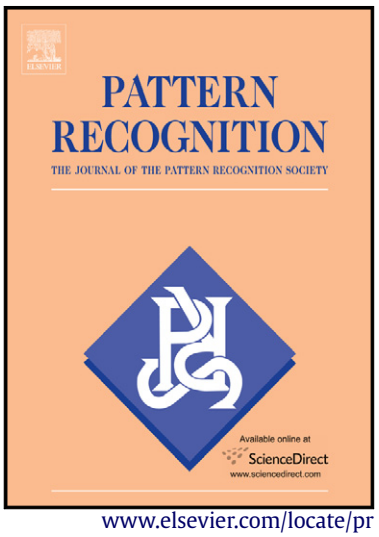


www.elsevier.com/locate/pr

Automatic classification of legumes using leaf vein image features

Mónica G. Larese, Rafael Namías, Roque M. Craviotto, Miriam R. Arango, Carina Gallo, Pablo M. Granitto

PII: S0031-3203(13)00264-1
DOI: http://dx.doi.org/10.1016/j.patcog.2013.06.012
Reference: PR4834

To appear in: *Pattern Recognition*

Received date: 2 January 2013
Revised date: 13 May 2013
Accepted date: 10 June 2013

Cite this article as: Mónica G. Larese, Rafael Namías, Roque M. Craviotto, Miriam R. Arango, Carina Gallo, Pablo M. Granitto, Automatic classification of legumes using leaf vein image features, *Pattern Recognition,* http://dx.doi.org/10.1016/j.patcog.2013.06.012

Automatic classification of legumes using leaf vein image features

Mónica G. Larese[a,*], Rafael Namías[a], Roque M. Craviotto[b], Miriam R. Arango[b], Carina Gallo[b], Pablo M. Granitto[a]

[a] *CIFASIS, French Argentine International Center for Information and Systems Sciences, UAM (France) / UNR-CONICET (Argentina) Bv. 27 de Febrero 210 Bis, 2000 Rosario, Argentina*

[b] *Oliveros Experimental Station, National Institute of Agricultural Technology (INTA) Ruta Nacional 11 km 353, 2206 Oliveros, Santa Fe, Argentina*

Abstract

In this paper, a procedure for segmenting and classifying scanned legume leaves based only on the analysis of their veins is proposed (leaf shape, size, texture and color are discarded). Three legume species are studied, namely soybean, red and white beans. The leaf images are acquired using a standard scanner. The segmentation is performed using the Unconstrained Hit-or-Miss Transform and adaptive thresholding. Several morphological features are computed on the segmented venation, and classified using four alternative classifiers, namely Support Vector Machines (linear and Gaussian kernels), Penalized Discriminant Analysis and Random Forests. The performance is compared to the one obtained with cleared leaves images, which require a more expensive, time consuming and delicate procedure of acquisition. The results are encouraging, showing that the proposed approach is an effective and more economic alternative solution which outperforms the manual expert's recognition.

Keywords: Leaf vein features, Leaf vein images, Legume classification, Leaf vein analysis, Unconstrained Hit-or-Miss Transform

1. Introduction

The automatic analysis of leaf images aimed at plant classification or plant image retrieval has been addressed by many researchers in the recent literature.

*Corresponding author. Tel.: +54-(0)341-4237248 ext.303; fax: +54-(0)341-4237248 ext.301

Email addresses: larese@cifasis-conicet.gov.ar (Mónica G. Larese), namias@cifasis-conicet.gov.ar (Rafael Namías), rcraviotto@correo.inta.gov.ar (Roque M. Craviotto), marango@correo.inta.gov.ar (Miriam R. Arango), cgallo@correo.inta.gov.ar (Carina Gallo), granitto@cifasis-conicet.gov.ar (Pablo M. Granitto)

Several approaches have been proposed, including leaf shape [12, 1, 4, 7, 20, 5], color information [11, 15] and leaf texture analysis [8, 2].

Although all these approaches are valid, they are not useful when dealing with species having similar leaf size, color, shape and texture features. For example, such is the case with individuals from different varieties of the same species, which have no clear visual differences in the previously mentioned leaf characteristics. Recently, some authors [6, 14, 22] stated that leaf venation properties may be of high importance to perform plant recognition. This hypothesis is also supported by recent studies [17, 18] which show correlations between venation networks and leaf properties (for example, drought and damage tolerance). Under these assumptions, it is feasible to think that the particular physiological characteristics of the plants are reflected in their leaf veins, even when the leaves have similar appearance.

In this work an automatic procedure exclusively based on the analysis of leaf vein morphological features is proposed for plant recognition. Leaf shape, texture, color and size are discarded. Leaf vein segmentation is performed resorting to the Unconstrained Hit-or-Miss Transform (UHMT)[19] and adaptive image thresholding applied to the gray scale leaf images. The UHMT is a mathematical morphology operator similar to template matching. It allows to extract all the pixels having a certain foreground and background neighboring configuration.

Simple morphological features are measured on the segmented veins, and four different state-of-the-art classifiers are compared to perform plant identification, namely Support Vector Machines (SVM) [23] with linear and Gaussian kernels, Penalized Discriminant Analysis (PDA) [10] and Random Forests (RF) [3].

The whole procedure was used to recognize three classes of legumes, namely soybean (*Glycine max (L) Merr*), red and white beans (*Phaseolus vulgaris*). Red and white beans belong to the same species, presenting similar leaves except for their vein color, which is dark for the red bean. However, color is not taken into account in this paper. Only vein morphological features are considered on gray scale images.

We report the quantitative performance of the whole procedure, discussing

the classification accuracies per class achieved by the automatic classifiers and the advantages of the proposed methodology. The procedure was developed searching also for simplicity and low cost. For this reason, the leaf images were acquired using a standard scanner, without any staining procedure. The results were compared to the performance achieved by human experts.

The proposed approach was also compared to the more sophisticated methodology of analyzing digital photographs of cleared leaves images. This alternative provides with enhanced high contrast leaf veins and higher orders of visible veins, but it is a much more expensive and time demanding procedure given the chemical staining process applied to the leaves.

The rest of the paper is organized as follows. In Sections 2.1 and 2.2 the proposed vein segmentation procedure is explained. The morphological measures computed on the segmented veins are summarized in Section 2.3. The employed classification algorithms are briefly described in Section 2.4. In Section 3, we describe the leaf images datasets and discuss the obtained results. Finally, some conclusions and future work are presented in Section 4.

2. Materials and methods

2.1. Unconstrained Hit-or-Miss Transform (UHMT)

The UHMT is an extension of the Hit-or-Miss Transform (HMT) for gray scale images [19]. 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

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

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.

2.2. Vein segmentation

The color information was removed by converting the RGB images to grayscale. The color information is discarded since there is interest in detecting vein patterns associated to vein morphology only.

The binary masks for the leaves were obtained via thresholding (automatic iterative threshold selection [21]), holes filling using morphological reconstruction [19] and removal of all the connected components except the largest one.

