Knowledge-Based Systems 33 (2012) 102–110

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

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

A new automatic identification system of insect images at the order level

Jiangning Wang [a,b], Congtian Lin [a], Liqiang Ji [a,*], Aiping Liang [b,*]

[a] Key Laboratory of Animal Ecology and Conservation Biology, Institute of Zoology, Chinese Academy of Sciences, 1 Beichen West Road, Beijing 100101, China
[b] Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, 1 Beichen West Road, Beijing 100101, China

ARTICLE INFO

Article history:
Received 4 May 2011
Received in revised form 13 March 2012
Accepted 14 March 2012
Available online 21 March 2012

Keywords:
Insect
Order
Feature
ANN
SVM

ABSTRACT

A new automatic identification system has been designed to identify insect specimen images at the order level. Several relative features were designed according to the methods of digital image progressing, pattern recognition and the theory of taxonomy. Artificial neural networks (ANNs) and a support vector machine (SVM) are used as pattern recognition methods for the identification tests. During tests on nine common orders and sub-orders with an artificial neural network, the system performed with good stability and accuracy reached 93%. Results from tests using the support vector machine further improved accuracy. We also did tests on eight- and nine-orders with different features and based on these results we compare the advantages and disadvantages of our system and provide some advice for future research on insect image recognition.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

The vast number of insect species is a challenge for insect identification, and undermines the bases of biodiversity, conservation and related research. The intricacy of traditional insect identification methods and a declining number of insect taxonomists seriously affect the efficiency of this task [6]. Taxonomists have been searching for efficient methods to meet real world insect identification requirements. Several assisted systems of insect identification based on computer technologies have been developed and tested in the last two decades including automatic bee identification system (ABIS), digital automated identification system (DAISY), BugVisux and But2fly. ABIS was constructed in 1995 and identifies bees in the field by analyzing images of wing veins [24]. DAISY is a prototype system applied for identifying insects using digital images based on fingerprint identification technology and is also being used to recognize and analyze museum collections [22,26,27]. BugVisux can identify 40 species of insects using morphologic features [32], and But2fly can identify 43 butterfly species through the color features of wings [15].

Although aforementioned systems above focus on the identification of insects at the species level, automatic insect identification at the order level is also important, especially in popular science and initial insect identification. A specimen usually needs to be identified to the order level or family level before a species name can be given by taxonomists and order level identification is of more use to the public and junior taxonomists. Therefore Identification of an insect at the order level is thus a key step in the entire insect identification process.

Little research about the order levels has been conducted [3] despite the importance of insect order identification. Difficulties surrounding insect order identification arise because of countless species and the complex classification system of insects. Many insect orders include thousands of species such as the Coleoptera, which includes more than 350,000 described species. Within each order, species may be highly different from each other, especially at the family level, while some species in different orders may seem similar. It is difficult to formulate taxonomic descriptions of orders into mathematical or other descriptions easy for computer understanding. In fact, only one study has analyzed the math-morphological features of insects at the order level [33].

Here we present a new system which can identify insects to the order level. To meet the public need for practical insect image identification we collected insect images covering various species across several common orders. For automatic insect image identification at the order level we designed a simple and well-performed preprocess solution, defined a range of new features and compared two pattern recognition methods (artificial neural networks and support vector machines). We discuss our experiments and draw important conclusions on automatic insect image identification at the order level.

* Corresponding authors.
E-mail addresses: wangjn@ioz.ac.cn (J. Wang), linct@ioz.ac.cn (C. Lin), ji@ioz.ac.cn (L. Ji), liangap@ioz.ac.cn (A. Liang).

0950-7051/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2012.03.014

Table 1
The number of families, species and images in each order used in tests.

Orders (or sub-orders)	Families	Species	Images
Coleoptera	4	25	25
Hemiptera			
Auchenorrhyncha	5	25	25
Heteroptera	9	22	25
Hymenoptera	10	25	25
Lepidoptera	8	25	25
Megaloptera	2	25	25
Neuroptera	8	25	25
Odonata	8	24	25
Orthoptera	10	25	25
Total	64	221	225

2. Materials and methods

2.1. Images

There are 29–38 insect orders, depending on the taxonomic system applied to the Insecta [30]. Only less than half of these orders are common or easy to find. We collected 225 specimen images from nine orders or sub-orders: 25 images from each order. The list of orders and number of images are shown in Table 1. The well placed position of specimens in these images facilitates the automatic extraction of features. However, the quality of some specimens was not very good. For instance, several specimens were incomplete or attached to objects such as pins. These factors may disrupt the extraction of features, so we reduced such interference via manual image processing, for example, we manually removed the obvious attachments.

2.2. Realization of the system

Based on the theory of pattern recognition [20] and the basic processing pathways in typical automated species identification systems [6], we designed a system for insect image identification at the order level (Fig. 1). The "preprocess" and "extraction" modules are shared with both the training and recognition process. Features of training images will be used to build a model of the classification progress pattern after feature extraction, and the features and trained model will be recorded in files or database. The recognition progress is then used to compute the identification result of the "test image"; this process will use two types of data, the model in the database, and the features extracted from recognition files. The following sections provide implementation details for each step in Fig. 1.

We implemented our system on .Net platform according to Fig. 1. Therefore it must be running on the .Net Framework (version 4.0). The current desktop version of the system only uses XML files to record the features and pattern model.

2.3. Image preprocess

In our system the aim of preprocessing is to acquire images with a pure background-color. For example, Fig. 2 takes one image (Fig. 2A) and produces an image (Fig. 2B) with the background that we need. New image segmentation methods such as JSEG [4] and edge flow [18] are available, but are unstable. To avoid instability, we adopted the following two simple and effective methods to finish preprocessing.

