Pattern Recognition 51 (2016) 225–241

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

A software system for automated identification and retrieval of moth images based on wing attributes

Linan Feng [a], Bir Bhanu [a], John Heraty [b]

[a] Center for Research in Intelligent Systems, Bourns College of Engineering, University of California at Riverside, Riverside, CA 92521, USA
[b] Entomology Department, University of California at Riverside, CA 92521, USA

ARTICLE INFO

Article history:
Received 23 July 2013
Received in revised form
12 September 2015
Accepted 14 September 2015
Available online 28 September 2015

Keywords:
Entomological image identification and retrieval
Semantically related visual attributes
Attribute co-occurrence pattern detection

ABSTRACT

Manually collecting, identifying, archiving and retrieving specimen images is an expensive and time-consuming work for entomologists. There is a clear need to introduce fast systems integrated with modern image processing and analysis algorithms to accelerate the process. In this paper, we describe the development of an automated moth species identification and retrieval system (SPIR) using computer vision and pattern recognition techniques. The core of the system is a probabilistic model that infers Semantically Related Visual (SRV) attributes from low-level visual features of moth images in the training set, where moth wings are segmented into information-rich patches from which the local features are extracted, and the SRV attributes are provided by human experts as ground-truth. For the large amount of unlabeled test images in the database or added into the database later on, an automated identification process is evoked to translate the detected salient regions of low-level visual features on the moth wings into meaningful semantic SRV attributes. We further propose a novel network analysis based approach to explore and utilize the co-occurrence patterns of SRV attributes as contextual cues to improve individual attribute detection accuracy. Working with a small set of labeled training images, the approach constructs a network with nodes representing the SRV attributes and weighted edges denoting the co-occurrence correlation. A fast modularity maximization algorithm is proposed to detect the co-occurrence patterns as communities in the network. A random walk process working on the discovered co-occurrence patterns is applied to refine the individual attribute detection results. The effectiveness of the proposed approach is evaluated in automated moth identification and attribute-based image retrieval. In addition, a novel image descriptor called SRV attribute signature is introduced to record the visual and semantic properties of an image and is used to compare image similarity. Experiments are performed on an existing entomology database to illustrate the capabilities of our proposed system. We observed that the system performance is improved by the SRV attribute representation and their co-occurrence patterns.

1. Introduction

Moths are important life forms on the planet with approximately 160,000 species discovered [1], compared to 17,500 species of butterflies [1], which share the same insect order with Lepidoptera. Although most commonly seen moth species have dull wings (e.g., the Tomato Hornworm moth, see Fig.1(a), there are a great number of species that are known for their spectacular color and texture patterns on the wings (e.g., the Giant Silkworm moth and the Sunset moth, see Fig. 1b and c respectively). As a consequence, much research on identifying the moth species from the entomologist side has focused on manually analyzing the taxonomic attributes on the wings such as color patterns, texture sizes, spot shapes, etc., in contrast with the counterpart biological research that classifies species based on DNA differences.

As image acquisition technology advances and the cost of storage devices decreases, the number of specimen images in entomology is growing at an extremely rapid rate both in private data collections and over the web [2–4]. Many real world tasks such as monitoring insects for agriculture and border control are very important because they can contribute to the analysis of environmental and land security crisis including spread of pollution, disease vector and area biodiversity change. These real world applications involving insect species identification rely on manual processing of images by entomologists and highly trained experts which is a very time-consuming and error-prone process. The demand for more automated methods to meet the requirements of accuracy and speed is increasing. Given the lack of manually

E-mail addresses: fengl@cs.ucr.edu (L. Feng), bhanu@cris.ucr.edu (B. Bhanu), john.heraty@ucr.edu (J. Heraty).

http://dx.doi.org/10.1016/j.patcog.2015.09.012
0031-3203/

Fig. 1. Moth wings have color and texture patterns at different levels of complexity based on their species: (a) Tomato Hornworm, (b) Giant Silkworm and (c) Saliana Fusta.

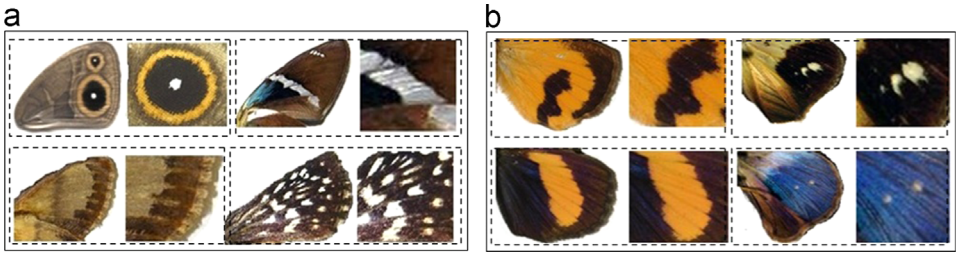

Fig. 2. Sample moth wings illustrate the semantically related visual (SRV) attributes. (a) Four sets of SRV attributes on the dorsal fore wings: eye spot (top left), central white band (top right), marginal cuticle (bottom left) and snowflake mosaic (bottom right). In each set, the right image is the enlarged version of the left image. (b) Four sets of SRV attributes on the ventral hind wings. Note that it is harder to describe the images in a semantic way with simple texts compared to the images in group (a).

annotated text descriptors to the images and the lack of consensus on the annotations caused by the subjectivity of the human experts, engines for archiving, searching and retrieving insect images in the databases based on keywords and textual metadata face great challenges.

The progress in computer vision and pattern recognition algorithms provides an effective alternative for identifying the insect species and many systems that incorporate these algorithms have been developed in the past two decades [5–9]. In the image retrieval domain, one of the common approaches introduced to complement the difficulties in text-based retrieval relies on the use of Content-Based Image Retrieval (CBIR) systems [10–12], where sample images are used as queries and compared with the database images based on visual content similarities [13,14] (color, texture, object shape, etc.). In both the identification and retrieval scenarios, visual features that are extracted to represent morphological and taxonomic information play an important role in the final performance. Context information is often used to help improve detection performance of the individual visual features [15].

These intelligent systems provide a number of attractive functions to entomologists, however, drawbacks have been revealed in several aspects:

- First, most systems only extract visual features that do not contain any *semantic* information. However, recent research [16] shows that human users want to access images at the *semantic* level. For example, users of a system are more likely to *find all the moths containing eye spots on the dorsal hind wings* rather than to *find all the moth containing a dark blue region near the bottom of the image*. An intermediate layer of image semantic descriptors that can reduce the gap between user information need and low-level visual features is absent in most existing systems.
- Second, most systems involve no human interaction and feedback. For example, the insect classification system introduced by Zhu et al. [17] works in an autonomous way on feature selection and classification. The retrieval systems [13,14,18] for butterfly images do no ask users to provide feedback and refine the results on the fly. However, the need for user-in-the-loop stems from the fact that intelligent systems are not smart enough to interpret images in the same way as humans. For example, two different species could be identified as the same based on their visual similarity. Without human intervention, the system will not be able to tune its parameters and correct the mistakes.
- Third, the current systems for species identification overlook the co-occurrence relationship among features. For example, in [5,19–21], the co-occurrence of features as contextual cues was not investigated to reduce or even remove the uncertainty in species identification. Intuitively, such information is helpful to better distinguish insect species. For example, in some species of Lepidoptera, a border “eye spot” feature may often be accompanied with a central “bands” feature on the wings, while other species may not have this combination of wing features. Such co-occurrence of features could be very useful to improve the performance of species identification.

In this paper, we present a new system for automated moth identification and retrieval based on the detection of visual attributes on the wings. The objective of our method is to mimic human behavior on differentiating species by looking at specific *visual contexts* on the wings. More specifically, the notion of “context” refers to discovering certain attribute relationships by taking into account their co-occurrence frequencies. The main motivation of our system relies on the conjecture that the attribute co-occurrence patterns encoded on different species can provide more information for refining the image descriptors. Unlike earlier work, we attempt to address all the above mentioned problems, and the contributions of this paper are summarized as follows:

1. We build image descriptors based on so-called *Semantically Related Visual (SRV) attributes*, which are the striking and stable physical traits on moth wings. Compared to a traditional visual feature used in many systems, our SRV attributes have human-designated names (e.g., blue preapical spot, white central bands, yellow eye spot, etc.) which makes them valuable as semantic cues. Some examples of SRV attributes are shown in Fig. 2. The probabilistic existence of these attributes can be discovered from images by trained detectors using computer vision and pattern recognition techniques. Compared to traditional image feature representations, which is usually a vector of numeric values denoting the visual properties, such as the curvature of a shape boundary, the color intensity of a region, etc., the SRV attribute based image descriptor provides a semantically rich

way which is much closer to the way that humans describe and understand images.

2. Our system detects and learns SRV attributes in a supervised way. The SRV-attributes are manually labeled by human experts on a small subset of the image database that is used for training the attribute detectors. The core of the detector is a probabilistic model that can infer SRV-attribute scores from the unlabeled test images. We characterize individual images by stacking the probabilistic scores of the present SRV attributes into a so-called *SRV-attribute signature*. The species identification and retrieval tasks are performed by comparing the SRV-attribute signatures. Specifically, in the image retrieval task, we incorporate a human relevance feedback scheme (often collected via user click-and-mark data) with the goal of retrieving more relevant images in future search sessions. We also consider ranking results based on the constraints of multi-attribute queries and the relative strengths of individual attributes to improve the effectiveness of attribute based image search.
3. We explicitly explore the co-occurrence relationship of SRV attributes. The underlying idea is that the attributes that appear together frequently across many images are likely to form a certain pattern. Moths from the same species often exhibit consistent patterns of SRV attributes on the wings. In this paper, we propose a novel approach that utilizes the external knowledge from human labeling in the training set to build a co-occurrence network of SRV attributes and further uncover the patterns of these attributes and use them as contextual cues to improve the individual attribute detection performance.

Our experimental evaluation shows that the proposed SRV attribute based image representation can improve moth species identification accuracy and image retrieval precision on different datasets. Experimental results also demonstrate that the proposed system can outperform state-of-the-art systems in the literature [14,22] in terms of effectiveness. We also evaluate other aspects of the proposed system (such as the impact of parameters) in the experiment section.

The remainder of this paper is organized as follows: Section 2 discusses related work. The technical approach is elaborated in Section 2.1. Experimental results are given in Section 2.2 followed by conclusions with future research directions in Section 2.3.

2. Related work

This section explains why automated systems are important for entomological research and how computer vision and pattern recognition techniques contribute to our understanding of images. In the following, we review approaches that are most relevant to our research along four directions: (i) Automated insect identification systems, (ii) Insect image retrieval systems, (iii) Visual words and attributes based image representations and (iv) Fusion techniques for image understanding.

2.1. Automated insect identification systems

Insect species identification has recently received great attention due to the urgent need for systems that can help in biodiversity monitoring [23], agriculture and border control [24,25], and conservation [26]. Likewise, identifying species is also the prerequisite to conducting advanced biological research on species evolution and developmental. However, the vast number of insect species and specimen images is a challenge for manual insect identification. The request for automated systems is only likely to grow in the future.

Several attempts have been made in the last two decades to design species identification systems for any type of available data. There have been sophisticated applications to solve problems in classifying orchard insects [5], recognizing the species-specific patterns on insect wings [6] and identification of insect morphologies on fossil images [7]. It has been recognized that these systems can overcome the manual processing time and errors caused by human subjectiveness.

Besides the above mentioned systems, there are other well-known systems: the SPecies IDentification Automated (SPIDA) system [9], the Digital Automated Identification SYstem (DAISY) [27], the Automated Bee Identification System (ABIS) [8] and DrawWing [28], a program for describing insect wings in a digital way. The first two systems use machine learning techniques such as neural network as the core of the classifier. SPIDA is designed for recognizing 121 spider species in Australia. The system keeps refining its learning accuracy as more user uploaded labeled images are available. DAISY is used not only for moth identification but also for any type of species identification, such as fish, pollen and plants. ABIS uses an idea similar to this paper based on finding attribute patterns from bee's wings to recognize their species. It utilizes an SVM-based discriminative classifier and the average performance reaches 95% in accuracy.

One common characteristic of these systems is that they all rely on images taken from carefully positioned target under controlled lighting conditions which reduces the difficulty of the task to some extent. One interesting aspect of automated species identification is that the data are not limited to images. For example, Ganchev et al. [29] describe the acoustic monitoring of singing insects and apply sound recognition technologies for insect identification tasks. Meulemeester et al. [30] report on the recognition of bumble bee species based on statistical analysis of the chemical scent extracted from the cephalic secretions. A challenge competition on multimedia life species identification [31] was recently held on identifying plant, bird and fish species using image, audio and video data.

The development of these systems have made great efforts in incorporating machine learning techniques like principal component analysis (PCA), linear discriminant analysis (LDA), artificial neural networks (ANNs), support vector machines (SVMs) and many other techniques.