In order to segment the veins in the scanned images, the UHMTs on 5 different sized versions of the images, namely at 100%, 90%, 80%, 70% and 60%, were computed. Each version is intended to highlight a different level of vein detail. Then, each resulting UHMT was resized back to its original size and added to obtain the combined UHMT, which highlights both small and large visible veins simultaneously. For this purpose, four composite structuring elements (foreground and background configurations) were used aimed at detecting leaf veins in 4 directions (vertical, horizontal, +45°and -45°). These structuring elements are shown in Fig. 1. After that, the contrast of the combined UHMT was enhanced and then binarized by means of a standard adaptive thresholding algorithm. All the connected components with less than 20 pixels were removed.

For the cleared images, the veins are already highlighted due to the staining procedure. For this reason, the segmentation was performed by simply applying adaptive histogram equalization followed by standard adaptive thresholding.

2.3. Vein measurements

In order to measure vein and areole features without the influence of the leaf shape, a central patch was extracted from each segmented scanned and cleared leaf, respectively. Ideally, we would like to work with the entire leaf

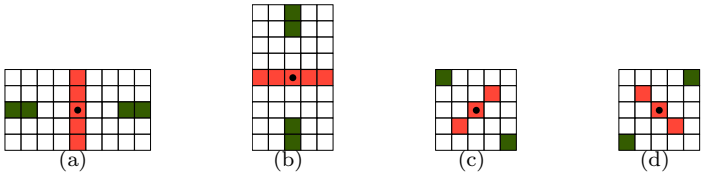


Figure 1: Flat composite structuring elements used for the UHMT to detect veins in four directions: (a) Vertical, (b) horizontal, (c) +45°and (d) -45°. Foreground and background pixel configurations are shown in red and green, respectively. The center of the composite structuring element is marked with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

venation network. However, in practice we cannot achieve this since we want to discard the leaf shape contour influence in order to analyze exclusively vein features. For this reason, in this paper we chose to extract a patch located at the center of each leaf, which we consider significant in order to capture primary and secondary order veins features, with a size big enough to include higher order veins. Another significant point of analysis could have been the union between the leaf blade and the petiole, or the leaf apex, but the vein characteristics at both locations are very much influenced by the leaf shape, so they were discarded. However, if the patch is too big we risk to touch the leaf contour and include it unintentionally. With these requirements in mind, we selected 100×100 pixel-sized patches for scanned leaves since this selection accomplished the goal, and 400×400 pixel-sized patches for cleared leaves, since the resolution of the latter is approximately 4 times higher than the former's. All the traits were computed on these patches, and the same traits were computed for the scanned and cleared leaves.

LEAF GUI measures [16] were adapted to extract a set of features of interest for the veins and areoles. For the particular problem of leaf classification, the individual vein/areole measures computed by LEAF GUI are not suitable. For this reason, the median, minimum and maximum measure values were computed for the veins and areoles where it was appropriate. An extra measurement not available in LEAF GUI, namely the vein orientation, was also considered

in this paper. Altogether, 52 measures were computed for each leaf patch. In AppendixA, the explanation and computing procedure for each measure is provided. Further information can be found in the work by Price et al. [16].

However, we found that 17 out of these 52 measures had a near zero variance across the leaves and therefore were discarded for classification purposes, namely: VmL, VmW, VmA, VmSA, VmV, VmO, AmP, AmA, AmCA, AmMaA, AmMia, AmE, AmEq, AmMD, AmVD, VMO and AMS. Thus, only 35 out of the 52 originally computed traits were effectively used as features for classification.

2.4. Classification methods

In this work we evaluated 4 different classifiers: Support Vector Machines with linear and Gaussian kernels, Penalized Discriminant Analysis and Random Forests. Each one of them is summarized in the following subsections.

2.4.1. Support Vector Machines (SVM)

Support Vector Machines (SVM) [23, 10] is a state-of-the-art classifier which assumes that applying an appropriate nonlinear mapping of the data into a sufficiently high dimensional space, two classes can be separated by an optimum hyperplane. This decision hyperplane is chosen in such a way that the distance between the nearest patterns of different classes (i.e., the margin) is maximized.

Given a dataset $D = \{(\mathbf{x}_i, y_i)\}$, formed by pairs of features-label examples, with $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$ and $i = 1, \ldots, n$, consider the case where the training examples can be linearly separated. In this case, the two classes can be separated by one of many possible hyperplanes given by:

$$f(\mathbf{x}_i) = \mathbf{w}^T \mathbf{x}_i + b = 0, \tag{2}$$

where $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$. A support vector classifier selects the hyperplane which maximizes the margin. This optimization problem can be posed as

$$\min_{\mathbf{w},b} ||\mathbf{w}||, \quad \text{subject to the constraint} \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1. \tag{3}$$

If the classes are not completely separable (there is overlap in feature space), some patterns might be allowed to be on the opposite side of the margin by introducing the slack variables $\xi = \{\xi_1, \xi_2, \ldots, \xi_n\}$, and converting the minimization problem in Eq. (3) into:

$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}\xi_i, \text{ subject to } \begin{cases} \xi_i \geq 0; \\ y_i(\mathbf{w}^T\mathbf{x} + b) \geq (1 - \xi_i), \forall i, \end{cases} \tag{4}$$

where C is a regularization constant.

If the decision surface is required to be nonlinear, a kernel function can be used to map the original features into a high dimensional space, where they can be separated by a linear boundary. The kernel κ is related to the transform θ following $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \theta(\mathbf{x}_i)\theta(\mathbf{x}_i)$. In this case, the problem can be stated as $f(\mathbf{x}_i) = \mathbf{w}^T\theta(\mathbf{x}_i) + b$, and an optimization problem similar to Eq. (4) can be derived.

In this work, we considered the linear kernel:

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i\mathbf{x}_j + 1 \tag{5}$$

and the Gaussian kernel:

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2}\right). \tag{6}$$

Both the standard deviation σ for the Gaussian kernel and the regularization parameter C were optimized using inner validation during the training.

Binary classification using SVM can be performed by means of the following classification rule

$$\hat{y}_i = \text{sign}(f(\mathbf{x}_i). \tag{7}$$

For the multiclass problem considered in this work, the *one-vs-one* strategy was followed. In this strategy, $k(k-1)/2$ binary classification problems are formulated between all pairs of the k classes. The final result is obtained using

a *max-wins* criterion: the example is preliminary assigned to one of two classes by each binary classifier, the corresponding class adds a vote, and the pattern is finally classified into the class with the maximum number of votes.

2.4.2. Penalized Discriminant Analysis (PDA)

Fisher's Linear Discriminant Analysis (LDA) [10] is a classical classifier and dimension reduction tool which searches for linear combinations of the features in such a way that the class means of the linear combinations are maximally separated relative to the intra-class covariance.

