VISION BASED ENTOMOLOGY – HOW TO EFFECTIVELY EXPLOIT COLOR AND SHAPE FEATURES

Siti Noorul Asiah Hassan, Nur Nadiah Syakira Abdul Rahman , Zaw Zaw Htike*and Shoon Lei Win

Department of Mechatronics Engineering, IIUM, Kuala Lumpur, Malaysia

ABSTRACT

Entomology has been deeply rooted in various cultures since prehistoric times for the purpose of agriculture. Nowadays, many scientists are interested in the field of biodiversity in order to maintain the diversity of species within our ecosystem. Out of 1.3 million known species on this earth, insects account for more than two thirds of these known species. Since 400 million years ago, there have been various kinds of interactions between humans and insects. There have been several attempts to create a method to perform insect identification accurately. Great knowledge and experience on entomology are required for accurate insect identification. Automation of insect identification is required because there is a shortage of skilled entomologists. We propose an automatic insect identification framework that can identify grasshoppers and butterflies from colored images. Two classes of insects are chosen for a proof-of-concept. Classification is achieved by manipulating insects' color and their shape feature since each class of sample case has different color and distinctive body shapes. The proposed insect identification process starts by extracting features from samples and splitting them into two training sets. One training emphasizes on computing RGB features while the other one is normalized to estimate the area of binary color that signifies the shape of the insect. SVM classifier is used to train the data obtained. Final decision of the classifier combines the result of these two features to determine which class an unknown instance belong to. The preliminary results demonstrate the efficacy and efficiency of our two-step automatic insect identification approach and motivate us to extend this framework to identify a variety of other species of insects.

KEYWORDS

Vision-based Entomology, Color Features, Shape Features, Machine Learning

1. INTRODUCTION

Vision based object recognition recently emerged as must explored field. This is due to the world are shifting to all computerized era. Previous technique concerning object identification rely too much on human expert thus do not tolerate expert absentee and consumed more time and human labor resources. Computerized system of image processing and object detection is an ideal method to improve the shortcomings of this traditional technique besides enhancing accuracy. There have been many successful attempts of using machine learning in automation of labour intensive tasks [8-10]. After succeeding on detection, researches becomes interested in doing classification of an image that come from similar family tree. The challenging part in classification is the object shares almost similar features and the variation points are limited.

DOI : 10.5121/cseij.2014.4201

Another difficulty of classifying an object is some species are difficult to distinguish visually since they differ in term of biological characteristics.

This paper shows our attempt to classify family of insects. Insects are most common living species on earth. There are more species of insect than any other land animal population. Insect can have different dietary nature such as predators, herbivores, host or decomposers, however their physical appearance dos not varied much. Insects comes from the arthropods group which their body can be divided into three sections, namely, head, thorax and abdomen. Thorax is the part that connect the head and their abdomen. In spite of this, the three section can fused together for certain type of insect. Another special trait of insect is they have wings and legs attached to the thorax. A large group of insect have antennae joined to the head that function as a sensor for smell and touch stimuli.

Three kinds of insects are selected for this project of insect classifications, namely, grasshopper and butterfly. Grasshopper and butterfly both have distinct color, shapes and size. Apparent and well defined trait make an easy detection and grouping.

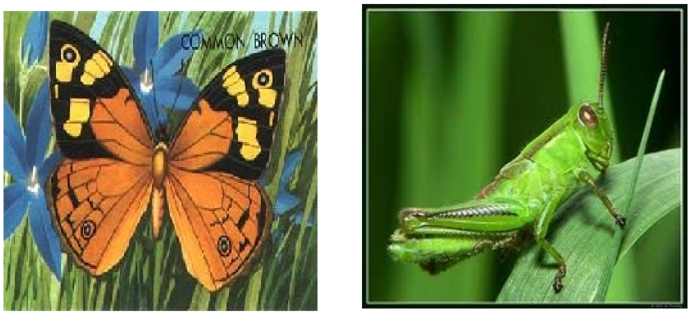


Figure 1: Insect Sample Data (butterfly and grasshopper)

The motivation of doing this project is that, first, aside from researches conducted in classification of animal species, there are not much research that deals specifically with insect classification. Therefore, a part of its flaw and limitation, this project provide a platform for the development of insect identification and classification. Second, automated vision based insect classification might replace entirely mundane task of manual recognition by the human labor. Third reasons would be enormous number of insect images available as open source in the internet. Hence, the process of obtaining images for training and testing data set is painless.

