Legume Identification by Leaf Vein Images Classification

Mónica G. Larese[1,2,⋆], Roque M. Craviotto[2], Miriam R. Arango[2], Carina Gallo[2], and Pablo M. Granitto[1]

[1] CIFASIS, French Argentine International Center for Information and Systems Sciences, UAM (France) / UNR-CONICET (Argentina)
Bv. 27 de Febrero 210 Bis, 2000, Rosario, Argentina
{larese,granitto}@cifasis-conicet.gov.ar

[2] Estación Experimental Oliveros, Instituto Nacional de Tecnología Agropecuaria
Ruta Nacional 11 km 353, 2206 Oliveros, Santa Fe, Argentina
{rcraviotto,marango,cgallo}@correo.inta.gov.ar

Abstract. In this paper we propose an automatic algorithm able to classify legume leaf images considering only the leaf venation patterns (leaf shape, color and texture are excluded). This method processes leaf images captured with a standard scanner and segments the veins using the Unconstrained Hit-or-Miss Transform (UHMT) and adaptive thresholding. We measure several morphological features on the veins and classify them using Random forests. We applied the process to recognize several legumes (soybean, white bean and red bean). We analyze the importance of the features and select a small set which is relevant for the recognition task. Our automatic procedure outperforms the expert manual classification.

Keywords: Leaf images automatic classification, Legume automatic recognition, Image analysis, Random forests, Unconstrained Hit-or-Miss Transform.

1 Introduction

In the recent literature, many works have been proposed aimed at automatically analyzing leaf images in order to classify them or to perform plant image retrieval. The most common approach consists in considering the leaf shape [1,2,3,4,5,6]. Other works, additionally, take into account the color information [7,8]. Moreover, the leaf texture can also be included [9,10].

However, in many practical situations there are not evident differences in the shape, size, color or texture features of the leaves for the plants under study. This is the case, for example, when identifying individuals from the same specie but which belong to different varieties or cultivars, so their leaves share the same visual properties. In recent studies [11,12,13], the authors highlighted the

⋆ Author to whom all correspondence should be addressed. MGL and PMG acknowledge grant support from ANPCyT PICT 237.

L. Alvarez et al. (Eds.): CIARP 2012, LNCS 7441, pp. 447–454, 2012.

importance of considering the vein architecture to perform leaf-based plant identification. As recent works in the literature demonstrate [14,15], there exists a correlation between the leaf venation characteristics and the leaf properties such as, for example, damage and drought tolerance. If the plants under study have different physiological characteristics, there is a chance that these properties can be reflected in their veins even if the leaves look similar. Then, the motivation of this work consists in developing an automatic procedure to perform leaf recognition exclusively analyzing the morphological traits from the leaf venation system, i.e., no leaf shape, size, color or texture information is considered.

We perform the leaf segmentation using the Unconstrained Hit-or-Miss Transform (UHMT)[18] and adaptive image thresholding in order to extract the veins from the gray scale images. The UHMT is a mathematical morphology operator useful to perform template matching in gray scale images, extracting all the pixels with a certain foreground and background neighboring configuration.

Then we measure simple morphological features and employ the Random forests algorithm [17] for the recognition task. This is a recent ensemble algorithm which uses a set of de-correlated trees as individual classifiers. It performs comparably well against the most powerful state of the art classifiers. It is also able to estimate the importance of the input variables, which we use to discuss the relevance of the venation features for our particular problem.

We apply the whole method to recognize three classes of legumes, namely soybean (*Glycine max (L) Merr*), red and white beans (*Phaseolus vulgaris*). Red and white beans are two varieties from the same specie, with very similar leaves. The only exception is the color of the veins, which are dark for the red bean. However, we do not consider color information in this work, but only vein morphological measures computed on the gray scale images.

We quantitatively assess the performance of the whole procedure by means of computing the classification accuracies for each class. We also report the accuracy obtained by expert manual classification for comparison, showing the improved results achieved by the automatic classifier.

The rest of the paper is organized as follows. In Section 2.1 we describe the leaf images dataset. Sections 2.2 and 2.3 explain the segmentation procedure that we employed to extract the leaf venation system. We summarize the measures that we computed on the segmented veins in Section 2.4. We briefly describe the Random forests classification algorithm in Section 2.5. We present and discuss the results in Section 3, where we assess the performance of the procedure and analyze the relevant features. Finally, we draw some conclusions in Section 4.