Let $D = \{(\mathbf{x}_i, y_i)\}$ be a labeled dataset with $i = 1, \ldots, n$ input/output examples. Every d-dimensional vector $\mathbf{x}_i$ is associated to one of K possible class labels $y_i \in \{1, 2, \ldots, K\}$. Let $\mathbf{m}_k \in \mathbb{R}^d$ be the centroid of class k (with $k = 1, \ldots, K$), $p_k \in \mathbb{R}$ be the estimated proportion of class k in the whole dataset, Σ_W be the pooled within-class covariance matrix of the inputs and $\bar{\mathbf{m}} = \sum_k p_k \mathbf{m}_k$ be the dataset mean. LDA finds $\boldsymbol{\beta} \in \mathbb{R}^d$ such that:

$$\boldsymbol{\beta}^T \Sigma_B \boldsymbol{\beta} = \boldsymbol{\beta}^T \sum_k p_k (\mathbf{m}_k - \bar{\mathbf{m}})(\mathbf{m}_k - \bar{\mathbf{m}})^T \boldsymbol{\beta} \tag{8}$$

is maximized subject to the constraint $\boldsymbol{\beta}^T \Sigma_W \boldsymbol{\beta} = 1$, where $\Sigma_B = \sum_k p_k (\mathbf{m}_k - \bar{\mathbf{m}})(\mathbf{m}_k - \bar{\mathbf{m}})^T$ denotes the inter-class covariance matrix.

Each $\boldsymbol{\beta}$ vector is a scaled eigenvector of $\Sigma_W^{-1} \Sigma_B$ representing each one of the directions in which the class means are most separable in the transformed space relative to the within-class covariance.

The classification of new observations is performed by assigning them to the closest centroid in the transformed space according to a distance metric (typically the Mahalanobis distance), as depicted by:

$$\hat{y}(\mathbf{x}_{test}) = \arg\min_k (\mathbf{x}_{test} - \mathbf{m}_k)^T \Sigma_W^{-1} (\mathbf{x}_{test} - \mathbf{m}_k) - 2\log\, p_k. \tag{9}$$

The first term is the Mahalanobis distance between the descriptor and every class mean, whereas the second term is the adjustment for the class size.

Fisher's LDA presents several advantages, such as robustness to non-Gaussian distributions and moderately different class covariances. However, it does not perform well when there is a large number of highly correlated variables, leading to overfitting.

In order to face this problem, Penalized Discriminant Analysis (PDA) was proposed by Hastie et al. [9]. PDA is a regularized version of LDA, which adds a penalty term to the intra-class covariance matrix. PDA is useful for image classification problems with large number of noisy features.

PDA proceeds exactly in the same manner as LDA, except for replacing Σ_W by the penalized within covariance matrix $\Sigma'_W = \Sigma_W + \Omega$, with $\Omega \in \mathbb{R}^{d \times d}$ such that $\boldsymbol{\beta}^T \Omega \boldsymbol{\beta}$ is large for $\boldsymbol{\beta}$'s having large Euclidean norm.

The penalty term can be defined as $\Omega = \lambda \mathbb{I}_d$, with $\mathbb{I}_d$ being the $d \times d$ identity matrix. In this definition, λ is a free parameter which controls the shrinkage level of the $||\boldsymbol{\beta}||$'s, similar to standard Ridge Regression [10]. The effect of adding a constant to the diagonal elements of Σ_W is to make rounder constraint ellipsoids in hyperparameter space and avoid their shape degeneration. In this work this parameter was automatically selected using a validation set in the training phase.

2.4.3. Random Forests (RF)

Random Forests (RF) [3] is a state-of-the-art 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 $D = \{(\mathbf{x}_i, y_i)\}$, with $i = 1, \ldots, n$. For each node of 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 and grow the decision tree. Given $\mathbf{x}_i$, the final classification result ($F(\mathbf{x}_i)$) is the class corresponding to the majority vote of the

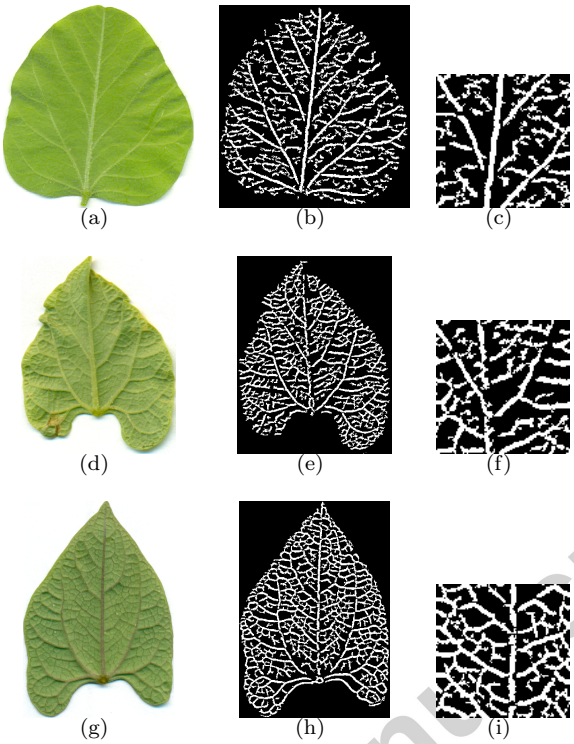


Figure 2: (a) Soybean leaf. (b) Vein segmentation for (a). (c) 100×100-pixel central patch from (b). (d) White bean leaf. (e) Vein segmentation for (d). (f) 100×100-pixel central patch from (e). (g) Red bean leaf. (h) Vein segmentation for (g). (i) 100×100-pixel central patch from (h).

ensemble of trees:

$$F(\mathbf{x}_i) = \text{majority vote } \{f_b(\mathbf{x}_i)\}_{b=1}^{B}. \tag{10}$$

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

3. Results and discussion

3.1. Leaf images datasets

Two datasets were used in this paper, provided by the National Institute of Agricultural Technology (INTA, Oliveros, Argentina). The images in both datasets correspond to the first foliage leaves (pre-formed in the seed) after 12 days of seedling grow. First foliage leaves were selected for the analysis since their characteristics are less influenced by the environment. The first dataset corresponds to the one employed in our previous work [13] and consists of leaf images scanned using a standard scanner. The second dataset is composed by images of chemically cleared leaves which were acquired with a fixed mounted digital camera. Next, both datasets are described in detail.

The scanned images dataset consists of a total number of 866 RGB first-foliage-leaf images. They correspond to 433 specimens (211 soybean plants, 136 red bean plants and 86 white bean plants). The images were obtained via a fast, inexpensive and simple imaging procedure (neither chemical nor biological procedures were used to physically enhance the leaf veins). The leaves were acquired using a Hewlett Packard Scanjet-G 3110 scanner, at a resolution of 200 pixels per inch and stored as 24-bit RGB TIFF images. This 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.

The cleared images dataset is composed by a total number of 150 RGB first-foliage-leaf images (50 soybean leaves, 50 red bean leaves and 50 white bean leaves). The images were clarified by immersion of the Petri dishes in boiling alcohol 96% for 16-18 hours. After that, staining was performed by immersion in saturated Safranin solution for one hour. The images were acquired using a fixed mounted Nikon D90 digital camera with a resolution of 900 pixels per inch, on a Hama Lightbox LP 554.

