Agricultural Research Foundation was solicited after a campus-wide competition screening letter-of-intent. Ours is one of eight proposals for which a full proposal was requested. This two-year, $100,000 proposal was submitted on April 01, 2009, by Hans Luh, Jeff Miller, and Andrew Moldenke as the co-PI's. The basis of the proposal is automated identification of insect pests in the Pacific Northwest. The NWACC grant to Miller and Luh provided the background for the OSU-ARF grant competition.

We are also planning to submit a letter-of-intent to the National AFRI Program (USDA) for rapid and accurate insect identification based on our image recognition website.

FUTURE RESEARCH ACTIVITIES

We plan to continue with the development and application of this project through a variety of activities, including the aforementioned OSU-ARF and USDA proposals.
1) Incorporate the LepID program into undergraduate and graduate classes covering technological integration into biological studies, and insect biology/taxonomy.
2) Teach graduate students conducting biodiversity studies how the website functions and demonstrate how it can assist them in their studies.
3) Add a function to click on any one of the five results images and be directed to an original species diagnosis/biosketch page.
4) In addition to a signature mark referenced as a box (rectangle or square) provide a tool for free drawing and circular signature marks.
5) Increase the number of images in the reference library to represent more species and wider ranges of variation within a species.
6) Acquire a faster server to support searching a larger reference library.
7) Test user behavior for patterns in species inquiry, signature mark locations, and interpretation of results.
8) Incorporate remotely captured real-time images programmed to obtain identification through a completely automated system.
9) Change template matching scores from a number with a maximum value of 765 to a number that ranges between 1 and 100.
10) Transform the search program to operate in CYMK instead of or in addition to RGB.

PATENTS:
None

PUBLICATIONS AND PUBLIC RELATIONS:
None

POSITIONS FILLED:
None requested.

EXPENDITURES:

PI Salary		
	Jeffrey C. Miller	$ 3,099.20
	Hans Luh	$ 2,982.93
Support Materials		
	Laptop computer	$ 1,926.24
	Wireless Camera	$ 357.30
	External Hard Drives	$ 359.97
	OSU Bookstore	$ 99.00
Software		
	Catenary Systems	$ 899.00
	Center Space Software	$ 78.00
	NIK Software	$ 115.41
TOTAL		$ 9,917.05

Remaining Balance Returned to NWACC = $ 82.95

Expenditure justification:

Salary – the salary budget line was capped at $3,000 per investigator and in combination we spent well over 375 hours on this project.

Computer – we required a portable, wireless, new computer for work that took place at a variety of locations, including, the OSU insect museum, three different offices on the OSU campus, and at two homes.

External Hard Drives – needed for primary and backup storage for files of the original images that have been placed into the reference library and images that are in the process of being added to the reference library.

Camera – this item was required to advance the acquisition of images from a studio set-up to automated field acquisition of images.

OSU Bookstore – purchase of miscellaneous office supplies such as notebooks, paper, batteries for cameras and flash units, and photographic supplies.

Software – Catenary Systems provided the source code for the image search program; Center Space provided a mathematical program for scoring the matched images; NIK Software provided a digital graphic editing program used to produce the final version of the images placed into the reference library.