Fourth reason lies in the application of insect detection and classification. Image-based insect recognition has wide range of applications especially in agriculture, ecology and environmental science [5]. It generally can be utilized in prevention of plant disease and insect pests, plant quarantine and as an essential part of eco-informatics research. This new intelligent system is very beneficial especially to laymen who do not possessed professional knowledge in distinguishing many species of insects [1].

In addition, this paper also can be a reference for development of classifying system of any other variety of species such as fruit, flower, cockroaches and other insects, fish, or even microorganism such as bacteria.

2. RELATED WORK

This part provides a review of the past literature in vision-based insect recognition and classifications. Zhu Le Qing and Zhang Zhen introduced their research in using color histogram and Gray Level Co-occurrence Matrix (GLCM) for insect recognition. Image preprocessing algorithm is used to segment out the region of interest (ROI) from the insect image. Then, color features which are represented by color histogram are extracted from ROI that can be used for coarse color matching [1]. The matching is done by comparing the correlation of the feature vectors with certain threshold [1]. The image samples that passed this stage will undergo the next stage that is fine level matching. The fine level features are represented by coefficients such as energy, entropy and correlation of GLCM of preprocessed image blocks [1].

In the research, one hundred species of insects are selected as samples. All the image samples are captured from the right wing of the insects. A part of them are forewing with backwing, the rest are forewing [1]. The image is resized to speed up the processing speed and then filtered using mean shift algorithm [2]. Then, it is converted to gray scale image and binarized with certain threshold.

To reduce the effect of rotation and tranlation during image acquisition, a universal coordinate must be located by connecting the centroid of the foreground region and rotate the image around the centroid until the connected line is horizontal [1]. Least-square method is applied to fit the upedge and downedge of the insect wing with straight lines. ROI is determined afterwards and the image is aligned using Gaussian before feature extraction.

The color image is transformed from RGB (Red-Green-Blue) space into HSV (Hue-Saturation-Value) space before construct the histogram. To minimize the effect of illumination variation, only hue and saturation is take into consideration [1]. The histogram for hue and saturation component is calculated and it shows that the histogram for same species is distributed similarly.

GLCM approximatescharacteristics of the image related to second order statistics such as energy, entropy, correlation and homogeneity. This technique can identify insects from low resolution wing images and the efficiency and accuracy is proven [1].

Another related work to insect classification is pattern recognition. According to Kim, Lim, Cho and Nam (2006), this research is focusing in identify butterflies and ladybugs. The image of the insects are acquired and noise rejection is processed using color information method. Edge detection technique is applied to RGB space after the pre-processing by top-hat filtering with certain threshold. The detected edge line is analyzed by chain code analysis. To improve the end result, the processed image undergo max/min filtering [11].

The computer will process the data and the program counts the butterfly and ladybug in a picture that is included with many kinds of insect [11]. By using this method, the researchers claimed that they attained recognition about 90% as a whole [3].

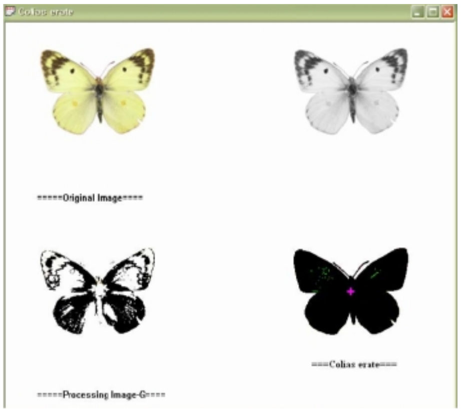


Figure 2: The program of insect discrimination [Adapted from Lim, Cho, Nam&Kim,2006]

