White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of York.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/75118/

Version: Published Version

Conference or Workshop Item:
Ayob, Mohd Zaki Bin and Chesmore, David (2012) Hybrid feature extractor for harlequin ladybird identification using color images. In: Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 09-12 May 2012.

https://doi.org/10.1109/CIBCB.2012.6217233

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

Hybrid Feature Extractor for Harlequin Ladybird Identification Using Color Images

M. Z. Ayob
UniKL-BMI
Batu 8 Jalan Sg. Pusu Gombak
53100 Selangor Malaysia
mohdzaki@bmi.unikl.edu.my

E.D. Chesmore
Department of Electronics
University of York
YO10 5DD UK
david.chesmore@york.ac.uk

***Abstract*—This paper describes research into the automation of the identification of harlequin and other ladybird species using color images. The automation process involves image processing and the use of an artificial neural network as a classifier. The ultimate aim is to reduce the number of color images to be examined by an expert by pre-sorting the images into correct, questionable and incorrect species. The ladybirds are 3-dimensional and the images have variable resolution. CIELAB has been useful as the color space in this research, as it provides good separation of chroma components from luminance on a color plane. Two major sets of features have been extracted from ladybird images: color and geometrical measurements. The system combination consisted of J48 decision trees which were used to filter out unnecessary features, and multilayer perceptron which was used for classification. Trials using ladybird images showed 92% class match for the species *Harmonia axyridis f. spectabilis* against *Exochomus 4-pustulatus*. The identification results are rotation and translation invariant. The methods allow quantitative data on both intra-species and inter-species variation for biodiversity studies.**

Keywords-color images; CIELAB; decision tree; multilayer perceptron; automated species identification

I. INTRODUCTION

Ladybirds are small and colorful insects which are commonly found in backyards and gardens. Ladybirds are beetles (Order: Coleoptera, family: Coccinellidae), and called ladybugs in the USA. Many species of ladybirds exist, in the UK alone there are over 46 species of ladybirds. Unfortunately, among these beetles are invasive species which present a threat to the ecological balance of ladybirds. One such species is the Harlequin ladybird (*Harmonia axyridis*), which is known to have invaded the UK since 2004 [1].

This alien species was introduced in Europe as a biological control agent [2], [3], [4], [5], [6]. In the USA it has been introduced since early 20th century [7]. It is a voracious ladybird species that feeds on aphids. Whenever aphids are scarce the Harlequin may attack the larvae of ladybird species. Harlequins over-winter in large groups to hibernate within sheds, attics and parts of buildings where the locations are dry and protected. When they are disturbed, they emit a foul secretion to deter predators, causing domestic and health issues like stained fabrics and skin irritations. The beetle has now spread to most parts of England and requires considerable attention due to its impact on ecological and biological balance through resource competition and intra guild predation [8]. To make things worse, there has been a general decline in the taxonomic workforce which in effect has been part of the taxonomic impediment to biodiversity studies [9], [10]. This taxonomic impediment will become serious unless solutions are explored to rectify it, and taxonomists need to work together with specialists in pattern recognition for getting more accurate and rapid results [11], [12]. Apart from constraints on human resources, the number of specimens will need to be very large in order to perform routine species identifications [13]. Partly in response to this, ladybird surveys were setup and launched in the UK. The Harlequin Survey and the UK Ladybird Survey are examples of web-based recording systems for the general public to send in their sightings and photographs [14a]. These color images will enable the physical details of the ladybird forewings (called 'elytra') and details of the thorax to be visually examined.

The use of automated identification of ladybirds may have enormous potential and has not been previously explored. In spite of the advantages, automating the identification process is not a trivial task. During a routine identification task an expert will manually obtain details from ladybird's elytra (and other features) for species identification. Dichotomous keys used are based on morphological criteria, such as color, shape and size. These are too small to examine and laborious considering some ladybird species have variety of color forms and spot patterns. Both intra-species and inter-species variations can be large, making species identification a difficult skill to master. There are many forms of *H. axyridis*; however, only three different forms are studied in this paper as shown in Fig. 1.

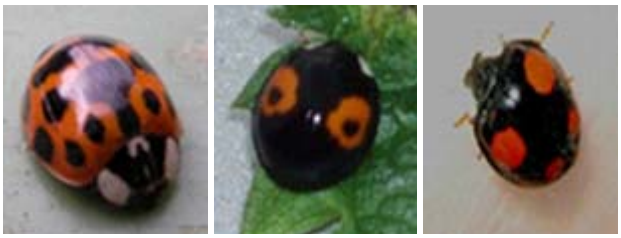

Fig. 1. Three different forms of Harmonia axyridis under study, from left to right: (a) form succinea (black spots & red-orange elytra) (b) form conspicua (pair of bull's eye spots & black elytra) and (c) form spectabilis (red-orange spots & black elytra)

978-1-4673-1191-5/12/$31.00 ©2012 IEEE

Table I describes an identification guide for harlequin ladybirds in the UK [15].