3.2. Performance evaluation

Three scanned exemplars corresponding to a soybean leaf, a white bean leaf and a red bean leaf, are shown in Figures 2(a), (d) and (g), respectively. The

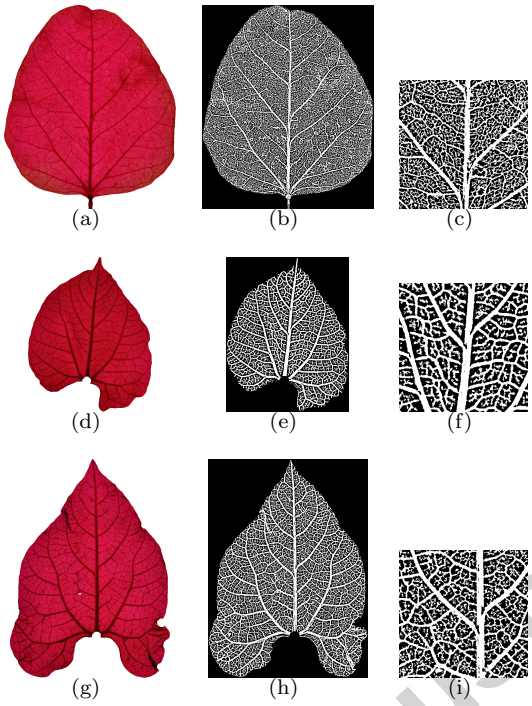


Figure 3: (a) Cleared soybean leaf. (b) Vein segmentation for (a). (c) 300×300-pixel central patch from (b). (d) Cleared white bean leaf. (e) Vein segmentation for (d). (f) 300×300-pixel central patch from (e). (g) Cleared red bean leaf. (h) Vein segmentation for (g). (i) 300×300-pixel central patch from (h).

combined UHMT images segmented according to Section 2.2 are depicted in Figures 2(b), (e) and (h). In Figures 2(c), (f) and (i) the 100×100-pixel central patches used for feature extraction are included. It can be noticed from these figures that only the primary order veins can be extracted for the scanned leaves. The higher order veins (e.g. terminal veins) are not segmented since they are not visible in plain sight (no staining nor amplification procedures were used for this set of images, as it was described previously).

Similarly, three cleared leaves of the same legumes are exemplified in Figures 3(a), (d) and (g), respectively. The segmentation results after adaptive histogram equalization and adaptive thresholding are depicted in Figures 3(b),

(e) and (h). Additionally, the 400×400-pixel patches extracted from the center of the segmented images are shown in Figures 3(c), (f) and (i). It is evident from these figures that veins of higher orders are clearly detected, in comparison with the scanned leaves, due to the help of the staining chemical procedure which highlights many more details of the venation structure.

Estimation errors at any step of the analysis affect the final accuracy of the whole process. For this reason, we evaluated the performance of the whole procedure (image segmentation, feature extraction and classification) by means of computing the final classification accuracies, which we discuss in the following paragraphs.

In Table 1, the average total accuracy obtained by the four considered classifiers are reported both for scanned and cleared leaves. The accuracies reported for each classifier were obtained by computing the mean over 10 independent runs of 10-fold cross validation (CV) of the percentages of the total number of correctly classified leaves using the 35 features described in Section 2.3 and AppendixA (standard deviations are also included).

The two datasets described in Section 3.1 were randomly sampled in order to generate two balanced datasets with number of examples per class equal to the one of the smallest class (172 specimens and 50 specimens per legume class for the scanned and cleared leaves, respectively). The best average performance corresponds to PDA, followed by SVM with linear kernel for scanned leaves and SVM with Gaussian kernel for cleared leaves. The lowest accuracy for scanned leaves corresponds to Random Forests, though it is over 85%. The usage of scanned leaves combined with PDA diminishes slightly the average performance against using cleared leaves. However, the standard deviation is almost the double in the last case. This may be due to the fact that less samples are considered for cleared leaves. The increase in the number of examples for scanned leaves allows to achieve lower standard deviations.

The reason for PDA to perform better is probably related to the characteristics of the features taken into consideration. Some of these features are linearly correlated to one another, as it can be noticed from the correlation matrices of

Table 1: Mean total accuracy and standard deviation for the four classifiers under consideration using scanned vs. cleared leaves, balanced datasets and 10 times 10-fold CV. The best results are highlighted in bold.

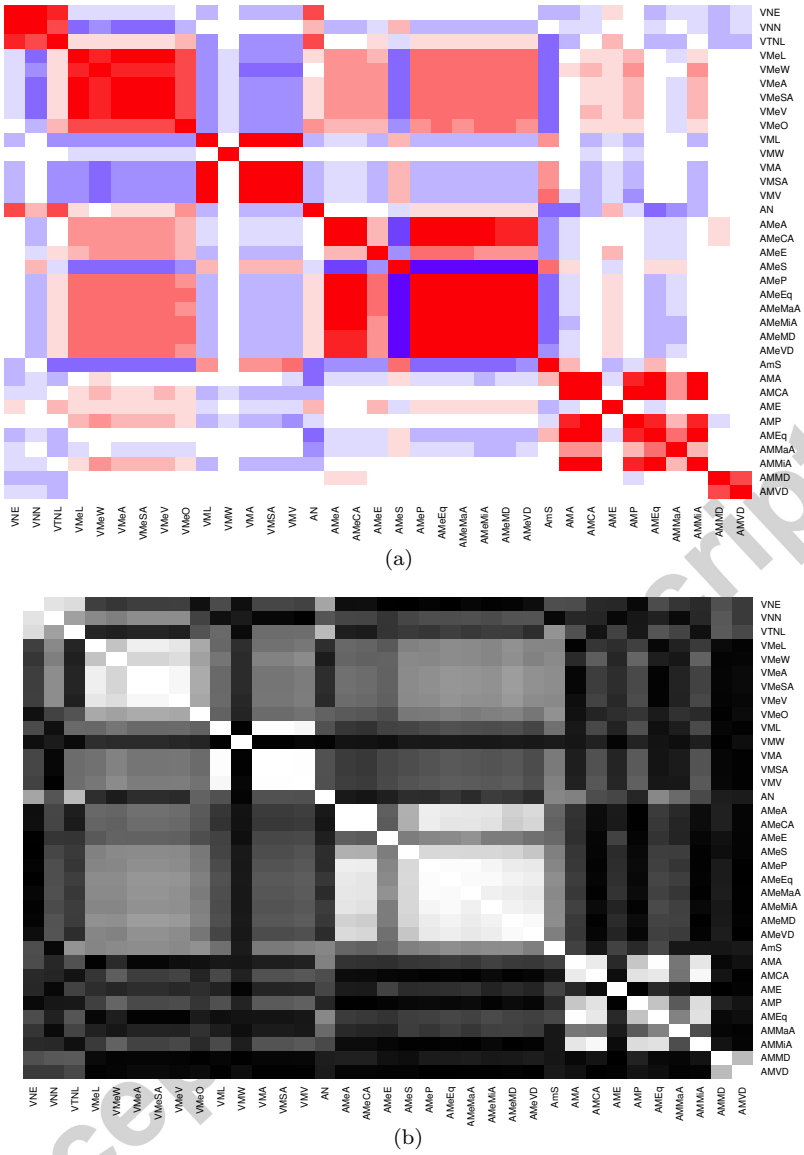
Classifier	**Mean accuracy ± Standard deviation**	
	Scanned leaves	**Cleared leaves**
SVM (Gaussian kernel)	$87.0 \pm 4.5\%$	$86.8 \pm 8.5\%$
SVM (Linear kernel)	$87.2 \pm 4.6\%$	$85.0 \pm 8.7\%$
Random Forests	$85.5 \pm 4.8\%$	$86.1 \pm 8.6\%$
PDA	$\mathbf{87.3 \pm 4.6}\%$	$\mathbf{89.1 \pm 8.6}\%$

Table 2: Mean accuracy per class for legume classification (scanned and cleared leaves) using different classifiers, balanced datasets and 10 times 10-fold CV. The best result per class is highlighted in bold.

