A Plant Recognition Approach Using Shape and Color Features in Leaf Images

Ali Caglayan, Oguzhan Guclu, and Ahmet Burak Can

Department of Computer Engineering,
Hacettepe University, Ankara/Turkey
{alicaglayan,oguzhanguclu,abc}@cs.hacettepe.edu.tr

Abstract. Recognizing plants is a vital problem especially for biologists, chemists, and environmentalists. Plant recognition can be performed by human experts manually but it is a time consuming and low-efficiency process. Automation of plant recognition is an important process for the fields working with plants. This paper presents an approach for plant recognition using leaf images. Shape and color features extracted from leaf images are used with k-Nearest Neighbor, Support Vector Machines, Naive Bayes, and Random Forest classification algorithms to recognize plant types. The presented approach is tested on 1897 leaf images and 32 kinds of leaves. The results demonstrated that success rate of plant recognition can be improved up to 96% with Random Forest method when both shape and color features are used.

Keywords: Leaf recognition, shape features, color features.

1 Introduction

Plants have many uses in industry, medicine, and foodstuff production. Recognizing plant species is an important process to obtain necessary raw materials from correct plants. Plant recognition is also important in environmental protection to correctly observe changes in plant species and population. However, recognizing plants is a difficult task and is generally done by human expert biologists. Designing automatic recognition systems for plants is useful, since it can facilitate fast classification of plants, and have applications in many scientific and industrial fields [10]. For instance, discovery of new species, plant resource surveys, population studies, and plant database management are demanding applications in biology, foodstuff, medicine, and agriculture. Automatic plant recognition may increase efficiency and speed in these fields, save time of human experts, and decrease cost of production stages.

A computer-based plant classification system can use various characteristics of plants such as leaves, flowers, fruits, branching styles, and outlooks. An easier and accurate way is using leaves to identify plants. Since leaves are considered as important features to characterize plant species, many studies on leaf image retrieval based on shape, venation, color, and texture information have been conducted in computer-aided plant identification systems.

A. Petrosino (Ed.): ICIAP 2013, Part II, LNCS 8157, pp. 161–170, 2013.

The aim of this work is to develop an approach to classify plants according to leaf features. The classification based on leaf images has an advantage such that sampling leaves (getting photos) is low-cost and convenient. Performance of a leaf recognition system depends on good feature selection and efficient recognition algorithm. In this work, we extended the method in [22] and studied new features and classification algorithms. In addition to shape features, we used color features of leaf images. A collection of machine learning techniques, k-Nearest Neighbor, Support Vector Machines, Naive Bayes, and Random Forest, are used to classify both shape and color features. Our results have shown that color features can improve the recognition performance. With Random Forest method, we obtained 96.32% classification accuracy rate in 32 plant species. To the best of our knowledge, the results are state of the art for such a large number of plant types.

Outline of the paper is as follows. Section 2 summarizes previous works on plant recognition systems. In section 3, we give an overview of commonly used features in previous works and explain new features of this work. Section 4 covers the experimental methods and results. Conclusions are summarized in section 5.

2 Related Work

Automated leaf recognition systems have been addressed by many researchers. Du et al. [10] proposed an automatic plant recognition method based on digital morphological features of leaf images, which generally include geometrical and invariable moment features. They also used Move Median Centers (MMC) hypersphere classifier as the classification method. Wu et al. [22] proposed a leaf recognition algorithm using Probabilistic Neural Networks (PNN). They define five basic geometric features: diameter, physiological length, physiological width, leaf area, and leaf perimeter. Then, they derive twelve digital morphological features from these basic geometric features. They finally use a PNN learner with two layers to classify these features. In [19], the authors used 7 morphological features in the literature and in addition to these features they introduced 3 new morphological features using half leaf images. PNNs were used as classification method. Gu et al. [12] used Wavelet Transform and Gaussian interpolation combination to extract the characteristics of leaf skeleton and veins. They used kNN classifier and radial basis PNN method for recognition. Chaki and Parekh [7] analyzed plant leaf images of three plant types using two different shape modeling techniques. The first is based on moments-invariant model and the second is based on the centroid-radii model. They used neural networks as classifier for discrimination. In [2], complex veins and contours of leaves are used as features. After collecting morphological features, mathematical models are constructed for classification of species. Valliammal and Geethalakshmi [20] proposed a new approach that combines a thresholding method and H-maxima transformation based method to extract leaf veins. In spite of the fact that they used a small data set for experimental evaluation, the results showed this hybrid approach is capable of extracting more accurate venation modality of the leaf for vein pattern classification. Lee et al. [15] presented a leaf recognition system for plant classification by using geometric features, vein features, contour and centroid of leaves. They perform Fast Fourier Transform on distance values between centroid and contour points of leaves for acquiring frequency domain data.