We used manual and automatic methods to acquire a normalized image whose background was set with one pure color not appearing in the insect specimen. The whole image preprocess procedure is shown in Fig. 3. The manual method depends on a single image processing tool such as Photoshop, while the automatic method is more complex and combines a series of simple image processing steps. The automatic method is suitable for processing mass amounts of images in similar photograph environments, but cannot meet all images. The manual method is more precise but much slower than the automatic method.

2.4. Feature extraction

Feature extraction plays an important role in the final result of identification. When extracting features they should represent taxonomic information and be feasibly acquired from given images.

The features in our system are used for automatic insect identification at the order level, and they are different from the features referred to in other systems such as ABIS and DAISY. ABIS uses vein features to identify bees [24], which are unsuitable for orders such as Coleoptera and Heteroptera. DAISY adopts a principal component analysis (PCA) method to acquire image features that contain nearly all the information of an image. PCA features are more suitable for species identification because the great amount of detailed information collected with PCA can weaken differences between high-level categories such as orders. For the same reason, some local features based on SIFT (scale-invariant feature transform) [16] such as CFH (concatenated feature histogram) [13], BOW (bag of words) [28], ScSPM (sparse coding spatial pyramid matching) [17] that are currently used in insect species recognition are not quite suitable in high-level insect identification.

A series of geometrical features including area, perimeter, holes' number, eccentricity and roundness have previously been tested [29,32]. All these features are intuitive because they can be directly measured or simply calculated from images. However, these precisely extracted features are easily affected by factors such as the posture of insects and shooting angle. Furthermore, it is usually difficult to compute the real size of insects from the images because of the lack of some shooting parameters such as object distance. So based on the features introduced in BugVisux, we carefully compared every order to find features with taxonomic context and selected or created features that could be efficiently

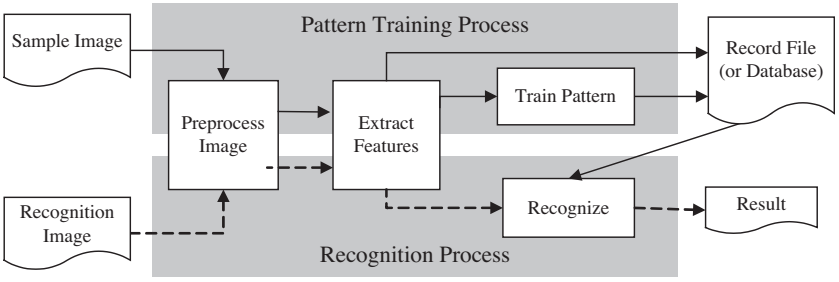

Fig. 1. Architecture of system.

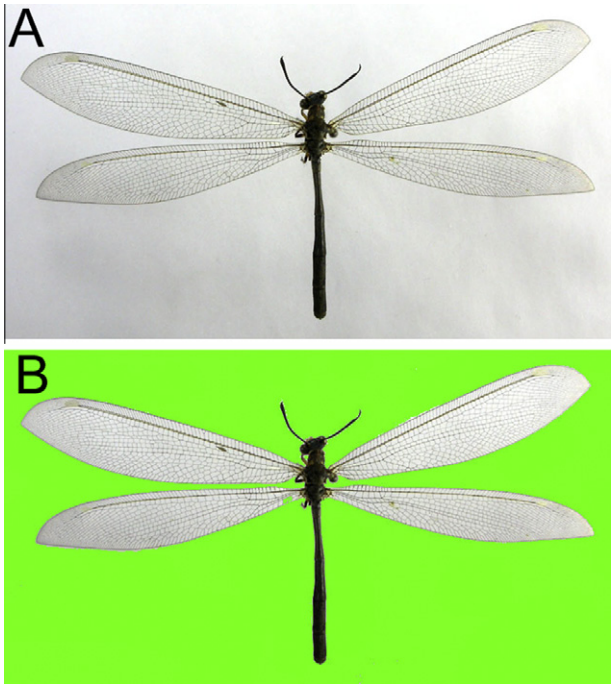

Fig. 2. Images produced in pre-processing. (A) Raw image, Neuropera, *Myrmeleon zanganus*, 800px × 443px. (B) Standard image with one color in background.

extracted with image processes. Before introducing the features we designed, several special terms should be defined:

(1) *Body*: in this paper, the body means the head, thorax (excluding the wings) and abdomen of the insect (the grey pixels within the white rectangle in Fig. 4). But if the wings are not outstretched, the body in the image will include wing.
(2) *Center of gravity:* the center of the insect (all grey pixels in Fig. 4), as shown in Fig. 4 by point *P*.
(3) *Upper part of body:* the upper part of body usually includes the head and chest of the insect, and is not exactly defined. In Fig. 4, the upper part of the body includes the grey pixels above the center of gravity (point *P*) in the body rectangle.

Fig. 4 also shows our new designing features introduced below:

(1) Body area ratio (*BAR*)

This feature is the ratio of the area of the body to the area of the whole insect, and can describe whether one individual has its wings outstretched. In Fig. 4, the body area ratio is the ratio of the number of grey pixels in *Re* (white rectangle) to all the grey pixels.