2.2. Insect image retrieval systems

With the increase of insect images, there is a growing tendency in the field of entomology to archive, organize and find images in an efficient manner using image retrieval systems. Content-based image retrieval [32] has been well studied and developed for many years in the computer vision and information retrieval domains. It examines the contents of the image itself by extracting certain pictorial features and use them to compute similarity between a pair of images based on a metric automatically. Significant efforts have been made using content-based image retrieval techniques to find the relevant images to a query based on the visual similarity. The prototype systems for retrieving Lepidoptera images include “butterfly family retrieval” [18], a web-based system “Butterfly Ecology” [14] and a part based system [13]. Most of these systems focus on extracting low-level features such as color, shape and texture and the image representation allows these systems to compare images based on these features.

These systems are attractive but still present a number of problems. For example, a powerful function of CBIR is the ability to integrate user interaction where retrieval precision is adjusted according to the user provided relevance feedback (RF) information [33–36]. However, none of the existing systems has adopted the RF scheme into the retrieval framework. Also, a common limitation of the

available systems is that they only cope with a comparatively small number of species or categories in the dataset.

2.3. Image representation: visual words vs. semantic attributes

A crucial step for identification or classification and retrieval is to describe images by extracting a set of local feature descriptors, encoding them into high-dimensional vectors and then fusing them into image-level signatures. Many local descriptors are built upon low-level visual features like HOG (Histogram of Oriented Gradients) [37] and SIFT (Scale Invariant Feature Transforms) [38]. Recently, the computer vision community has found histograms of local features, also known as "*bag-of-visual-words*" [39–43] to be a powerful image representation [22,44–47]. The visual words paradigm usually contains three steps [48]: automatically selecting regions-of-interest, extracting visual features locally (reviewed in [48]), and vector-quantizing regional feature vectors into prototypes and using the histogram of prototypes as the image-level signature [39]. The prototypes or the visual words represent statistical information of repetitive image regions. A common extension to the approach is to adopt weighting schemes [39,41] on the visual words to distinguish their strengths.

The original visual words framework loses all the spatial relations of the regions in an image. However, the region locations could bring structural cues for classifying an image. Therefore, much recent work [39,40,49,50] considers the spatial distribution of regions that could possibly form or contain objects in an image. A further extension that includes the geometric information is to partition the images into a grid pattern. The image similarity is then computed from the sum of all the corresponding grid similarities. Lazebnik et al. [51] generalized this idea into spatial pyramid matching where the images are partitioned into a sequence of increasingly coarser grids and image similarity is calculated as a weighted sum over the matched grids at each level of resolution. Fisher vectors [52] were introduced as an alternative to aggregate local descriptors into a single global descriptor and this state-of-the-art vector has been demonstrated to be more effective than the bag-of-features representation for the same dimensionality [53].

Whereas the visual words model has been successfully used in many image understanding tasks, it has two major drawbacks. First, the performance of visual words model is highly dependent on the selection of vocabulary size. A local feature descriptor could have more than one neighboring visual word in a large vocabulary due to the information redundancy which causes the visual word "uncertainty" problem [43]. To overcome the visual vocabulary redundancy and over-completeness problem, in [**?**] a visual word pruning technique has been introduced to generate more meaningful visual words. Also a feature descriptor could be assigned to a visual word in a small vocabulary without a suitable candidate, which is known as the visual word "plausibility" [43]. Second, although the bag-of-visual-words model is analogous to the term-frequency model of text documents, visual words have limited semantic meanings. Therefore, it is hard for humans to contribute domain knowledge into the image understanding process.

Attribute-based representations have become a very promising direction in image classification [55,56] and visual recognition [57,58] due to their intuitive way in interpreting images and the cross-category generalization ability [59]. Unlike visual words, semantic attributes are sharable discriminative *visual properties* that are machine-detectable and human nameable (e.g., "square" as a shape property, "silk" as a texture property, "has wing" as a sub-component property, and "can fly" as a functionality property). One of the unique advantage of semantic attributes are that they naturally reduce the gap between low-level visual features and high-level concepts. In other words, semantic attributes can be used to answer not only "how" two images are similar in a human interpretable way [60], but also "why" an image is identified to belong to a specific category [61].

Attributes are also used frequently in the multimedia retrieval community as an intermediate level semantic description [62–64]. In our moth image retrieval system, the user's search intention does not simply emphasize appearance similarity between the query and the database images, but more the semantic closeness (e.g., the species category). This implies that the retrieved images should contain similar semantic attributes. In this paper, we represent images by using the proposed SRV attribute signature and compute the distance between images based on this new image representation.

2.4. Fusion approaches for image understanding and analysis

Multimodal fusion has been used by many researchers for various multimedia analysis tasks. Multimodal fusion refers to the integration of multimedia, associated features and other intermediate results to perform the decision making process in multimedia analysis and understanding [65]. The fusion of these multimodal data can provide extra knowledge of the multimedia content and improve the accuracy of the overall system.

Many techniques have been introduced to effectively fuse multimodal data, e.g., linear weighting, or using a weighted strategy for evaluating feature scores [66]. Visual and semantic information are fused in many approaches dealing with the ImageCLEF dataset [67] and the results show that fusion based approaches outperform the single modality approaches. An overview of the state-of-the-art fusion approaches is given by Atrey et al. in [65]. In our system, we apply late fusion of the visual feature vector and SRV attribute signature in a weighted manner, and this provides more flexibility than fusion at an early stage.

3. Technical approach

3.1. Moth image dataset

The dataset used in this study is collected from an online library of moth, butterfly and caterpillar specimen images created by Dr. Dan Janzen [68] over a long-term and ongoing project started in 1977 in northwestern Costa Rica. The goal of the inventory is to have records for all the $12,500+$ species in the area. As of the end of 2009, the project had collected images of 4500 species of moths, butterflies and caterpillars. We use a subset of the adult moth images from the 2009 collection with the permission of Dr. Dan Janzen. The dataset is publicly available at http://janzen.sas.upenn.edu.

The images are available for both the dorsal and ventral aspects of the moths. Each image was resized into 600×400 pixels in resolution, and is in RGB colors. Our complete dataset contains 37,310 specimen images covering 1580 species of moth, but a majority of the species have less than twenty samples. Because our feature and attribute analysis are based on regions on the wings, and some specimens show typical damage ranging from age-dependent loss of wing scales (color distortion), missing parts of wings (incomplete image), or uninformative orientation differences in the wings or antennae, this makes the number of qualified samples even less, and we have carefully selected fifty species across three family groups and six sub-family groups: *Hesperiidae* (*Hesperinae*, *Pyrginae*), *Notodontidae* (*Dioptinae*, *Nystaleinae*) and *Noctuidae* (*Catolacinae*, *Heterocampinae* [=*Rifargiriinae*]) from the original dataset. This new sub-collection has a total of 4530 specimens of good quality (see Table 1 for the distribution of the species used in our work).

Table 1
Families, species and the number of samples in each species used in our work.

Sub-families	Species	Images	Sub-families	Species	Images
Catolacinae	*Ceroctenaamynta*	101	*Nystaleinae*	*Bardaximaperses*	74
Catolacinae	*Eudocimamaterna*	85	*Nystaleinae*	*Dasylophiabasitincta*	78
Catolacinae	*Eulepidotisfolium*	76	*Nystaleinae*	*Dasylophiamaxtla*	98
Catolacinae	*Eulepidotisrectimargo*	57	*Nystaleinae*	*Nystaleacollaris*	85
Catolacinae	*Hemicephalisagenoria*	121	*Nystaleinae*	*Tachudadiscreta*	112
Catolacinae	*Thysaniazenobia*	79	*Pyrginae*	*Atarnessallei*	101
Dioptinae	*Chrysoglossanorburyi*	75	*Pyrginae*	*Dyscophellusphraxanor*	86
Dioptinae	*Erbessaalbilinea*	98	*Pyrginae*	*Tithraustesnoctiluces*	96
Dioptinae	*Erbessasalvini*	117	*Pyrginae*	*Entheusmatho*	99
Dioptinae	*Nebulosaerymas*	69	*Pyrginae*	*Hyalothyrusneleus*	82
Dioptinae	*Tithrausteslambertae*	87	*Pyrginae*	*NascusBurns*	94
Dioptinae	*Polypoetesharuspex*	92	*Pyrginae*	*Phocidesnigrescens*	104
Dioptinae	*Dioptislongipennis*	92	*Pyrginae*	*Quadruscontubernalis*	69
Hesperiinae	*Methionopsisina*	122	*Pyrginae*	*Urbanusbelli*	88
Hesperiinae	*Neoxeniadesluda*	107	*Pyrginae*	*MelanopygeBurns*	76
Hesperiinae	*SalianaBurns*	70	*Pyrginae*	*Myscelusbelti*	103
Hesperiinae	*Salianafusta*	97	*Pyrginae*	*Mysoriaambigua*	93
Hesperiinae	*TalidesBurns*	70	*Rifargiriinae*	*Dicentriarustica*	78
Hesperiinae	*Vettiusconka*	96	*Rifargiriinae*	*Farigiasagana*	84
Hesperiinae	*Aromaaroma*	135	*Rifargiriinae*	*Hapigiodessigifredomarini*	93
Hesperiinae	*Carystoidesescalantei*	88	*Rifargiriinae*	*Malocampamatralis*	100
Nystaleinae	*Lirimirisguatemalensis*	95	*Rifargiriinae*	*MeragisaJanzen*	65
Nystaleinae	*Isostylazetila*	99	*Rifargiriinae*	*Naprepahoula*	74
Nystaleinae	*Oriciadomina*	101	*Rifargiriinae*	*Pseudodryaspistacina*	83
Nystaleinae	*Scoturaleucophleps*	117	*Rifargiriinae*	*Rifargiadissepta*	69

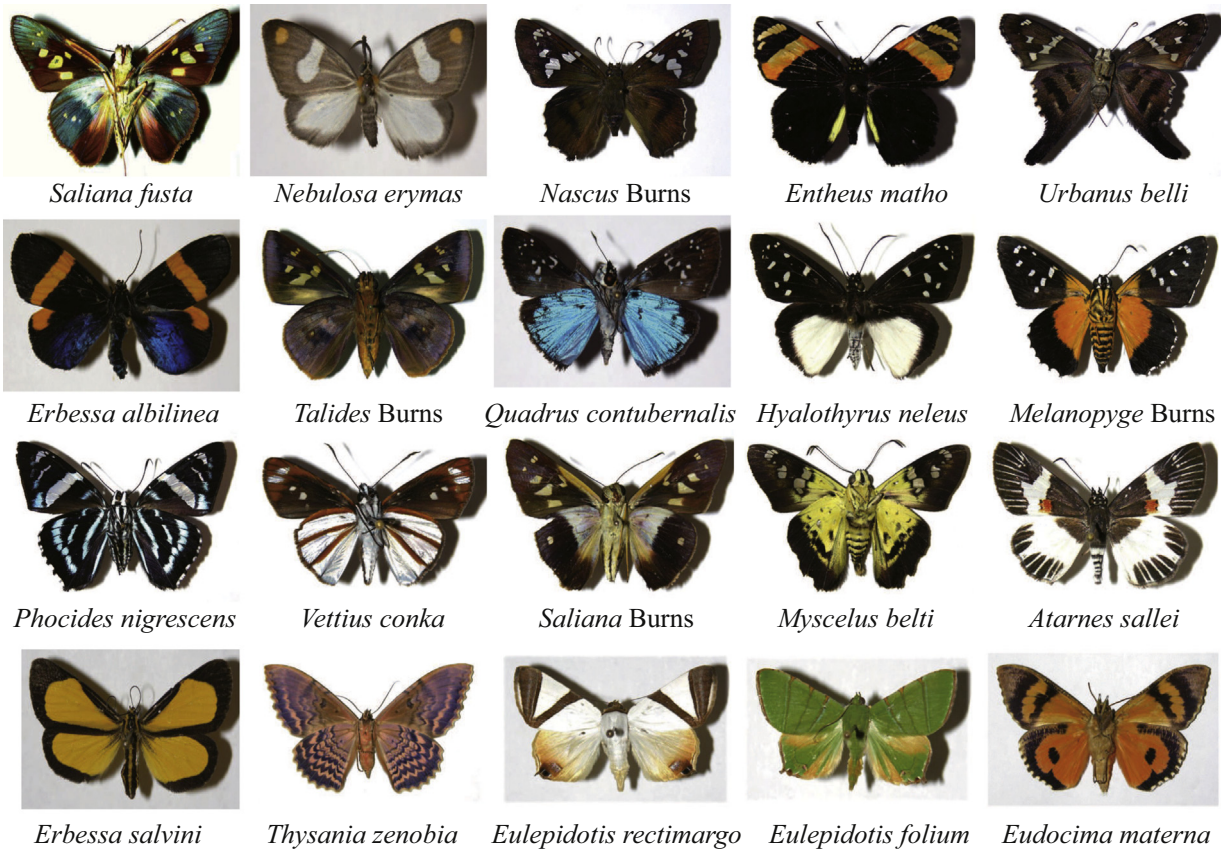

Fig. 3. Sample images for twenty moth species selected from all the species used in this work. We do not show all the species due to space limitations.

