Tropical Agricultural Research Vol. 23 (1): 42 – 50 (2011)

Automatic Detection System for the Identification of Plants Using Herbarium Specimen Images

D. Wijesingha and F.M.M.T. Marikar[1*]

Sri Lanka Institute of Information Technology
Colombo 03
Sri Lanka

ABSTRACT. *An automatic leaves image identification system for endemic plants in Sri Lanka using neural networks is described in this study. Stemonoporus, a genus of Dipterocarpaceae which has about 30 species of plants was selected for the proposed system. National Herbarium specimens were used to obtain the images. Digital pictures of leaves were enhanced, segmented, and a set of features were extracted from the image. The most discriminating set of features were selected and then used as inputs to a Probabilistic Neural Network (PNN) which is used in MATLAB classifier and tests were performed to identify the best system. Several classification models were assessed via cross-validation in order to select the leaves in an image and identify the correct one. The results suggested that, leaf width, length, perimeter and area related features can be used as factors for prediction, and that machine vision systems lead to successful prediction of targets when fed with appropriate information. The overall classification accuracy utilizing the proposed technique for the test set was 85 %, whereas that feature extraction obtained was 95 %.*

Keywords: *Automatic detection, Probabilistic Neural Network, Stemonoporus, herbarium specimens*

INTRODUCTION

Precision Botany (PB) refers to the application of new technologies in plant identification. Computer vision can be used in PB to distinguish plants from its species level, so that an identification can be applied on the size and number of plants detected for the classification purpose. This work is focused on the application of computer vision for identification purposes of species in *Stemonoporus* genus. Surveys reveal that there are 3711 flowering plant species in Sri Lanka (Dassanayake, 2003). Out of these, 926 are endemic (Ashton *et al.*, 1997; Senarathne, 2001). Since some of these have minute variations, identification of these species has become difficult. Accurate and speedy identification of plants has become a time consuming and a fuzzy work due to non-availability of a computerized scientific plant identification system. Design and implementation of image-based plant classification system is a long felt need in Sri Lanka.

The current electronic devices for capturing images have been developed to a point where there is little or no difference between the target and its digital counterpart. The success of machine learning for image recognition also suggests applications in the area of identification of plant by herbarium specimens. Once the image of a target is captured

[1]Faculty of Medicine and Allied Science, University of Rajarata, Saliyapura, Anuradhapura, Sri Lanka
[*] Author for correspondence: faiz.marikar@fulbrightmail.org

digitally, a myriad of image processing algorithms can be used to extract features from it. The use of each of these features will depend on the particular patterns to be highlighted in the image. The automatic classification by computer vision of plants has received increasing attention in the recent past For instance; some relevant machine vision algorithms can classify plants into either crop or weeds (Yang *et al.,* 2000; Burks *et al.,* 2005). The classification of crops and weeds was studied by Brivot and Marchant (1996), who used infrared images in low-light conditions to study transplanted crops. Lee *et al*. (1999) and Hemming and Rath (2001) conducted similar experiments with tomato crops and weeds, respectively, using controlled artificial lighting to identify morphologic features of plants.

Leaf patterns are particular features for identification of plants. Maximum-Likelihood Estimation (MLE) to leaf features, such as area and length, the latter which is a non-invariant feature (Petersen *et al*., 2002; Asnor *et al*., 2009). Gebhardt *et al.* (2007) and Gebhardt and Kuhbauch (2007) added the features of shape, color and texture and used them in the MLE classification, still using controlled lighting. The advantage of image classification by feature assessment is that the patterns remain identical even if preliminary conditions are changed. Another approach is to use a small number of features that are known to be more representative of the target, and to rely on the classifier (Tang *et al*., 2000; Guo *et al*., 2001; Camargoa & Smith, 2009). In summary, the success of any pattern recognition system will depend not only on the particular feature but also on the quality of the information in the system.

With the increasing use of innovative computer technology, machine vision systems have become a possibility for plant identification. Tang *et al*. (1999) performed an image-based weed classification using a Gaborwavelet to classify images into broadleaf and grass categories for real time selective herbicide application. A quick glance at various images of leaves will reveal that a more robust technique for plant identification and classification based on an image by herbarium sample is a timely need. Neural networks appear in identification and classification tasks. For instance, Huang (2007) presented an application of neural networks and image processing for classifying seedling diseases. He successfully used color and texture features in all, in a classical Multi Layer Probability (MLP) with the back-propagation learning algorithm, though his process did not include feature selection. In our case, we have decided to use a more recently developed neural-genetic network architecture based on Probabilistic Neural Network (PNN) which is used in MATLAB classifier algorithm (Arribas & Cid-Sueiro, 2005). PNN network has proven to obtain better classification results than the MLP.