Another similar research was done by the Yang et al [7] where the process of insect identification is focused on pattern recognition technology. Pattern recognition can be specifically defined as “the act of taking in raw data and taking an action based on the category of the pattern” (Yang et al, 2010). They further explained that the process involve three stages which are, input data and information collected from the sample using sensor, feature extraction mechanism that compute numeric or symbolic data from the sample, and classification scheme that determine the group of the sample[7]. Meanwhile, image segmentation is defined as the process on which the image is segregate into several portions where each portions grouped together pieces that share similar characteristic such as intensity or texture in order to generate non-overlapping image [7].

For success in insect recognition, they had collected several features of insect consists of rectangularity, elongation, roundness, eccentricity, sphericity, lobation, compactness and seven Hu moment invariants[7]. There are three classes of image features i.e. colour, texture and edges. However, edge extraction seems to be the best approach since it allowed minimal interference of light and noise[7]. This type of process employed the procedure of machine learning to train the algorithm to match with the image loaded and the result were saved in database. There are variety of classifiers that offer different approaches to sort an image. And each classifiers were tested on the samples to determine the best classifier that can successfully fit the data into respective group. In this study, the researchers claimed that Random trees algorithm is the best choice since it surpassed other classifiers in the test of insects’ image data samples [7].

The flowchart in Figure 3 shows the summary of training and recognition process involve in insect’s classification. The solid line represent the flow of data signifies the process of training the classifiers whereas the dotted line signifies the path taken by a new loaded image to test the corresponding classifier.

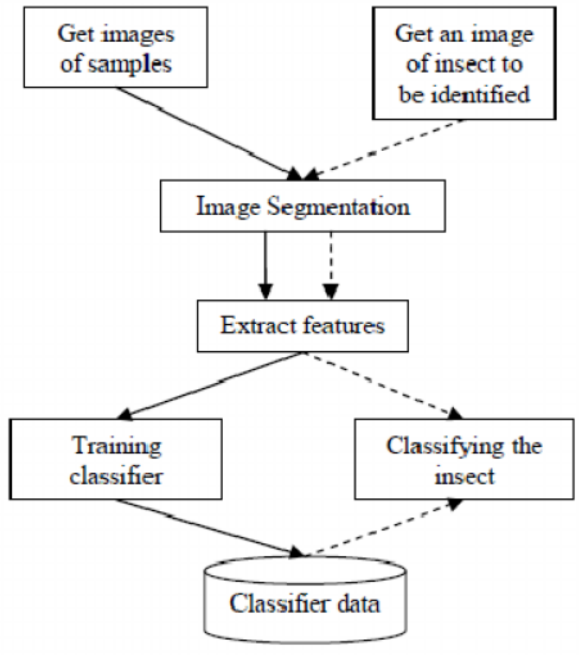


Figure 3: Main Flowchart of Training and Recognition [Adapted from Yang et al, 2010]

For image segmentation, they have been exercising a seven steps for classifying an insect. The seven procedures are: First, the image is loaded from a file. Then, image is converted to grayscale. Third, the edged is extracted using Canny edge algorithm. Close operator is used on the edges image. The next step would be finding an image contour by using contour algorithm. The biggest contour shows the contour of the insect that previously been loaded.

Contour feature extraction benefitted in the processing image with different image sizes an angles. However, the disadvantage would be obvious if the species have similar contour since the system cannot differentiate those contours. Hence, the solution to this problem is by exploiting image color and texture.

Another technique used in insect classification is extension theory. Extension theory is a method that constructed the matter element matrix of the insect based on mean and variance of the image features [4]. The matter-element matrix to be recognized includes the object itself, the feature and the feature value of the object in this case is an insect[4]. The first step to be done in extension method for classification of insects is feature extraction of insects. The image will be captured in real time and converted to gray scale image, segmented by the adaptive method so that the binary insect image is separated from the grain and the background[4]. Seventeen morphological features are extracted from the binary image and then are normalized [13].