TABLE I
IDENTIFICATION GUIDE FOR *H. AXYRIDIS F. SUCCINEA, F. CONSPICUA* AND *F. SPECTABILIS*

Character	*Form succinea*	*Form conspicua*	*Form spectabilis*
Color	Orange with 0 to 21 black spots	Black with 2 red/orange spots, and inner black spots	Black with 4 red/orange spots
Shape	Round, domed	Round, domed	Round, domed
Size	6-8 mm	6-8 mm	6-8 mm

One of the objectives of this research is to get specific features that would make harlequin ladybirds identifiable through automation. These features are simplified through the use of decision trees. Once image processing and classification have been performed, species identification can then be enhanced using 'local' information such as species distribution, biogeography, etc.

The rest of the paper is organized in four sections; a literature review on similar work in insect identification is discussed next, section III describes the image processing techniques and artificial neural networks (ANNs) and presents results. The paper then concludes with recommendations for future work.

II. BACKGROUND

Most of the research to date in automated identification has concentrated on the use of image processing and artificial neural networks. An image analysis system was developed for use on the diagnosis of the teliospores of *Tilletia indica* (Karnal bunt) [16]. The image-processing automatically locates spores on a given image and calculates morphological characters. Neural networks have long been established to perform rapid reliable identification on suitable data sets [17]. Image-based automated identification systems such as DAISY (digital automated identification system) and ABIS (automated bee identification system) [18] blend human expertise in taxonomy and computing power to improve insect identification. They use image processing, statistical analysis and/or intelligent systems. ABIS focuses on identifying bee using linear discriminant analysis. DAISY is a holistic system and uses a plastic self-organizing map (PSOM) as the classifier and attempts to identify every possible species [19]. Dai and Chesmore have used semi-automatic wing venation descriptions using multilayer perceptron (MLP) and learning vector quantization (LVQ) for the identification of flying insects (Diptera and Hymenoptera). The system achieved 90.9% accuracy for hoverflies and 95.6% for bumblebees using LVQ, while getting only 60% for hoverflies and 30% for bumblebees using multilayer perceptron (MLP) [20]. While these systems have been successful, there exists plenty of room for further improvement.

The system described here is the first known system for identifying ladybirds. It has been developed and modified from the concept of a generic automated taxon identification system proposed by Chesmore [21]. The system obtains geometric and color data from macro images of ladybirds. The features are then used by an artificial neural network acting as a classifier. The outputs are three categories: whether a ladybird is a harlequin, non-harlequin, or unknown. For the latter case, human expertise will be required to do manual sorting, bearing in mind that it is not the system intention to replace human expertise as it sorts the input images into categories. It is meant for utilizing the computing power available to simplify identification, hence reducing fatigue and many errors imposed by human abilities when compared to machines. It also substantially reduces the number of images the expert has to identify.

III. METHODOLOGY

Fig. 2 shows the block diagram of the system.

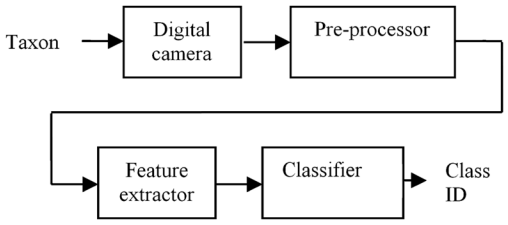

Fig. 2. Block diagram of automated ladybird identification system (ALIS)

Raw images were provided by the Centre for Ecology & Hydrology (CEH), Wallingford, England, which have been supplied by members of the public via a web site. The color images are fed into a pre-processing stage. At present backgrounds are subtracted manually and the size adjusted to 640x480 pixels. Next, the image is average filtered and then de-correlation stretched, when necessary. Thresholding produces a binary image. Specific markings on the insect body such as spots should appear significantly distinct from noise through correct selection of the threshold since the spots are usually significantly larger. Then, morphological operations such as dilation and opening are performed to produce a cleaner binary image. All the image processing steps previously mentioned are used to obtain the binary version of the original color image, so that proper measurement of geometrical features can be performed. The primary geometrical features are area, perimeter, major axis length and minor axis length of the insect. The derived geometrical features are area ratio and aspect ratio. Two sets of geometrical measurements were taken per sample ladybird; one for elytra and another for spots, if there is any.

Another vital feature in use is color. Color is selected due to its ability to naturally represent the natural characteristics of the ladybird, apart from the number of spots. For instance, *H. axyridis f. succinea* is fairly easy to identify due to its orange elytra color and 19 black spots. In contrast, *H. axyridis f. spectabilis* can be identified through the number of spots, spot color and elytra color. However, it could still be confused with a pine ladybird (*Exochomus 4-pustulatus*), as their colors are

similar. In this case, physical measurements are employed to confirm the identity.

The color space to represent pixel colors of both the spots and the elytra is CIELAB (or CIEL*a*b*). CIELAB is an approximate uniform color scale to represent visual difference in the form of color plane [22]. CIELAB is known for its ability to represent color separately from luminance [23]. It is also able to represent the chroma values in the form of color plane, therefore permitting visualization of the 'clusters' of colors. This is evident from a GretagMacbeth color chart converted into CIELAB color plane [24]. CIELAB is used since the majority of images are ladybirds photographed in natural scenes, not in the laboratory. These ladybird photographs were taken out door in various illumination settings, which is indeed a variable that is very difficult to control. The spot color and the elytra color values are taken per pixel and averaged over a 'pixel capture box'. A fixed size box was not used due to the limitations on the actual image resolution. This has been based on tests using ladybird images of low resolution, where some level of magnification made border pixels indistinct and blurred, hence affecting the actual CIELAB values after averaging. Fig. 3 shows CIELAB color planes for some ladybird species. Both x and y axes are normalized values of a^* and b^*.