	Classifier	**SB**	**WB**	**RB**
Scanned leaves	Manual classification	98.3%	66.4%	69.4%
	SVM (Gaussian kernel)	**95.5%**	83.2%	82.3%
	SVM (Linear kernel)	**95.5%**	82.9%	83.2%
	Random Forests	92.8%	82.8%	80.7%
	PDA	94.4%	**83.6%**	**84.1%**
Cleared leaves	SVM (Gaussian kernel)	95.8%	80.8%	83.8%
	SVM (Linear kernel)	97.6%	80.6%	76.8%
	Random Forests	97.4%	77.8%	83.2%
	PDA	**98.0%**	**81.0%**	**88.4%**

the 35 features for the scanned and cleared leaves, which are depicted in Figures 4 and 5, respectively. In both figures, in part (a) the Pearson correlation variations are represented in colors from blue (-1) to red (1), whereas in part (b) the absolute values of the correlation coefficient are depicted, ranging from 0 (black) to 1 (white).

In Table 2, the classification accuracy per class for the different alternative classifiers, the scanned and the cleared leaves for the three different legume species are reported, considering balanced datasets and 10 times 10-fold CV.



(a)

(b)

Figure 4: Correlation matrix for the 35 features computed on the scanned leaves. (a) Correlation values ranging from -1 (blue) to 1 (red). (b) Absolute values of the correlation matrix, ranging from 0 (black) to 1 (white).

This accuracy is the percentage of successful classifications in every class relative to the number of examples in the corresponding class. The mean over the 100

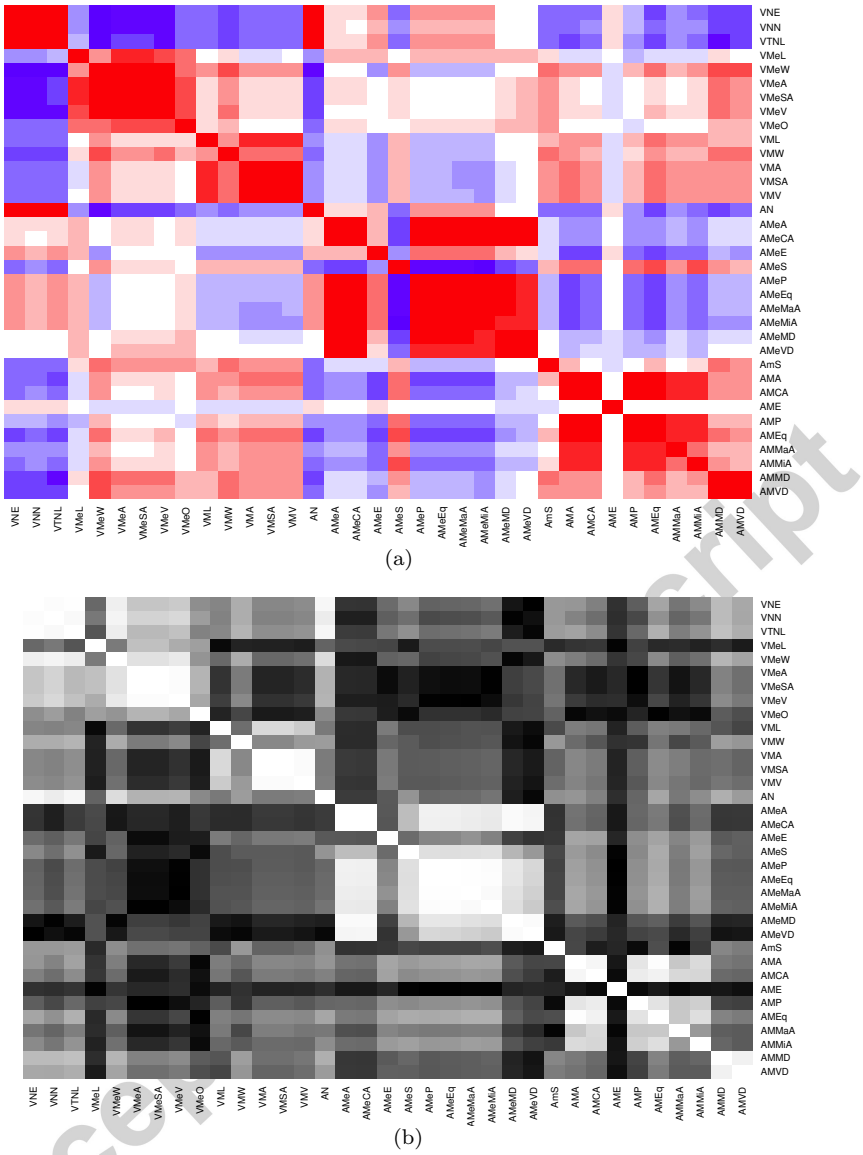


Figure 5: Correlation matrix for the 35 features computed on the cleared leaves. (a) Correlation values ranging from -1 (blue) to 1 (red). (b) Absolute values of the correlation matrix, ranging from 0 (black) to 1 (white).

runs is reported. The average classification accuracy obtained by 5 experts who manually classified the same central patches for the scanned leaves is also

included for reference. The manual classification for the cleared leaves is not relevant since the proposal of this work is to use scanned leaves instead of the cleared ones, and to compare its performance against the more costly procedures of using cleared leaves or doing manual classification.

From the accuracies described in Table 2, it is noticeable that for all the classification algorithms under consideration and using both cleared and scanned leaves, the recognition for soybean is much easier than for red and white beans (the accuracies rise to, at least, 95.5% for scanned leaves and 98.0% for cleared leaves). When using cleared leaves, the identification of red bean leaves seems to be a slightly easier problem than for white bean leaves for all the classifiers, except for SVM with linear kernel which solves the white bean recognition better. However, when considering the scanned leaves, the four classifiers obtain similar performances both for red and white beans. Additionally, this table shows that there is little advantage in using cleared leaves over scanned leaves: the accuracies are slightly higher for soybean, but are lower for white bean. In the case of the red bean, the performance is also slightly better when using cleared leaves, except for the SVM with linear kernel which improves 6.4% with scanned leaves.

The usage of cleared leaves partially improves the accuracies over their counterparts obtained with the scanned leaves, but at the expense of an increase in the time and cost of the image acquisition process. It requires the leaves to be chemically treated in a laboratory for several hours, and it cannot be implemented directly in the field. The best results for the cleared leaves are obtained by PDA for all the legume species, and it provides an improvement of 2.5% and 4.3% over the best results for soybean and red bean for the scanned leaves (obtained using SVM both with Gaussian and linear kernels in the first case, and PDA for the second), respectively. For white bean, the best performance is obtained by PDA with cleared and scanned leaves, but an improvement of 2.6% is achieved by using scanned leaves.