We show sample images of twenty representative species out of the fifty species used in our work in Fig. 3. The moth specimens were photographed against an approximately uniform (usually white or gray) background, but often with shadow artifacts. The specimens are curated in a uniform way with the wings horizontal and generally with the hind margin of the forewing roughly perpendicular to the longitudinal axis, which facilities the subsequent image processing and feature extraction steps.

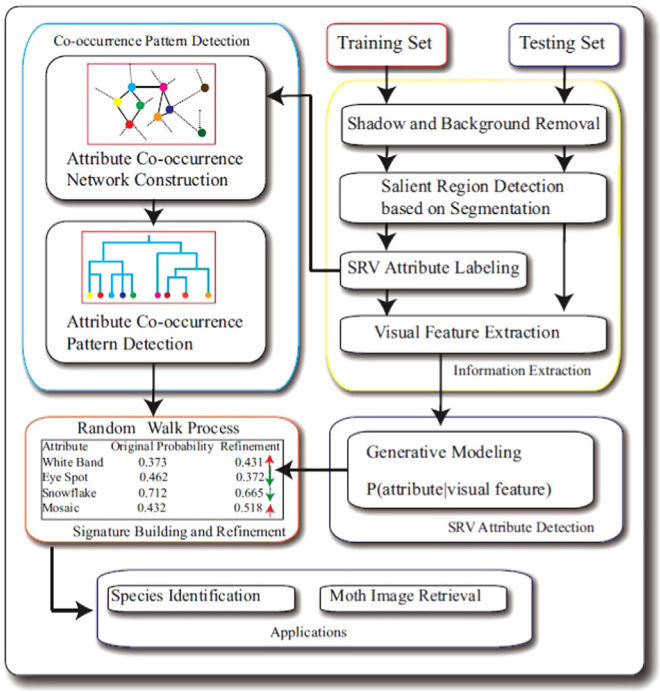

Fig. 4. The flowchart of the proposed moth species identification and retrieval system. It consists of (1) information extraction, (2) SRV attribute detection, (3) attribute co-occurrence pattern detection, (4) signature building and refinement, and (5) moth identification and retrieval applications.

3.2. System architecture

The flowchart of the proposed moth identification and retrieval system is shown in Fig. 4. The system architecture contains five major parts: (1) information extraction of moth images, (2) SRV attribute detection on moth wings, (3) co-occurrence network construction and co-occurrence pattern detection for the SRV attributes, (4) image signature building and refinement based on SRV attributes and their co-occurrence patterns, and finally (5) applications in moth species identification and retrieval. We give the details about each part in the following sections.

The information extraction module consists of several steps including background and shadow removal, salient region detection by segmentation, SRV attribute labeling for the training set and visual feature extraction.

In order to train the attribute detectors, we use a small subset of the image collection as the training set. Each training image is segmented manually into regions and the attributes labeled manually to the corresponding regions. The SRV attribute detector is learned from extracted local visual features and the SRV attribute labels by modeling the joint probability of occurrence. After the joint distribution is obtained, we infer the posterior probabilities of attributes from the visual features of the test images without attribute labeling. The output of the detectors is a pool of the posterior probability scores of each attribute. These are combined into the attribute signature representation of the images.

As the attribute detection relies on the effectiveness of the low-level features to some extent, and in order to improve the detection accuracy by narrowing the semantic gap, we propose a novel approach to explore the contextual information of the attributes. Specifically, the co-occurrence pattern recognition module is aimed at uncovering the explicit co-occurrence relationship between attributes in images and utilizing it to further improve the individual attribute detection performance. A random walk process is integrated in this module to maximize the agreement on appearance of individual attributes in an image with respect to co-occurrence.

Relevance feedback is a crucial strategy in image retrieval systems for retrieval result refinement. In our system, we provide the application interface with functions like marking the relevance decisions on the retrieved images. However, as the users of the system may have different levels of professional knowledge, we evaluate their expertise by requiring them to participate in a sample species identification test and authorizing them different levels of permissions to submit feedback based on their scores. The following sections provide the implementation details of each part shown in Fig. 4.

3.3. Feature extraction

3.3.1. Background removal

It is important to partition the images into “background” and “foreground” because the background usually contains disturbing visual information (such as shadows created by the lighting device, bubbles and dirt on the specimen holder, etc.) that can affect the performance of the detector. We adopted the image reflection symmetry based approach [69] for background and shadow removal. The moth image dataset used in this paper has the high reflection symmetry property of moth wings (true for all moths in general) (Fig. 5(a)). Because the shadows have the most salient influence on the following processing steps, and they are not symmetric in the images, we use symmetry as the key constraint to remove shadows.

The SIFT points of the image are detected (Fig. 5(b)) and symmetric pairs of the points are used to vote for a dominant axis of symmetry (Fig. 5(c)). Based on the axis, a symmetry-integrated region growing segmentation scheme is used to remove the white background from the moth body and shadows (Fig. 5(d)), and the same segmentation process is run with smaller thresholds to partition the image into shadows and small local parts of the moth body (Fig. 5(e)). Finally, symmetry is used again to separate the shadows from the moth body by computing a symmetry affinity matrix. Since the shadows are always asymmetric with the axis of reflection, their symmetry affinity will have higher values than the parts of moth body. This is used as the criterion to remove the shadows (Fig. 5(f)).

3.3.2. SRV attribute labeling

A sub-region of the moth wing is considered an SRV attribute if: (1) it repeatedly appears on moth wings across many images, (2) it has salient and unique visual properties and (3) it can be described by a set of textual words that are descriptive for the sub-region.

We scan the moth images and manually pick a group of SRV attributes. Similar methods have been utilized for designing “concepts” or “semantic attributes” in image classification and object recognition tasks. An example is building nameable and discriminative attributes with a human-in-the-loop [70,71]. However, as compared to their semantic attributes, our SRV attributes cannot be described with concise semantic terms (e.g., “A region with scattered white dots on the margin of the hind wing on the dorsal side”). Therefore, we propose to index the SRV attributes by numbers, e.g., “attribute_1”, “attribute_2” and so forth. We also explicitly incorporate the positions of the SRV attributes into the attribute index. Each moth has two types of wings: the forewing and the hindwing, and each type of wing has two views: the ventral view and the dorsal view, the SRV attribute index is finally defined in an unified format “attribute_No./wing_type/view”, e.g., “attribute_1/forewing/dorsal”, “attribute_5/hindwing/ventral”, etc. Furthermore, as the moths are symmetrical to the center axis (axis of symmetry), we only label one side of the moth with the index of SRV attributes.

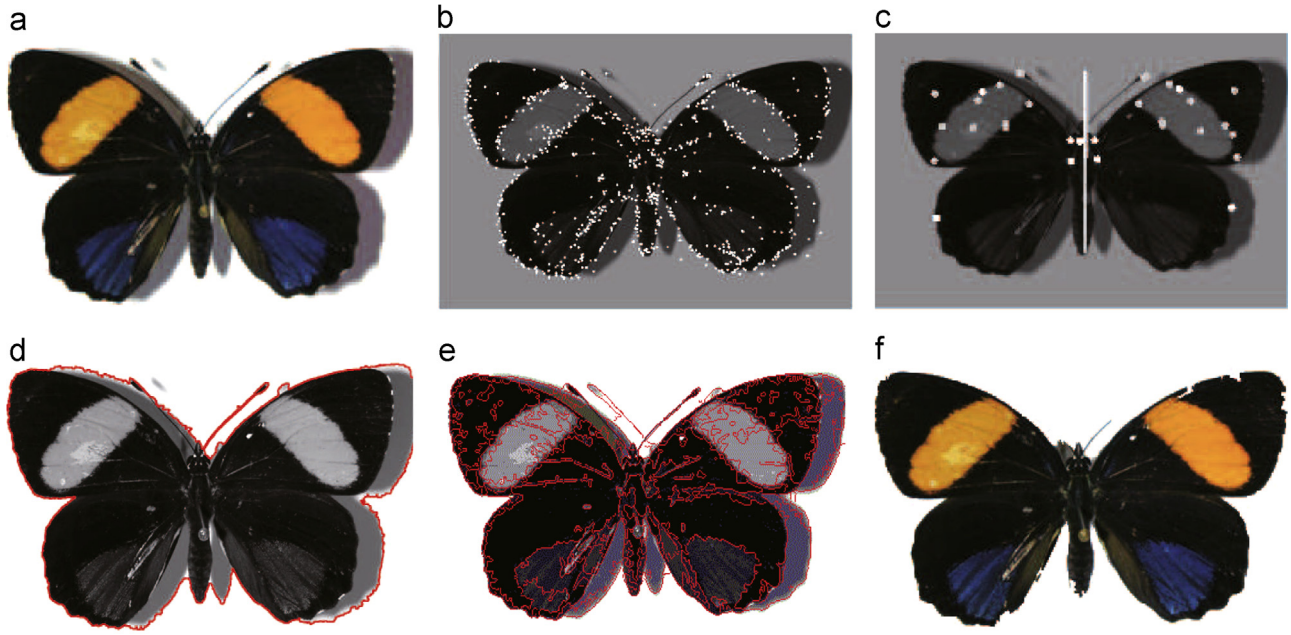

Fig. 5. Steps for background and shadow removal. (a) Original image (with shadow), (b) detected SIFT points, (c) detected symmetry axis, (d) background removed image, (e) segmentation for small parts, and (f) image after shadow removal.

Fig. 6. Results from salient region detection. (a) symmetry based segmentation, (b) segmentation without using symmetry. Two more results are shown in (c) and (d) by using symmetry based segmentation.

In order to acquire reliable attribute detectors, SRV attributes are labeled by human experts to the regions in the training images. The regions are represented by the minimum bounding rectangles (MBRs) which are produced by using the on-line open source image labeling tool "LabelMe" [72].

3.3.3. *Salient region detection by segmentation*

For the test images, we use the *salient region detector* to extract small regions or patches of various shapes that could potentially contain the interesting SRV attributes. A good region detector should produce patches that capture salient discriminative visual patterns in images. In this work, we apply a hierarchical segmentation approach based on reflection symmetry introduced in [69] to jointly segment the images and detect salient regions.

We apply symmetry axis detection on moth images to compute a symmetry affinity matrix, which represents the correlation between the original image and the symmetrically reflected image. Each pixel has a continuous symmetry affinity value between 0 (perfectly symmetric) and 1 (totally asymmetric), which is computed by the Curvature of Gradient Vector Flow (CGVF) [73]. The symmetry affinity matrix of each image is further used as the symmetry cue to improve the region-growing segmentation. The original region-growing approach considers aggregating pixels into regions by pixel homogeneity. In this paper, we modified the aggregation criterion to integrate the symmetry cue. More details about the approach are explained in [69].

Comparison between Figs. 6(a) and (b) indicates that by using symmetry, more complete and coherent regions are partitioned. The result in Fig. 6(b) is obtained by using the same region

growing, but without symmetry, so it has many noisy and incomplete regions. The improvements are obtained by using the symmetry cue only. Two more results on salient region detection by using symmetry based segmentation are shown in Figs. 6 (c) and (d).

3.3.4. Low-level feature extraction

We represent the above detected salient regions by the minimum bounding rectangles (MBRs). The local features of each bounding rectangle are extracted and pooled into numeric vector descriptors. We have three different types of features used to describe each region: (a) color feature, (b) texture feature, and (c) SIFT keypoint feature.

(1) *HSV* (*Hue-Saturation-Value*) *color feature*: The color feature is insensitive to changes of size and direction of regions. However, it suffers from the influence of illumination variations. For the color feature extraction, the original RGB (Red-Green-Blue) color image is first transformed into HSV (Hue-Saturation-Value) space, and only the hue and saturation components are used to reduce the impact from lighting conditions. We then divide the interval of each component into 36 bins, the image pixels inside the salient region are counted for each bin, and the histogram of the 72 bins is concatenated and normalized into the final color feature vector.
(2) *Gray Level Co-occurrence Matrix* (*GLCM*) *based texture feature*: Texture features are useful to capture the regular patterns of the spatial arrangement of pixels and the intrinsic visual property of regions. We adopt the gray level co-occurrence matrix (GLCM) proposed by Haralick in [74] to extract the texture features. The GLCM is a pixel-based image processing method.

The co-occurrence matrices in GLCM are calculated based on second order statistics as described in [75]. Each element $P(i,j,d,\varphi)$ in the matrix represents the frequency of co-occurrence of the gray levels of the pixel pair (i, j) along a specific direction φ (e.g., horizontal, diagonal, vertical, etc.) at a distance d (e.g., one to six pixels) between the pixels.