(2) Body eccentricity (*BE*)

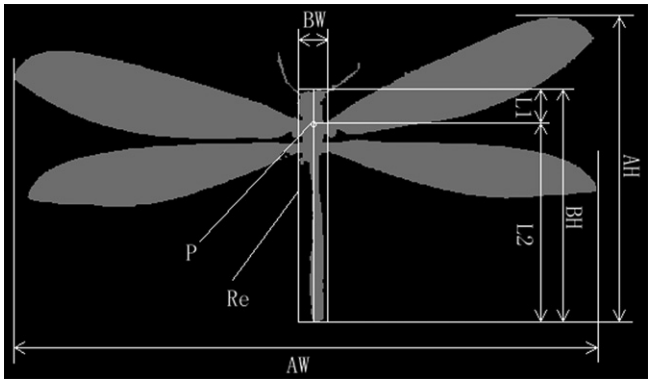

Fig. 4. Hints for features.

This feature describes the shape of the body of the insect and is defined as:

$$BE = BW/BH \tag{1}$$

BW and *BH* are shown in Fig. 4. *BW* is usually shorter than *BH*, but if *BW* is longer than *BH*, then *BE* will be set as 1.

(3) Upper body length ratio (*ULR*)

ULR is the ratio of the length of the upper part of the body to the length of the whole body (*BH*). As shown in Fig. 4, *ULR* is defined as:

$$ULR = L1/BH \tag{2}$$

(4) Width ratio (*WR*)
This feature is the ratio of the width of the insect (*AW*) to the sum of the height (*AH*) and width (*AW*) of the insect. *WR* is defined as:

$$WR = AW/(AH + AW) \tag{3}$$

(5) Upper body area ratio (*UAR*)
UAR is defined with respect to Fig. 4 as the ratio of the grey area above point *P* within the white rectangle to the entire grey area within the white rectangle. According to this definition, *UAR* describes the relationship of the head and chest area of an insect with the area of its tail.

(6) Body shape parameter (*BSP*)

We first define full body rows (*FBRs*) as rows that are longer than 95% of *BW* (the width of *Re*). Then *BSP* can be defined as the ratio of *FBR* and *BH* (the height of body rectangle).

(6) Color complexity (CC)

Color complexity describes the color diversity of an insect. It is defined as the ratio of the colors of the insect contained in the

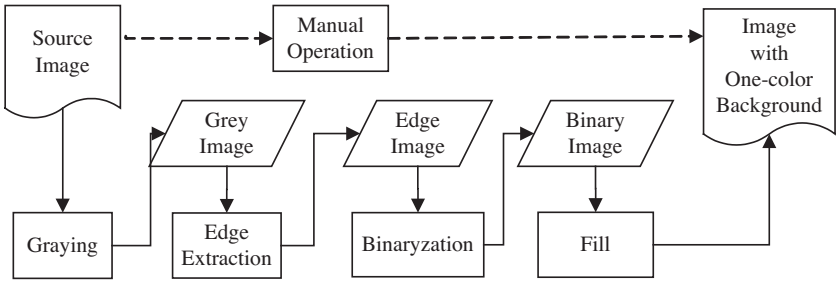

Fig. 3. Flow chart of image processing.

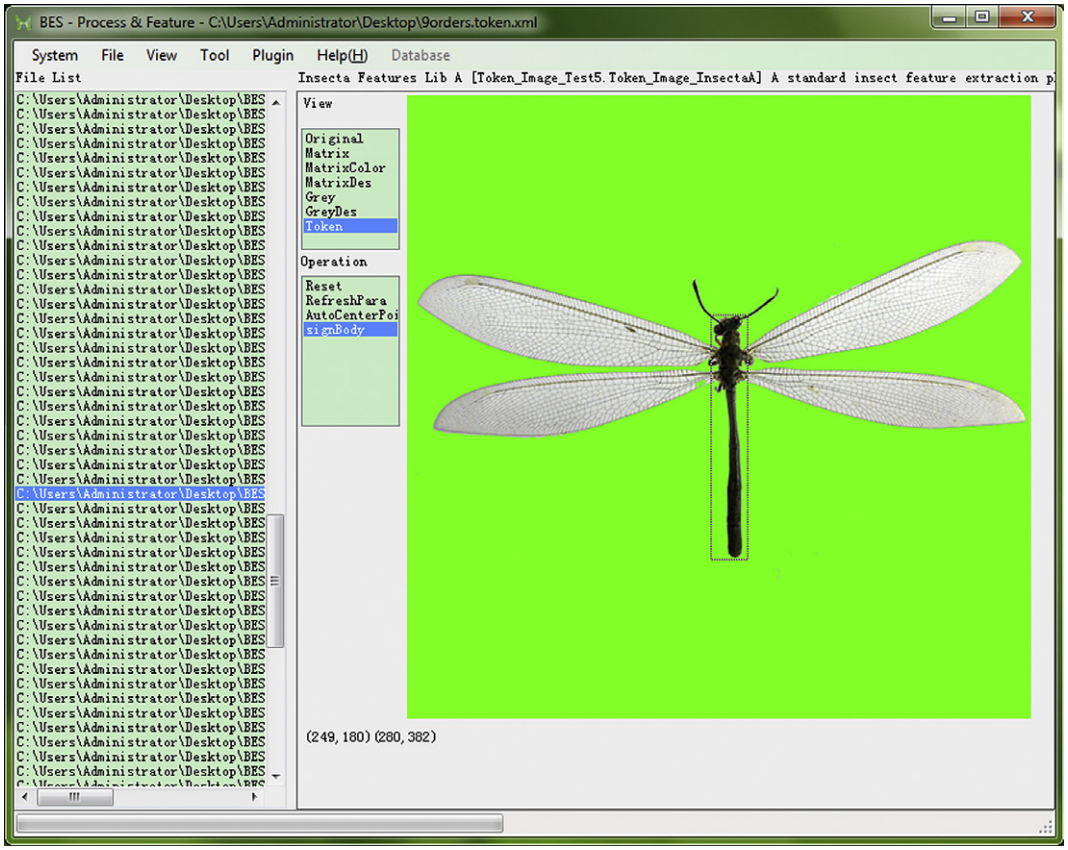

Fig. 5. Screen shot of the user interface of feature extraction in our system. The body rectangle can be both computed automatically (select "autoCenterPoint" in the list box) and selected manually (select "signBody" in the list box).

image to the colors of the color model. Here, we designed a 64-color RGB model (4 levels in each channel of R, G, and B) based on the standard RGB model (256 levels in each channel of R, G and B). This simple color model is more rough but also more useful than the standard model here as it divides R, G and B more simply.