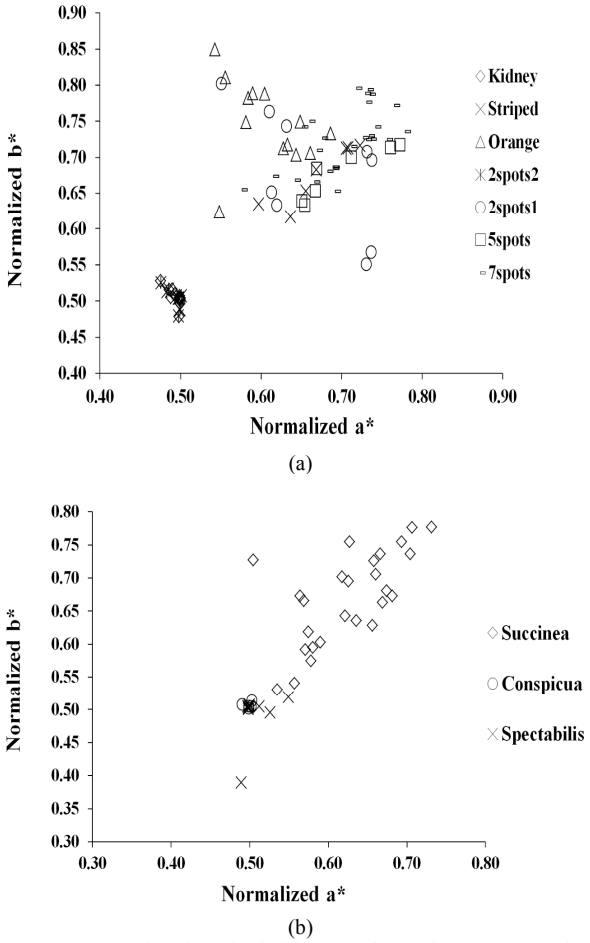

Fig 3. CIELAB color plane for ladybirds under study: (a) UK species (b) Invasive species (Harlequin forms)

The normalization was based on the following formulae:

$$\text{Normalized } L^* = (L^*/100) \quad (1)$$
$$\text{Normalized } a^* = (a^*+120 / 240) \quad (2)$$
$$\text{Normalized } b^* = (b^*+120 / 240) \quad (3)$$

Fig. 3(a) depicts the 'clusters' of normalized chroma values for some UK species, and Fig. 3(b) shows the distribution of chroma values for Harlequin ladybirds (*form succinea, form conspicua and form spectabilis*). The graph tells that a reduced feature space in terms of chroma values have been obtained from the original three-dimensional RGB image. From a design perspective, this information is vital as future baseline reference as it would be useful when utilized as part of the knowledge base of an expert system.

Training data was obtained after successive operations of image pre-processing and feature extraction on all ladybird images, as summarized in Table II.

TABLE II
STEPS TO OBTAIN TRAINING DATASET

Step 1	Read input image and average filtered
Step 2	Crop region of interest
Step 3	Perform morphological operations to minimize noise and unwanted segments
Step 4	Get geometrical measurements as the first feature set
Step 5	RGB to CIELAB conversion
Step 6	Get a* and b* for spots and base region (the second feature set)
Step 7	Normalize both feature sets
Step 8	Repeat for next input

A total of 12 normalized features have been selected, as listed in Table III. Both chroma channels (a^* and b^*) were used in addition to the hue angle.

TABLE III
FEATURES

Geometrical features	Color features
Spot area	Spot color a*
Spot perimeter	Spot color b*
Spot max axis length	Spot hue angle
Spot min axis length	Elytra color a*
Spot area ratio	Elytra color b*
Spot aspect ratio	Elytra hue angle

The images were tested for compliance to scale and rotation invariance. The features were normalized to [-1,1]. It is worth mentioning that no image scaling has been implemented, hence the size of the insects is not known before image processing. This is an advantage since users supplied images generally have no scale. All the spots' physical features were normalized against their corresponding values for the insect's body.

To provide an unbiased estimate of classification accuracy, the data was randomly sampled and stratified 10-fold cross validation was used. From the 12 features, further feature reductions were applied in order to minimize complexity. The

more features will mean more hyper planes constructed in the feature space; hence this may make identification difficult. For this application there could be instances where only a small number of samples has been gathered, thus potentially causing difficulty for the following stages. There are many choices of statistical method to deal with reducing dimensionality; the most popular ones are principal component analysis (PCA), singular value decomposition (SVD), independent component analysis (ICA), Fisher's linear discriminant analysis (LDA) and kernel PCA. In light of the low number of samples, pruned decision tree has been used for deriving a reduced number of features, using J48 developed under WEKA 3.6.4 environment [25]. J48 is an open source Java implementation of the C4.5 algorithm in WEKA. C4.5 was derived from ID. Both are Ross Quinlan's algorithms for generating classification models, better known as decision trees [26]. The system is also tested using PCA as feature reducer to stand as a comparison of fitness to the problem. This is highlighted in the next section.