Let $I(x,y)$ denote a two-dimensional digital image of size $M \times N$, and suppose the maximum gray level is G, hence $i,j \in [0,G]$, an element in the GLCM representing the co-occurrence value of two pixels $(x_1,y_1),(x_2,y_2)$ in the image I at angle φ and distance d is expressed in the following equation:

$$P(i,j,d,\varphi) = \sum_{d,\varphi} \Delta[(x_1,y_1),(x_2,y_2)] \tag{1}$$

where $\Delta = 1$, if $(x_1,y_1) = i$ and $(x_2,y_2) = j$, else $\Delta = 0$. In the original approach, the author [74] computed 14 statistical features from the matrix. We use 256 gray levels for quantization. The resulting GLCMs can be sparse, and computing statistics looping through each of the GLCMs can result in a very inefficient procedure since many of the matrix entries are zeros. We use a subset of patches containing the SRV attributes with ground-truth labels. The 14 GLCM features are extracted for each patch. We conduct a classification task for each patch using each of the features. The best features that have higher discriminative power and lower computation time for all the patches are selected (by plotting the error rate vs. computation time and selecting the optimum point located within a certain radius range to the origin where the error is low and the computation time is also low). This results in the four most effective and efficient features listed below:

- Energy (Angular Second Moment):

$$ASM = \sum_i \sum_j P(i,j)^2 \tag{2}$$

- Energy measures the image gray-level distribution and the texture uniformity. *ASM* is relatively large when the distribution of $P(i,j)$ is more concentrated on the main diagonal.
- Entropy:

$$ENT = -\sum_i \sum_j P(i,j) \log P(i,j) \tag{3}$$

- Entropy measures the disorder of an image. ENT is larger when the value of $P(i,j)$ is more dispersed and it achieves its largest value when all the $P(i,j)$ s are equal.
- Correlation:

$$COR = \frac{\sum_i \sum_j (ij) P(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{4}$$

- Correlation measures the gray tone linear dependencies in an image. $\mu_x, \mu_y, \sigma_x, \sigma_y$ are the means and standard deviations of $P_x(i) = \sum_j P(i,j)$ and $P_y(j) = \sum_i P(i,j)$.
- Homogeneity (Inverse Difference Moment):

$$IDM = \sum_i \sum_j \frac{1}{1+(i+j)^2} P(i,j) \tag{5}$$

Homogeneity is inversely proportional to the image contrast feature at constant energy. Smaller gray tone difference in pair elements contribute to larger value of homogeneity.

By using only four components of the GLCM feature instead of the entire fourteen components, the trained attribute detection model achieved comparatively the same performance in image identification while saved a lot of computation time. This has been demonstrated on a subset of the training image patches for evaluation. We set the distance between the pair of pixels at 4 scales (1, 2, 4, 8) and set the directions at 4 angles (0°, 45°, 90°, 135°). These scale and orientation parameters were examined as the most appropriate setting by applying Chi-square test on the optimal GLCM computed with the selected four features of the training patches. The final GLCM texture feature vector is of length 64 (4 feature types $\times$ 4 direction $\times$ 4 distances).

(3) SIFT (Scale Invariant Feature Transform) based keypoint feature. SIFT [38], proposed by Lowe, is a very popular feature used in computer vision and pattern analysis. SIFT features have the advantage that they are invariant to changes in scale, rotation, and intensity. The major issues related to extracting SIFT features include selecting the keypoints and calculating the gradient histogram of pixels in a neighboring rectangular region. In this work, we apply the Difference-of-Gaussians (DoG) operator to extract the keypoints. For each keypoint, the 16×16 pixels in its neighborhood region are used. We divide a region into 16 4×4 subregions. For each pixel in a subregion, we calculate the direction and magnitude of its gradient. We quantize the directions into 8 bins, and build a histogram of gradient directions for each subregion. The magnitude of the gradient is used to weight the contribution of a pixel. Finally, the 8-dimensional feature vectors from the eight-bin direction histogram of each subregion are combined and weighted into a 128-dimensional vector to record local information around the keypoint.

3.4. SRV attribute detector learning module

In this module, the SRV attribute detector is trained by using a generative approach based on probability theory. To illustrate the basic idea, consider a scenario in which an image region depicted by an N-dimensional low-level feature vector $\overrightarrow{X^N}$ is to be assigned into one of the K SRV attributes $k=1,\ldots,K$ at a higher level of semantics. From probability theory we know that the best solution is to achieve the *a posterior probabilities* $p(k|X)$ for a given X and each attribute category k, and assign the attribute with the largest probability score to the region. In the generative model, we model the joint probability distribution $p(k,X)$ of image region features and attributes, and Bayes'theorem provides an alternative to derive $p(k|X)$ from $p(k,X)$:

$$p(k|X)=\frac{p(k,X)}{p(X)}=\frac{p(X|k)p(k)}{\sum_{i=1}^{K}p(X|i)p(i)} \tag{6}$$

As the sum in the denominator takes the same value for all the attribute categories, it can be viewed as a normalization factor over all the attributes. Eq. (6) can be rewritten as

$$p(k|X)\propto p(k,X)=p(X|k)p(k) \tag{7}$$

which means we only need to estimate the attribute prior probabilities $p(k)$ and the likelihood $p(X|k)$ separately. The generative model has the advantage that it can augment the large amount of unlabeled data in a dataset from a small portion of the labeled data.

As defined earlier K denotes the pool of SRV attributes. Let k_i be the ith attribute in K. According to the previous section, k_i is assigned to a set of image regions $R_{k_i}=\{r^i_1,r^i_2,\ldots,r^i_{n_{k_i}}\}$ along with the corresponding feature vectors $X_{k_i}=\{x^i_1,x^i_2,\ldots,x^i_{n_{k_i}}\}$, where n is the number of regions in an image. We assume that the feature vectors are sampled from an underlying multi-variate density function $p_X(\cdot|k_i)$. We use a non-parametric kernel-based density estimate [76] for the distribution p_X. Assuming region r_t to be in the test image with feature vector x_t, we estimate $p_X(x_t|k_i)$ by using a Gaussian kernel over the feature vectors X_{k_i}:

$$p_X(x_t|k_i)=\frac{1}{n}\sum_{j=1}^{n}\frac{\exp\{-(x_t-x_j)^T\Sigma^{-1}(x_t-x_j)\}}{\sqrt{2^n\pi^n|\Sigma|}} \tag{8}$$

Σ is the covariance matrix of the feature vectors in X_{k_i}.

$p(k_i)$ is estimated by using Bayes estimators with a prior beta distribution, the probability distribution of $p(k_i)$ is given by

$$p(k_i)=\frac{\mu\delta_{k_i,r}+N_{k_i}}{\mu+N_r} \tag{9}$$

where μ is the smoothing parameter estimated from the training set, $\delta_{k_i,r}=1$ if attribute k_i occurs in the training region r and 0 otherwise. N_{k_i} is the number of training regions that contain attribute k_i and N_r is the total number of training regions.

Finally, for each test region with feature vector x_t, the *posterior probability* of observing attribute k_i in K given x_t, $p(k_i|x_t)$ is given by multiplying the estimates of the two distributions:

$$p(k_i|x_t)=\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\exp\{-(x_t-x_j)^T\Sigma^{-1}(x_t-x_j)\}}{\sqrt{2^n\pi^n|\Sigma|}}\right)\times\left(\frac{\mu\delta_{k_i,r}+N_{k_i}}{\mu+N_r}\right) \tag{10}$$

For each salient region extracted from a test image I, the frequency of each attribute in that region is inferred by (10). The probabilities for all attributes are combined into a single vector which is called *region SRV attribute signature*. For a test image with several salient regions, we combine the region SRV attribute signature into a final vector by choosing the max score for each attribute. We name this vector as the *image SRV attribute signature* and it is used as the semantic descriptor for an image.

3.5. SRV attribute co-occurrence pattern detection module

Attribute labels given by human experts as ground-truth semantic descriptions across the entire training image set are used to learn the contextual information based on the attribute label co-occurrences. In this section, we devise a novel approach to discover the co-occurrence patterns of the individual attributes based on network analysis theories. More specifically, we construct an attribute co-occurrence network to record all the pairwise co-occurrence between attributes. The patterns are detected as the communities in a network structure. A similar concept is used in social networks to describe a group of people that have tightly established interpersonal relationships.

3.5.1. SRV attribute co-occurrence pattern detection

We first introduce the notion of community structure from the network perspective. One way to understand and analyze the correlations among individual items is to represent them in a graphical network. The nodes in the network correspond to the individual items (attributes in our case), the edges describe the relationships (attribute co-occurrence in our case), and the edge weights denote the relevant importance of the relationship (co-occurrence frequency in our case).

A very common property of a complex network is known as the community structure, i.e., groups of nodes may have tight internal connections in terms of a large number of internal edges, while they may have fewer edges connecting each other. These groups of nodes constitute the communities in the network. The existence of community structure reflects underlying dependencies among elements in the target domain. If a group of individual attributes always occurs together in the training image set, then an underlying co-occurrence pattern can be defined by these attributes, and this pattern can be used as a priori knowledge in the attribute detection for the test images.

The approach we adopted to detect the communities in the network is modularity optimization [77]. Suppose attributes a_i and a_j in A are represented as two nodes i and j, and suppose i belongs to community C_i and j belongs to community C_j in a partition. The modularity Q is defined as a qualitative measure of a particular partition on the network in the form of

$$Q=\frac{1}{2d}\sum_{i,j}\left[w_{ij}-\frac{w_iw_j}{2d}\right]\delta(C_i,C_j) \tag{11}$$

where d equals to half of the summation of all the edge weights in the network, w_{ij} is the edge weight between i and j, $w_i(w_j)$ equals the summation of the edge weights attached to node $i(j)$, $\delta(C_i,C_j)=1$ if $C_i=C_j$ and 0 otherwise.

We consider iteratively merging the nodes into communities based on the criterion that the merge of nodes generates a positive modularity gain at each iteration. The modularity gain of moving an outside node i into a community C is evaluated by

$$\Delta Q=\left[\frac{\Sigma_{in}+k_{i,C}}{2d}-\left(\frac{\Sigma_{out}+w_{i,C}}{2d}\right)^2\right]-\left[\frac{\Sigma_{in}}{2d}-\left(\frac{\Sigma_{out}}{2d}\right)^2-\left(\frac{w_{i,C}}{2d}\right)^2\right] \tag{12}$$

where Σ_{in} represents the sum of edge weights inside C, $w_{i,C}$ equals the sum of weights of edges that link i to C, d is the same as defined in Eq. (20), Σ_{out} is the sum of weights of edges that link outside nodes to nodes in C, w_i is the sum of weights of the edges incident to i. Based on modularity optimization, we propose the following two phase algorithm, given as Algorithm 1 in the

following, to detect the attribute co-occurrence patterns in the network.

Algorithm 1. SRV attribute co-occurrence pattern detection

Input: SRV attribute co-occurrence network.

Output: Hierarchical SRV attribute co-occurrence patterns.

Partitioning phase:
do
  Assign each node a different community tag $C_i, i = 1, ..., N$;
  foreach *node i in Community* C_i **do**
    Remove i from its original community C_i;
    Add i into each of its neighboring nodes j's community C_j;
    if $\Delta Q > 0$ *computed by* (12) *from placing* i *to* C_j **then**
      Examine the value of Q_{C_i} and Q_{c_j} with i assigned to each neighboring community by (11);
      if $Q_{C_i} \geq 0.3$ && $Q_{c_j} \geq 0.3$ **then**
        Attribute i is shared by the two communities C_i and C_j;
        Split i into i and i', put them into C_i and C_j;
        Copy the edges of i incident to other nodes for i';
      else
        Place i into C_j;
    else
      No node will be moved;
while *Every node has been traversed && no increase can be achieved for* ΔQ;
Coarsening phase:
foreach *Existing community* C_i **do**
  Replace the entire community C_i by a single node i in the network;
  Replace the edges between community C_i and its neighboring communities by single edges;
  Compute the weight for a single edge as the sum of old edge weights;
  Represent internal edges as a self-looped edge with weight equals the sum of internal edge weights;
Iteration: Repeat $1 \rightarrow 23$ until no positive ΔQ can be achieved;

3.5.2. SRV attribute signature refinement with the co-occurrence patterns

The co-occurrence patterns are utilized for refining the detection results on each individual SRV attribute by performing a random walk process [78] over the patterns. We define the distance between two attributes a_i and a_j as

$$D_{a_i,a_j} = \frac{2 \times \#\text{ of } CP\{a_i, a_j\}}{\#\text{ of } CP\{a_i\} + \#\text{ of } CP\{a_j\}} \tag{13}$$

where $\#\text{ of } CP\{a_i, a_j\}$ is the number of co-occurrence patterns containing both attribute a_i and a_j. Suppose initially the frequency of attribute a_i in the *image attribute signature* is $s(a_i)$ (given by the generative model), then in the mth iteration the new value of the probability is formulated by the following random walk process:

$$s_m(a_i) = \alpha \sum_j s_{m-1}(a_j) \cdot D_{a_i,a_j} + (1-\alpha) \cdot s(a_i) \tag{14}$$

where α is a weight parameter that takes a value between $(0, 1)$. The above formula can strengthen the occurrence probabilities of the attributes in the same patterns and weaken the isolated ones. The controlling parameter is determined by using the training sets.