The main objective of this study is to identify and classify individual species of *Stemonoporus*, a genus of Dipterocarpaceae, by image processing using Probabilistic Neural Network (PNN) to implement a machine vision system (Gonzalez and Woods, 1992; Dassanayake, 2003). In addition, a plant identification system was constructed by using combination of four features; leaf length, width, perimeter and leaf area feature in selection step, in order to select the most distinct features. This helped to keep our neural-genetic classifier from producing overly-generalized results when an excessive number of input features are entered into the neural network. Finally, we believe that this system might potentially be of help in the field of PB applications in Sri Lanka.

MATERIALS AND METHODS

Experiment samples

A sample of 30 species was selected from the genus *Stemonoporus* which is endemic to Sri Lanka. Photographs of the specimens in the herbarium were taken in triplicates. Whole specimen sheet consisted of a branch, upper surface of a single leaf and the lower surface of the same leaf. Seventeen species were selected for the study and the images that were taken are listed in Table 1.

Table 1. List of *Stemonoporus* species used in the study and the number of images taken

Species	No of images taken
S. acuminatus (Thw) Beddome.	5
S. affinis Thw.	2
S. angustisepalum Kosterm.	4
S. bullatus Kosterm.	5
S. canaliculatus Thw.	6
S. cordifolius (Thw.) Alston.	11
S. elegans (Thw.) Alston.	9
S. gardneri Thw.	4
S. gilimalensis Kosterm.	6
S. kanneliyensis Kosterm.	6
S. laevifolius Kosterm.	4
S. latisepalum Kosterm.	1
S. oblongifolius Thw.	5
S. reticulatus Thw.	3
S. rigidus Thw.	3
S. scalarinervis Kosterm.	4
S. wightii Thw.	2

Image set

The set of 79 images of *Stemonoporus* species used in this study was obtained from the National Herbarium at the Royal Botanic Garden, Sri Lanka. Color images of the leaf samples were acquired and saved digitally using a Sony Cybershot T77 digital camera and the images were uploaded to a laptop computer. The packaging was removed from the herbarium and the samples were exposed to the atmosphere for ten minutes. The surface moisture was then blotted with a paper towel and the samples were imaged in an enclosed chamber with warm white deluxe fluorescent lighting. The light bulbs were mounted all around the steak sample at about a 45^{o} incidence angle from the steak surface imaged and they were heavily diffused. The same exposure and focal distance were used for all the images. In all cases, the image format used was JPEG, 24 bits and photographs of the specimens were taken in triplicates (Fig. 1)

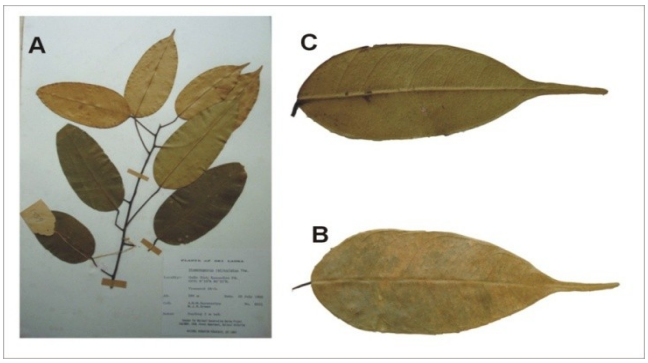

Fig. 1. Images of the *Stemonoporus elegans* (A) herbarium sample of (B) upper surface of the leaf (C) lower surface of the leaf Image format conversion

The most often used color coordinate systems include the gray scale to binary and the hue, saturation, and intensity systems (Gonzalez & Woods, 1992). Examination and preliminary analysis of the different image formats indicated that the saturation component, which gives a monochromatic image, revealed the leaf image texture most clearly (Fig. 2). The Gray Scale to binary was selected for training part of the system.

Fig. 2. Image format conversion (A) image without processing (B) gray scale processing (C) gray scale to binary system

Feature extraction

The last stage is the feature extraction in terms of individual and overall characteristics of the leaf. The output of this stage is a description of each leaf candidate in the image. This represents a great reduction in image information and ensures that the subsequent classification of species of the degree of accuracy. In the present work, the features extracted are the leaf length, width, area and perimeter. At this stage, the procedure generates an input vector (4 components) for each leaf candidate. Length and width of the leaf images were extracted using distance tools and, the shape of the leaf was extracted using edge detection tool in MATLAB IPT.