In addition, the recognition accuracies achieved by using the scanned leaves are superior in more than 11% to manual identification for the two bean vari-

eties, considering any of the accounted classifiers (improvements of 17.2% for white bean and 14.7% for red bean using PDA in both cases). In the case of soybean, the classification accuracy slightly decreases (2.8%) versus manual classification when using SVM with Gaussian and linear kernels (the best automatic recognition rates). The other automatic classifiers have slightly lower recognition rates in this particular case. However, for soybean and all the considered classifiers the accuracies are at least 92.8%. It is noticeable that Random Forests, which was recently proposed in our previous work [13] to discriminate between soybean, white bean and red bean, is the classifier with the lowest performance.

A more detailed analysis about the performance can be developed by analyzing the confusion matrices for the scanned and cleared leaves using each one of the considered alternative classifiers (Table 3), where the true labels are represented in the rows. It is noticeable that for both the scanned and cleared leaves, the percentages of correct classifications are high for the three legume species, and most of the misclassifications are between the white and red bean classes (two varieties from the same species). This fact repeats for all the classifiers. Also, a smaller number of mistakes appears between white bean and soybean, showing that the considered classifiers tend to confound these two classes. Minor errors are reported between red bean and soybean, indicating that the classifiers find stronger differences in the measures computed on the venation of these two species. This indicates that the white and red bean classes are partially superimposed on each other in feature space, whereas soybean is more separated.

The inclusion of more examples per class helps to improve the overall classification accuracy, as shown in Table 4, where the whole dataset of scanned leaves was taken into account. Again, 10 runs of 10-fold CV were implemented and the per-class and average $\pm$ standard deviation accuracies were computed. The results depicted in this table can be compared to the ones reported in Table 1 and Table 2 for scanned leaves. Both the soybean and red bean classes (the two major classes) are the species which experiment the largest improvements

Table 3: Confusion matrices for the scanned and cleared leaves using the four classifiers under analysis (Rows correspond to the true labels). Percentage values are reported. Percentages of correct classification are highlighted in bold.

Scanned leaves			
SVM (Gaussian kernel)			
	SB	WB	RB
SB	**95.5%**	4.5%	0.0%
WB	3.9%	**83.2%**	12.9%
RB	0.1%	17.6%	**82.3%**
SVM (Linear kernel)			
	SB	WB	RB
SB	**95.5%**	4.4%	0.1%
WB	4.8%	**82.9%**	12.3%
RB	0.9%	15.9%	**83.2%**
Random Forests			
	SB	WB	RB
SB	**92.8%**	6.6%	0.6%
WB	3.1%	**82.8%**	14.1%
RB	1.3%	18.0%	**80.7%**
PDA			
	SB	WB	RB
SB	**94.4%**	4.7%	1.0%
WB	3.8%	**83.6%**	12.6%
RB	0.0%	15.9%	**84.1%**

Cleared leaves			
SVM (Gaussian kernel)			
	SB	WB	RB
SB	**95.8%**	3.8%	0.4%
WB	2.0%	**80.8%**	17.2%
RB	0.0%	16.2%	**83.8%**
SVM (Linear kernel)			
	SB	WB	RB
SB	**97.6%**	1.8%	0.6%
WB	3.2%	**80.6%**	16.2%
RB	1.8%	21.4%	**76.8%**
Random Forests			
	SB	WB	RB
SB	**97.4%**	1.6%	1.0%
WB	1.6%	**77.8%**	20.6%
RB	0.0%	16.8%	**83.2%**
PDA			
	SB	WB	RB
SB	**98.0%**	1.6%	0.4%
WB	0.0%	**81.0%**	19.0%
RB	0.0%	11.6%	**88.4%**

for all the classifiers: between 1.6% and 3.5% for soybean and between 2.3% and 4.4% for red bean. On the contrary, the classification performance of the white bean class (which has the lowest number of examples) diminishes between 3.1% and 10.9% no matter which classifier is used. However, the best classifier (PDA) achieves accuracies of over 80% for the three classes. The best average accuracy in combination with the lowest standard deviation is also achieved by PDA ($89.9 \pm 2.7\%$) followed by SVM with linear kernel ($89.7 \pm 2.7\%$), reinforcing the idea that the linear dependency between the features influences the better performances of these methods. The worst result is obtained by Random

Table 4: Accuracies for legume classification using scanned leaves, different classifiers, non-balanced datasets and 10 times 10-fold CV. The arrows ↑ and ↓ indicate an increment or decrement in the accuracies, respectively, relative to the accuracies depicted in Table 2 for the scanned leaves. The best result is highlighted in bold.

Classifier	**Scanned leaves**			
	SB	**WB**	**RB**	**Average**
	(422 images)	(172 images)	(272 images)	**accuracy**
SVM (Gaussian kernel)	97.4% (↑ 1.9%)	74.4% (↓ 8.8%)	86.7% (↑ 4.4%)	$89.5 \pm 2.8\%$
SVM (Linear kernel)	97.1% (↑ 1.6%)	77.3% (↓ 5.6%)	86.1% (↑ 2.9%)	$89.7 \pm 2.7\%$
Random Forests	96.3% (↑ 3.5%)	71.9% (↓ 10.9%)	84.7% (↑ 4.0%)	$87.8 \pm 3.2\%$
PDA	96.0% (↑ 1.6%)	80.5% (↓ 3.1%)	86.4% (↑ 2.3%)	$\mathbf{89.9 \pm 2.7}\%$

Forests, which also presents the highest standard deviation ($87.8 \pm 3.2\%$).

The CPU times associated to the different stages of the proposed procedure are the following: 3.95 seconds both for leaf and vein segmentation of a scanned leaf (in MatLab), 0.32 seconds to compute the whole set of 52 features on the central patch (in MatLab), and 0.01 seconds to classify the exemplar (in R).

4. Conclusions

In this work, an automatic procedure aimed at recognizing legume species is proposed. The procedure discards any leaf shape, size, color or texture information, since the interest is focused exclusively in detecting differences in the leaf vein morphology. The images are acquired using a standard scanner, which is an economic and easy procedure which requires no delicate manipulation of the exemplars. The veins are segmented resorting to the UHMT and adaptive thresholding. A set of 35 morphological measures are obtained from a central patch extracted from the segmented veins, and these features are used to perform classification using four alternative classifiers, namely SVM with linear and Gaussian kernels, PDA and RF. The central patch is used to exclude the leaf shape from the analysis.

The performance of the proposed procedure is compared to the one obtained with cleared leaves, whose veins are stained and more clearly visible in higher order levels. Even though the classification accuracies achieved by using the cleared leaves are slightly higher, the acquisition procedure is more expensive and time consuming, requiring much more care in the handling of the leaves. In contrast, the proposed procedure using scanned leaves is much more simple, could be extendable to field work, and obtains very good average classification accuracies of over 87% with the PDA classifier. Even more, this accuracy can be easily improved in field work using measurements over several leaves of the same batch/field.

