Malaysian Medicinal Plant Leaf Shape Identification and Classification

Mohd Shamrie Sainin[1], Taqiyah Khadijah Ghazali[2] and Rayner Alfred[3]

[1,2]*Universiti Utara Malaysia, Malaysia, shamrie@uum.edu.my*
[3]*Universiti Malaysia Sabah, Malaysia, ralfred@ums.edu.my*

ABSTRACT

Malaysian medicinal plants may be abundant natural resources but there has not been much research done on preserving the knowledge of these medicinal plants which enables general public to know the leaf using computing capability. This study proposes a framework to identify and classify tropical medicinal plants in Malaysia based the extracted patterns from the leaf. The extracted patterns from medicinal plant leaf are obtained based on several angle features. Five classifiers, obtained from WEKA and an ensemble classifier, called Direct Ensemble Classifier for Imbalanced Multiclass Learning (DECIML), are used to compare their performance accuracies over this data. In this experiment, five species of Malaysian medicinal plants are identified and classified in which each species will be represented by using 65 images. This study is important in order to assist local community to utilize the knowledge discovery and application of Malaysian medicinal plants for future generation.

Keywords: Malaysian medicinal plant, leaf, shape, identification, classification, image processing.

I   INTRODUCTION

The Convention on Biological Diversity (CBD) which consists of 188 countries signed and adopted the documentation of Global Strategy for Plant Conservation (GSPC) for conserving plant diversity Krupnick (2005). In order to successfully implement this plan, there are 16 targets (Convention on Biological Diversity, 2007) which are grouped into five major headings for the target namely: (1) understanding and documenting plant diversity (UDPD); (2) conserving plant diversity (CPD); (3) using plant diversity sustainably (UPDS); (4) promoting education and awareness about plant diversity (EAPD); and (5) building capacity for the conservation of plant diversity (CCPD).

It has been reported that the efforts to understand and document plant diversity continue to grow where there are a number of projects held in order to document the flora diversity around the globe. The documentation includes various data and images of all kinds of plants. These processes are mostly related to the knowledge management, where information about leaves are gathered and managed in order to provide useful knowledge to the community. Taking this as part of this paper's motivations, plant image recognition is very much required to further support the conservation efforts as specified in UPDS. A plant image recognition system can be used to assist possible endangered plant trafficking activities by using image recognition technology where it applies the shape matching algorithm to identify certain specimen from the identified database of plant images.

Since in the early 1990s, the efforts to identify plant from images have attracted various studies on different techniques for image processing, feature extraction and identification. Leaf identification can be categorized into three types, which includes shape-based, venation-based and combination of both approach.

In this paper, a preliminary study is conducted to illustrate the full scale leaf shape identification technique in order to identify leaf species.

This paper is organized as follows. Section 2 briefly discusses the medicinal plant identification and classification, plant leaf shape based features, image processing and feature extraction for plant leaf and classifier for medicinal leaf images. The experimental setup is discussed in section 3 and Section 4 discusses the results of six classifiers including the DECIML. Finally section 5 concludes this paper.

II   RELATED WORKS

The apparent tasks in any automated identification and classification has been discussed in previous section. This section brings several works and important attention in the focus of this research which is plant species identification and classification.

A. **Related works in medicinal plant leaf identification and classification**

To the best of researcher knowledge, no attempts have been made on identification and classification of Malaysia medicinal plant leaves using computing capabilities. Studies on Malaysian medicinal plant are mostly on physical scientific characteristics and

study for consumption as seen in Kadir (1998), Jantan (2004), and Ong et al. (2011).

Few attempts were done in global mainly in Unites States, India, China, Thailand and Indonesia. Pornpanomchai et al. (2011) presented their leaf recognition system called Thai herb leaf image recognition system which consists of four main components: 1) image acquisition, 2) image preprocessing, 3) recognition and 4) display of results. The system applied several image-processing techniques and extracted 13 features from the leaf image and uses a k-nearest neighbor (k-NN) algorithm for recognition process. The experiment involved 32 species of Thai herbs, with more than 1,000 leaf images and they reported that the classification performance is 93.29%.

Gao and Lin (2012) have studied and discussed an automatic recognition system of medicinal plants. With the leaf image recognition of medicinal plants as its core, the system applied the up-to-date technologies of image processing and neural network. There was no reported performance of the automatic recognition in their paper.