Pixel-value run lengths

A pixel-value run is a connected set of pixels in a specific direction having the same pixel values. Pixel-value run lengths can be used to characterize the spatial variation of pixel values in an image. All the textural primitives for each sample image were combined to form an approximation image and pixel-value run lengths in the horizontal direction were computed from the approximation image by MATLAB IPT. Leaf area was calculated by counting the number of pixels of binary image and leaf perimeter was calculated by counting

pixels on the leaf margin. The run length histogram made simple and the best composition to detect leaf area and perimeter.

Probabilistic Neural Networks

PNN derived from Radial Basis Function neural network is one of the most influential neural network models which consists of several layers of nodes (Haykin, 1999). They include an input layer, an output layer and a hidden layer, which contain input node(s), output node(s), and hidden node(s), respectively. The second layer sums these contributions for each class of inputs to produce a vector of probabilities as its net output. Finally, a complete transfer function on the output of the second layer picks the maximum of these probabilities and produces 1 for that class and 0 for the other classes. Moreover, a well known concern with neural networks is ‘‘overtraining.” To ease this problem, Roiger and Geatz (2003) suggested that the experiments could be continually conducted by different parameters. .

Statistical analysis

The performance of a trained network can be measured to some extent by the errors on the test sets but it is often useful to investigate the network response in more detail. One option is to perform a regression analysis between the network response and the corresponding targets. The correlation coefficient obtained between the outputs and targets is a measure of how well the variation in the output is explained by the targets. If this number equals to 1, then there is a perfect correlation between targets and outputs indicating a perfect fit. SPSS version 13 was used for this purpose. During the performance of classification of leaves, before training and testing, all features were normalized and the PNN was used for estimated principal components analysis.

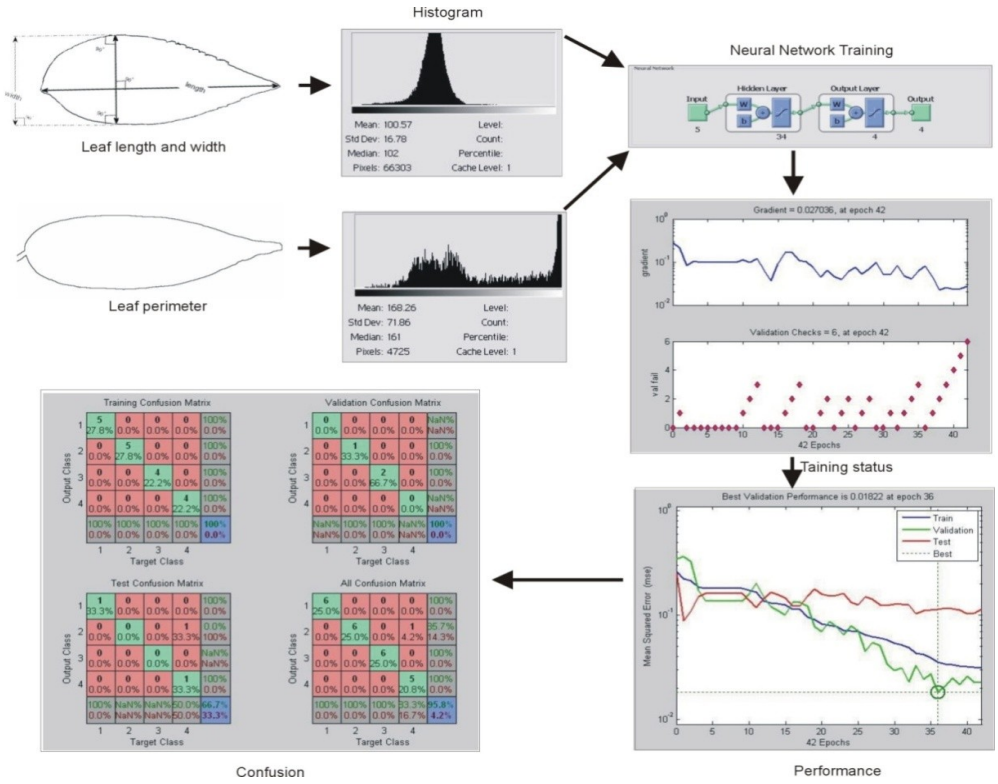

Fig. 3. Process of determining the leaf area, perimeter and training and outcome

RESULTS AND DISCUSSION

To validate the proposed technique for the automatic detection of plant species by leaf image processing, a set of 79 leaf images of the herbarium collection from the Peradeniya Royal Botanical Gardens were used. PNN is a popular classifier and has been applied for training and testing purpose in this work. In the training process, four categories of features were assigned based on only using leaf images with image sizes of 120×120 pixels.