The design involved two phases. The first phase is the selection of features when subjecting *H. axyridis f. spectabilis* against known negative samples (local species) using all 12 features; apply J48 to generate pruned decision tree and saving the resultant features. The first phase covered the following species: *Adalia 2-punctata* (2-spot), *Coccinella 5-punctata* (5-spot), *Coccinella 7-punctata* (7-spot) and *E. 4-pustulatus* (pine ladybird). The first phase involved all four species been identified pairwise, for instance, training set A contains samples from *H. axyridis f. spectabilis* against *E. 4-pustulatus*, etc. This makes the binary identification possible, if not easier to manage. Some examples of the local ladybird images are shown in Fig. 4.

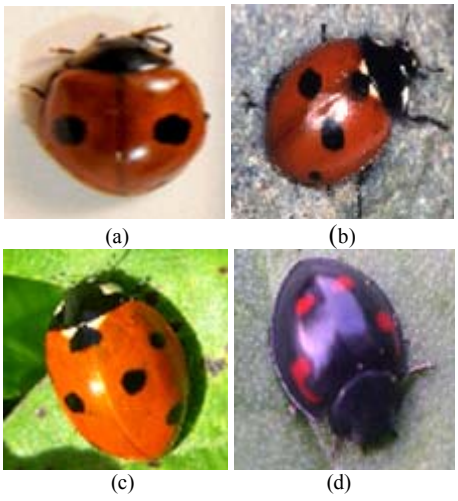

Fig. 4. Images of (a) *A. 2-punctata* (2-spot) (b) *C. 5-punctata* (5-spot) (c) *C. 7-punctata* (7-spot) (d) E. 4-pustulatus (Pine)

Learning rate for MLP was 0.3, the neurons used sigmoid activation function, and momentum was 0.2. The 'resultant' features were then supplied to the MLP utilizing back propagation algorithm (BP) for classification. As the system involves binary (2-class) pattern classification at a time, a One-against-one (OAO) model was implemented [27]. Only one hidden layer was used, as increasing the number of hidden layers is not economically feasible in terms of computation time and generalization ability of the network [28]. The number of hidden neurons has been varied to find the optimal learning curve, and it has been found by experiment. Consider K features and c output classes. The number of neurons at the hidden layer will equal $M = cK$ neurons [29]. The number of neurons for the output layer was varied depending on the number of target classes.

The second phase is the test when using the reduced features for testing against sample data taken from known positive samples (same species but different forms). The results from BP are then analyzed to determine the success rate of classification. One of the objectives of the research is to determine how much intra-species misclassification would occur and what features contribute to the misclassification errors. Confusion matrix and ROC curves can be used for this purpose, whereby the accuracy is obtained by the following formulae [30]:

$$\text{Accuracy} = 1 - \text{error} = (Tp+Tn) / (Tp+Fp+Tn+Fn) \qquad (4)$$

where

Tp = true positive rate
Tn = true negative rate
Fp = false positive rate
Fn = false negative rate

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The 12 features have been narrowed down to lower number of features, most of the time to single feature only. This has been achieved through J48 decision tree involving H. axyridis f. spectabilis (first phase), as shown in Fig. 5.

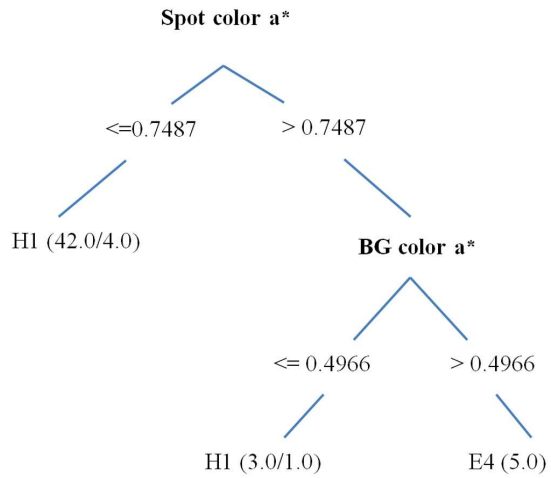

Fig. 5. Example of J48 pruned decision tree

The pruned tree shows which features and the minimum number of features required to separate the two classes. In this example the test was to identify between *E. 4-pustulatus* (pine ladybird) and *H. axyridis f. spectabilis*. It also shows the threshold values for decision at each node, which makes statistical inference possible without delving into the

mathematical rigor. At the terminal nodes, for instance, the lower right box contains the label E4 for *E. 4-pustulatus* as a class, and numbers in brackets indicate the correct classifications and the apparent misclassifications. For the tests conducted in the first phase, the corresponding feature selection against other 'negative' species is shown in Table IV.