3.6. Identification module

The attribute detector learned from the training data is used in the identification module for the test images. The inputs to the detector are the detected salient regions from the test images as well as the extracted low-level visual features. The output of the detector is the so-called "image SRV attribute signature". The species identification of test images is performed by comparing test image signatures with the training image signatures. Therefore, we also build the attribute signatures for the training images. For a training image I, the attribute signature is $S^{|A|}$ with each element $s(a_i) \in \{0, 1\}$ and $s(a_i) = 1$ when image I has regions labeled with attribute a_i and $= 0$ otherwise. We further divide the training images into groups based on their scientific species designation. The element values are averaged across the signatures within each species group for each individual attribute and the obtained signature is called the *species prototype signature*.

The test image of a species is identified by comparing its image attribute signature with the species prototype signatures of the

fifty species. The distance between the two signatures is calculated by the Euclidean distance. The test image is finally identified as the species with the smallest distance value. If several species have very similar distances to the test image, we assign all the species labels to that image, and let the image retrieval system give the final decision on the species based on the feedback from the users who are experts.

3.7. Retrieval & relevance feedback module

We implement a query-by-example (QBE) paradigm for our retrieval system. QBE is widely used in conventional content-based image retrieval (CBIR) systems when the image meta-data, such as captions, surrounding texts, etc. are not available for keyword based retrieval.

3.7.1. Image retrieval using query-by-example

In the QBE mode, the user is required to submit a query in terms of an example specimen image to the system. Finding an appropriate query example, however, is still a challenging problem in the research area of CBIR [35]. In our system, we provide an image browsing function in the user interface, and the user is allowed to browse all the images in the database and submit a query. Images are compared by their content similarity. Each image in the database is represented by a low-level visual feature vector F and a high-level SRV attribute signature S, for a query image Q and a database image Y. The distance between them is calculated by fusing the Euclidean distance over the visual feature vectors and the Earth Mover's distance [79] over the SRV attribute signatures:

$$Dist(Q,Y)=\eta D_{Euc}(F_Q,F_Y)+(1-\eta)D_{EMD}(S_Q,S_Y) \tag{15}$$

where η is the adjusting parameter between the two distance measures and is determined by the long-term cross-session retrieval history working on the subset of training images [36]. If the precision for a particular query is increased when more importance is put on the feature distance, then η is adjusted to a larger value, otherwise it becomes smaller.

The Earth Mover's Distance (EMD) is used as a proper measure for comparing signatures given the pre-defined ground distances for pairs of attributes. The underlying idea of the Earth Mover's distance is that given two signatures of attributes, one can be seen as a mass of earth spread in the attribute space and the other as a collection of holes in the same attribute space. EMD is defined as the least amount of work needed to fill the holes with the earth. The ground distance between a pile of earth (an attribute element in the first signature) and a hole (an attribute element in the second signature) corresponds to the amount of work needed to move that pile of earth to the hole. The base metric is defined in the attribute space and used to compute the distance between two attributes. In our setting, the ground distance can be obtained by taking the reciprocal of the edge weights between the two attributes in the co-occurrence network which reflects the hardness that two attributes occur together in the images. Let $d(S_Q(a_i),S_D(a_j))$ denote the ground distance between attribute a_i in the query signature and attribute a_j in the database image signature. The Earth Mover's Distance between their signatures is defined as

$$D_{EMD}(S_Q,S_D)=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}f_{ij}d(S_Q(a_i),S_D(a_j))}{\sum_{i=1}^{m}\sum_{j=1}^{n}f_{ij}} \tag{16}$$

where f_{ij} is called a flow that is transferred from one signature to the other. The EMD is computed by solving all the f_{ij} using linear programming [80]. The EMD can be viewed as a measure of the least amount of work needed to transfer one signature into the other, a unit of work in the process is evaluated by the ground distance.

3.7.2. Relevance feedback

The Relevance feedback (RF) scheme has been verified as a performance booster for our retrieval system. The reason is that RF can capture more information about a user's search intention, which can be used to refine the original image descriptors from feature extraction and attribute detection [33].

Our RF approach follows the Query Point Movement (QPM) paradigm as opposed to the Query Expansion (QEX) paradigm. We move the query point in both the feature space and the attribute space toward the center of the user's preference by using both the relevant and irrelevant samples marked by the user at each retrieval iteration. However, before the users' decisions are used to refine the descriptors, their expertise in identifying moth species are evaluated by sample tests when they first enter the system. If a user has 90% accuracy in identifying the species, and his/her relevance feedback will take effect.

Suppose in each retrieval iteration the system returns N images. Let $F=\{f_1,f_2,\ldots,f_N\}$ denote the visual feature vectors and $S=\{s_1,s_2,\ldots,s_N\}$ denote the attribute signatures of the retrieved images and let f_Q and s_Q represents the query descriptors accordingly. The refinement of the descriptors is equivalent to learning projection matrix W_f that transforms $\{f_1,f_2,\ldots,f_N,f_Q\}$ into $\{f'_1,f'_2,\ldots,f'_N,f'_Q\}$, as well as W_s that transforms $\{s_1,s_2,\ldots,s_N,s_Q\}$ into $\{s'_1,s'_2,\ldots,s'_N,s'_Q\}$, by which the query and the relevant images resemble as much as possible in the feature and attribute spaces and deviate from the irrelevant ones.

Let $\mathcal{P}$ and $\mathcal{N}$ denote the sets of positive and negative results. We build a pairwise relevant descriptor set Λ_f,Λ_s and pairwise irrelevant descriptor set Ω_f,Ω_s in the following way:

$$\begin{cases}\Lambda_f=\{(f_Q,f_i)|f_i\in\mathcal{P}_f\}\cup\{(f_i,f_j)|f_i,f_j\in\mathcal{P}_f\}\\ \Lambda_s=\{(s_Q,s_i)|s_i\in\mathcal{P}_s\}\cup\{(s_i,s_j)|s_i,s_j\in\mathcal{P}_s\}\\ \Omega_f=\{(f_Q,f_i)|f_i\in\mathcal{N}_f\}\cup\{(f_i,f_j)|(f_i\in\mathcal{P}_f\cap f_j\in\mathcal{N}_f)\cup(f_i\in\mathcal{N}_f\cap f_j\in\mathcal{P}_f)\}\\ \Omega_s=\{(s_Q,s_i)|s_i\in\mathcal{N}_s\}\cup\{(s_i,s_j)|(s_i\in\mathcal{P}_s\cap s_j\in\mathcal{N}_s)\cup(s_i\in\mathcal{N}_s\cap s_j\in\mathcal{P}_s)\}\end{cases} \tag{17}$$

After the transformation W_f, the sum of the squared distances of the visual feature pairs in Λ_f is comupted as

$$\begin{aligned}&\sum_{(f_i,f_j)\in\Lambda_f}(W_f^Tf_i-W_f^Tf_j)^T(W_f^Tf_i-W_f^Tf_j)\\&=\sum_{(f_i,f_j)\in\Lambda_f}Tr[W_f^T(f_i-f_j)(f_i-f_j)^TW_f]\\&=Tr(W_f^TX_{\Lambda_f}W_f),\end{aligned} \tag{18}$$

where $X_{\Lambda_f}=\sum_{(f_i,f_j)\in\Lambda_f}(f_i-f_j)(f_i-f_j)^T$ and Tr is the trace of the matrix. Similarly, we have $\mathrm{Tr}(W_s^TX_{\Lambda_s}W_s)$, $\mathrm{Tr}(W_f^TX_{\Omega_f}W_f)$ and $\mathrm{Tr}(W_s^TX_{\Omega_s}W_s)$. We would like to have the sum of distances from Λ as small as possible and the sum of distances from Ω as large as possible, so have the following objective functions:

$$\begin{cases}\min\limits_{W_f^TW_f=I}\mathrm{Tr}(W_f^TX_{\Lambda_f}W_f),\ \max\limits_{W_f^TW=I}Tr(W_f^TX_{\Omega_f}W_f)\\ \min\limits_{W_s^TW_s=I}Tr(W_s^TX_{\Lambda_s}W_s),\ \max\limits_{W_s^TW=I}Tr(W_s^TX_{\Omega_s}W_s)\end{cases} \tag{19}$$

where I is the identity matrix, the purpose of having the constraints $W_f^TW_f=I, W_s^TW_s=I$ is to prevent arbitrary scaling of the projection. The minimization and maximization problems in (19) is usually formulated as a *trace ratio* optimization problem [81]:

$$\begin{cases}\max\limits_{W_f^TW_f=I}\dfrac{\mathrm{Tr}(W_f^TX_{\Omega_f}W_f)}{\mathrm{Tr}(W_f^TX_{\Lambda_f}W_f)}\\ \max\limits_{W_s^TW_s=I}\dfrac{\mathrm{Tr}(W_s^TX_{\Omega_s}W_s)}{\mathrm{Tr}(W_s^TX_{\Lambda_s}W_s)}\end{cases} \tag{20}$$

Wang et al. [81] proposed an iterative algorithm to conduct trace

ratio optimization, which is adopted in our work to solve the problem in (20) and is summarized in Algorithm 2.

Algorithm 2. Trace ratio optimization [81]

Input: The sum of descriptor distances in the positive and negative sets: $X_{\Lambda_f}, X_{\Lambda_s}, X_{\Omega_f}, X_{\Omega_s}$.

Output: The transformation matrices W_f and W_S

Initialize W_f^0, W_s^0 as arbitrary columnly orthogonal matrices such that $(W_f^0)^T W_f^0 = I$ and $(W_s^0)^T W_s^0 = I$.

Set iteration counter $n = 1$.

repeat

Compute λ_f^n, λ_s^n defined as follows:

$$\begin{cases} \lambda_f^n = \frac{Tr((W_f^{n-1})^T) X_{\Omega_f} W_f^{n-1}}{Tr((W_f^{n-1})^T) X_{\Lambda_f} W_f^{n-1}} \\ \lambda_s^n = \frac{Tr((W_s^{n-1})^T) X_{\Omega_s} W_s^{n-1}}{Tr((W_s^{n-1})^T) X_{\Lambda_s} W_s^{n-1}} \end{cases} \tag{21}$$

Solve the following trace difference maximization problem to obtain W_f^n and W_s^n by performing eigen-decomposition of $(X_{\Omega_f} - \lambda_f^n X_{\Lambda_f})$ and $(X_{\Omega_s} - \lambda_s^n X_{\Lambda_s})$:

$$\begin{cases} W_f^n = \underset{W_f^T W_f = I}{argmax}\, Tr[W_f^T (X_{\Omega_f} - \lambda_f^n X_{\Lambda_f}) W_f] \\ W_s^n = \underset{W_s^T W_s = I}{argmax}\, Tr[W_s^T (X_{\Omega_s} - \lambda_s^n X_{\Lambda_s}) W_s] \end{cases} \tag{22}$$

Set $n = n + 1$.

until *convergence*;

return W_f^n and W_s^n.

4. Experimental results

We implemented the system on a Microsoft Windows platform using C# ṅet with the Windows Presentation Foundation application development framework. The image database with relevant features and attributes are deployed on MySQL server. The database is set up by importing .txt files with numeric values of the attributes and features, and textual information describing the image properties of the moth images. We show the screenshot of the application in Fig. 7. We report here the results in two application scenarios: (i) moth species identification based on SRV attributes; (ii) Moth image retrieval with relevance feedback based on visual features and SRV attributes.

4.1. Image source and system parameters

Examination of the moth image collection used in this study is introduced in Section 5. All 4530 specimen images used in our experiments were manually labeled with SRV attributes with MBRs by using the tool introduced in Section 3.3.2. The species labels are provided by human experts (Dr. Janzen and his colleagues). The labels of the training images are used in the training process. The labels of the test images are used as ground-truth for validation.

4.2. Species identification results

We randomly divided the images into training and testing sets, which contain 80% and 20% of the entire data seperately. We then sampled the training set into 10 subsets, one subset was held out for validation and the remaining subsets were used for training the model. This process was repeated ten times for tunning the parameters. The final results on the testing set is reported in Table 3 and the tuned parameters based on the cross-validation in the training process are summarized in Table 2.

4.2.1. Evaluation criteria

The performance of the automated species identification is evaluated by the *accuracy* measure. A test image is assigned to the species category for which prototype signature has the smallest distance to the image's SRV attribute signature. The accuracy measure is defined for each species as the number of correctly identified individuals divided by the total number of specimens of that species in the testing set. A testing image is considered as a correct identification if the species label generated by the program matches with the ground-truth label.