Using the above seven ratio features has two advantages. Firstly, the features can express the structure of the insect body which is important for identification at the order level. Secondly, the results of the feature extraction is little affected by the image quality. Furthermore, our features were elaborately chosen to avoid using feature optimization methods like adapted fuzzy reasoning [12].

We designed and realized automatic extraction algorithms to compute the values of these features so that all variables and features can be calculated automatically. To correct the errors caused with automated method, we provided user interfaces for setting the key variables manually, which are integrated in the feature extraction module in the system (Fig. 5). For example, the body rectangle (*Re*) in Fig. 4 was corrected by manually selection operation.

2.5. Pattern recognition methods

Pattern recognition methods are now used in so many fields such as character recognition [1,23], face affective detect [9], leaf recognition [31] and hyper-spectral image classification [14]. In insect image identification, some of these methods such as the nearest neighbor classifier have been used [19]. We realized several methods in our system and chose two for experimental purposes, artificial neural networks (ANNs) and support vector machines (SVMs). ANNs and SVMs have been used in insect image identification [5,21], so we provide only a brief introduction here.

2.5.1. Artificial neural network

An artificial neural network is a mathematical model inspired by the structure of biological neural networks, and has been widely applied in many scientific fields [7]. There are quite a few common types of ANNs, such as feed forward network, self organizing map (SOM) and Hopfield network, these ANNs are optimized for different applications.

Here we implement a back-propagation neural network, a common type of supervised three layers (input layer, one hidden layer and output layer) artificial neural networks that adopts the back-propagation learning method to adjust the network. When training an ANN, several functions can be used as the core function in neural nodes, and here we chose the sigmoid function, which is the most common one:

$$f(x) = 1/(1 + e^{-x}) \tag{4}$$

We usually use the default or recommended values when choosing important parameters such as *alpha*, *moment* and *learning rate*. These important parameters will be given in the results section. The random initiation process of ANNs result in differences between two ANN models with the same parameters, therefore we did more than one test for each group of parameters.

2.5.2. Support vector machines

Support vector machines (SVMs) are a set of supervised learning methods that can be used for classification. They are based on the structural risk minimization principle from computational learning theory and are universal learners [25]. The standard SVM is a non-probabilistic binary linear classifier and can be used like three-layer sigmoid neural networks with the appropriate kernel function.

Table 2
Results of ANN tests on 9 orders with 7 features.

Accuracy	Images for training/recognition (percentage of recognition images)			
	117/108 (48%)	153/72 (32%)	171/54 (24%)	180/45 (20%)
Range	63–86%	68–93%	68–88%	64–86%
Average	75%	76%	76%	76%

C-Support Vector Classification (C-SVC) is one of the most popular types of SVMs and it was usually used with the radial basis kernel function (RBF) for classification. This kernel maps samples into a higher dimensional space nonlinearly, and the linear kernel is regarded as a special case of it [10]. There are two parameters (c, g) of the combination of C-SVC and RBF. C is the penalty parameter which balances the structural risk and the empiric risk, while g is the kernel parameter that defines the function range.

We used LibSVM [2] to analyze the feature data exported from our system. LibSVM (version 2.89) supplies several types of SVMs, and we employed the C-SVC formulation. It is important to find out good c and g base on training dataset before prediction, so we utilized the tool provided by LibSVM which was intended for parameters selection with the methods of grid-search and cross-validation to pick up a good (c, g) in our study.

3. Results

All features were extracted from images before tests according to the methods introduced in Section 2.4. We tested different orders, pattern recognition methods and features. The results of the main tests with different test policies are listed below along with some details and extra test results to be given in the discussion section.

3.1. Results based on the ANN method

In ANN experiments all 225 images (Table 1) were divided into two groups: images in the "training images" group were used for building the classification model; the "test images" group was used for the reorganization test on the existing model. We randomly selected some images from each category as training images at a fixed rate; leftover were used for the reorganization test. We did the experiment more than ten times for each combination of training and test images to observe the stability of ANN models.

When training an ANN model, we set parameters according to the following rules: (1) we set 0.1 as *learning rate*, 1.0 as *alpha*, 0.0 as *moment*, 100 000 as the least amount of training *steps*, and 0.00001 as *convergence error*; (2) we set the number of inputs to equal the number of features, double numbers of inputs as the nodes of hidden layer, and number of outputs to equal the number categories (orders or sub-orders).

We first selected all nine orders (or sub-orders, listed in Table 1) for testing. Table 2 provides the accuracy range and average accuracy value for each group of tests with seven features and different image divisions (48%, 32%, 24% and 20% as the test images). The results did not have a very high average accuracy, thus we selected eight orders at one time for tests to check whether some orders acutely influence the results. In these tests, we took the same parameters as the tests on all nine orders when building the ANN model (excluding the number of output neuron nodes, which equal the number of categories). We did nine groups of tests by removing one order in each group, the results of which are listed in Table 3. Each "Exclude order" cell in Table 3 represents the order that is not in the eight orders of that test group. Obviously, results from tests on eight orders were better than those on nine orders.

Table 3
Results of ANN tests with eight orders tests with seven features.