Results from Table IV show that *H. axyridis f. spectabilis* and *E. 4-pustulatus* can be identified with reduced features by applying the J48 algorithm. The system is able to discriminate the two species based on the geometric features better than using color features. However, the identification of other species (*A. 2-punctata*, *C. 5-punctata* and *C. 7-punctata*) against *H. axyridis f. spectabilis* relies more on color features. This sensible choice by the pruned tree can be verified with the actual physical comparison of the ladybirds. Training of the MLP using BP algorithm utilized random weight initializations per run, and training outcome is provided in Table V.

TABLE IV
FIRST PHASE FEATURE SELECTION FOR *H. AXYRIDIS F. SPECTABILIS* AGAINST OTHER SIMILAR SPECIES

Features		2-spot	5-spot	7-spot	Pine
Geometrical	*Spot area*				X
	Spot perimeter				
	Spot max axis length				X
	Spot min axis length				
	Spot area ratio				X
	Spot aspect ratio				
Color	*Spot color a**				
	*Spot color b**	X			
	Spot hue angle				
	*Elytra color a**		X	X	
	*Elytra color b**				
	Elytra hue angle				

TABLE V
RESULTS OF TRAINING MLP

Run	Q	N	M	J	I	RMS error		% Accuracy	
						BP	With J48	BP	With J48
1	50	12	24	2	500	0.45	0.2	80	80
					1000	0.45	0.29	80	92
					1500	0.45	0.29	80	92
2	50	12	22	2	500	0.25	0.45	94	80
					1000	0.2	0.29	96	92
					1500	0.2	0.29	96	92
3	50	12	12	2	500	0.2	0.29	96	92
					1000	0.2	0.32	96	90
					1500	0.2	0.29	96	92
4	50	12	8	2	500	0.2	0.29	96	92
					1000	0.2	0.29	96	92
					1500	0.2	0.29	96	92
5	50	12	4	2	500	0.2	0.29	96	92
					1000	0.2	0.29	96	92
					1500	0.2	0.32	96	90

Note:
Q = no. of exemplar vectors
N = input features
M = no. of hidden neurons
J = no. of output neurons
I = no. of iterations

Table VI shows the confusion matrix, where only 80% cross-validation accuracy was obtained for *H. axyridis f. spectabilis* and none *for E. 4-pustulatus*. The resultant features obtained from J48 decision tree were applied to a MLP and BP algorithm was used. For identifying H. axyridis f. spectabilis against E. 4-pustulatus three geometrical features were used (max axis, area ratio and area). The number of hidden neurons was set to 6.

TABLE VI
CONFUSION MATRIX SHOWING CROSS-VALIDATION ACCURACY FOR *H. AXYRIDIS F. SPECTABILIS* AND *E. 4-PUSTULATUS* (FIRST PHASE)

	E. 4-pustulatus	*H. axyridis f. spectabilis*
E. 4-pustulatus	0	10
H. axyridis f. spectabilis	0	40

Results for the J48 operation are shown in the form of confusion matrix, as in Table VII.

TABLE VII
CONFUSION MATRIX SHOWING IMPROVED BP CROSS-VALIDATION ACCURACY FOR *H. AXYRIDIS F. SPECTABILIS* AND *E. 4-PUSTULATUS* APPLYING J48 (FIRST PHASE)

	E. 4-pustulatus	*H. axyridis f. spectabilis*
E. 4-pustulatus	9	1
H. axyridis f. spectabilis	3	37

The true positive rates were 0.9 and 0.925 for *H. axyridis f. spectabilis* and *E. 4-pustulatus* respectively. The result in Table VII indicated that it is possible to identify the two species with selectively distinct geometrical features. It also shows that the system is able to perform without the effect of rotation and scaling, despite the similarities between species. Fig. 6 shows the significance of the J48 implementation, by having a decision tree reducing the number of features and feeding only selected features into BP algorithm. Cross-validation accuracy increases to 92%, instead of 80% when using BP alone.

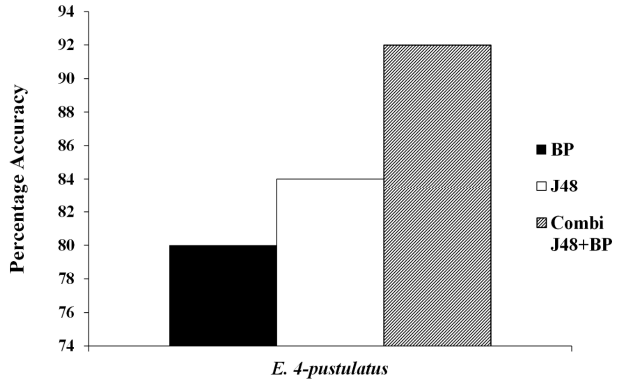

Fig. 6. Cross-validation identification accuracies for *H. axyridis f. spectabilis* against E.4-pustulatus using BP, J48 and a combination of the two

Results for the identification of *H. axyridis f. spectabilis* against other local species are plotted in Figure 7. It shows that significant improvement to BP cross-validation accuracy can be achieved for identifying some ladybird species using J48 pruned tree as feature pre-processor.