There are some works which consider texture and color information. In [1], a volumetric fractal dimension approach is used and texture information of leaves was considered. 30 texture samples from each leaf were extracted and some of them which better describe the texture were chosen by the proposed algorithm. The authors of [5] proposed Gabor Wavelet Filters for texture analysis of leaves. They used a set of filters each of which has different scale and rotation parameters to extract texture features. Linear Discriminant Analysis (LDA) was used as classifier in both [1] and [5]. The authors of [17] proposed a semi automatic plant identification algorithm that combines color features, shape features, cell features and volume fraction. They used an unsupervised learning method based on a multistage comparison technique for classification.

Despite the above methods for general plant recognition, there are some works which are applicable to some certain species. Gouveia et al. [11] used a semi-automatic method for characterization of chestnut tree leaves. Some works are proposed to identify damaged leaves due to diseases [9], [14]. Degrees of damage on tomato leaves were measured in [9] and cucumber leaf diseases were detected in [14]. In both [9] and [14], Support Vector Machine (SVM) algorithm is used to recognize infected leaves.

Most of the studies ignore difficulties caused by natural background and use pictures with smooth and clean background. There are some approaches that work independent from the change in background. Wang et al. [21] proposed a framework for leaf images with complicated background. They first applied segmentation to distinguish leaves from background. Then, shape features were extracted and finally the Moving Center Hypersphere (MCH) approach was used for classification. Manh et al. [16] proposed a new method of weed leaf segmentation based on the use of deformable templates. This method has been used to segment one single specie and produce good results even with partial occlusions and overlaps. Cerutti et al. [6] presented a method designed to perform segmentation of a leaf in a natural scene. It's based on the optimization of a polygonal leaf model, which is used as a shape prior for active contour segmentation. They classified nearly 50 tree species by using global shape descriptors. Instead of giving a certain classification result, this approach returns five possible types for each leaf image. With a high probability, the leaf image is from one of the possible types.

3 Feature Extraction

We used two groups of features from leaf images in this work. The first group of features is based on shapes of leaf images (SF-Shape Features) and the other group is based on color values of leaf images (CF-Color Features).

3.1 Shape Features

We used the morphological features in [22] as the shape features, which are common shape features used in the literature. There are five fundamental features: the longest distance between any two points on a leaf border (L), the length of main vein (lengthwise-L_v), the widest distance of a leaf (crosswise-W), the leaf area (A) and the leaf perimeter (P). Then, twelve features are constructed using these five fundamental features by some mathematical operations:

- smoothness of a leaf image
- aspect ratio (L_v/W)
- form factor, the difference between a leaf and a circle ($4\pi A/P^2$)
- rectangularity ($L_v W/A$)
- narrow factor (L/L_v)
- ratio of perimeter to longest distance (P/L)
- ratio of perimeter to the sum of main vein length and widest distance ($P/(L_v+W)$)
- and five structural features obtained by applying morphological opening on grayscale image.

3.2 Color Features

In our experiments, we realized that some leaf images in Flavia dataset [22] have very similar shapes. Thus classification accuracy is greatly affected by this similarity. However, even shapes are similar in some leaves, there are some differences in colors of leaves. Therefore, in addition to the shape features, we extracted color based features from leaf images. When calculating these features, we eliminated background color of leaf images.