Exclude order	Accuracy	
	Range (%)	Average (%)
Auchenorrhyncha	65–100	89
Coleoptera	83–100	94
Heteroptera	78–97	91
Hymenoptera	77–100	89
Lepidoptera	73–98	88
Megaloptera	66–97	82
Neuroptera	66–97	85
Odonata	65–100	81
Orthoptera	66–95	82

3.2. Results based on the SVM method

We also used SVM to compare with ANN. In the experiments with SVM, we conducted 10-fold cross validation (which is commonly accepted [11]) tests for all images using the default configuration of LibSVM. The results are listed in Table 4. Additionally, the 'CV Rate' column is the best result of the cross validation training, and the 'accuracy' column is the result of all data on the last trained model. Like tests using the ANN method we selected different orders or features for SVM tests, which produced results that seem much better than those of the ANN tests.

4. Discussion

4.1. System advantages

According to suggestions from taxonomists and entomologists the processes of preprocessing and feature extraction should be completed both automatically and professionally. However, realizing these processes is difficult because the features that computers can automatically extract and those that taxonomists describe are often quite different. Our system tried to solve this problem by designing features with taxonomic characters that are also easy to acquire through a computer, and the results seem favorable.

Although the steps involved from preprocessing images to extracting features could be finished automatically, we supplied an interface for users to manually control each step. When designing these manual operations we attempted to create a convenient and effective system that resulted in precise feature extraction.

Our system performs with strong stability. Although we did ANN tests with random training images each time, the average accuracy for tests on nine orders was above 75% and the lowest accuracy was just above 63% (Table 2). Furthermore, results of ANN tests on eight orders shown in Table 3 were better than those on nine orders, and the SVM tests show the same difference (Table 4). These advantages mentioned above indicate that our system can identify the majority of images in our tests. They also confirm the feasibility of the idea of insect identification at the order level, therefore prove that this system could be used more widely in the future.

Table 4
Results of 10-fold cross validation with SVM (c-svc).

Tests and conditions	CV rate (%)	Accuracy (%)
Nine orders, six features (exclude CC)	81.78	95
Nine orders, seven features	84.44	92
Eight orders, seven features, without		
Auchenorrhyncha	83	98
Coleoptera	89.5	96
Heteroptera	88	98
Hymenoptera	82.0	97
Lepidoptera	84	95
Megaloptera	84	97
Neuroptera	85.5	100
Odonata	83	94
Orthoptera	84	91

4.2. Order relationships

The more categories the system contains, the poorer the test results are. The accuracies of nine-order test results (Table 2) were lower than those of eight orders (Table 3). Our earlier tests on Lepidoptera and Coleoptera actually produced perfect results (accuracy of 100%). Based on this evidence, we conclude that our test results could be improved if we split the nine-order identification system into several two-order systems. In its default configuration, LibSVM solves multi-class problems using several two-class problems, and our test results (Table 4) further prove this conclusion.

We selected one group of tests (Table 2) with the worst results (accuracy range 63–86%) to analyze the identification results of each image to find out how those orders may have influenced the test results (especially those of the ANN tests). The confusions matrix of Table 5 shows that some images of low-accuracy orders or suborders such as Auchenorrhyncha, Heteroptera and Hymenoptera were always wrongly identified as Coleoptera, and the images of Megaloptera and Neuroptera were easily confused with those of Lepidoptera. Then we selected some of the images that were always not identified (Figs. 6–8). Fig. 6 shows the first situation, the images in Heteroptera that are always wrongly identified. We found that some species indeed look like beetles in Coleoptera, especially regarding those shape features described in Fig. 6. Meanwhile, it is evident that some members in Heteroptera are quite different from each other. The situation concerning Auchenorrhyncha is different from that of Heteroptera: most species of Auchenorrhyncha in tests look like *Formotosena seebohmi* (Fig. 7C) and only 3–4 images such as Fig. 7A and B look more like a species of Coleoptera or Orthoptera than Fig. 7C. Images of Hymenoptera pose the third situation: two classes of species like Fig. 8A and B both have quite a few images that all seem entirely different from each other. Therefore it is difficult for classifiers such as ANN and SVM to compute and provide a single pattern for this order containing two diverse classes.

Biodiversity and innumerable large number of species result in inter-class similarities and intra-class differences which commonly occur in insect order categories. Natural phenomena bring more difficulties for insect identification at the order level than that at the species level. From the analysis of Figs. 6 and 7 and the results of Table 5, we found that this nature law is just the main factor causing those identification errors described above. The ideal method to solve this problem is to collect all species from each order, but it is impossible to implement because currently there are immeasurable number of unknown species. Actually, it is most often found that the similarities in species within a single family and the differences among species of separate families are prominent enough to identification. So if we could collect enough images that delegate all families for each order, the system would be much more improved. Obviously realizing this aim would be much easier than collecting all species in existence.

For future tasks, especially during data collection and tests, greater focus on the family coverage of each order is needed. More attention should also be given to similar orders, and to extra images for supplementing the data of orders.

4.3. Advantages and disadvantages of features

Stable and high-accuracy test results proved that the features we designed are effective enough to differentiate the majority image members of orders in our system.

Features play different roles in identification. We conducted experiments for each feature (or feature combinations) separately to observe their respective performance (Table 6). We found that all the results from tests with only one feature were very low, but the produced accuracy order results indicated that features could be ordered by importance. According to Table 6, we can order the features as *ULR*, *BAR*, *BE*, *UAR* and *WR*|*BSP*|*CC*. Although we can put features in order of importance, each of them is essential, since we found that the fusion of all seven features led to the best result, and the results of tests with only *ULR* (*BAR*, *BE* or *UAR*) were poor.