Afterward, the matter element of features of the insects are constructed[4]. According to “3δ ” property of the normal distribution, if the object belonged to a class, then each feature of the object should be within the threefold variance with 99.7% probability[4]. Otherwise some features should be outside of the threefold variance [4].

The system is trained to identify the nine species of the insects given 135 samples[4]. Classical field matter element and controlled field matter element matrix of each species is formed [4]. The correlation degree is calculated. The higher correlation degree between the object to be recognized and a class, it means that it is much closer to that class. The correlation degree was simple to calculate, easy to use, fast and efficient. The accuracy of the classifier is increased compared to fuzzy classifier and k nearest neighbor classifier[4].

In this research, An Lu, Xinwen Hou, Cheng-Lin Liu and Xiaolin Chen try to use a hybrid approach in classifying insects called discriminative local soft coding (DLsoft) [5]. In soft coding strategies, a feature vector of the insect image is encoded by a kernel function of distance. As for the discriminative part, Multiple Kernel Learning (MKL) is used for the classification to improve the drawbacks of the soft coding alone.

This method is test on the fruit fly, Tephritidae that consists of different species. The dataset is divided into two; training set and testing set. Sample used for both dataset is an image of Tephritidae. First, all the images are transformed into gray scale. SIFT features of patches are extracted by densely sampled from each image. This is done so that each image will be represented by SIFT features. SIFT is done by dividing the images into several section of squares and each section will compute its own histogram. The local soft codes and discriminative is calculated. Subsequently, max pooling over all the vectors of patches in the same image is calculated. The final presentation of an image sample is spatial pyramid pooled vector which can be used by any machine learning method for classification [6].Zhang et al. [12] proposed exploiting color and shape features to detect cat head. In this paper, we investigate whether color and shape features can also be used to classify insects.

3. METHODOLOGY

3.1. Data collection

The first step that have been done in this project is to collect samples of data from different type of insects. In our work, we are classifying two type of insects which are butterfly and grasshopper. We collected 15 samples of data for each species and the total number of samples are 30 images for the data training. Meanwhile, for the data testing, we are selecting 4 images for each species to be tested later on the system. Most of the dataset is downloaded from google images and also some are obtained from Caltech 101 dataset. These two type of insect is chosen to be tested because of their wide range of numbers and easy to obtain certain features from feature extraction that will be further discussed in the next sub-topic.

3.2. Feature extraction

The next stage in image classification is extraction of features in an image. There many type of feature extractor or descriptor that have been used nowadays. In our work, we chose to use feature extraction based on color. Before the samples underdo feature extraction process, they will be preprocesses to several stages. First, the samples will be resize to a fixed size i.e. [100 100] dimension in matrix form. Usually the resize image is smaller to speed up the processing speed. After that, the samples will be converted to grayscale image and will be transformed to double format.