We defined two sets of color features. In the first set, we used mean (μ) and standard deviation (σ) of intensity values of red, green, blue channels and average of these channels. So that, the first set of color features contain eight features. Mean and standard deviation of each component are calculated as follows:

$$\mu = \frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N} p_{xy} \tag{1}$$

$$\sigma = \sqrt{\frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N} (p_{xy} - \mu)^2} \tag{2}$$

In (1) and (2), M and N are dimensions of an image and p_{xy} is the intensity value of pixel at (x, y) coordinate.

The second set of color features consists of color histograms in red, green, and blue channels. RGB histograms provide us an efficient representation of color distribution. Thus, we are able to effectively analyze color information in an image by using these histograms. After studying several bin sizes, we obtained the best results with 10 bins in each histogram. Since we are calculating three histograms for red, green and blue channels, there are thirty new features in the second set of color features.

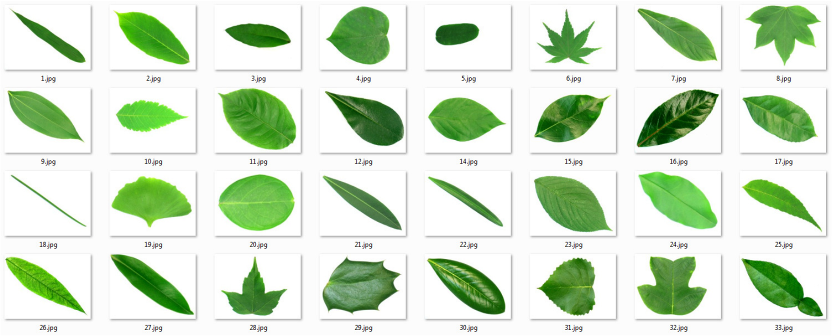


Fig. 1. Screenshot of all 32 leaf types

4 Experiment

To test our feature set, we used the leaf images in Flavia Project data set [22].[1] The data set contains leaf images of 32 plant types, which are common plants in Yangtze Delta, China. Figure 1 lists sample pictures of each leaf type. In our experiments, we used 1897 leaf image samples in total. After extracting shape and color features from leaf images, we classified leaf images using four different classification algorithms: k-Nearest Neighbor, Support Vector Machines, Naive Bayes, and Random Forest.

- Support Vector Machines (SVM): In general, SVM is shown as a case in a two-dimensional space with linearly separable data points. However, it also handles a high dimensional space and data points that are not linearly separable. For a comprehensive information, we suggest looking at these references [8], [3].
- k-Nearest Neighbor (kNN): kNN is a non-parametric classification algorithm [18]. In this algorithm, unknown samples are classified according to their nearest neighbors. For classifying an unknown sample, k closest training samples are determined. The most frequent class among these k neighbors is chosen as the class of this sample. The k value can be determined according to error rates and it should not be too big or too small. There is no training time in kNN. However, its testing time is long because all computations are made at that time.
- Naive Bayes: Bayesian classifiers [13] are statistical models and able to predict probability that an unknown sample belongs to a specific class. It is a practical learning approach and based on Bayes Theorem. A disadvantage of this learning algorithm is that conditional independence may decrease accuracy. It is a constraint over attributes because there may not be dependencies between them.

[1] The data set can be downloaded from `http://flavia.sourceforge.net`

Table 1. Classification Accuracy of Four Classification Methods

(a) Using Only Shape Features

Method	**CV**	**RS**
SVM	0.7185	0.7289
kNN	0.8234	0.8158
Naive Bayes	0.7992	0.8026
Random Forest	0.8761	0.8605

(b) Using Shape Features and the First Set of Color Features

Method	**CV**	**RS**
SVM	0.8650	0.8474
kNN	0.9246	0.9158
Naive Bayes	0.8877	0.8842
Random Forest	0.9388	0.9395

- Random Forest: This approach is based on classification tree approach [4]. For a data set, predictions of multiple classification trees are aggregated. Each tree in the forest is grown using bootstrap samples. At prediction, classification results are taken from each tree and that means trees in the forest use their votes for the target class. The class which has most votes is selected by the forest. Random forests can be efficient on large data sets with high accuracy. However, it has some constraints on memory and computing time.