4.2.2. Baseline approaches

To demonstrate the effectiveness of our proposed framework for the moth species identification application, we compare with the following approaches as baselines:

- *Baseline-I*: The most basic model that only uses the visual features extracted from Section 3.3.4. No SRV attributes and the signature representation were used. The images are identified purely based on the visual feature vector similarity calculated by using the Euclidean distance.
- *Baseline-II*: Our generative model for individual attribute detection unified with the attribute signature representation serves as the Baseline-II model. However, this model does not include attribute co-occurrence pattern detection and random walk refinement on the SRV attribute signatures.
- *VW-MSI*: We implemented a visual words based model based on the work by Lazebnik et al. [51] and name it as “Visual Words based Moth Species Identification” (VW-MSI). This technique works by partitioning the image into increasingly fine sub-

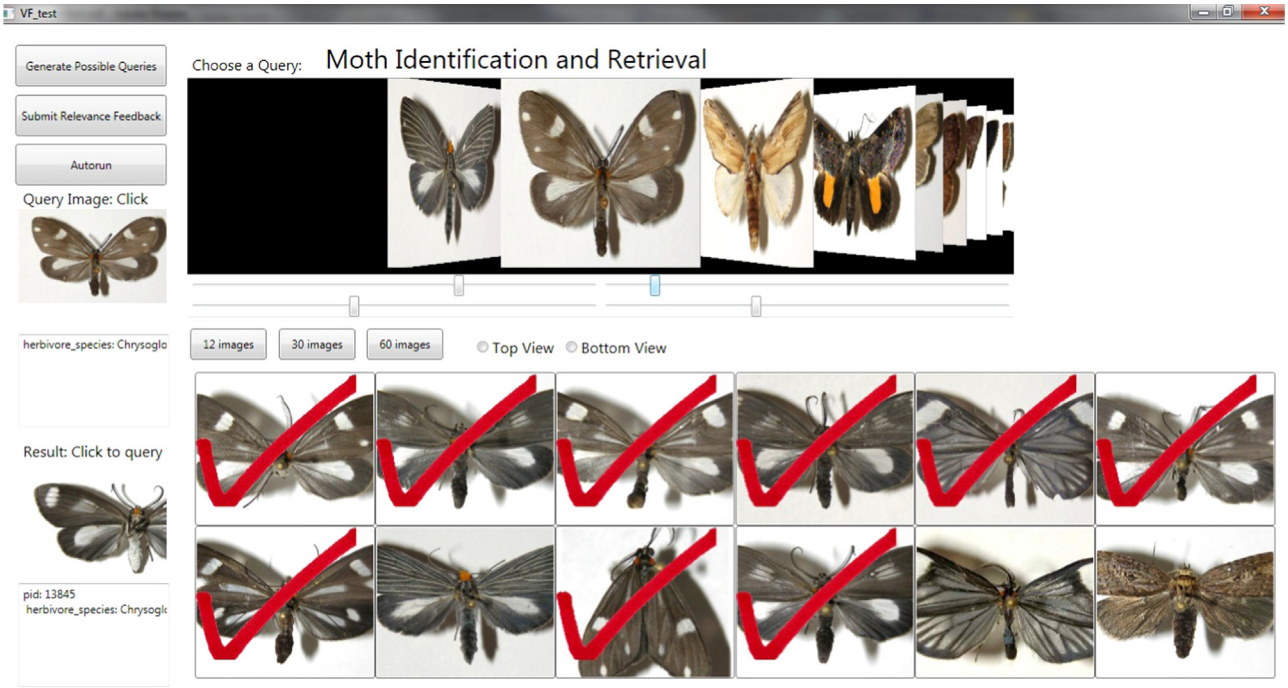

Fig. 7. The screen shot of the system. The images can be browsed in the display window and selected as queries. The "Submit Relevance Feedback" button is used for manual submissions and the "Autorun" button is used for automated queries. The species labels are shown in the text area. The user can click to mark the images as relevant, and the rest are used as irrelevant samples automatically. We can show up to 60 retrieved images in dorsal and ventral views.

regions and compare image similarity based on the histogram of local features.

- *SRV-MSI*: Our proposed approach integrated with co-occurrence pattern detection and SRV attribute signature refinement. We name it as "SRV attribute based Moth Species Identification" (SRV-MSI).

We compared the species identification results of the proposed approach with the other three approaches in Table 3. The best performance as well as the worst performance are bold in the table. The mean and standard deviation of the accuracy are shown for the fifty species. As we can observe from Table 3, our system performs the best for almost all the fifty species except that VW-MSI outperforms ours in five species: *Neoxeniades luda*, *Isostyla zetila*, *Atarnes sallei*, *Nascus* Burns and *Mysoria ambigua*. This demonstrates the effectiveness of SRV attributes and the co-occurrence patterns used for signature refinement.

The range of the mean identification accuracy of our system on the fifty species is between 0.3455 and 0.7764. The identification accuracy of some of the species is relatively low (e.g. *Hemicephalis agenoria*, *Neoxeniades luda*, *Dasylophia basitincta*, *Dasylophia maxtla* and *Nascus* Burns). When we visually examined the samples from these species, we found that the moth has less discriminative visual patterns or SRV attributes on the wings. This phenomenon reflects that our system may lose the power in identifying moth species with dull wings. Specifically, our system achieved low performance in two species categories: *Dasylophia basitincta* and *Dasylophia maxtla*, which have very similar visual appearances. The confusion matrix (we do not show it in the paper for the reason of space limitation) shows that our system mis-identifies the samples from one species into the other species. However, we observe that VW-MSI and other baselines also lose the effectiveness when dealing with moth images with very similar physical appearances. Based on the values of the standard deviation, our system still gives the most stable results across all the species categories compared to the other three approaches. The total number of SRV attributes manually given to the images by the human experts is 450. As a result, the maximum length of the SRV attribute signature for the images is 450.

4.3. Image retrieval results

To test the performance of our SRV attribute based approach for image retrieval with the proposed relevance feedback scheme, like for species identification in Section 4.2, we divided the entire image dataset into 10 folds. The parameters are determined using the same scheme as described in Section 4.2.1. We set the number of attributes to 300. In order to reduce the amount of work of submitting relevance feedback that are required to be given by users, we propose to simulate the user interaction by launching queries and submitting feedback automatically by the system. The launching of automated queries works in the following way: the system compares the ground-truth species labels of the retrieved images with the query image, if the species label matches the label of the query image, the system will mark the image as relevant, otherwise, the image is marked as irrelevant. By doing this, we assume the relevance feedback provided by the users will always be correct and complete (i.e., users will only mark the relevant images as those from the same species category as the query and all the relevant images will be marked). The reason we have this automated query process is because we want to use each image in the database as a query image and collect the relevance feedback. However, considering the large number of images in the database, manually launching queries and providing relevance feedback by clicking through the images will take a great amount of time. Therefore, we simulate the whole process by launching automated queries behind the scene. In this way, we are simulating expert users who will always provide correct and complete answers to the system. Note that the ground-truth is only used by the system to judge the relevance of the retrieved images. It is not involved in comparing image similarity in the retrieval procedure. For each query, we request the users or the system to provide five iterations of relevance feedback. We have half of the queries in each species

Table 2
The system parameters for the experiments.

Parameter	Description	Section	Setup
Q	The threshold for determining whether a community is a good partition in the network.	Section 3.5.1	The value is in the range of $[-1, 1]$, we set to 0.3 based on 10 cross-fold validation.
α	The weighting parameter in the random walk process.	Section 3.5.2	The value is in the range of $[0, 1]$, we set the value to 0.6 based on 10 cross-fold validation.
η	The adjusting parameter between two image distance measures.	Section 3.7.1	The value is in the range of [0,1], the value is set 0.5 based on 10 cross-fold validation.

Table 3
Identification accuracy for the fifty species. The performance of SRV-MSI is better than all other approaches except for *Neoxeniades luda*, *Isostyla zetila*, *Atarnes sallei* and *Nascus Burns*.

	Baseline I		Baseline II		VW-MSI		SRV-MSI	
Species	*Mean*	*Std*	*Mean*	*Std*	*Mean*	*Std*	*Mean*	*Std*
Ceroctena amynta	0.2965	0.0321	0.4176	0.0169	0.4347	0.0184	0.4582	0.0174
Eudocima materna	0.4968	0.0257	0.5483	0.0275	0.5772	0.0279	0.5944	0.0209
Eulepidotis folium	0.3910	0.0279	0.4141	0.0264	0.4241	0.0213	0.4482	0.0371
Eulepidotis rectimargo	0.5561	0.0246	0.5875	0.0236	0.5982	0.0211	0.6134	0.0163
Hemicephalis agenoria	0.3314	0.0268	0.3349	0.0302	0.3764	0.0315	0.3931	0.0236
Thysania zenobia	0.4102	0.0327	0.4329	0.0236	0.4675	0.0218	0.4971	0.0356
Chrysoglossa norburyi	0.5472	0.0225	0.5553	0.0253	0.5693	0.0214	0.5752	0.0205
Erbessa albilinea	0.6048	0.0365	0.6324	0.0336	0.6564	0.0112	0.6755	0.0174
Erbessa salvini	0.3562	0.0468	0.3634	0.0425	0.3894	0.0313	0.4143	0.0345
Nebulosa erymas	0.5432	0.0312	0.5647	0.0291	0.5722	0.0219	0.5935	0.0225
Tithraustes noctiluces	0.5438	0.0214	0.5624	0.0331	0.5948	0.0215	0.6086	0.0251
Polypoetes haruspex	0.5247	0.0216	0.5369	0.0234	0.5699	0.0226	0.5906	0.0202
Dioptis longipennis	0.5621	0.0281	0.5746	0.0212	0.6013	0.0124	0.6154	0.0175
Methionopsis ina	0.4721	0.0375	0.4835	0.0367	0.5056	0.0317	0.5102	0.0425
Neoxeniades luda	0.3742	0.0374	0.3852	0.0432	0.4183	0.0345	0.3975	0.0457
Saliana Burns	0.5042	0.0364	0.5356	0.0256	0.5523	0.0227	0.5731	0.0234
Saliana fusta	0.6480	0.0247	0.6597	0.0275	0.6993	0.0205	0.7346	0.0134
Talides Burns	0.5437	0.0256	0.5572	0.0247	0.5872	0.0158	0.6352	0.0176
Vettius conka	0.6417	0.0334	0.6782	0.0148	0.7364	0.0153	0.7544	0.0169
Aroma aroma	0.5437	0.0273	0.6035	0.0245	0.6244	0.0144	0.6461	0.0211
Carystoides escalantei	0.5326	0.0324	0.5487	0.0264	0.5873	0.0212	0.6033	0.0254
Lirimiris guatemalensis	0.3975	0.0421	0.4129	0.0256	0.4635	0.0216	0.4930	0.0249
Isostyla zetila	0.5248	0.0363	0.5392	0.0365	0.5482	0.0231	0.5364	0.0357
Oricia domina	0.4964	0.0368	0.5175	0.0316	0.5391	0.0376	0.5632	0.0195
Scotura leucophleps	0.5014	0.0378	0.5246	0.0217	0.5574	0.0238	0.5757	0.0221
Bardaxima perses	0.4764	0.0371	0.4954	0.0314	0.5337	0.0276	0.5551	0.0307
Dasylophia basitincta	0.3842	0.0457	0.3976	0.0351	0.4031	0.0216	0.4344	0.0275
Dasylophia maxtla	0.3683	0.0416	0.3754	0.0363	0.3948	0.0314	0.4113	0.0278
Nystalea collaris	0.4173	0.0285	0.4326	0.0291	0.4861	0.0243	0.5021	0.0274
Tachuda discreta	0.3647	0.0321	0.4056	0.0249	0.4314	0.0327	0.4512	0.0269
Atarnes sallei	0.6084	0.0372	0.6396	0.0278	0.7072	0.0127	0.7059	0.0187
Dyscophellus phraxanor	0.5483	0.0364	0.5731	0.0381	0.6295	0.0331	0.6494	0.0362
Tithraustes lambertae	0.6053	0.0271	0.6056	0.0374	0.6314	0.0249	0.6713	0.0285
Entheus matho	0.6153	0.0490	0.6273	0.0411	0.6319	0.0263	0.6534	0.0279
Hyalothyrus neleus	0.6472	0.0394	0.6717	0.0285	0.6961	0.0184	0.7106	0.0168
Nascus Burns	0.3258	0.0173	0.3394	0.0314	0.3549	0.0387	**0.3455**	0.0372
Phocides nigrescens	0.6138	0.0442	0.6359	0.0321	0.6789	0.0174	0.6797	0.0171
Quadrus contubernalis	0.6432	0.0316	0.6572	0.0257	0.6942	0.0268	0.7096	0.0263
Urbanus belli	0.5276	0.0164	0.5713	0.0268	0.5953	0.0182	0.6132	0.0254
Melanopyge Burns	0.6261	0.0255	0.6527	0.0275	0.6799	0.0134	0.6930	0.0214
Myscelus belti	0.6438	0.0354	0.6765	0.0241	0.7564	0.0182	**0.7764**	0.0158
Mysoria ambigua	0.5537	0.0341	0.5864	0.0213	0.6141	0.0275	0.6247	0.0218
Dicentria rustica	0.3497	0.0354	0.3764	0.0252	0.4431	0.0309	0.4546	0.0277
Farigia sagana	0.3647	0.0387	0.4145	0.0262	0.4744	0.0265	0.4854	0.0254
Hapigiodes sigifredomarini	0.4126	0.0264	0.4352	0.0225	0.4553	0.0321	0.4894	0.0196
Malocampa matralis	0.4832	0.0346	0.5167	0.0374	0.5382	0.0314	0.5893	0.0217
Meragisa Janzen	0.5654	0.0246	0.5987	0.0320	0.6187	0.0211	0.6375	0.0212
Naprepa houla	0.4264	0.0257	0.4583	0.0315	0.4832	0.0247	0.5126	0.0276
Pseudodryas pistacina	0.4126	0.0354	0.4323	0.0267	0.4654	0.0209	0.4879	0.0219
Rifargia dissepta	0.5836	0.0321	0.5917	0.0289	0.6283	0.0217	0.6412	0.0365

category launched by the users and the other half simulated by the system. The results are computed based on the combination of the two methods.