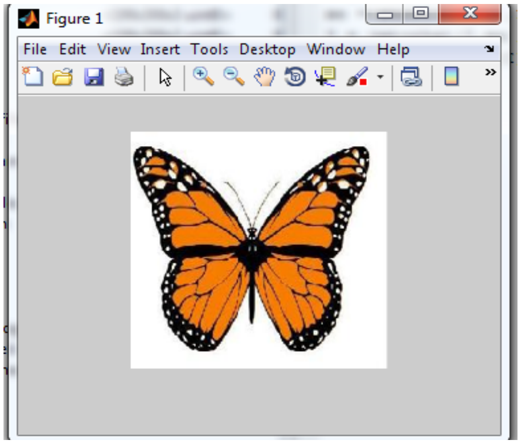


Figure 10: Resize image for Samples

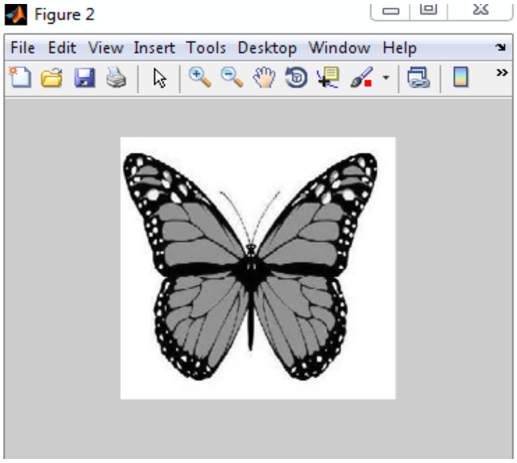


Figure 11: Grayscale Image of Sampled Data

After grayscale image is obtained, certain features will be extracted from the image sample. Basically, RGB is used to extract the color feature. The image is made up of pixel of matrices that in range from 0 to 256 pixel. According to Zhu Le Qing and Zhan Zhen, the color histogram of the same spesies is similar in distribution. Hence, it is easy to classify the insects based on the color distribution as well as other feature such as length and size.

3.2. Feature classification

Image classification analyzes the numerical properties of various image features and organizes data into categories. Classification algorithms typically employ two phases of processing: *training* and *testing*. In the initial training phase, characteristic properties of typical image features are isolated and, based on these, a unique description of each classification category, *i.e. training class*, is created. In the subsequent testing phase, these feature-space partitions are used to classify image features [14]. Similar to feature extraction, in image classification, there are different type of classifier that can be used. For example, K-Nearest Neighbour, Naïve Bayes, Kmeans and also Support Vector Machine (SVM). The classifier that we select are SVM. This is due to its simplicity and easy for the data set training. However, SVM only can classify two types of samples. Multi class SVM should be used to classify a higher number of type of samples.

SVM is chosen to be used in this work because of it regularization parameter and over-fitting may be avoided during classification. Besides, SVM uses kernel trick that receives the input data or images and transform it.

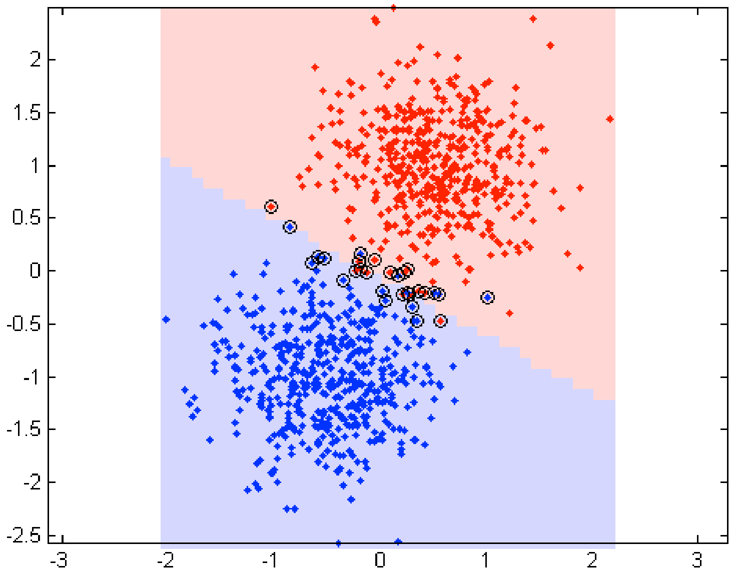


Figure 12: Example of SVM classifier (Linear) [14]

After the classifier is chosen, the samples collected is trained using 10 samples each for butterfly and grasshopper. The result of the system is in tabulate form and will be further discussed in experiments section.

4. EXPERIMENTS

We collected 300images of insects for training using the Google search engine. 150 of the images contain butterflies while the rest of the 150 images. We also collected 100 testing samples consisting of 50 butterfly images and 50 grasshoppers images. All of the samples image consists of adult specimens of butterfly and grasshopper. All the tests are run on the Intel Core i3, 2.16GHz, 2Gb RAM with Matlab r2009a as a software developing tools.

At the first place, we are classifying the insects using feature extractor, vl_sift under vl_feat open source software. Vl_sift function is used to extract the important features from the samples of images. Vl_sift computes SIFT frames (keypoints) of the image where the image must be in grayscale and in single precision. All of the samples are undergo several preprocessing stages; resize, grayscale and top-hat filtering.