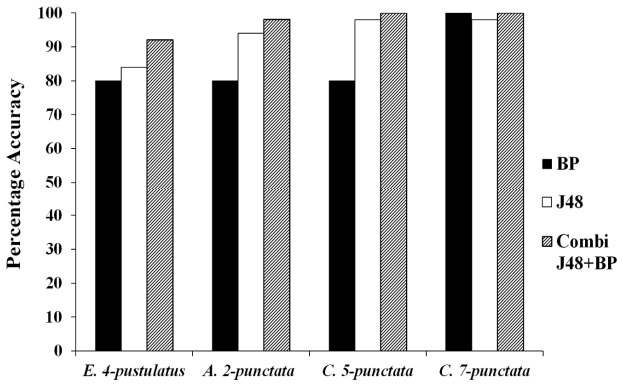

Fig. 7. Cross-validation accuracies for *H. axyridis f. spectabilis* against other species

Fig. 7 shows considerable improvement obtained in the identification of *H. axyridis f. spectabilis* against *A. 2-punctata*, *C. 5-punctata*, and *C. 7-punctata*. On average, using BP alone did not give satisfactory identification accuracy against the three species even though a 98% cross-validation accuracy was obtained for the identification of *H. axyridis f. spectabilis* against *C. 7-punctata*. For the other species, applying J48 pruned decision tree has improved the cross-validation identification accuracy by 18% difference. In contrast, there was not much change for *H. axyridis f. spectabilis* against *C. 7-punctata* when using J48 alone and the combination with BP, as evident from the group of bars recorded on the far right.

In the second phase, reduced features were used for testing against sample data taken from known positive samples with similar color characteristics in order to check for intra-species variations. *H. axyridis f. spectabilis* was paired with *H. axyridis f. conspicua* and *H. axyridis f. succinea*, trained using the same MLP network using BP algorithm and tested. They were also tested using J48 alone, and results compiled. The same set was also tested using a combination of J48 and BP algorithms. Results are shown in Fig. 8.

By using J48 the finalized feature obtained was 'elytra color *a**'. There was not much improvement when subjecting the identification of *H. axyridis f. spectabilis* with *H. axyridis f. conspicua*. It could be due to the color similarity of the elytra. This is evident from close inspection of both the elytra color and its corresponding CIELAB color value. Both species have dark elytra. The difference in color is quite undetectable to human eye, and to a taxonomist a series of typical investigations may include visual referencing to reference collections. An image processing system, however, can quantify the measurements through feature vector operation, and through an intelligent system like a trained artificial neural network, should be able to provide a solution. A contrasting result is obtained for the identification of *H. axyridis f. spectabilis* against *H. axyridis f. succinea*, where significant improvement is noticeable up to 97.5% accuracy. Referring to the ladybird images in Fig. 1, their elytra obviously differ in color and patterns (spots). This confirms the initial idea that contrasting (or similar) elytra colors or spot colors pose significant contribution to the identification rate of *H. axyridis*.

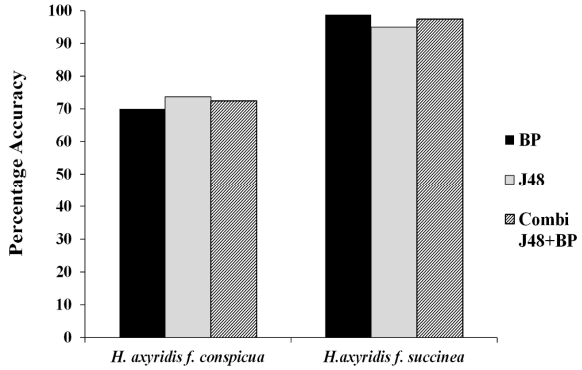

Fig. 8. Intra-species cross-validation accuracies for identification of *H. axyridis form spectabilis* against *H. axyridis form conspicua* and *H. axyridis form succinea*

A further test was conducted where each of *H. axyridis f. spectabilis*, *H. axyridis f. conspicua* and *H. axyridis f. succinea* was paired with *E.4-pustulatus* to make up a new database for training. They were trained using the same MLP network using BP algorithm and tested. Then, the same set was re-trained with principal component analysis (PCA) applied to the 12 features with 95% variance. Fig. 9 shows the cross-validation accuracies. In effect the features have been minimized through PCA, but information seems preserved. PCA has effectively reduced dimensionality by removing data redundancies (correlations) and extracting only the most significant features.

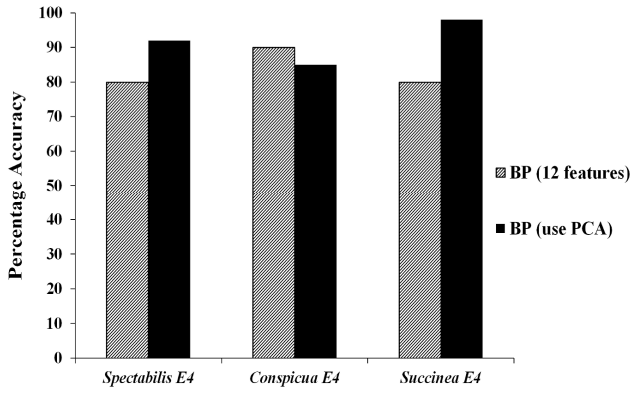

Fig. 9. Comparison of cross-validation accuracies for *H axyridis* against *E.4-pustulatus* between results of typical BP algorithm using 12 features and BP using PCA as feature-reducer