4.3.1. Evaluation criteria

In each iteration, the retrieval precision is evaluated by the rank of the relevant images. Further statistical evaluation of the

Table 4
Comparison of the retrieval performance for the fifty species.

Geometric mean average precision							
Species	BL-I	BL-II	SRV-IR	*Species*	BL-I	BL-II	SRV-IR
Ceroctenaamynta	0.4032	0.3856	0.4533	*Bardaximaperses*	0.2819	0.3047	0.3218
Eudocimamaterna	0.4511	0.4142	0.4738	*Dasylophiabasitincta*	0.3517	0.4102	0.4619
Eulepidotisfolium	0.3794	0.4105	0.4743	*Dasylophiamaxtla*	0.3598	0.3692	0.4107
Eulepidotisrectimargo	0.5563	0.5051	0.6117	*Nystaleacollaris*	0.3408	0.3726	0.3819
Hemicephalisagenoria	0.4132	0.3947	0.4693	*Tachudadiscreta*	0.2872	0.2935	0.3084
Thysaniazenobia	0.4104	0.3933	0.4705	*Atarnessallei*	0.5832	0.6224	0.6778
Chrysoglossanorburyi	0.5856	0.5710	0.6786	*Dyscophellusphraxanor*	0.5324	0.5799	0.6128
Erbessaalbilinea	0.6045	0.5972	0.7153	*Tithrausteslambertae*	0.4846	0.4472	0.5315
Erbessasalvini	0.4529	0.4297	0.5428	*Entheusmatho*	0.4748	0.4876	0.5432
Nebulosaerymas	0.5219	0.5346	0.5857	*Hyalothyrusneleus*	0.6042	0.6584	0.6971
Tithraustesnoctiluces	0.5486	0.5148	0.5749	*NascusBurns*	0.2396	0.2846	0.3167
Polypoetesharuspex	0.5745	0.5237	0.5964	*Phocidesnigrescens*	0.5755	0.5942	0.6398
Dioptislongipennis	0.4816	0.4754	0.5048	*Quadruscontubernalis*	0.6492	0.7047	0.7168
Methionopsisina	0.3581	0.3847	0.3994	*Urbanusbelli*	0.5693	0.5480	0.5724
Neoxeniadesluda	0.3625	0.3827	0.4117	*MelanopygeBurns*	0.6454	0.6845	0.6992
SalianaBurns	0.5298	0.5446	0.5829	*Myscelusbelti*	0.6894	0.7008	**0.7631**
Salianafusta	0.6046	0.5917	0.6459	*Mysoriaambigua*	0.4917	0.4802	0.5746
TalidesBurns	0.5154	0.5308	0.5742	*Dicentriarustica*	0.3969	0.4105	0.4453
Vettiusconka	0.6296	0.6115	0.7135	*Farigiasagana*	0.2946	0.3072	0.3418
Aromaaroma	0.4537	0.4425	0.5289	*Hapigiodessigifredoma*	0.3634	0.3728	0.4051
Carystoidesescalantei	0.5046	0.4672	0.5274	*Malocampamatralis*	0.4746	0.4869	0.5537
Lirimirisguatemalensis	0.3234	0.3456	0.3753	*MeragisaJanzen*	0.5643	0.5756	0.6683
Isostylazetila	0.5924	0.5483	0.6175	*Naprepahoula*	0.3748	0.4245	0.4886
Oriciadomina	0.4641	0.4547	0.5044	*Pseudodryaspistacina*	0.2975	0.3174	0.3531
Scoturaleucophleps	0.5179	0.5357	0.5678	*Rifargiadissepta*	0.5648	0.5247	0.6190
				Overall Mean (50*species*)	0.4743	0.4771	0.5312
				Overall Std (50*species*)	0.1088	0.1072	0.1143

averaged precision for each species relies on standard image retrieval measure: *Mean average precision of top D retrieved images* over all the query images from a specific species category. Let D be the number of retrieved images and R be the relevant ones with size $|R|$. Given a query Q, the average precision is defined as $AP(Q) = 1/|R| \sum_{i=1}^{|R|} i/Rank(R_i)$, and the geometric mean average precision (*GMAP*) which is defined as $GMAP = |Q| \sqrt{\prod_{i=1}^{|Q|} AP}$ is the measure for the system performance over all the species.

4.3.2. Baseline approaches

To demonstrate the effectiveness of our proposed retrieval framework, we use the following approaches as the baselines to compare the results:

- *Baseline-I*: The proposed image retrieval framework without relevance feedback scheme.
- *Baseline-II*: We reimplemented an insect image identification approach [14] and integrated it into our retrieval framework with five iterations of relevance feedback process. The features used are a combination of color, shape and texture features and there are no higher level image descriptors like our SRV attributes.
- *SRV-IR*: Our proposed retrieval framework with relevance feedback scheme based on the SRV attributes.

We show the top twelve retrieved images in the application interface. However, the application can be adjusted to show more images upon request. Table 4 summarizes the geometric mean average precision from the three approaches for all the fifty species. As we can observe, when the RF scheme is applied (Baseline-II and SRV-IR), the geometric mean averaged precision is increased compared to the retrieval without RF (Baseline-I), which demonstrates the effect of human interaction in improving the retrieval performance. When more retrieval iterations (with relevance feedback) are involved in the search process, the system can find more relevant images matching the user's search intention. In the two approaches that adopt a relevance feedback scheme, our approach which uses SRV attribute based image descriptors outperforms Baseline-II for all the species categories. The system response time for each individual query for a database of 1000 images is approximately 150 ms. For a database of 4000 images the response time for each individual query is approximately 500 ms.

5. Conclusions

This paper describes a novel insect species identification and retrieval system based on wing attributes of moth images. The purpose of the research is to design adn develop computer vision and pattern recognition system for conducting automated image analysis that can be used by the entomologists for insect studies. We demonstrated the effectiveness of our system in species identification and image retrieval for fifty moth species.

There are two major processes for species identification and retrieval: preprocessing images for attribute extraction and learning the co-occurrence relationships of the attributes. Current systems try to make the first process automatic while overlooking the importance of the second process which could bring much contextual information. Our identification and retrieval system based on CBIR architecture is fully automatic. For example, many current systems require manual separation of foreground and background in the preprocessing step while in our system we have automated segmentation of moth from the background with shadow removal.

The dataset that we used contains 4530 images which could be easily extended to larger sizes in the future to test the scalability of the system. Overall, our system achieves a better performance compared to the baseline approaches in identification and retrieval. The identification accuracy reaches 70% for some species in the image collection while the majority of the species has the identification accuracy between 40% and 60%. The lowest accuracy

comes from species Nascus Burns and it is approximately 34%. The mean average precision for the image retrieval task also reaches 70% for some of the species, and the majority of the species has the GMAP above 40%. Eight species have the retrieval precision lower than 40%. By examining the images, we found that the differences in performance for different species are caused by the different level of visual properties. Some of the species have easily distinguishable visual attributes on the wings while others may not have them. This demonstrates that using our co-occurrence pattern based attribute learning and detection can achieve a better performance by bringing the relationship of attributes into consideration. However, it is still confined by the visual properties of images and it is a challenge for all the approaches.

A significant difference between our work and similar work in insect identification is that we provide an intermediate-level feature, namely, the SRV attributes, which function to narrow the semantic gap between machine understanding and human interpretation of images. We are excited to see that SRV attributes successfully capture the visual patterns on the wings of moth at a higher semantic level and generate better results consequently.

However, the discriminative power of our system drops when the moth species contain highly similar visual properties. This could cause reduction in both identification and retrieval once more images are included in the dataset that belong to different species categories but share strong visual patterns on the wings. These cases would be difficult for humans as well. Also, our research belongs to the category that generates human-designated attributes and learn the relationship between these attributes and image samples. All the approaches that belong to this category suffer from scalability problems especially for applications that are deal with images from general domains such as natural scenes, outdoor/indoor scenes, etc. In our case, we are working in the specific domain of moth images where the SRV attributes may not increase significantly even for a different and larger dataset. Therefore, compared to applications in other general domains, our system and approach could mitigate the issues associated with scalability.

Our future research includes investigations on more effective features and attributes [82], including deep learning based features [83] which could address both the scalability and discrimination issues.

Conflict of interest

No conflict of interest.

Acknowledgments

This work was supported in part by US National Science Foundation (NSF) grants 0641076, 0727129, and 0905671. The authors thank Dr. Dan Janzen, University of Pennsylvania, for supplying test images (http://janzen.sas.upenn.edu).