When testing with these classification algorithms, we used two sampling approaches. In the random sampling approach, we used 80% of images for training and the remaining 20% for testing. In the second approach, we used 10-fold cross validation method. In the 10-fold cross-validation, the dataset is partitioned into 10 equal sized subsamples. A single subsample is used for testing, and the remaining 9 subsamples are used as training data. This process is repeated 10 times and a different subsample is used in each run. The average of 10 runs is used as the final result.

In the experiments, we run each experiment 5 times and presented the average of experiments in the paper. However, due to randomness of classification algorithms, these results may have minor changes after every run.

4.1 Experimental Results

We performed three different experiments with various sets of features. In the first experiment, we made classification with the shape features. In the second experiment, we used the shape and the first set of color features together in classification. In the last experiment, we used all shape and color features (both first and second set of color features) together.

For the first experiment, the best classification accuracy results using only the shape features are shown in Table 1(a). In the table, cross validation and random sampling methods are abbreviated as CV and RS. The worst success rate was nearly 72% and obtained with SVM approach. The best result was approximately 88% with the Random Forest algorithm.

Table 2. Classification Accuracy of Four Classification Methods Using All Shape and Color Features (Including Both First and Second Set of Color Features)

Method	**CV**	**RS**
SVM	0.9283	0.9289
kNN	0.9304	0.9421
Naive Bayes	0.8925	0.8895
Random Forest	0.9610	**0.9632**

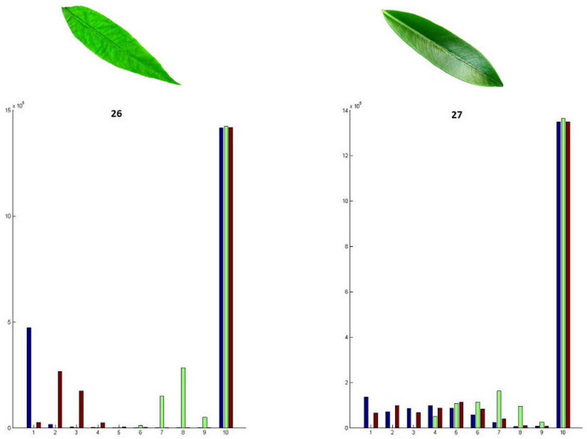
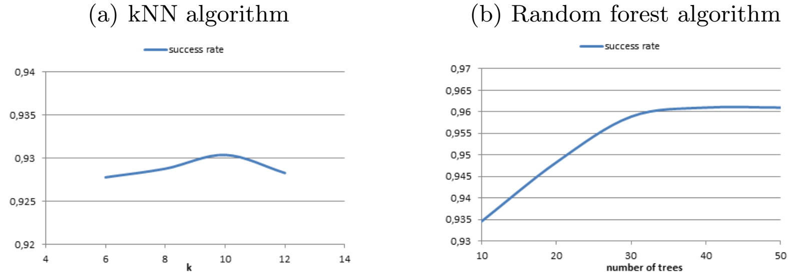


Fig. 2. Sample leaves for 26. and 27. species and their color histograms for red, green, and blue channels.

For the second experiment, the results with shape features and the first set of color features are shown in Table 1(b). As it is clear from Table 1(a) and 1(b), using shape and color features together increases the classification accuracy significantly. The greatest increase was observed in SVM algorithm, in which the classification accuracy was improved about 15%. As in the first experiment, the best result of this experiment is obtained with the Random Forest algorithm, which is 93.95% classification accuracy.