Herdiyeni and Santoni (2012) have proposed a new method for Indonesian medicinal plants identification using the combination of some leaf features, i.e. texture, shape, and color. The feature extraction was performed based on the Local Binary Pattern Variance and the classification was performed by using a Probabilistic Neural Network classifier. In this research, they reported that the average accuracy of medicinal plant identification was 72.16%. The data used comprises of 51 different species of Indonesian medicinal plants, in which 48 different images were used to represent each species.

Furthermore, Herdiyeni and Wahyuni, (2012) proposed a new mobile application based on Android operating system for identifying Indonesian medicinal plant images based on texture and color features of digital leaf images. The research investigated the effectiveness of the fusion between the Fuzzy Local Binary Pattern (FLBP) and the Fuzzy Color Histogram (FCH) in order to identify medicinal plants. This research used Probabilistic Neural Network (PNN) classifier for classifying medicinal plant species. The experimental results showed that the fusion between FLBP and FCH can be used to improve the average accuracy of medicinal plants identification. The accuracy performance of the identification process using the fusion of FLBP and FCH was reported to be 74.51% considering 51 different species of Indonesian medicinal plants with 48 images used to represent each species.

The development of Indonesian leaf recognition system is further studied by Pravista & Herdiyeni (2013). They have proposed a system called MedLeaf as a new mobile application for medicinal plants identification based on leaf image texture. Previous methods described in Herdiyeni and Wahyuni, (2012) were applied for the development in which 30 species of Indonesian medicinal plants with 48 leaf images each were used in the experiment. They reported that the accuracy performance of the medicinal plant identification process, based on leaf texture, was 56.33%.

Meanwhile in India, Arun et al. (2013) presented an automated system that was able to recognize and classify medicinal plant leaves. This automated system comprises of 250 different leaf images obtained for five different species. The types of leaf features obtained include grey textures, grey tone spatial dependency matrices (GTSDM), and Local Binary Pattern (LBP) operators which generates statistical values. It was reported that the accuracy performance of the proposed automated system based on 70% training and 30% testing set was 94.7%.

Ananthi et al. (2014) proposed a recognition approach using a MATLAB based Neural Network algorithm as a classifier that identifies the shape and texture features of the medicinal leaves. They discussed the identification and classification comparison of several leaf data that includes Hibiscus, Betel, Castor and Manathakali leaves. However, the number of leaves in the data used in their experiments was not presented.

B. **Plant Leaf Shape based Features**

Leaf image features are extracted mainly from shape information. Other features that can be extracted are vein patterns, colour and textures. Most of the previous and current leaf identification literatures utilize the whole leaf for feature extraction and to be used in the leaf identification process. Shape-based is the most popular approach for feature extraction as many of the researches show that this approach provides not only speed-up image processing but low cost and its conveniences (Langner, 2006).

Since in the early 1990s, the efforts to identify plant from images have attracted various studies on different techniques for image processing, feature extraction and identification. Most of the studies are concentrating on full scale leaf features, thus research on partly visible leaf for identification is available. Prior to this study, leaf identification can be categorized into three types, which are shape-based, venation-based and the combination of both approaches.

Shape-based is one of the popular approaches used for feature extraction as it provides rich information for classification (Xiao et al., 2010). Efficient shape feature extraction should present several essential properties such as identifiability, scalability, affinity and occultation invariance, noise resistance, statistically independent and reliable (Mingqiang et al., 2008). The earliest work in leaf shape-based automated identification on specific leaf was started by Heymans' group (Heymans et al., 1991), which involved the task of extracting the shape of the leaf (represented as grid) using a neural network algorithm for identification purposes.

Accelerated from the early shape-based, researchers began to introduce other techniques such as shape and centroid contour distance (Hong et al., 2004; Wang et al., 2003; Zhiyong et al., 2002). One of the most successful leaf identifications to date which was able to produce systematic leaf identification was designed based on the shape-based leaf identification. This approach used the inner-distance shape context approach (Agarwal et al., 2006; Ling and Jacobs, 2005; White et al., 2006; White et al., 2007). Some of the refined works related to this approach have been conducted by Belhumeur et al. (2008). Moving Median Center Hypersphere (MMC) was also introduced in the plant leaf identification technique (Du et al., 2007; Guo-Jun et al., 2004). Recently, leaf identification based on shape and MMC has been used in solving leaf identification problem with a complicated background as described in (Wang et al., 2008). Chaki (2011), discussed a plant leaf recognition using shape based features which provides accuracies ranging from 90%-100%. Recent study on shape-based leaf images recognition is presented by Mouine, et al. (2012) and Mouine et al. (2013).