References

[1] D. Carter, Butterflies and moths, Eyewitness Handbooks, 1992.
[2] P. Kerr, E. Fisher, M. Buffington, Dome lighting for insect imaging under a microscope, Am. Entomol. 54 (2008) 198–200.
[3] M. Buffington, M. Gates, Advanced imaging techniques ii: using a compound microscope for photographing point-mount specimens, Am. Entomol. 54 (2008) 222–224.
[4] M. Buffington, R. Burks, L. McNeil, Advanced techniques for imaging parasitic Hymenoptera (Insecta), Am. Entomol. 51 (2005) 50–56.
[5] C. Wen, D.E. Guyer, W. Li, Local feature-based identification and classification for orchard insects, Biosyst. Eng. 104 (3) (2009) 299–307.
[6] T.M. Francoy, D. Wittmann, M. Drauschke, S. Müller, V. Steinhage, M.A. F. Bezerra-Laure, D.D. Jong, L.S. Goncalves, Identification of africanized honey bees through wing morphometrics: two fast and efficient procedure, Apidologie 39 (5) (2008) 488–494.
[7] U. Jean, D. Arne, F. Simon, G. Malcolm, A stile project case study: the evaluation of a computer-based visual key for fossil identification, Assoc. Learn. Technol. J. 4 (2) (1996) 40–47.
[8] S. Schroder, W. Drescher, V. Steinhage, B. Kastenholz, An automated method for the identification of bee species, in: Proceedings of the International Symposium on Conserving European Bees, 1995, pp. 6–7.
[9] M.T. Do, J.M. Harp, K.C. Norris, A test of a pattern recognition system for identification of spiders, Bull. Entomol. Res. 89 (3) (1999) 217–224.
[10] J. Yue, Z. Li, L. liu, Z. Fu, Content-based image retrieval using color and texture fused features, Math. Comput. Model. 54 (2011) 1121–1127.
[11] K. Bunte, M. Biehl, M. Jonkman, N. Petkov, Learning effective color features for content based image retrieval in dermatology, Pattern Recognit. 44 (2011) 1892–1902.
[12] N. Singhai, S. Shandilya, A survey on: content based image retrieval systems, Int. J. Comput. Appl. 4 (2010) 22–26.
[13] B. Bhanu, R. Li, J. Heraty, E. Murray, Automated classification of skippers based on parts representation, Am. Entomol. (2008) 228–231.
[14] J. Wang, C. Lin, L. Ji, A. Liang, A new automatic identification system of insect images at the order level, Knowl.-Based Syst. 33 (2012) 102–110.
[15] S.K. Divvala, D. Hoiem, J.H. Hays, A.A. Efros, M. Hebert, An empirical study of context in object detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1271–1278.
[16] A. Hanjalic, R. Lienhart, W.Y. Ma, J.R. Smith, The holy grail of multimedia information retrieval: so close or yet so far away? Proc. IEEE, 96 (4) (2008) pp. 541–547.
[17] L. Zhu, Z. Zhang, Auto-classification of insect images based on color histogram and GLCM, in: 7th International Conference on Fuzzy Systems and Knowledge Discovery, 2010, pp. 971–980.
[18] J. Wang, L. Ji, A. Liang, D. Yuan, The identification of butterfly families using content-based image retrieval, Biosyst. Eng. 111 (2011) 24–32.
[19] M. Mayo, A.T. Watson, Automatic species identification of live moths, Knowl.-Based Syst. 20 (4) (2007) 195–202.
[20] S.-H. Kang, W. Jeon, S.-H. Lee, Butterfly species identification by branch length similarity entropy, J. Asia-Pacific Entomol. 15 (3) (2012) 437–441.
[21] Y. Kaya, L. Kayci, Application of artificial neural network for automatic detection of butterfly species using color and texture features, Vis. Comput. 30 (1) (2014) 71–79.
[22] J. Sivic, B. Russell, A. Efros, A. Zisserman, W. Freeman, Discovering object categories in image collections, in: International Conference on Computer Vision, 2005, pp. 1543–1550.
[23] H.M. Pereira, S. Ferrier, M. Walters, G.N. Geller, R.H.G. Jongman, R.J. Scholes, M.W. Bruford, N. Brummitt, S.H.M. Butchart, A.C. Cardoso, et al., Essential biodiversity variables, Science 339 (1) (2013) 277278.
[24] S.J. Bacon, S. Bacher, A. Aebi, Gaps in border controls are related to quarantine alien insect invasions in Europe, PLoS One 7 (10) (2012) 0047689, http://dx.doi.org/10.1371/journal.pone.
[25] S. Kumschick, S. Bacher, W. Dawson, J. Heikkil, A conceptual framework for prioritization of invasive alien species for management according to their impact, NeoBiota 15 (10) (2012) 69–100.
[26] P.R. Steele, J.C. Pires, Biodiversity assessment: state-of-the-art techniques in phylogenomics and species identification, Am. J. Bot. 98 (3) (2011) 415–425.
[27] M. ONeill, I.G. nd K.J. Gaston, P. Weeks, Daisy: an automated invertebrate identification system using holistic vision techniques, in: Inaugural Meeting of the BioNET-International Group for Computer-aided Taxonomy, 2000.
[28] A. Tofilski, Drawwing, a program for numerical description of insect wings, J. Insect Sci. 4 (2004) 17–22.
[29] T. Ganchev, I. Potamitis, N. Fakotakis, Acoustic monitoring of singing insects, in: IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 4, 2007.
[30] T.D. Meulemeester, P. Gerbaux, M. Boulvin, A. Coppe, P. Rasmont, A simplified protocol for bumble bee species identification by cephalic secretion analysis, Int. J. Study Soc. Arthropods 58 (5) (2011) 227236.
[31] A. Joly, H. Goau, H. Glotin, C. Spampinato, P. Bonnet, W. Vellinga, R. Planque, A. Rauber, R. Fisher, H. Müller, Lifeclef 2014: multimedia life species identification challenges, in: Proceedings of the LifeCLEF, 2014, pp. 229–249.
[32] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, IEEE Trans. Pattern Anal. Mach. Intell. 22 (12) (2000) 13491380.
[33] A. Dong, B. Bhanu, Active concept learning in image databases, IEEE Trans. Syst. Man Cybern. Part B 35 (2005) 450–456.
[34] J. Peng, B. Bhanu, S. Qing, Probabilistic feature relevance learning for content-based image retrieval, Comput. Vis. Image Underst. 75 (2005) 150–164.
[35] P. Yin, B. Bhanu, K. Chang, A. Dong, Integrating relevance feedback techniques for image retrieval using reinforcement learning, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 1536–1551.
[36] P.Y. Yin, B. Bhanu, K.C. Chang, Long-term cross-session relevance feedback using virtual features, IEEE Trans. Knowl. Data Eng. 20 (3) (2008) 352–368.
[37] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 886–893.
[38] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.

[39] J. Yang, Y.G. Jiang, A.G. Hauptmann, C.W. Ngo, Evaluating bag-of-visual-words representations in scene classification, in: Proceedings of the International Workshop on Multimedia Information Retrieval, 2007, pp. 197–206.
[40] V. Viitaniemi, J. Laaksonen, Spatial extensions to bag of visual words, in: Proceedings of the ACM International Conference on Image and Video Retrieval, 2009, pp. 197–206.
[41] Y.G. Jiang, A.H.J. Yang, C.W. Ngo, Representations of keypoint-based semantic concept detection: a comprehensive study, IEEE Trans. Multimed. 12 (1) (2010) 42–53.
[42] K.E.A.V.D. Sande, T. Gevers, C.G.M. Snoek, Evaluating color descriptors for object and scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1582–1596.
[43] J.C.V. Gemert, C.J. Veenman, A.W.M. Smeulders, J.M. Geusebroek, Visual word ambiguity, IEEE Trans. Pattern Anal. Mach. Intell. 32 (7) (2010) 1271–1283.
[44] J. Zhang, M.M. lek, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: a comprehensive study, Int. J. Comput. Vis. 73 (2) (2007) 213238.
[45] Y.G. Jiang, C.W. Ngo, J. Yang, Towards optimal bag-of-features for object categorization and semantic video retrieval, in: Proceedings of the 6th ACM International Conference on Image and Video Retrieval, 2007, pp. 494–501.
[46] J.R.R. Uijlings, A.W.M. Smeulders, R.J.H. Scha, Real-time bag of words, approximately, in: Proceedings of the 8th ACM International Conference on Image and Video Retrieval, 2009, pp. 494–501.
[47] P. Tirilly, V. Claveau, P. Gros, Language modeling for bag-of-visual words image categorization, in: International Conference on Content-based Image and Video Retrieval, 2008, pp. 249–258.
[48] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, IEEE Trans. Pattern Anal. Mach. Intell. 27 (10) (2005) 1615–1630.
[49] J. Philbin, O. Chum, M. Isard, J. Sivic, Object retrieval with large vocabularies and fast spatial matching, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[50] Y. Yang, S. Newsam, Bag-of-visual-words and spatial extensions for land-use classification, in: Proceedings of the International Conference on Advances in Geographic Information Systems, 2010, pp. 270–279.
[51] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 2169–2178.
[52] F. Perronnin, C. Dance, Fisher kernels on visual vocabularies for image categorization, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[53] F. Perronnin, Y. Liu, J. Sanchez, H. Poirie, Large-scale image retrieval with compressed fisher vectors, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 3384–3391.
[55] C.H. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 951–958.
[56] L. Bourdev, S. Maji, J. Malik, Describing people: a poselet-based approach to attribute classification, in: International Conference on Computer Vision, 2011, pp. 1543–1550.
[57] A. Farhadi, I. Endres, D. Hoiem, D. Forsyth, Describing objects by their attributes, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1778–1785.
[58] N. Kumar, A.C. Berg, P.N. Belhumeur, S.K. Nayar, Attribute and simile classifiers for face verification, in: International Conference on Computer Vision, 2009, pp. 365–372.
[59] A. Farhadi, I. Endres, D. Hoiem, Attribute-centric recognition for cross-category generalization, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2352–2359.
[60] Y. Wang, G. Mori, A discriminative latent model of object classes and attributes, in: European Conference on Computer Vision, 2010, pp. 155–168.
[61] D. Parikh, K. Grauman, Relative attributes, in: International Conference on Computer Vision, 2011, pp. 801–808.
[62] M. Douze, A. Ramisa, C. Schmid, Combining attributes and fisher vectors for efficient image retrieval, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 745–752.
[63] X. Wang, K. Liu, X. Tang, Query-specific visual semantic spaces for web image re-ranking, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 857–864.
[64] B. Siddiquie, R.S. Feris, L.S. Davis, Image ranking and retrieval based on multi-attribute queries, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 801–808.
[65] P.K. Atrey, M.A. Hossain, A.E. Saddik, M.S. Kankanhalli, Multimodal fusion for multimedia analysis: a survey, Multimed. Syst. 16 (6) (2010) 345–379.
[66] P. Wilkins, P. Ferguson, A.F. Smeaton, Using score distributions for query-time fusion in multimedia retrieval, Multimed. Inf. Retr., 2006.
[67] H. Müller, T. Clough, P. Deselaers, B. Caputo, ImageCLEF Experimental Evaluation in Visual Information Retrieval, Springer, London, 2010.
[68] D.H. Janzen, W. Hallwachs, Dynamic database for an inventory of the macrocaterpillar fauna, and its food plants and parasitoids, of area de conservacion guanacaste (acg), northwestern costa rica (nn-srnp-nnnnn voucher codes), 2009. Available from ⟨http://janzen.sas.upenn.edu⟩.
[69] Y. Sun, B. Bhanu, Reflection symmetry-integrated image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 34 (9) (2012) 1827–1841.
[70] K. Duan, D. Parikh, D. Crandall, Discovering localized attributes for fine-grained recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3474–3481.
[71] D. Parikh, K. Grauman, Interactively building a discriminative vocabulary of nameable attributes, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1681–1688.
[72] B.C. Russell, A. Torralba, K.P. Murphy, W.T. Freeman, Labelme: a database and web-based tool for image annotation, Int. J. Comput. Vis. 77 (2008) 157–173.
[73] V.S.N. Prasad, B. Yegnanarayana, Finding axes of symmetry from potential fields, IEEE Trans. Image Process. 13 (12) (2004) 1559–1566.
[74] R.M. Haralick, Statistical and structural approaches to texture, Proc. IEEE, 67, 1979, pp. 786–804.
[75] C.C. Gotlieb, H.E. Kreyszig, Texture descriptors based on co-occurrence matrices, Comput. Vis. Graph. Image Process. 51 (1) (1990) 70–86.
[76] G. Fu, F. Shih, H. Wang, A kernel-based parametric method for conditional density estimation, Pattern Recognit. 44 (2) (2011) 284–294.
[77] M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2) (2004) 84–99.
[78] W.H. Hsu, L.S. Kennedy, S.-F. Chang, Video search reranking through random walk over document-level context graph, in: Proceedings of the 15th International Conference on Multimedia, 2007, pp. 971–980.
[79] Y. Rubner, C. Tomasi, L.J. Guibas, The earth movers distance as a metric for image retrieval, Int. J. Comput. Vis. 40 (2) (2000) 99–121.
[80] H. Ling, K. Okada, An efficient earth movers distance algorithm for robust histogram comparison, IEEE Trans. Pattern Anal. Mach. Intell. 29 (5) (2007) 840–853.
[81] H. Wang, S. Yan, D. Xu, X. Tang, T. Huang, Trace ratio vs. ratio trace for dimensionality reduction, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[82] R. Li, B. Bhanu, A. Dong, Coevolutionary feature synthesized em algorithm for image retrieval, in: Proceedings of the 13th ACM International Conference on Multimedia, 2005, pp. 696–705.
[83] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.

Linan feng received the BSc degree in Electrical Engineering in 2006 and the M.E. degree in Software Engineering in 2009 both from Shanghai Jiao Tong University, Shanghai, China. Since 2009, he has been a Ph.D. candidate in Computer Science, at the University of California, Riverside, CA. His reseach interests are in computer vision, pattern recognition and machine learning, with emphasis on automatated image annotation and concept-based image retrieval. He is the student member of IEEE.

Bir bhanu received the S.M. and E.E. degrees in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology; the Ph.D. degree in Electrical and Computer Engineering from the Image Processing Institute, University of Southern California and the M.B.A. degree from the University of California, Irvine. Dr. Bhanu is the Distinguished Professor of Electrical Engineering and Cooperative Professor of Computer Science and Engineering, Mechanical Engineering and Bioengineering, Director of the Center for Research in Intelligent Systems (CRIS), and the Visualization and Intelligent Systems Laboratory (VISLab) at the University of California, Riverside (UCR). He also serves as the Director of NSF IGERT program on Video Bioinformatics and the Interim Chair of the Department of Bioengineering. His current research interests are Computer Vision, Pattern Recognition and Data Mining, Machine Learning, Artificial Intelligence, Image Processing, Image and Video Database, Graphics and Visualization, Robotics, Human-Computer Interactions, Biological, Medical, Military and Intelligence applications. He has been the principal investigator of various programs from NSF, DARPA, NASA, AFOSR, ONR, ARO and other agencies and industries. He is Fellow of IEEE, AAAS, IAPR, AIMBE and SPIE.

John heraty received his Ph.D. degree in Biology in 1990 from Texas A&M University, USA. He is a Professor of Entomology at the University of California at Riverside. His research focuses on the systematics, phylogeny and biogeography of the Chalcidoidea (Hymenoptera). Chalcidoid wasps rank numerically among the largest groups of insects, with estimates of as many as 100,000 species; however, the fauna is poorly known. Most are specialized parasites, and the majority of successful biological control projects have utilized these minute wasps to achieve partial or complete control of insect pests. All of his studies incorporate morphological, biological or molecular information into analyses that are used to formulate hypotheses of phylogenetic relationships and the evolution of behavioral patterns.