Classification results using PNN are presented in Table 2 and the *Stemonoporus* genus identification results in Table 3. In addition, to investigate the plant identification method, different lighting conditions were considered using two sets of data. The leaf width, length, area and perimeter results are tabulated in Table 2 with 95 % accuracy. By comparing those with the results of the PNN trained with the *Stemonoporus* regularization, it is clear that the discrimination performance was improved, especially in the correct recognition of the species. A slight improvement is needed in case of correct recognition of *S. acuminatus* with respect to *S. oblongifolius* which is significant (Table 3). The classification accuracy for using the curvature value feature analysis is 85 %. It is important to note that without proper optimization of the leaf width, length, area, and perimeter detection process, such good results would not be possible.

Table 2. Leaf feature of expected and actual results in selected images

Leaf Sample	Expected results	Actual results

	W	L	A	P	W	L	A	P
DSC00995.JPG	980	2474	5180579	21712	980	2474	5180580	21710
DSC00996.JPG	896	2444	1384842	20894	896	2444	1384840	20890
DSC00998.JPG	709	2190	1059613	21970	709	2190	1059610	21970
DSC00999.JPG	735	2371	1158688	19141	735	2371	1158680	19140
DSC01001.JPG	788	2332	1207093	19868	788	2332	1207090	19870
DSC01003.JPG	856	2296	1291313	19505	856	2296	1291310	19500

W=width; L=length; A=area and P=perimeter

Fig. 3 shows the results of the whole process of segmentation. The classifier was trained with the proposed system as mentioned above. In this example, the leaf model was trained to acquire and convert into shadowed histogram and it has been correctly detected, since the normalized transformation from RGB into grayscale conversion makes the segmentation independent of lighting.

Since only 79 images were used in training; we gained classifier 85 % accuracy. This is mainly due to the unbalanced output patterns especially when input patterns are highly unbalanced (Fig. 3). Nevertheless, this is a major issue, related to the number of input samples (less than 100 in our study), the number of classes, the statistics behind those input sample sets, and the optimal size of the network while learning (estimated through the PNN algorithm) since by providing that balanced dataset, one is indeed not following the true prior probability distribution of input data samples. Since the neural network was trained in PNN mode, we divided the input data patterns into two disjointed data sets; the training set to properly train the network under the supervised, known-label mode and the testing set to test the generalization capabilities of the network after the training step (Fig. 3). Thus the images were classified manually by a human expert in order to provide supervised learning to the neural network classifier (Table 3).
Consequently, further development is required to increase the accuracy. Nevertheless, the findings indicate that the proposed feature extraction algorithm has a great potential for feature representation of leaf images in this classification task. Finally, we computed for 79 leaf images and the four selected input features, and the results are promising.

Table 3. Test scenarios for product testing

Image	Expected results	Actual results	Success rate (%)
DSC00995.JPG	*S. acuminatus*	*S. acuminatus*	99.38
DSC00996.JPG	*S. acuminatus*	*S. acuminatus*	100.00
DSC00998.JPG	*S. acuminatus*	*S. acuminatus*	98.53
DSC01008.JPG	*S. angustisipalum*	*S. angustisipalum*	95.12
DSC01009.JPG*	*S. angustisipalum*	*S. acuminatus*	78.05*
DSC01011.JPG	*S. angustisipalum*	*S. angustisipalum*	92.10
DSC01089.JPG	*S. elegans*	*S. elegans*	98.00
DSC01090.JPG	*S. elegans*	*S. elegans*	95.43
DSC01091.JPG	*S. elegans*	*S. elegans*	99.23

DSC01201.JPG	*S. oblongifolius*	*S. oblongifolius*	95.34
DSC01202.JPG	*S. oblongifolius*	*S. oblongifolius*	94.50
DSC01204.JPG*	*S. oblongifolius*	*S. acuminatus*	75.34*

* Failures of product testing

CONCLUSIONS

This study confirms the importance of leaf length, width, area and perimeter since the results obtained by the feature selection method selected these features as the most discriminant ones and combined them with other morphological features increased the results to 85 %. Considerable deviations were observed in the *S. angustisipalum* and *S. oblongifolius* suggesting needs for further improvement of the system. However, as the study was based on a limited sample size, reconfirmation of findings is needed with an adequate sample size. As the automated system is a novel method of identification of plants from herbarium specimens, we believe that the performance, accuracy and results obtained are at least promising and have a potential in real plant identification application. In order to further improve, the numerical results need to find the best back propagation models.