2 Materials and Methods

2.1 Leaf Images Dataset

Our dataset consists of a total number of 866 RGB leaf images provided by Instituto Nacional de Tecnología Agropecuaria (INTA, Oliveros, Argentina). The dataset is divided in the following way: 422 images correspond to soybean leaves, 272 images to red bean leaves and 172 to white bean leaves. They are the images

of the first foliage leaves (pre-formed in the seed) of 433 specimens (211 soybean plants, 136 red bean plants and 86 white bean plants). First foliage leaves were selected for the analysis, after 12 days of seedling grow, since their characteristics are less influenced by the environment. We did not use any chemical or biological procedure to physically enhance the leaf veins. Instead, a fast, inexpensive and simple imaging procedure was used: the leaves were scanned using a Hewlett Packard Scanjet-G 3110 scanner, at a resolution of 200 pixels per inch and stored as 24-bit RGB TIFF images.

2.2 Unconstrained Hit-or-Miss Transform (UHMT)

The UHMT is an extension of the Hit-or-Miss Transform (HMT) for gray scale images [18]. It extracts all the pixels matching a certain foreground and background neighboring configuration. A composite structuring element $\mathbf{B}$ is employed, which is a disjoint set formed by one structuring element that specifies the foreground configuration, B_{fg}, and one structuring element for the background setting, B_{bg}. The origin of the composite structuring element matches the foreground.

The UHMT is defined as

$$UHMT_{\mathbf{B}}(Y)(y) = \max\left\{\varepsilon_{B_{fg}}(Y)(y) - \delta_{B_{bg}}(Y)(y), 0\right\}, \tag{1}$$

where Y is a gray scale image with set of pixels y and $\mathbf{B}$ is a composite structuring element. It can be computed as the difference between an erosion with B_{fg}, $\varepsilon_{B_{fg}}(Y)(y)$, and a dilation with B_{bg}, $\delta_{B_{bg}}(Y)(y)$, if $\delta_{B_{bg}}(Y)(y) < \varepsilon_{B_{fg}}(Y)(y)$. Otherwise it equals 0.

2.3 Vein Segmentation

Since we are only interested in the vein morphology, we removed all the color information by converting the RGB images to gray scale. We thresholded the gray scale image Y and filled its holes using morphological reconstruction [18]. After deleting all the connected components except the largest one, we obtained a binary mask for the leaf.

On the other hand, we computed the UHMT on 5 different sized versions of Y, namely at 100%, 90%, 80%, 70% and 60%. Each version is intended to highlight a different level of vein detail. Next, we resized back to the original size each resulting UHMT and summed them to obtain the combined UHMT, which highlights both small and large visible veins simultaneously. We used the 4 composite structuring elements (foreground and background configurations) shown in Fig. 1 to detect leaf veins in 4 directions (vertical, horizontal, +45°and -45°). After that, we enhanced the contrast of the combined UHMT and binarized the result by means of a standard adaptive thresholding algorithm. We removed all the connected components with less than 20 pixels.

Finally, we computed the product between the result of the segmentation and the previously computed leaf binary mask.

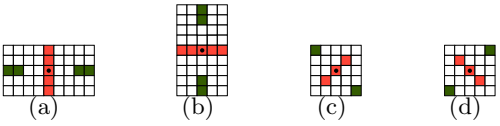

Fig. 1. The four pairs of flat composite structuring elements used for the UHMT to detect veins in four directions: (a) Vertical, (b) horizontal, (c) +45°and (d) -45°. Foreground and background pixel configurations are shown in red and green, respectively. The center of the composite structuring element is marked with a black dot.

2.4 Vein Measurements

In order to avoid the influence of the leaf shape we cropped a centered 100 × 100-pixel patch from the combined UHMT for each leaf, and measured all the features on these patches. We adapted LeafGUI [16] measures to extract a set of features of interest for veins and areoles. For our particular problem aimed at leaf classification, the individual vein/areole measures computed by LeafGUI are not suitable. For this reason, we computed the median, minimum and maximum feature values for veins and areoles where it was appropriate. The interested reader can found in the paper by Price et al. [16] a detailed explanation of each feature as well as the computation procedure. An exception is the edge orientation (not available in LeafGUI), which we computed as the angle (in the range [-90°, 90°]) between the x-axis and the major axis of the ellipse with the same second order moments as the vein.