Among all of the above approaches especially in shape-based leaf recognition, a full scale leaf was taken as the main source for feature extraction. Leaf in general has a near symmetrical shape geometry as shown in Figure 1.
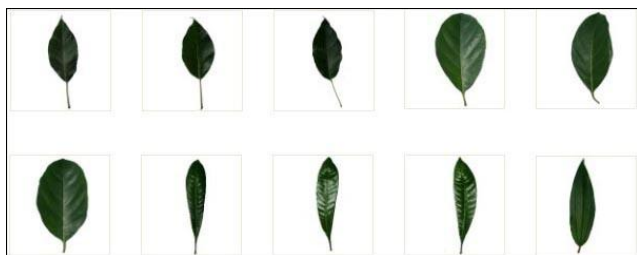

Fig. 1. Leaf shapes as reference to symmetrical characteristic.

III   METHOD

The method for Malaysian medicinal plant images identification and classification is discussed in this section.

A. **Image processing and feature extraction for plant leaf**

Image processing is considered as one of the important steps in the digital leaf image processing scheme (Figure 2), in which it involves edge detection and thinning processes (Langner, 2006). These techniques used in image processing are employed as it is much simpler due to less number of features extracted and processed.

In the edge detection phase, the Prewitt Edge detection algorithm is applied to the image. This simple edge detection algorithm computes the root mean square of two matrices of pixels with the template size 3x3 where a vertical edge component of X is calculated with the horizontal value in Y (Wu et al., 2007). It will produce an image where higher grey-level values indicate the presence of an edge between two objects. Grey-level values are measured by the number between 0-255, where 0 indicates no presence of edge (e.g. white) and 255 is a solid presence of edge (e.g. black).
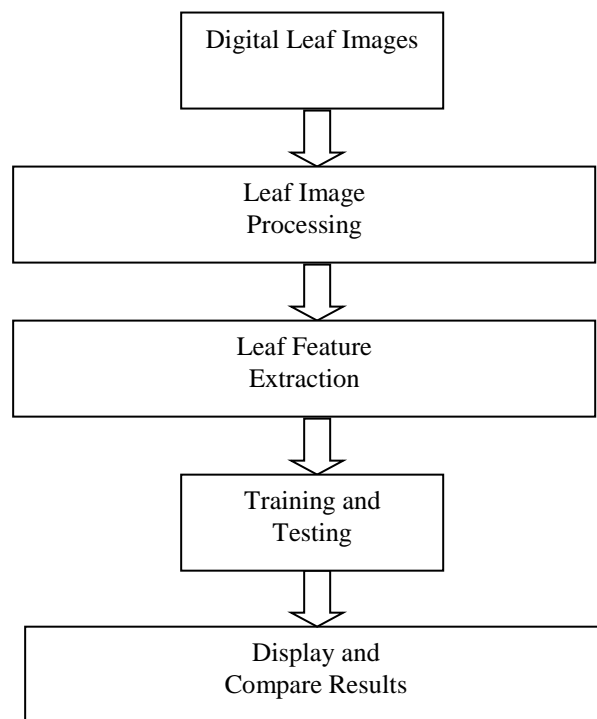

Fig. 2. Flow diagram of the leaf image processing scheme.

In the next step, thinning is applied to minimize the boundary of the leaf to one pixel only by comparing the actual pixel situation with specific patterns. Thinning is essential for the boundary of the leaf

because edge detection specifies only the intensity of the gradient of the pixel based on preconfigured threshold.