Even though applying PCA seems promising, Fig. 10 shows the identification result when comparing the cross-validation accuracy for the identification of *H. axyridis f. spectabilis* against *H. axyridis f. conspicua* and *H. axyridis f. succinea* due to applying J48 and PCA as feature pre-processors. An interesting observation is that both techniques have similar effects on the respective species. The group of

bars on the left indicates percentage correct identification of *H. axyridis f. spectabilis* against *H. axyridis f. conspicua*, which shows only 68.75% accuracy for PCA. This is a slight reduction of about 4% from J48, while same accuracies for both techniques were observed at 97.50% for the identification of *H. axyridis f. spectabilis* against *H. axyridis f. succinea*. This means using PCA may not be productive for the H. axyridis' intra-species identification, which leaves room for realizing J48's strength. PCA is a common statistical technique to reduce the dimensionality of input patterns. PCA has already been used by many researchers in pattern recognition and other research areas. The test result shows that J48 pruned decision tree is also a potential candidate as a feature extractor prior to ANN operations.

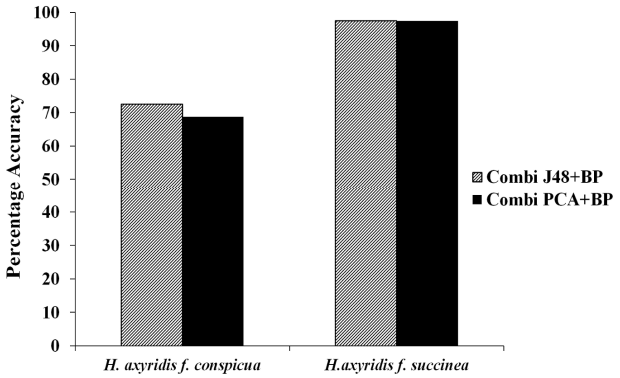

Fig. 10. Comparison between combination of J48 & BP, and combination of PCA & BP for identification of *H. axyridis f. spectabilis* against *H. axyridis f. conspicua* and *H. axyridis f. succinea*

This paper has demonstrated the usefulness of using a J48 pruned decision tree in tandem with multilayer perceptron neural network employing back propagation algorithm as a classifier in ladybird identification. The pruned decision tree indicates decision path for identification, and this is useful when coding a binary classification path in an expert system. MLP is clever due to the learning it acquires during training process. A caveat of using MLP in this domain is the network will need retraining every time new taxa are to be identified. It can be very time consuming, considering the vast amount of taxon and samples to work with. A workaround to this is the integration of an expert system into the loop, since it would acquire user inputs such as number of spots, geographical location etc. that the system could not acquire due to visual limitations. For instance, the expert system is in a better position in identifying a 2-spot ladybird among 5-spot ladybirds or 7-spot ladybirds by querying users whether they can provide the number of spots to the system. This is because the number of spots will sometimes be harder to determine from the many angles of a 3-dimensional ladybird using an image capturing device. In a way the expert system should provide considerable reasoning to the user, which is lacking in a typical ANN classifier. A possible scheme is to use J48 in the feature extractor, saving the number of features required to process by the next stage by pruning the features. Once the extracted features are obtained, they can be fed into MLP ANN for training the classifier. The expert system may interact with the user for extra inputs to improve identification accuracy. Another possible solution is to use other pattern recognition techniques, like support vector machine (SVM), radial basis function (RBF) network, learning vector quantization (LVQ), etc. This is reserved for future work.

V. Conclusion

A hybrid feature extractor for the automation of ladybird identification system is proposed. It is based on the application of J48 pruned decision tree, and MLP neural network using BP algorithm. The pruned tree is a useful reference when creating an expert system because it provides a simplified, but vital decision path. The system has correctly identified *A. 2-punctata*, *C. 5-punctata*, *C. 7-punctata* when color features are used. The system has also identified *H. axyridis f. spectabilis* from *E. 4-pustulatus* to 92% accuracy when geometrical features are presented to the MLP neural network, which is significant considering their physical similarities. The use of CIELAB as color space has reduced various illumination effects, and has improved character measurements. Future work will investigate the automated identification of intra species (*H. axyridis*) and other UK ladybird species using expert system, SVM, RBF and LVQ.

Acknowledgment

The authors would like to thank staff at the Centre for Ecology & Hydrology (UK) for the provision of ladybird images and expertise.

References

[1] M. E. N. Majerus, H. E. Roy, and P. Mabbott, "The harlequin ladybird, *Harmonia axyridis* (Pallas) (Col., Coccinellidae) arrives in Britain," in *Entomologists' Monthly Magazine*, vol. 42, pp.87–92, 2006.

[2] P. Katsoyannos et al., "Establishment of Harmonia axyridis on citrus and some data on its phenology in Greece," in *Phytoparasitica* vol. 25, pp.183–191, 1997.

[3] T. Adriaens, E. Branquart and D. Maes, "The multicoloured Asian ladybird Harmonia axyridis Pallas (Coleoptera: Coccinellidae), a threat for native aphid predators in Belgium?," *Belgian J Zool* vol. 133, pp.195–196, 2003.