The comparison against manual classification is also performed. The results show an improvement over the expert's performance for two of the three legumes (white bean and red bean). Even though soybean vein characteristics are better recognized by humans, the automatic algorithm achieves at least 92% of accuracy for this species and all the classifiers under analysis when considering the balanced datasets. Overall, the proposed automatic method based on scanned leaves improves the results manually obtained by the experts, with clear advantages in confiability and repeatability.

Current work is being developed to extend this procedure to the identification of different cultivars from single species.

Appendix A

The considered measures encompass traits computed both on the veins (edge segments in the segmented vein image) and the areoles (background spaces enclosed by veins). In Figures A.6(a) and A.6(b), schemes including some two-dimensional vein and areole measures are respectively presented. These schemes are provided as reference to better understand the measurement calculation procedure.

According to Price et al. [16] and their supplemental material, the skeleton (medial axis) of the binary image is first computed for vein measurement.

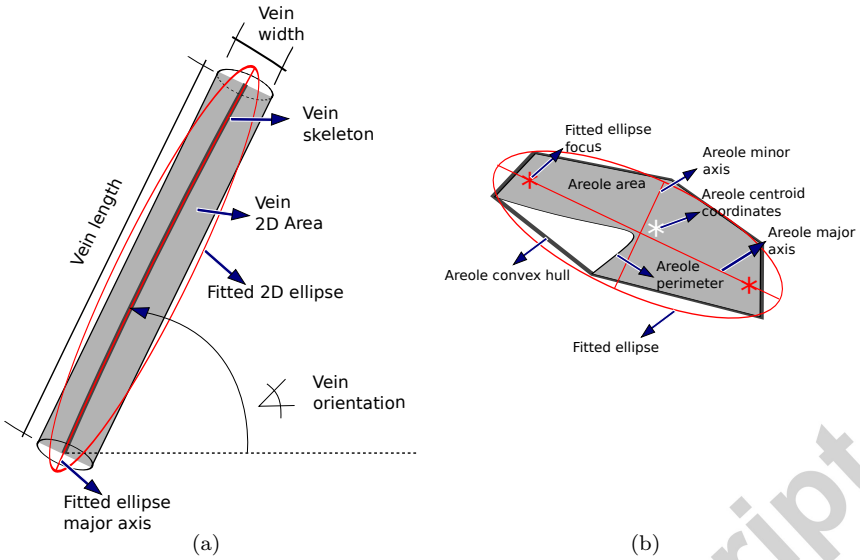


Figure A.6: (a) Two-dimensional vein measures computed on a disconnected edge. (b) Areole measures.

The skeleton branchpoints and tips are obtained, and the disconnected edges are extracted by removing the branchpoint pixels and performing connected components labeling. Next, a distance transform is computed on the original binary image in order to measure the Euclidean distance between each vein pixel (foreground) and its nearest areole pixel (background). The distance transform is multiplied pixel by pixel to the skeleton, in order to obtain the Euclidean distance of each skeleton pixel to the nearest areole.

For the individual measures of vein/areoles computed by LEAF GUI, we calculated the minimum, median and maximum values of all the vein/areoles in the patch. We also measured the minimum, median and maximum orientation of edges in each image patch (not available in LEAF GUI). Following, the explanation for each one of the considered 52 measures is provided. We suggest the reader to use Fig. A.6 as reference. Additional information can be found in the paper about LEAF GUI [16].

Total number of edges (VNE), computed as the count of existing vein segments.

Each vein segment is internally assigned a unique code number from 1 to the number of disconnected edges. Therefore, VNE can be computed as the maximum vein code number.

Total number of nodes (VNN). A node is a branchpoint of two or more veins. Each node is also assigned a unique code number from 1 to the number of nodes. VNN is computed as the maximum code number.

Total network length (VTNL). Total distance along the skeleton of the vein image patch. The image resolution is used to scale the total distance in pixels in order to obtain VTNL expressed in mm.

Median/min/max vein length (VMeL/VmL/VML). The length of a single vein is the distance along its skeleton. The distance between two 4-neighbor pixels of the skeleton is 1, whereas the distance between two diagonal neighbor pixels is $\sqrt{2}$. The length in pixels is next scaled to obtain the measure expressed in mm.

Median/min/max vein width (VMeW/VmW/VMW). Each vein is modeled as a series of connected cylinders, each cylinder having length equal to 1 pixel and diameter (width) equal to two times the value of the distance transform at each vein skeleton pixel. The width for a single vein is obtained as the mean of the widths computed at each vein skeleton pixel. The vein width in pixels is appropriately scaled in mm.

Median/min/max vein 2D area (VMeA/VmA/VMA). The vein 2D area (in mm^2) is the sum of the widths computed at every skeleton pixel of the current edge times the length of one pixel.

Median/min/max vein surface area (VMeSA/VmSA/VMSA). The surface area (SA, in mm^2) of the cylinder centered at the edge skeleton is computed as the sum of the individual surface areas for each skeleton pixel of the current edge, as $\sum_i SA_i = 2\pi(d_i/2)l_i$, where d_i is the diameter (width) and l_i is the length for a skeleton pixel i.

Median/min/max vein volume (VMeV/VmV/VMV). The vein volume (V, in mm^3) corresponds to the volume of the same cylinder as in surface area, and is computed as $\sum_i V_i = \pi(d_i/2)^2 l_i$.

Median/min/max vein orientation (VMeO/VmO/VMO). A 2D ellipse having the same second central moments as the vein is constructed, and the angle (in the range [-90°, 90°]) between the x-axis and the major axis of this ellipse is measured. This angle corresponds to the orientation of a single vein.

For areoles, the complement of the original binary image is computed. Connected components labeling is developed to isolate each areole.

Total number of areoles in the image patch (AN). Labeling of areoles assigns a unique code to each areole from 1 to the number of areoles. AN can be computed as the maximum areole code number.

Median/min/max areole perimeter (AMeP/AmP/AMP). The areole perimeter is calculated as the distance (in pixels) along the pixels of the border of the areole. This value is next scaled to express it in mm.

Median/min/max areole area (AMeA/AmA/AMA). The areole area is computed as the number of pixels in each areole times the area of one pixel (in mm^2).

Median/min/max areole convex area (AMeCA/AmCA/AMCA). The convex area is measured as the area of the convex hull (smallest convex polygon which encloses the areole). This value is scaled by the image resolution in order to express it in mm^2.

Median/min/max areole solidity (AMeS/AmS/AMS). The solidity is a dimensionless parameter between 0 and 1 which measures the proportion of the pixels in the convex hull that are also in the area. It is computed as the ratio between the areole area and the convex area.

Median/min/max areole major axis (AMeMaA/AmMaA/AMMaA). An ellipse having the same normalized second moments as the areole is constructed. The major axis of this ellipse is expressed in mm.

Median/min/max areole minor axis (AMeMiA/AmMiA/AMMiA). The minor axis (in mm) of the same ellipse explained in the preceding paragraph.

Median/min/max areole eccentricity (AMeE/AmE/AME). The eccentricity is a dimensionless parameter between 0 (a circle) and 1 (a line), which measures the ratio of the distance between the foci of the previously fitted ellipse and its

major axis.

Median/min/max areole equivalent diameter (AMeEq/AmEq/AMEq). The equivalent diameter (in mm) is the diameter of a circle having the same area as the areole. It is computed as $\sqrt{4 \cdot Area/\pi}$.