Although we obtained 93.95% classification accuracy with the first set of color features, we tried to increase accuracy with more color features. In the last experiment, we used shape features with both first and second set of color features. Table 2 shows results of the last experiment. Adding color histograms in red, green, and blue channels improved the classification accuracy and produced the best results in our experiments. The best classification accuracy in our experiments is 96.32% when Random Forest algorithm is used with all shape and color features. In all experiments, both CV and RS methods produce close results.


(a) kNN algorithm

(b) Random forest algorithm

Fig. 3. Effect of k and tree number parameters on the success ratio in kNN and random forest algorithms

4.2 Discussion

As it can be seen from Table 1(a), 1(b), and 2, using only shape features increases classification errors. The classification accuracies are increased significantly by adding color features. When only shape features are used, false classification rates increase for similar shaped leaves of some species. As an example from our data set, leaf samples of 26. and 27. species are given in Figure 2. These leaf species are often misclassified when only shape features are used. Although these leaves have similar shapes, color histograms of them are different. It is clear that using a combination of shape and color features increases the classification accuracy. Furthermore, in Flavia dataset, the leaf samples have different orientations (left-oriented or right-oriented) but the proposed approach handled this situation very well. In our experiments, for kNN and Random Forest algorithms, we studied various k value and tree number parameters. For SVM classifier, we used a linear kernel function. Figure 3(a) shows the success ratio related to the neighbor number (k) for kNN method. The results belong to the last experiment (when both shape features and all color features are used). When k number increases, the success ratio also increases up to $k = 10$. After this point, the success ratio decreases. In Figure 3(b), the success ratio with respect to the tree number in the Random Forest algorithm is shown. The results in this graph also belong to the last step of the experiment. The success rate improves until number of trees reaches 40 and then stabilizes. However, we obtained the best results with 50 trees and presented the results with 50 trees in the paper. Finally, we compared the performance of our proposed work to the other multi-species recognition methods in the literature. We consider only leaf-based plant recognition methods. We also ignore methods developed for only certain species. The comparison results are listed in Table 3. As it can be seen from these results, the proposed work produces the best results with a maximum number of different species. The referenced papers use various testing approaches and thus our results may not be directly compared with other papers. However, we tested our approach with two testing methods (cross validation and random sampling) and both methods produced more than 96% classification accuracy for the random forest algorithm. Therefore, the proposed approach can be considered as a robust approach to classify plant leaves.

Table 3. Comparison of the proposed method with the previous methods

Refs	Accuracy Rate	Number of Species
Backes et al. [1]	89.1%	10
Casanova et al. [5]	84%	20
Du et al. [10]	93%	20
Gu et al. [12]	93.2%	20
Wu et al. [22]	90.3%	32
Uluturk et al. [19]	92.5%	32
Lee et al. [15]	95.4%	32
Proposed Method	96.3%	32

5 Conclusion

In this paper, we proposed an approach to recognize 32 plant species by using leaf images. Both shape and color features extracted from leaf images are used to classify plants. For classification, Support Vector Machines, k-Nearest Neighbor, Naive Bayes, and Random Forest algorithms are studied. The best results are obtained with the Random Forest algorithm. Experiments confirmed that using only shape features does not produce good results when classifying similar shaped leaves. Classification accuracy can be improved by using shape and color features together. However, seasonal changes of leaf color may reduce classification accuracy. Therefore, textural features may also be studied to classify plants as a future work. Data sets that provide leaf image samples from various seasons are needed to verify that textural features are helpful to recognize leaves independent from seasonal changes.