After analyzing the order of feature importance we found that the four most important features (*BAR*, *BE*, *ULR* and *UAR*) are all related to the body of the insect. These features are designed to reflect the relationship between the body and the whole insect, connecting the microfeature (from body) with macrofeature (from the whole insect), an important principle for research of feature extraction for insect identification. In insect taxonomy a key factor is the final ratio of the body to the wings, so features based on this principle are supported by taxonomy and have evolutionary grounding. However, features used in our tests only take into

Table 5
Confusion matrix of detailed results of 15 tests on nine orders with seven features (using 117 images for training, whose accuracy range is 63–86% in Table 2).

Order (sub-order) name	Actual orders									
	ID	1	2	3	4	5	6	7	8	9
Auchenorrhyncha	1	145	0	0	1	0	0	0	0	0
Coleoptera	2	34	180	136	93	0	0	5	32	10
Heteroptera	3	0	1	36	1	0	0	0	0	0
Hymenoptera	4	1	0	1	74	0	0	0	0	0
Lepidoptera	5	0	0	0	11	180	31	40	0	3
Megaloptera	6	0	0	0	0	0	147	0	0	0
Neuroptera	7	0	0	0	0	0	2	130	0	0
Odonata	8	0	0	0	0	0	0	5	148	0
Orthoptera	9	0	0	7	0	0	0	0	0	173
Total of test images		180	180	180	180	180	180	180	180	180
Error identified times		35	0	144	106	0	33	20	32	3
Error identified images		6	0	24	24	0	8	10	17	1

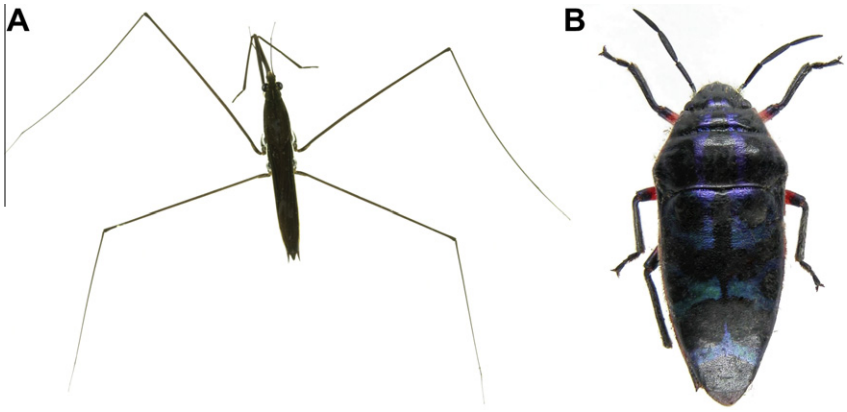

Fig. 6. Images always wrongly identified in Heteroptera. (A) Unknown species in Gerridae. (B) *Scutellera fasciata*.

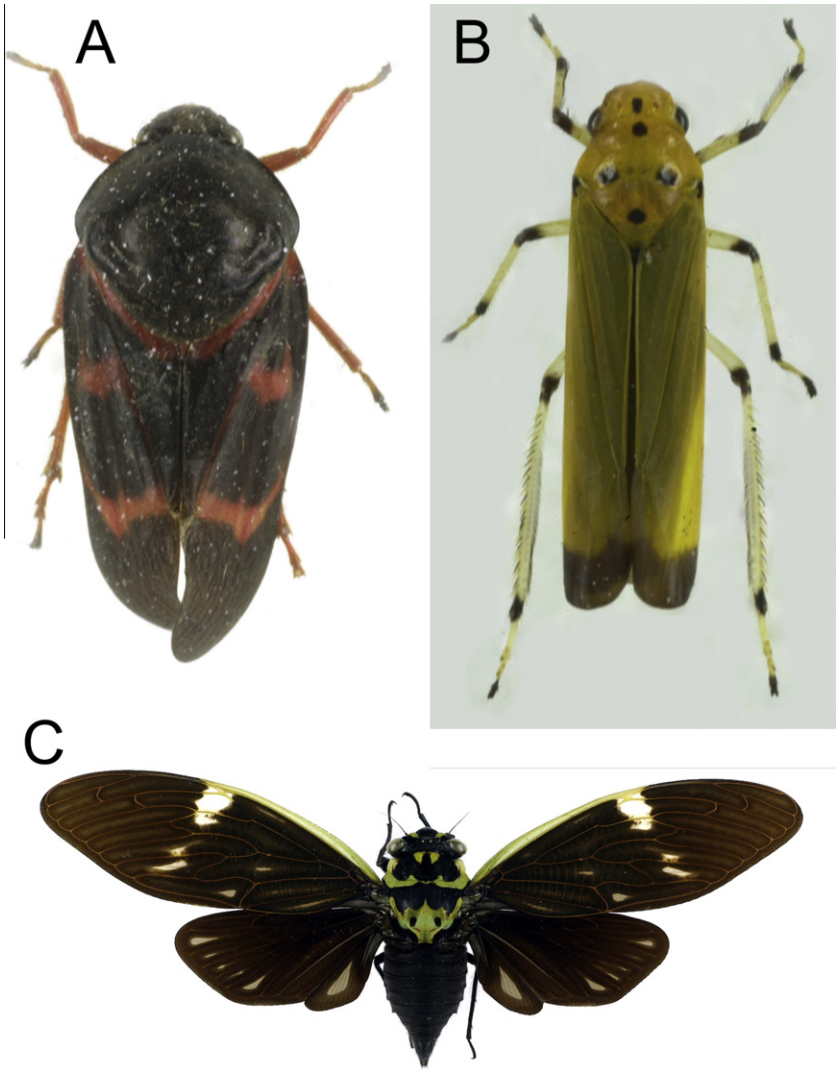

Fig. 7. Images always wrongly identified in Auchenorrhyncha. (A) *Cosmoscarta uchidae*. (B) *Nephotettix bipunctatus*. (C) *Formotosena seebohmi*.

account some parts of the insect, so more effective features are required.