[4] M. Kenis, H. E. Roy, R. Zindel, M.E.N. Majerus, "Current and potential management strategies against Harmonia axyridis," *BioControl,* vol. 53, doi:10.1007/s10526-007-9136-7, 2007.

[5] P. M. J. Brown, "Harmonia axyridis in Europe: spread and distribution of a non-native coccinellid," *BioControl,* vol. 53, doi:10.1007/s10526-007-9132-y

[6] H. E. Roy and A. Migeon, "Ladybeetles (Coccinellidae)," *BioRisk,* vol. 4, no. 1, pp. 293–313, 2010.

[7] R. D. Gordon, "The Coccinellidae (Coleoptera) of America north of Mexico," *J. New York Entomol Soc,* vol. 93, pp.1–91, 1985.

[8] R. L. Koch and T. L. Galvan, "Bad side of a good beetle: the North American experience with Harmonia axyridis," *BioControl.* doi:10.1007/s10526-007-9121-1, 2008.

[9] G. W. Hopkins and R. P. Freckleton, "Decline in the numbers of amateur and professional taxonomists: implications for conservation," in *Anim. Conserv.,* vol. 5, pp. 245-249, 2002.

[10] D. E. Walter and S. Winterton, "Keys and the crisis in taxonomy: extinction or reinvention?," *Annu. Rev. Entomol.* vol. 52, pp. 193–208, 2007.

[11] N. MacLeod, M. Benfield and P. Culverhouse, "Time to automate identification," *Nature,* 467, pp. 154–155, 2010.

[12] Q. D. Wheeler, P.H. Raven and E.O. Wilson, Taxonomy: impediment or expedient? *Science,* vol. 303, pp. 285, 2004.

[13] M. A. O'Neill, DAISY: A Practical Tool For Semi-Automated Species Identification. www.fao.org/ag/AGP/AGPS/C-CAB/Castudies/pdf/3-001.pdf, retrieved Oct 2009.

[14] UK Ladybird Survey [Online]. Available: http://www.ladybird-survey.org/

[15] UK Harlequin Survey [Online]. Available: http://www.harlequin-survey.org/

[16] D. Chesmore, T. Bernard, A. J. Inman and R. J. Bowyer, "Image analysis for identification of the quarantine pest *Tilletia indica*," *Bulletin OEPP/EPPO Bulletin,* vol. 33, pp. 495–9, 2003.

[17] L. Boddy, C. W. Morris and A. Morgan, "Development of artificial neural networks for identification," in *Information Technology, Plant Pathology and Biodiversity*, P. Bridge, P. Jeffries, D.R. Morse and P.R. Scott Eds., Oxon: CAB International, 1998.

[18] V. Steinhage et al., "Automated Extraction and Analysis of Morphological Features for Species Identification," in *Automated Taxon Identification in Systematics: Theory, Approaches and Applications*, CRC Press, Ch.8, pp.115-130, 2007.

[19] M. A. O'Neill et al., "DAISY: an automated invertebrate identification system using holistic vision techniques," in *Proc. Inaugural Meeting BioNET-INTERNATIONAL Group for Computer-Aided Taxonomy (BIGCAT)*, D. Chesmore, L. Yorke, P. Bridge & S. Gallagher, Eds. Egham: BioNET-INTERNATIONAL Technical Secretariat , pp. 13-22, 2000.

[20] J. Dai, *"Automated Identification of Insect Taxa Using Structural Image Processing,"* Ph.D. Thesis, Dept. Electron., University of York, 2006.

[21] E. D. Chesmore, "The Automated Identification of Taxa: Concepts and Applications," in *Automated Taxon Identification in Systematics: Theory, Approaches and Applications*, CRC Press, Ch.6, pp.83-100, 2007.

[22] CIEL*a*b* color scale [Online]. Available: http://www.hunterlab.com/appnotes/an07_96a.pd

[23] CIELAB colour models – Technical Guides [Online]. Available: http://www.dba.med.sc.edu/price/irf/Adobe_tg/models/cielab.html

[24] C. S. McCamy, H. Marcus and J. G. Davidson, "A Color-Rendition Chart," in *Journal of Applied Photographic Engineering*, vol. 2, no. 3, pp. 95-99, Summer 1976.

[25] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques.* 2nd ed., San Francisco: Morgan Kaufmann 2005.

[26] Building Classification Models: ID3 and C4.5 [Online]. Available: http://www.cis.temple.edu/~giorgio/cis587/readings/id3-c45.html

[27] G. Ou and Y.L. Murphey, "Multi-class pattern classification using neural networks," in *Pattern Recognition Society*, *vol. 40*, Elsevier Ltd., pp. 4-18, 2007.

[28] C. M. Bishop, *Neural networks for pattern recognition*. Oxford: Oxford University Press, 2000.

[29] C. G. Looney, *Pattern recognition using neural networks: theory and algorithms for engineers and scientists*. New York: Oxford University Press, 1997.

[30] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters, vol. 27*, Elsevier Ltd., pp.861-874, 2006.