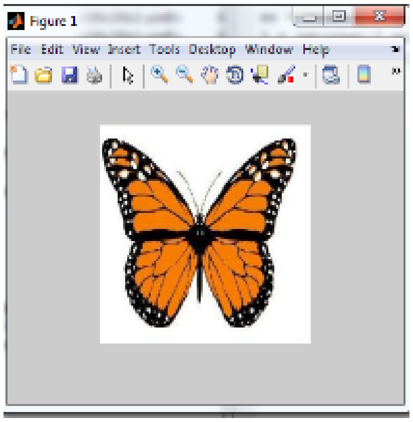


Figure 13: Sample Image of Adult Butterfly

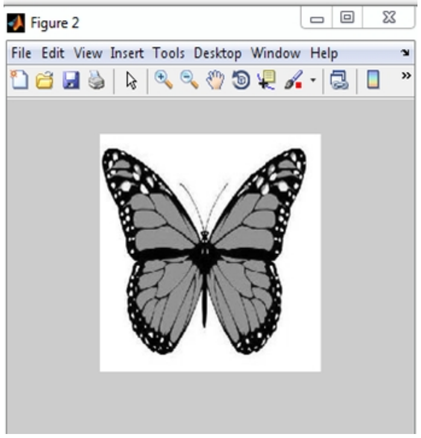


Figure 14: Grayscale Image

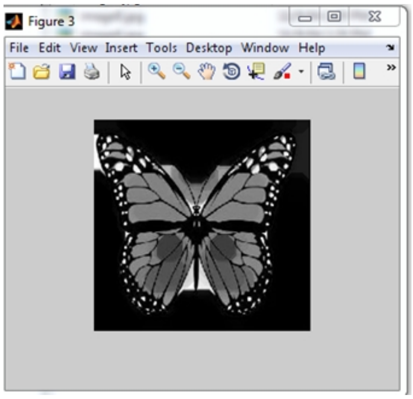


Figure 15: Filtered Image

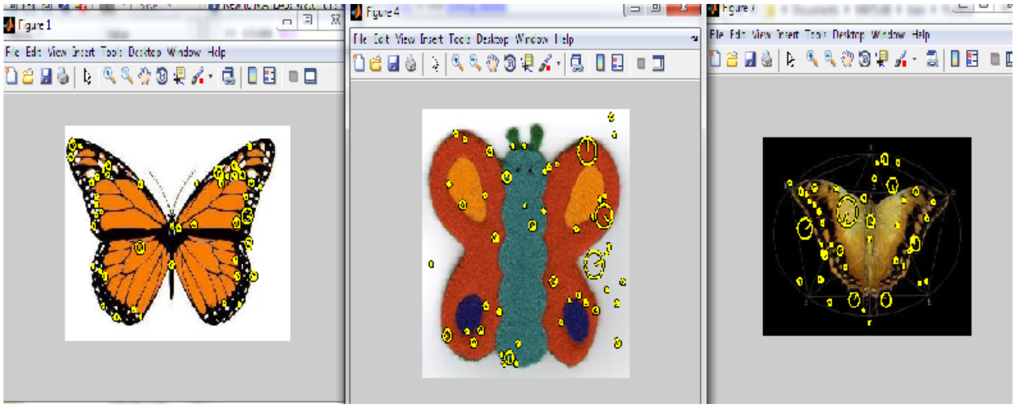
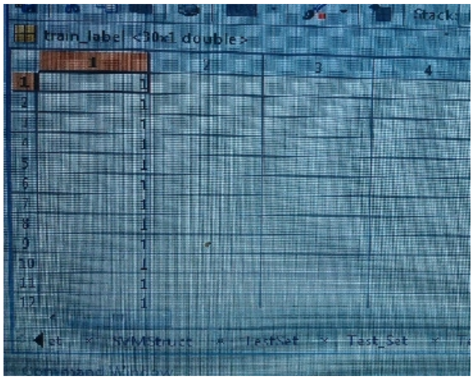