We measured the following features: Total number of veins (VNE); Total number of nodes (VNN), i.e., number of connecting nodes between veins; Total network length (VTNL; in mm); Median/min/max edge length (VMeL/VmL/VML; in mm); Median/min/max edge width (VMeW/VmW/VMW; in mm); Median/min/max edge 2D area (VMeA/VmA/VMA; in mm^2); Median/min/max edge surface area (VMeSA/VmSA/VMSA; in mm^2); Median/min/max edge volume (VMeV/VmV/VMV; in mm^3); Median/min/max edge orientation (VMeO/VmO/VMO; in degrees); Total number of areoles (AN) in the image patch; Median/min/max areole perimeter (AMeP/AmP/AMP; in mm); Median/min/max areole area (AMeA/AmA/AMA; in mm^2); Median/min/max areole convex area (AMeCA/AmCA/AMCA; in mm^2); Median/min/max areole solidity (AMeS/AmS/AMS; dimensionless in [0,1]); Median/min/max areole major axis (AMeMaA/AmMaA/AMMaA; in mm); Median/min/max areole minor axis (AMeMiA/AmMiA/AMMiA; in mm); Median/min/max areole eccentricity (AMeE/AmE/AME; dimensionless between 0 -a circle- and 1 -a line-); Median/min/max areole equivalent diameter (AMeEq/AmEq/AMEq; in mm); Median/min/max areole mean distance (AMeMD/AmMD/AMMD; in mm); Median/min/max areole variance distance (AMeVD/AmVD/AMVD; in mm). Altogether, these measures make a feature vector of 52 components.

2.5 Random Forests

Random forests [17] is a recent ensemble algorithm where the individual classifiers are a set of de-correlated trees. They perform comparably well to other state of the art classifiers and are also very fast. Random forests also allows to estimate the importance of input variables (in their original dimensional space).

The algorithm constructs a set of unpruned trees from B random samples with replacement (bootstrap versions) of the original training dataset. For each random forest tree f_b, a random sample of m variables from the full set of p variables ($m \leq p$) is selected to split the data at each node and grow the decision tree. The final classification result $F(\mathbf{d}_i)$ is the class corresponding to the majority vote of the ensemble of trees,

$$F(\mathbf{d}_i) = \text{majority vote } \{f_b(\mathbf{d}_i)\}_{b=1}^{B} \tag{2}$$

Random forests has a built-in procedure to estimate the relevance of the input variables. After training the model, the features are shuffled one at a time. An out-of-bag estimation of the prediction error is made on this permuted dataset. Intuitively, a feature that is not relevant to the model will not alter significantly the classification performance when shuffled. On the other hand, if the model made strong use of a given feature, modifying its values will produce an important decrease in performance. The relative loss in performance between the original dataset and the shuffled dataset is therefore related to the relative relevance of the feature affected by the process.

In this work we used 500 trees and a standard value of $m = \sqrt{p}$ for the number of variables randomly sampled as candidates at each split.

3 Results and Discussion

We show in Fig. 2 an example of the segmentation results for a soybean leaf, a white bean leaf and a red bean leaf, as well as the 100×100-pixel central patches used for feature extraction. Figures 2(b), (e) and (h) are the combined UHMT images segmented according to Section 2.3. As it can be noticed from this figure, mainly primary order veins are extracted. Higher order veins (e.g. terminal veins) are not possible to segment since they are not visible (the images were scanned with no clearing or amplification procedures, as explained in Section 2.1).

We computed the 52 features described in Section 2.4 for each leaf patch. However, we had to discard 13 features since they presented a constant value for all the leaves (except for a few outliers), namely: VmL, VmO, VMO, AmA, AmCA, AmE, AmEq, AmMaA, AmMia, AmP, AmMD, AMS and AmVD. Thus, we used for classification 39 out of the 52 originally computed features.