There are two types of features described as results of feature extraction step, shape tokens, which determine the number of features to be extracted and the angle feature for the final data construction. The important part of this phase is the process of extracting tokens from the boundary line of the leaf image. Tokens are assigned to the boundary line of the leaf image based on the predefined distance between tokens. The shorter the distance between the tokens are the more tokens will be assigned to the boundary of the leaf image as shown in Figure 3.
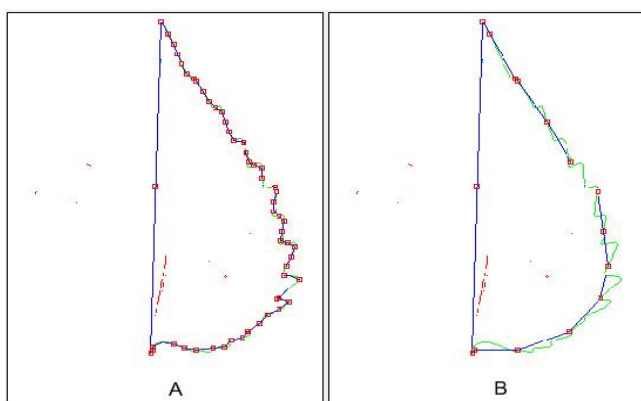

Fig. 3. Different distance effects on the number of tokens assigned to the boundary of the leaf image. (A) More tokens if distance is 1 (configured as 10) and (B) less tokens if distance is 3 (configured as 30).
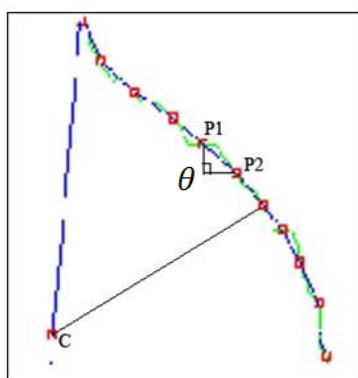

Fig. 4. Portion of the processed leaf image and representation of feature extraction, θ, hypotenuse angle for token P1 and P2.

Angles of the tokens are the features where the values for cosines and sinus are computed according to the direction of the angle. As shown in the portion of the processed leaf image in Figure 4, the two adjacent tokens (P1 and P2) are used to define angles based on the direction of hypotenuse from both tokens. According to Figure 4, P1 and P2 are

tokens and θ is the hypotenuse and C is the leaf centroid.

In this paper, features from five species with 65 leaves were extracted to produce 564 statistical values as feature (attributes) according to token distance at 1.

B. **Classifier Selection**

Several classifiers for leaf classification have been described in previous section, ranging from neural networks and k-nearest neighbor. In this paper, several classifiers from WEKA (Witten and Frank, 2000) were used to compare their classification accuracy over the leaf data. The selected classifiers obtained from WEKA, includes SMO, Decision Tree (J48), Naïve Bayes (NB), kNN and Random Forest (RF), using WEKA default settings.

In addition to that, an ensemble based classifier called a Direct Ensemble Classifier for Imbalanced Multiclass Learning (DECIML) is used to investigate the classifier performance. The classifier was proposed by Sainin and Alfred (2012) specifically to address the problem of multiclass classification with imbalance data. An imbalanced data with multiclass labels refers to a dataset with target class, which is skewed in distribution and poses a significant effect to classifier performance (Yang and Wu, 2006). This is due to the fact that medicinal leaf data is not often balanced for all collected species, where some of the leaf samples are limited for specimen purpose. Thus, imbalance data in term of the number of samples for some leaves will exist. The researchers reported that the average accuracy using the DECIML on 16 imbalanced multiclass benchmark data was higher than the other tested single classifiers.

IV **EXPERIMENT**

The dataset for the experiment is obtained from villages situated in Perlis state where, 65 leaf samples are randomly selected from specified leaf species for the experimental data. The leaf sample size is selected in this preliminary study due to enormous time required to process the images without specific automated image processing. Table 1 is the list of leaf species selected in this paper.

In order to create the preliminary experimental data, basic image processing and feature extraction as described in previous section were implemented. Full leaf features based on token angle (cosines and sinus) are extracted. Initially, the image processing and feature extraction processes will produce different number of features. Thus, further data pre-processing is applied to find the largest number of features and leaf with fewer features will fill several dummy values to the remaining features (-1). Table 2 shows the description of the experimental data.

Based on Table 2, #Examples is the number of leaf image, #Attribute in Description column shows the largest number of leaf features using full leaf image processing, #Training and #Testing are number of leaf images as training and testing. The #Majority and #Minority represent the imbalance for the data.

Table 1. Selected leaf species for the experimental data.