ACKNOWLEDGEMENT

We thank Dr. D.S.A. Wijesundara, Mr. N.M. Peramunagama, and Ms T. Subhani for providing samples at National Herbarium, Pereadeniya Botanical Gardens.

REFERENCES

Arribas, J.I. and Cid-Sueiro, J. (2005). A model selection algorithm for *a posteriori* probability estimation with neural networks. IEEE Transactions on Neural Networks *16* (4), 799–809.

Ashton, M., Gunathilleke, S., Zoysa, N., Dassanayake, M.D., Gunathilleke, N. and Wijesundera, S. (1997). A field guide to the common trees and shrubs of Sri Lanka, Wild life Heritage Trust Publications, Sri Lanka; 95-106.

Asnor, J.I., Hussain A. and Mustafa, M.M. (2009). Weed image classification using Gabor wavelet and gradient field distribution. Computer Electronic Agriculture, *66*, 53–61.

Brivot, R. and Marchant, J.A. (1996). Segmentation of plants and weeds for a precision crop protection robot using infrared images. In: IEEE Proceedings Vision, Image & Signal Processing, vol. 143, pp. 118–124.

Burks, T.F., Shearer, S.A., Heath, J.R. and Donohue, K.D. (2005). Evaluation of neural network classifiers for weed species discrimination. Biosystems Engineering *91*, 293–304.

Camargoa, A., and Smith, J.S. (2009). Image pattern classification for the identification of disease causing agents in plants. Computer Electronic Agriculture Doi:10/1016/j.compag.200901.003

Dassanayake, M. D. (2003). A revised handbook to the flora of Ceylon. CRC Press, USA, Volume 4, 404-418,

Gebhardt, S. and Kuhbauch, W. (2007). A new algorithm for automatic *Rumex obtusifolius* detection in digital images using colour and texture features and the influence of image resolution. Precision Agriculture *8* (1), 1–13.

Gebhardt, S., Schellberg, J., Lock, R. and Kuhbauch, W. (2007). Identification of broadleaved dock (*Rumex obtusifolius* L.) on grassland by means of digital image processing. Precision Agriculture *7 (3)*, 165–178.

Gonzalez, R. C. and Woods, R. E. (1992). Digital image processing. Addison-Wesley Publishing Company, U.S.A. pp 125-151.

Guo, G., Li, S.Z. and Chan, K.L. (2001). Support vector machines for face recognition. Image Vision Comput. *19*: 631–638.

Haykin, S. (1999). Neural Networks: A Comprehensive Foundation. Prentice Hall publishers, U.K. pp 108-125.

Hemming, J. and Rath, T. (2001). Computer-Vision-based Weed Identification under Field Conditions using Controlled Lighting. Journal of Agricultural Engineering Research *78 (3),* 233–243.

Huang, K.Y. (2007). Application of artificial neural network for detecting Phalaenopsis seedling diseases using color and texture features. Computers and Electronics in Agriculture *57 (1)*, 3–11.

Lee, W.S., Slaughter, D.C. and Giles, D.K. (1999). Robotic Weed Control System for Tomatoes. Precision Agriculture *1*, 95–113.

Petersen, M.W., Ridder, D. and Handels, H. (2002). Image processing with neural networks. Pattern Recognition *35*, 2279–2301.

Roiger, R. J. and Geatz, M.W. (2003). Data mining: A tutorial-based primer. Addision Wesley publishers, U.S.A. pp 160-168.

Senarathne, L. K. (2001). A checklist of the flowering plants of Sri Lanka. National Science Foundation publishers, Sri Lanka, 62-67.

Tang, Y.Y., Tao, Y. and Lam, E.C.M. (2000). New method for feature extraction based on factorial behavior. Pattern Recognition *35*, 1071–1081.

Tang, L., Tian, L.F., Steward, B.L. and Reid, J.F. (1999). Texture-based weed classification using gaborwavelets and neural network for real-time selective herbicide application.Trans. ASAE, St. Joseph, MI, No. 99-3036.

Yang, C.C., Prasher, S.O., Landry, J.A., Ramaswamy, H. and Ditommaso, A. (2000). Application of artificial neural networks in image recognition and classification of crop and weeds. Canadian Agriculture Engineering *42*, 147–152.