We report in Table 1 the results for the classification of the 3 different legume species performing leave-one-out cross validation (LOOCV). The accuracies we obtained using 39 features are depicted in the first row of Table 1, showing that the identification rates are very good: in average, only 18 leaves out of 422 are misclassified for soybean and 39 out of 272 are misclassified for red bean. The

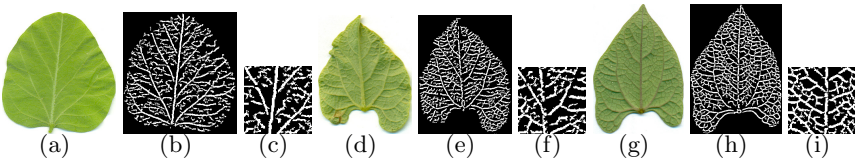

Fig. 2. (a) Soybean leaf. (b) Vein segmentation for (a). (c) 100×100-pixel central patch from (b). (d) White bean leaf. (e) Vein segmentation for (d). (f) 100×100-pixel central patch from (e). (g) Red bean leaf. (h) Vein segmentation for (g). (i) 100×100-pixel central patch from (h).

classification of white bean leaves seems to be the most difficult, although the results are quite satisfactory (in average 49 out of 172 images are misclassified).

Our automatic procedure with 39 features achieves an improvement of 16.22% over manual classification based on the same central patches (average of 5 experts; shown in the third row of Table 1) for red bean recognition, and 5.08% for the white bean.

However, the computation of the correlation matrix between the features (not included because of lack of space) shows that some of these features are strongly correlated one to each other (Pearson coefficient $|r| \geq 0.9$), e.g., the vein median volume (VMeV) with the vein median length (VMeL), the vein median area (VMeA) and the vein median surface area (VMeSA). This means that some of the features are redundant, and could lead to some overfitting of the data.

In order to find a small set of non-redundant relevant features able to describe the vein differences between soybean, white bean and red bean, we report in Fig. 3 the 30 most relevant traits according to Random forests. From this figure, it can be noticed that there is a differentiated small group of 7 features which achieve a mean decrease accuracy of more than 0.7 [1]. These features are, in order of priority, the vein median width (VMeW), the areole minimum solidity (AmS), the vein median orientation (VMeO), the vein median volume (VMeV), the number of areoles (AN), the vein maximum volume (VMV) and the total network length (VTNL).

In order to analyze the degree of independence between these 7 features we also computed the correlation matrix for them and concluded that there were not strong correlations ($|r| < 0.9$). Performing classification with only this small subset achieves the accuracies shown in the second row of Table 1.

This small set of 7 uncorrelated features achieves a performance similar to the one obtained by using the 39 original features (the accuracy improves slightly for the white bean while it diminishes a little for the red bean, still providing an improvement of 12.55% over manual classification), having the advantage that they are a reduced number of non-redundant properties and much easier to handle. In both cases the experts achieve a better recognition for soybean, although the automatic procedure has a high degree of accuracy (over 95%).

[1] The procedure followed for feature selection is the one suggested by Breiman at `http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#varimp`.

Table 1. Accuracy for legume detection using different numbers of relevant features

Number of features	Soybean	White bean	Red bean
39	95.74%	71.51%	85.66%
7	95.50%	72.67%	81.99%
Manual classification	98.29%	66.43%	69.44%

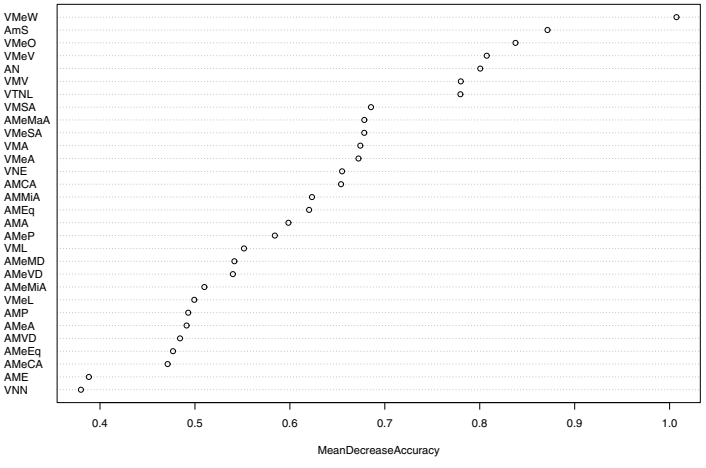

Fig. 3. Variable importance showing the 30 most relevant features

4 Conclusions

In this work we propose an automatic legume recognition procedure based only on the analysis of leaf vein morphology. The leaves are captured using a standard scanner and do not require any clearing or staining treatment. In order to perform segmentation, we used the UHMT and adaptive thresholding. We measured several morphological features on the segmented venation network from a small central patch of the images, and classified them with Random forests.