Figure 1
File Edit View Insert Tools Desktop Window Help
Figure 4
File Edit View Insert Tools Desktop Window Help
File Edit View Insert Tools Desktop Window Help

Figure 16: Detected features using vl_sift

However, after the features of the images are detected by the yellow circle as shown in the figure above, we did not managed to proceed with further steps which is to classify the images using a certain classifier. We did not able to store the extracted features in a form a tabulated data. Hence, we figured other solutions on how to classify the insects based on their groups or type. We are finalizing our classification system by using SVM classifier with certain training data. The samples data we collected is 30 in total and are undergo this system tobe classified as in butterfly or grasshopper group.

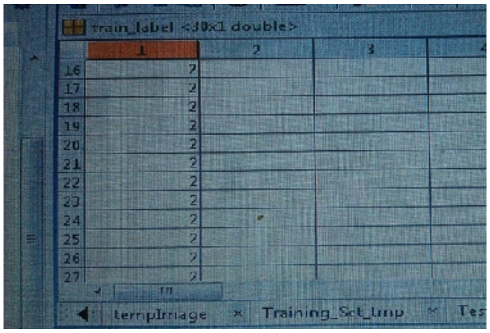
The samples image is processed into grayscale and then be resize to 100x100 matrices. The processed sample image has been generated and is stored in database as shown in Figure 13. The reading and storing process accomplished successfully. Since we are using grasshopper and butterfly, these two items constitute two label namely ‘1’for butterfly and ‘2’ for grasshopper.



train_label <30x1 double>

	1	2	3	4
1	1			
2	1			
3	1			
4	1			
5	1			
6	1			
7	1			
8	1			
9	1			
10	1			
11	1			
12	1			

Figure 17: Data Input for Sample Image (Butterfly)

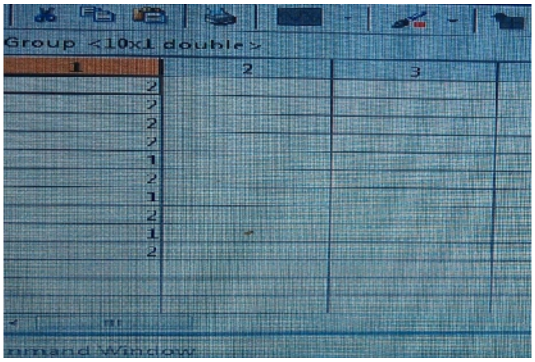


train_label <30x1 double>

	1	2	3	4
16	2			
17	2			
18	2			
19	2			
20	2			
21	2			
22	2			
23	2			
24	2			
25	2			
26	2			
27	2			

Figure 18: Data Input for Sample Image (Grasshopper)

By recognizing these two classes, the database for the reference classes has been generated. The system then is ready to prepare numeric matric for SVM training. After the first run of SVM, the trained data then stored is stored in the Trainingset matrix. The system then load the test image to assess the reliability of the classifier. The result of the test is stored in the Group variable. By opening the variable workshop, we can view the result of the test image. Since the image is being read from left to right in folder containing the test image, one can obviously figure out which image does the sequence of number belong to.



Group <10x1 double>

1	2	3
2		
2		
2		
2		
1		
2		
1		
2		
1		
2		

Figure 19: Result of Classifier is presented in Group variable.

The accuracy of the system is defined as the total number of correct predictions over the total number of predictions. The system acquired a satisfactory accuracy score of 92%. However, we found out that grasshopper has higher classification accuracy with a record of 98% while butterfly has an accuracy of 84%. This might be due the vast variations of the butterflies color and shape that can resulted in classifier being confused. Whilst the classifier did not acquire 100% accuracy, the SVM classifier was totally unoptimized. We only used the default SVM parameters. If we optimize the SVM parameters, the accuracy can improve. SVM has several significant advantages. First, SVM has regularisation parameter that is not sensitive to outliers. Another advantage is SVM can be applied in a wide range of classification problem even in high dimensional space with non-linear relationship. Despite the good sides, the employment of SVM also comes with several drawbacks. SVM shows a poor performance when especially when dealing with data that require more number of features rather than the given samples.