Class	Example	Name	Train	Test
1		Cemumar	11	4
2		Kapal Terbang	12	4
3		Kemumur Itik	11	4
4		Lakom	5	4
5		Mengkudu	6	4
		Total	45	20

Fundamentally, the data is small but it depicts a fairly high dimension where it poses the challenge of possible problem such as: 1) not enough data, 2) the "curse of dimensionality". Thus, this data can be used to investigate the effects of using imbalance data on the accuracy performances of all the classifiers used in this experiment. Therefore, the dataset constructed in this preliminary study is designed to show how the classifiers work on the available experimental data. Real tasks on image processing and feature extraction optimization will be left out for future work in this domain.

Table 2. Information of the experimental leaf data.

Description	Value #
#Examples	65
#Attributes	564
#Training	45
#Testing	20
#Majority	12
#Minority	5

V RESULTS

There are five classifiers used which can be obtained from WEKA (SMO, J48, NB, kNN (k=3) and RF) and the ensemble classifier called a DECIML (DE) were experimented and their accuracies are compared over the data. Table 3 shows the average accuracy (precision evaluation metric will be further discussed in future work).

Table 3. Average performance of the classifiers.

Classifier	SMO	J48	NB	kNN	RF	DE
Accuracy %	50	40	35	40	45	**65**

The results shown in Table 3 indicate that the experimental data is very challenging and in another point of view (leaf classification and identification community), the samples and extracted features from the leaf (angle of the leaf shape tokens) are not enough to describe the domain problem. In addition, the data has not been further preprocessed (discretization, etc.). However, the objective of the preliminary study in this domain is to create an experimental data based on leaf features is achieved in testing the performance of the several classifiers. The DECIML classifier, an ensemble approach to classify multiclass data with imbalance performs slightly better than the other classifiers. Closer to the DECIML, SMO which is the implementation of SVM in Weka produce 50% accuracy. This shows that the SVM is able to work with the high dimensional imbalanced multiclass data. In contrast, NB performs worst over the data. These due to the probabilistic estimation over the large features degrade the performance of NB.

VI CONCLUSION

Malaysian medicinal leaf image recognition and classification is considered important for its

knowledge utilization. Future generation could easily forget the plants and benefits from these traditional remedies. This preliminary paper will motivate more research in this domain in order to preserve the knowledge with the help of computing technologies. In this paper, a basic image processing and feature extraction have been performed to prepare the Malaysian medicinal leaf data. Several classifiers including the DECIML have been applied to the data and show that the DECIML provides the promising result. In future works, more Malaysian medicinal plant leaves will be collected to create large pool of leaf data. Advanced image processing and other feature extraction approach will also be investigated. Improvement of the DECIML classifier will be studied to match with other leaf data classification.

REFERENCES

Agarwal, G., Ling, H., Jacobs, D., Shirdhonkar, S., Kress, W. J., Russell, R., White, S. (2006). First steps toward an electronic field guide for plants. Taxon, 55(3), 597-610.

Ananthi, C., Periasamy, A., & Muruganand, S. (2014). Pattern Recognition Of Medicinal Leaves Using Image Processing Techniques Journal of NanoScience and NanoTechnology, 2(2), 214-218.

Arun, C. H., & Emmanuel, W. R. S. (2013). Texture Feature Extraction for Identification of Medicinal Plants and comparison of different classifiers. International Journal of Computer Applications, 62(12).

Belhumeur, P. N., Chen, D., Feiner, S., Jacobs, D. W., Kress, W. J., Ling, H., . . . Zhang, L. (2008). Searching the World's Herbaria: A System for Visual Identification of Plant Species. Paper presented at the European Conference on Computer Vision, Marseille-France.

Chaki, J., & Parekh, R. (2011). Plant Leaf Recognition using Shape based Features and Neural Network classifiers International Journal of Advanced Computer Science and Applications, 2(10), 41-47.

Du, J., Huang, D., Wang, X., & Gu, X. (2005). Shape Recognition Based on Radial Basis Probabilistic Neural Network and Application to Plant Species Identification. Paper presented at the International Symposium of Neural Networks, ser. LNCS 3497.

Gao, L., & Lin, X. (2012). A Study on the Automatic Recognition System of Medicinal Plants. Paper presented at the 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet), Yichang.

Guo-Jun, Z., Xiao-Feng, W., De-Shuang, H., Zheru, C., Yiu-Ming, C., Ji-Xiang, D., & Yuan-Yuan, W. (2004). A hypersphere method for plant leaves classification. Paper presented at the International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.