We found a small subset of 7 uncorrelated features ($|r| < 0.9$) which improve expert manual classification for two of the three classes (white bean and red bean). Although the recognition for soybean is better done by humans, the automatic algorithm achieves over 95% of accuracy. Overall, the automatic method improves the results obtained by human experts, with clear advantages in repetibility, confiability and economy. The 7 relevant features are, in order of priority: the vein median width, the areole minimum solidity, the vein median orientation, the vein median volume, the number of areoles, the vein maximum volume and the total network length.

Since this procedure shows to be successful to identify different legume species, we are currently working on extending this method to separate different cultivars from a single specie.

References

1. Im, C., Nishida, H., et al.: Recognizing plant species by leaf shapes-a case study of the Acer family. In: ICPR, vol. 2, pp. 1171–1173 (1998)
2. Agarwal, G., Ling, H., et al.: First steps toward an electronic field guide for plants. Taxon, J. of the International Association for Plant Taxonomy 55, 597–610 (2006)
3. Camargo Neto, J., Meyer, G.E., et al.: Plant species identification using Elliptic Fourier leaf shape analysis. Comput. Electron. Agric. 50, 121–134 (2006)
4. Du, J.X., Wang, X.F., et al.: Leaf shape based plant species recognition. Applied Mathematics and Computation 185(2), 883–893 (2007), special Issue on Intelligent Computing Theory and Methodology
5. Solé-Casals, J., Travieso, C.M., et al.: Improving a leaves automatic recognition process using PCA. In: IWPACBB, pp. 243–251 (2008)
6. Chaki, J., Parekh, R.: Designing an automated system for plant leaf recognition. Int. Journal of Advances in Engineering & Technology 2(1), 149–158 (2012)
7. Horgan, G.W., Talbot, M., et al.: Towards automatic recognition of plant varieties. In: British Computer Society Electronic Workshops in Computing: The Challenge of Image Retrieval (1998)
8. Perez, A., Lopez, F., et al.: Colour and shape analysis techniques for weed detection in cereal fields. Comput. Electron. Agric. 25, 197–212 (2000)
9. Golzarian, M.R., Frick, R.A.: Classification of images of wheat, ryegrass and brome grass species at early growth stages using principal component analysis. Plant. Methods 7(28) (2011)
10. Bama, B.S., Valli, S.M., et al.: Content based leaf image retrieval (CBLIR) using shape, color and texture features. Indian Journal of Computer Science and Engineering 2(2), 202–211 (2011)
11. Clarke, J., Barman, S., Remagnino, P., Bailey, K., Kirkup, D., Mayo, S., Wilkin, P.: Venation Pattern Analysis of Leaf Images. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Remagnino, P., Nefian, A., Meenakshisundaram, G., Pascucci, V., Zara, J., Molineros, J., Theisel, H., Malzbender, T. (eds.) ISVC 2006. LNCS, vol. 4292, pp. 427–436. Springer, Heidelberg (2006)
12. Park, J., Hwang, E., et al.: Utilizing venation features for efficient leaf image retrieval. J. Syst. Softw. 81(1), 71–82 (2008)
13. Valliammal, N., Geethalakshmi, S.: Hybrid image segmentation algorithm for leaf recognition and characterization. In: PACC 2011, pp. 1–6 (July 2011)
14. Sack, L., Dietrich, E.M., et al.: Leaf palmate venation and vascular redundancy confer tolerance of hydraulic disruption. PNAS USA 105, 1567–1572 (2008)
15. Scoffoni, C., Rawls, M., et al.: Decline of leaf hydraulic conductance with dehydration: relationship to leaf size and venation architecture. Plant Physiology 156, 832–843 (2011)
16. Price, C.A., Symonova, O., et al.: Leaf extraction and analysis framework graphical user interface: Segmenting and analyzing the structure of leaf veins and areoles. Plant Physiology 155, 236–245 (2011)
17. Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001)
18. Soille, P.: Morphological Image Analysis: Principles and Applications. Springer (1999)