5. CONCLUSION

This paper proposes an automatic insect identification framework that can identify grasshoppers and butterflies from colored images. Two classes of insects are chosen for a proof-of-concept. Classification is achieved by manipulating insects' color and their shape feature since each class of sample case has different color and distinctive body shapes. The proposed insect identification process starts by extracting features from samples and splitting them into two training sets. One training emphasizes on computing RGB features while the other one is normalized to estimate the area of binary color that signifies the shape of the insect. SVM classifier is used to train the data obtained. Final decision of the classifier combines the result of these two features to determine which class an unknown instance belong to. Although the proposed insect classification system produced satisfactory results, it was built merely for a proof-of-concepts. There are many rooms for improvement. As future work, we would like to collect more training samples. We also would like to optimize the system parameters using optimization techniques such as genetic algorithm. After that, we would like to apply this framework to recognize a variety of other species of insects.

REFERENCES

[1] L. Q. Zhu and Z. Zhang, "Auto-classification of Insect Images Based on Color Histogram and GLCM", Seventh International Conference on Fuzzy Systems and Knowledge Discovery, 2010.

[2] Comaniciu, D., Meer, P. , "Mean shift: a robust approach toward feature space analysis". IEEE Transactions on Pattern Analysis and Machine Intelligence,. 24(5): p. 603-619, 2002.

[3] J. H. Lim, J. M.Cho, T. W. Nam and S. H. Kim, "Development of a Classification Algorithm for Butterflies and Ladybugs", 2006.

[4] H. T. Zhang and Yuxia Hu, "Extension Theory for Classification of the Stored-Grain Insects", International Conference on Machine Vision and Human-machine Interface, 2010.

[5] An Lu, Xinwen Hou, C.L Liu and Xiaolin Chen, "Insect Species Recognition using Discriminative Local Soft Coding", 21st International Conference on Pattern Recognition, 2012.

[6] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification". Computer Vision and Pattern Recognition(CVPR), 2009.

[7] H. Yang, L. Wei, K. Xing, J. Qiao, X. Wang, L. Gao, Z. Shen, "Research on Insect Identification Based on Pattern Recognition Technology". Sixth International Conference on Natural Computation, 2010.

[8] Z.Z. Htike, S.L. Win "Recognition of Promoters in DNA Sequences Using Weightily Averaged One-dependence Estimators", Procedia Computer Science, Volume 23, 2013, Pages 60-67, ISSN 1877-0509.

[9] Z.Z. Htike, S.L. Win "Classification of Eukaryotic Splice-junction Genetic Sequences Using Averaged One-dependence Estimators with Subsumption Resolution", Procedia Computer Science, Volume 23, 2013, Pages 36-43, ISSN 1877-0509.

[10] Z.Z. Htike, S. Egerton, Y.C. Kuang, "A Monocular View-Invariant Fall Detection System for the Elderly in Assisted Home Environments," 7th International Conference on Intelligent Environments (IE), 2011, vol., no., pp.40,46, 25-28 July 2011.

[11] J. Lim; J. Cho; T. Nam; S. Kim, "Development of a classification algorithm for butterflies and ladybugs," TENCON 2006. 2006 IEEE Region 10 Conference, vol., no., pp.1,3, 14-17 Nov. 2006.

[12] W.Zhang, J. Sun, X. Tang, "Cat Head Detection - How to Effectively Exploit Shape and Texture Features" Lecture Notes in Computer Science Volume 5305, 2008, pp 802-816.

[13] H. Zhang; H. Mao, "Feature Selection for the Stored-grain Insects Based on PSO and SVM," Second International Workshop on Knowledge Discovery and Data Mining, 2009. WKDD 2009, vol., no., pp.586,589, 23-25 Jan. 2009.

[14] http://homepages.inf.ed.ac.uk/rbf/HIPR2/classify.htm, Retrieved on 10 December 2013.