References

1. Backes, A.R., Casanova, D., Bruno, O.M.: Plant leaf identification based on volumetric fractal dimension. IEEE PAMI 23, 1145–1160 (2009)
2. Bruno, O.M., de Oliveira Plotze, R., Falvo, M., de Castro, M.: Fractal dimension applied to plant identification. Information Sciences 178(12), 2722–2733 (2008)
3. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. 2, 121–167 (1998)
4. Caiola, G., Reiter, J.P.: Random forests for generating partially synthetic, categorical data. Trans. Data Privacy 3, 27–42 (2010)
5. Casanova, D., de Mesquita Sá Junior, J.J., Bruno, O.M.: Plant leaf identification using gabor wavelets. Int. J. Imaging Syst. Technol. 19(3), 236–243 (2009)
6. Cerutti, G., Tougne, L., Mille, J., Vacavant, A., Coquin, D.: Guiding active contours for tree leaf segmentation and identification. Amsterdam, Netherlands. Cross-Language Evaluation Forum (2011)
7. Chaki, J., Parekh, R.: Plant leaf recognition using shape based features and neural network classifiers. International Journal of Advanced Computer Science and Applications 2 (2011)
8. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, 1st edn. Cambridge University Press (2000)

9. Dake, W., Chengwei, M.: The support vector machine (SVM) based near-infrared spectrum recognition of leaves infected by the leafminers. In: Proceedings of International Conference on Innovative Computing, vol. 3, pp. 448–451 (2006)
10. Du, J.X., Wang, X., Zhang, G.-J.: Leaf shape based plant species recognition. Applied Mathematics and Computation 185(2), 883–893 (2007)
11. Gouveia, F., Filipe, V., Reis, M., Couto, C., Bulas-Cruz, J.: Biometry: the characterisation of chestnut-tree leaves using computer vision. In: IEEE International Symposium on Industrial Electronics, Guimaraes, Portugal (1997)
12. Gu, X., Du, J.X., Wang, X.-F.: Leaf recognition based on the combination of wavelet transform and Gaussian interpolation. In: Huang, D.-S., Zhang, X.-P., Huang, G.-B. (eds.) ICIC 2005. LNCS, vol. 3644, pp. 253–262. Springer, Heidelberg (2005)
13. Islam, M.J., Wu, Q.M.J., Ahmadi, M., Sid-Ahmed, M.A.: Investigating the performance of naive- bayes classifiers and k-nearest neighbor classifiers. JCIT (2007)
14. Jian, Z., Wei, Z.: Support vector machine for recognition of cucumber leaf diseases. In: International Conference on Advanced Computer Control (ICACC) (2010)
15. Lee, K.-B., Chung, K.-W., Hong, K.-S.: An implementation of leaf recognition system based on leaf contour and centroid for plant classification. In: Han, Y.-H., Park, D.-S., Jia, W., Yeo, S.-S. (eds.) Ubiquitous Information Technologies and Applications. LNEE, vol. 214, pp. 109–116. Springer, Netherlands (2013)
16. Manh, A.-G., Rabatel, G., Assemat, L., Aldon, M.-J.: Weed leaf image segmentation by deformable templates. J. Agric. Engng. Res. 80(2), 139–146 (2001)
17. Mishra, P.K., Maurya, S.K., Singh, R.K., Misra, A.K.: A semi automatic plant identification based on digital leaf and flower images. In: International Conference on Advances in Engineering, Science and Management (ICAESM), pp. 68–73. IEE (2012)
18. Singh, S.: Nearest-neighbour classifiers in natural scene analysis. Pattern Recognition 34(8), 1601–1612 (2001)
19. Uluturk, C., Ugur, A.: Recognition of leaves based on morphological features derived from two half-regions. In: International Symposium on Innovations in Intelligent Systems and Applications (INISTA), pp. 1–4. IEEE (2012)
20. Valliammal, N., Geethalakshmi, S.N.: Hybrid image segmentation algorithm for leaf recognition and characterization. In: International Conference on Process Automation, Control and Computing (PACC), pp. 1–6 (2011)
21. Wang, X.-F., Huang, D.-S., Du, J.-X., Hu, H., Heutte, L.: Classification of plant leaf images with complicated background. Mathematics and Computation, 916–926 (2008)
22. Wu, S.G., Bao, F.S., Xu, E.Y., Wang, Y.X., Chang, Y.F., Xiang, Q.I.: A leaf recognition algorithm for plant classification using probabilistic neural network. In: IEEE 7th International Symposium on Signal Processing and Information Technology (2007)