Huang has tested many kinds of popular features on insect identification at the species level, some of which worked [8]. We also tried a majority of those features but they did not produce favorable results during our identification tests. This was mainly caused by the difference of test datasets and the identification level. Thus, any new features should be considered with both the identification aim and the objects' identification characters. Taking insect order identification for example, we should select features in accordance with those described in taxonomy.

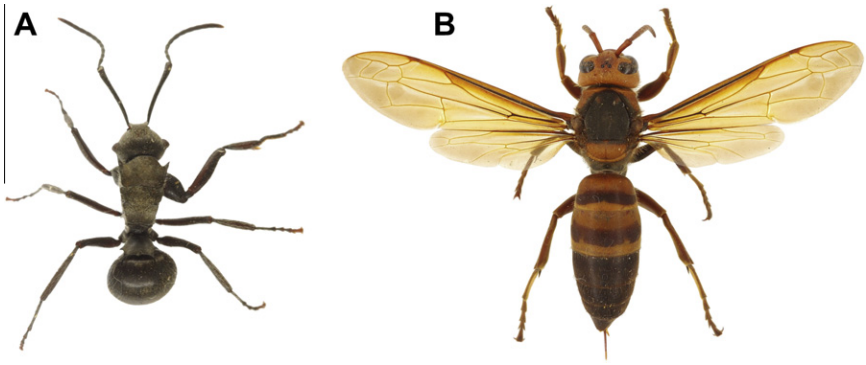

Fig. 8. Images always wrongly identified in Hymenoptera. (A) an unknown species in Formicidae. (B) *Vespa ducalis*.

Table 6
Results of ANN tests on nine orders with different features (accuracy is the average accuracy of each group of tests and the full names of features are listed in Section 2.4).

Feature abbreviation	Accuracy (%)
BAR	16
BE	13
ULR	22
WR	11
UAR	15
BSP	11
CC	11
BAR, BE, ULR, UAR	48
BAR, BE, ULR, WR, UAR	66
BAR, BE, WR, UAR, BSP	58
BAR, BE, ULR, WR, UAR, BSP	67
BAR, BE, ULR, WR, UAR, BSP, CC	76

Automatic feature extraction is also an important topic. During the long process of feature design, we found that the key to realizing the automatic feature extraction process was to find stable and attainable points of all images. Center of gravity is an example of this, and some of our other features are based on this point. Because there are only a few such points, some better systems used today are semi-automatic.

4.4. About pattern recognition methods

The ANN method is a black-box model, meaning it is hard to explain the result using biological theory. Unlike the ANN method, the SVM method is based on statistical theory, allowing some results to be supported with statistical rules. However, we discuss ANN more fully because the differences among the test results using the SVM are too insignificant and unsuitable for deep analysis.

We note that the average accuracies of tests on different divisions of the training image set barely differentiated from each other, such as the tests with six features in Table 2: all average accuracies are around 67%. From this we conclude that the division of the training image set has little influence on the ANN test result, which is also strongly supported by our other tests. We can therefore do tests without concern for the division of image training sets and focus on other aspects such as the efficiency of features.

The SVM performed better than the ANN, especially in tests on nine orders. Despite changing the parameters of ANN to improve results, the results of the SVM were still better. We conducted 10-fold cross validations on each test, and the SVM selects best results. However, the best test results using the ANN are similar when using the SVM.

Both of these two pattern recognition methods produced fairly good results in tests with seven features. According to their performance in Table 2 and 6, it would be most useful to use an ANN to test new features and choose a SVM when deploying the system to real applications. Actually, related researches also show that the SVM classifier is better than some other classifiers such as 1-NN [31].

5. Conclusions

In this paper we present a new system focusing on insect image identification at the order level. We designed seven features following basic geometrical features for automatic identification as well as insect taxonomy and morphology. We conducted various experiments and found that this system performed with good stability at the order level and resulted in good user experiences.

Parts of members in each some order were always confused with other orders, thus more research should be done on specific features of those orders of concern. The ANN performed with good stability but the SVM results were better. According to these conclusions we can improve this insect order identification system by focusing on feature extraction and designing newer and more effective features from the insect order. Although our system is able to achieve automatic insect order identification of orders at a small scale we have to test it on a dataset with more categories of images before it is fully rolled out.

The data and software in this paper can be downloaded from http://159.226.67.82/pubs.htm.

Acknowledgements

This work was supported by the National Basic Research Program of China (973 Program) (Grant no. 2011CB302102) and the National Natural Science Foundation of China (Grant nos. 30970400, 31172128) (all awarded to AL). We wish to thank the Agriculture Digital Museum of China Agriculture University, Insect Museum of Taiwan University, National Zoological Museum of China and Forum of Insect Fans for providing images used in this study. We are grateful to Huijie Qiao, Bengui Xie and Ji Yang for advice on this research and suggestions in laboratory work.

References

[1] J.H. AlKhateeb, O. Pauplin, J. Ren, J. Jiang, Performance of hidden Markov model and dynamic Bayesian network classifiers on handwritten Arabic word recognition, Knowledge-Based Systems 24 (2011) 680–688.
[2] C.C. Chang, C.J. Lin. 2001. LIBSVM: a library for support vector machines <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
[3] X.L. Chen, X.W. Hou, C.L. Liu, X.Q. Liu, Z.B. Zhang, Advances in the automated insect image identification, Chinese Bulletin of Entomology 45 (2008) 317–322.
[4] Y.N. Deng, B.S. Manjunath, Unsupervised segmentation of color-texture regions in images and video, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (2001) 800–810.