Herdiyeni, Y., & Santoni, M. M. (2012). Combination of Morphological, Local Binary Pattern Variance and Color Moments Features for Indonesian Medicinal Plants Identification. Paper presented at the International Conference onAdvanced Computer Science and Information Systems (ICACSIS2012), Depok.

Herdiyeni, Y., & Wahyuni, N. K. S. (2012). Mobile Application for Indonesian Medicinal Plants Identification using Fuzzy Local Binary Pattern and Fuzzy Color Histogram. Paper presented at the International Conference onAdvanced Computer Science and Information Systems (ICACSIS2012), Depok.

Heymans, B. C., Onema, J. P., & Kuti, J. O. (1991). A neural network for opuntia leaf-form recognition. Paper presented at the IEEE International Joint Conference on Neural Networks.

Hong, F., Zheru, C., Dagan, F., & Jiatao, S. (2004). Machine learning techniques for ontology-based leaf classification. Paper presented at the Control, Automation, Robotics and Vision Conference, 2004. ICARCV 2004 8th.

Jantan, I. (2004). Medicinal Plant Research in Malaysia: Scientific Interests and Advances. Jurnal Sains Kesihatan Malaysia, 2(2), 27-46.

Kadir, A. A. (1998). Biodiversity Prospecting Of Tropical Plants For Medicinal Uses. Paper presented at the Third National Congress on Genetics.

Krupnick, G. A., & Kress, W. J., (eds.) (2005). Plant Conservation: A Natural History Approach. The Plant Press, 346.

Langner, J. (2006). Neuronal Network based recognition system of leaf images. Retrieved 24 Febuary 2009, 2009, from http://www.jens-langner.de/lrecog/

Ling, H., & Jacobs, D. W. (2005). Using the inner-distance for classification of articulated shapes. Paper presented at the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005.

Mingqiang, Y., Kidiyo, K., & Joseph, R. (2008). A survey of shape feature extraction technique. Pattern Recognition, 43(90).

Ong, H. C., Zuki, R. M., & Milow, P. (2011). Traditional Knowledge of Medicinal Plants among the Malay Villagers in Kampung Mak Kemas, Terengganu, Malaysia. Ethno Med, 5(3), 175-185.

Pornpanomchai, C., Rimdusit, S., Tanasap, P., & Chaiyod, C. (2011). Thai Herb Leaf Image Recognition System (THLIRS). Nat. Sci., 45(551-562).

Prasvita, D. S., & Herdiyeni, Y. (2013). Medleaf Mobile Application for Medicinal Plant Identification Based on Leaf Image. Advanced Science Enginering Information Technology, 2(2), 5-8.

Wang, X.-F., Huang, D.-S., Du, J.-X., Xu, H., & Heutte, L. (2008). Classification of plant leaf images with complicated background. Applied Mathematics and Computation, 205(2), 916-926.

Wang, Z., Chi, Z., & Feng, D. (2003). Shape based leaf image retrieval. Vision, Image and Signal Processing, IEE Proceedings -, 150(1), 34-43.

White, S., Marino, D., & S., F. (2006). LeafView: A User Interface for Automated Botanical Species Identification and Data Collection. Paper presented at the ACM UIST 2006 Conference Companion, Montreux, Switzerland.

White, S. M., Marino, D., & Feiner, S. (2007). Designing a mobile user interface for automated species identification. Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems, San Jose, California, USA.

Witten, I., & Frank, E. (2000). Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations: Morgan Kaufmann.

Wu, S. G., Bao, F. S., Xu, E. Y., Yu-Xuan, W., Yi-Fan, C., & Qiao-Liang, X. (2007). A Leaf Recognition Algorithm for Plant Classification Using Probabilistic Neural Network. Paper presented at the Signal Processing and Information Technology.

Xiao, X.-Y., Hu, R., Zhang, S.-W., & Wang, X.-F. (2010). HOG-Based Approach for Leaf Classification. In D.-S. Huang, X. Zhang, C. Reyes García & L. Zhang (Eds.), Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence (Vol. 6216, pp. 149-155): Springer Berlin / Heidelberg.

Yang, Q., & Wu, X. (2006). 10 challenging problems in data mining research. International Journal of Information Technology and Decision Making, 5(4), 597-604.

Zhiyong, W., Zheru, C., & Dagan, F. (2002). Fuzzy integral for leaf image retrieval. Paper presented at the Fuzzy Systems, 2002. FUZZ-IEEE'02.