Median/min/max areole mean distance (AMeMD/AmMD/AMMD). In order to obtain this measure, a distance transform is computed which measures the Euclidean distance (in pixels) between the pixels of each areole and the nearest vein pixel. The areole mean distance for a single areole is the mean value of all these distances, appropriately scaled in mm.

Median/min/max areole variance distance (AMeVD/AmVD/AMVD). The variance distance for a single areole (in mm) is obtained as the variance of the Euclidean distances computed in the same way as explained in the previous paragraph.

[1] Agarwal, G., Ling, H., Jacobs, D., Shirdhonkar, S., Kress, W., Russell, R., Belhumeur, P., Dixit, N., Feiner, S., Mahajan, D., Sunkavalli, K., , White, S., 2006. First steps toward an electronic field guide for plants. Taxon, Journal of the International Association for Plant Taxonomy 55, 597–610.

[2] Bama, B. S., Valli, S. M., Raju, S., Kumar, V. A., 2011. Content based leaf image retrieval (CBLIR) using shape, color and texture features. Indian Journal of Computer Science and Engineering 2 (2), 202–211.

[3] Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

[4] Camargo Neto, J., Meyer, G. E., Jones, D. D., Samal, A. K., 2006. Plant species identification using Elliptic Fourier leaf shape analysis. Computers and Electronics in Agriculture 50, 121–134.

[5] Chaki, J., Parekh, R., 2012. Designing an automated system for plant leaf recognition. Int. Journal of Advances in Engineering & Technology 2 (1), 149–158.

[6] Clarke, J., Barman, S., Remagnino, P., Bailey, K., Kirkup, D., Mayo, S., Wilkin, P., 2006. Venation pattern analysis of leaf images. Advances in

Visual Computing, Lecture Notes in Computer Science (ISVC2006) 4292, 427–436.

[7] Du, J.-X., Wang, X.-F., Zhang, G.-J., 2007. Leaf shape based plant species recognition. Applied Mathematics and Computation 185 (2), 883–893, special Issue on Intelligent Computing Theory and Methodology.

[8] Golzarian, M. R., Frick, R. A., 2011. Classification of images of wheat, ryegrass and brome grass species at early growth stages using principal component analysis. Plant Methods 7:28.

[9] Hastie, T., Buja, A., Tibshirani, R., 1995. Penalized discriminant analysis. Ann. Statist. 23 (1), 73–102.

[10] Hastie, T., Tibshirani R., Friedman, J., 2009. The Elements of Statistical Learning, Second Edition. Springer.

[11] Horgan, G. W., Talbot, M., Davey, J. C., 1998. Towards automatic recognition of plant varieties. British Computer Society Electronic Workshops in Computing: The Challenge of Image Retrieval.

[12] Im, C., Nishida, H., Kunii, T. L., 1998. Recognizing plant species by leaf shapes-a case study of the Acer family. ICPR 2, 1171–1173.

[13] Larese, M., Craviotto, R., Arango, M., Gallo, C., Granitto, P., 2012. Legume identification by leaf vein images classification. In: Alvarez, L., Mejail, M., Gomez, L., Jacobo, J. (Eds.), Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Vol. 7441 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 447–454. URL http://dx.doi.org/10.1007/978-3-642-33275-3_55

[14] Park, J., Hwang, E., Nam, Y., 2008. Utilizing venation features for efficient leaf image retrieval. J. Syst. Softw. 81 (1), 71–82.

[15] Perez, A., Lopez, F., Benlloch, J., Christensen, S., 2000. Colour and shape analysis techniques for weed detection in cereal fields. Computers and Electronics in Agriculture 25, 197–212.

[16] Price, C. A., Symonova, O., Mileyko, Y., Hilley, T., Weitz, J. S., 2011. Leaf extraction and analysis framework graphical user interface: Segmenting and analyzing the structure of leaf veins and areoles. Plant Physiology 155, 236–245.

[17] Sack, L., Dietrich, E. M., Streeter, C. M., Sanchez-Gomez, D., Holbrook, N. M., 2008. Leaf palmate venation and vascular redundancy confer tolerance of hydraulic disruption. PNAS USA 105, 1567–1572.

[18] Scoffoni, C., Rawls, M., McKown, A. D., Cochard, H., Sack, L., 2011. Decline of leaf hydraulic conductance with dehydration: relationship to leaf size and venation architecture. Plant Physiology 156, 832–843.

[19] Soille, P., 1999. Morphological Image Analysis: Principles and Applications. Springer-Verlag.

[20] Solé-Casals, J., Travieso, C. M., Alonso, J. B., Ferrer, M. A., 2008. Improving a leaves automatic recognition process using PCA. In: IWPACBB. pp. 243–251.

[21] Sonka, M., Hlavac, V., Boyle, R., 2008. Image processing analysis and machine vision. Thomson.

[22] Valliammal, N., Geethalakshmi, S., July 2011. Hybrid image segmentation algorithm for leaf recognition and characterization. In: PACC 2011. pp. 1–6.

[23] Vapnik, V., 1995. The nature of statistical learning theory. Springer-Verlag.

Mónica G. Larese received the Ph.D. degree in Informatics in 2011 from National University of Rosario (UNR), Argentina, where she is also an Auxiliary Professor since 2008. She is a full-time Researcher at CIFASIS, CONICET (Argentina). Her research interests include image analysis, computer vision, machine learning and pattern recognition.

Rafael Namias is a Ph.D. student at National University of Rosario (UNR), Argentina, and CIFASIS, CONICET (Argentina). He researches on Digital Imaging Processing. His current interests include segmentation in Medical Imaging and in biological problems.

Roque M. Craviotto is a Ph.D. Agricultural Engineer. He researches on seed science technology, seed quality control techniques and development of seed quality control instruments. He is Chief of Seed Technology Team and Co-Director of INTA Oliveros Seed Laboratory, Argentina. Staff and editorial responsible of Seed Analisis Magazine (Argentina).

Miriam R. Arango is a M.Sc. Agricultural Engineer. She researches on seed science technology, seed quality control techniques and development of seed quality control instruments. She is Director of INTA Oliveros Seed Laboratory, Argentina. Staff and editorial responsible of Seed Analisis Magazine, Argentina. Member of National Seed Institute (INASE), Argentina.

Carina Gallo is a M.Sc. Agricultural Engineer. She researches on seed science technology, seed quality control techniques and development of seed quality control instruments. She is a member of the Vigor Commitee of the International Seed Testing Association (ISTA). Staff and editorial responsible of the Seed Analisis Magazine (Argentina).

Pablo M. Granitto received a Ph.D. in Physics from UNR, Argentina, and was a Post-Doc at FEM-IASMA, Italy. He is a full-time Researcher at CONICET-UNR. He leads the Machine Learning Group (CIFASIS). His interests include applications of modern machine learning techniques to agroindustrial and biological problems.

- We develop an automatic procedure to classify legume species using scanned leaves.
- The method is based exclusively on the analysis of the leaf venation images.
- We analyze the advantages over the usage of cleared leaves.
- Different state-of-the-art classifiers are compared.
- The proposed method outperforms human expert classification.