[5] M.T. Do, J.M. Harp, K.C. Norris, A test of a pattern recognition system for identification of spiders, Bulletin of Entomological Research 89 (1999) 217–224.
[6] K.J. Gaston, M.A. O'Neill, Automated species identification: why not?, Philosophical Transactions: Biological Sciences 359 (2004) 655–667
[7] M.T. Hagan, H.B. Demuth, M.H. Be ale, Neural Network Design, PWS Publishing Company, 1995.
[8] S.G. Huang, Research on the key techniques of image-based insects recognition. Nothwest University, Xi'an, 2008.
[9] K.A. Hwang, C.H. Yang, Assessment of affective state in distance learning based on image detection by using fuzzy fusion, Knowledge-Based Systems 22 (2009) 256–260.
[10] S.S. Keerthi, C.J. Lin, Asymptotic behaviors of support vector machines with Gaussian kernel, Neural Computation 15 (2003) 667–1689.
[11] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: International Joint Conference on Artificial Intelligence, 1995, pp. 1137–1145.
[12] L. Lancieri, L. Boubchir, Using multiple uncertain examples and adaptative fuzzy reasoning to optimize image characterization, Knowledge-Based Systems 20 (2007) 266–276.
[13] N. Larios, H. Deng, W. Zhang, M. Sarpola, J. Yuen, R. Paasch, A. Moldenke, D.A. Lytle, S.R. Correa, E.N. Mortensen, L.G. Shapiro, T.G. Dietterich, Automated insect identification through concatenated histograms of local appearance features: feature vector generation and region detection for deformable objects, Machine Vision and Applications 19 (2008) 105–123.
[14] S. Li, H. Wu, D. Wan, J. Zhu, An effective feature selection method for hyperspectral image classification based on genetic algorithm and support vector machine, Knowledge-Based Systems 24 (2011) 40–48.
[15] F. Liu, Z.-R. Shen, J.-W. Zhang, H.-Z. Yang, Automatic insect identification based on color characters, Chinese Bulletin of Entomology 45 (2008) 150–153.
[16] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2004) 91–110.
[17] A. Lu, X. Hou, X. Chen, C. Liu, Insect species recognition using sparse representation, in: 21th British Machine Vision Conference, 2010, pp. 1–10.
[18] W.Y. Ma, B. Manjunath, EdgeFlow: a technique for boundary detection and image segmentation, IEEE Transaction on Image Processing 9 (2002) 1375–1388.
[19] N. MacLeod, Automated taxon Identification in Systematics: Theory, Approaches and Applications, CRC Press, 2007.
[20] J.P. Marques de Sá, Pattern Recognition: Concepts, Methods and Applications, Springer, 2001.
[21] M. Mayo, A.T. Watson, Automatic species identification of live moths, Knowledge-Based Systems 20 (2007) 195–202.
[22] M.A. O'Neill, I.D. Gauld, K.J. Gaston, P.J.D. Weeks, Daisy: an automated invertebrate identification system using holistic vision techniques, in: Inaugural Meeting of the BioNET-International Group for Computer-aided Taxonomy, 2000, pp. 13–22.
[23] M.I. Razzak, F. Anwar, S.A. Husain, A. Belaid, M. Sher, HMM and fuzzy logic: a hybrid approach for online Urdu script-based languages' character recognition, Knowledge-Based Systems 23 (2010) 914–923.
[24] S. Schröder, W. Drescher, V. Steinhage, B. Kastenholz, An automated method for the identification of bee species (Hymenoptera: Apoidea), in: Proceedings of the International Symposium on Conserving Europe's Bees, International Bee Research Association & Linnean Society, 1995, pp. 6–7.
[25] V.N. Vapnik, The Nature of Statistical Learning Theory, second ed., Springer Verlag, 2000.
[26] A.T. Watson, M.A. O'Neill, I.J. Kitching, Automated identification of live moths (macrolepidoptera) using Digital Automated Identification SYstem (DAISY), Systematics and Biodiversity 1 (2004) 287–300.
[27] P.J.D. Weeks, I.D. Gauld, K.J. Gaston, M.A. O'Neill, Automating the identification of insects: a new solution to an old problem, Bulletin of Entomological Research 87 (1997) 203–211.
[28] C. Wen, D.E. Guyer, W. Li, Local feature-based identification and classification for orchard insects, Biosystems Engineering 104 (2009) 299–307.
[29] X.W. Yu, Z.R. Shen, S. Ninomiya, Measuring geometrical features of insect specimens using image analysis, in: Proceedings of the Third Asian Conference for Information Technology in Agriculture, 2002, pp. 591–595.
[30] F. Yuan, X.Q. Yuan, Research advances on phylogeny of hexapoda with a new classification system, Entomotaxonomia 28 (2006) 1–12.
[31] S. Zhang, Y.-K. Lei, Y.-H. Wu, Semi-supervised locally discriminant projection for classification and recognition, Knowledge-Based Systems 24 (2011) 341–346.
[32] H.Q. Zhao, Z.R. Shen, X.W. Yu, On computer-aided insect identification through math-morphology features, Journal of China Agricultural University 7 (2002) 38–42.
[33] H.Q. Zhao, Z.R. Shen, X.W. Yu, Use of math-morphological features in insect taxonomy. I. At the order level, Acta Entomologica Sinica 46 (2003) 45–50.