IFSC/USP at ImageCLEF 2011: Plant identification task

Dalcimar Casanova*, João Batista Florindo**, and Odemir Martinez Bruno

USP - Universidade de São Paulo
IFSC - Instituto de Física de São Carlos, São Carlos, Brasil
bruno@ifsc.usp.br

Abstract. The leaves are one of the most important main sources used for plant identification. Because of this the ImageCLEF 2011 proposed a challenge based on leaf analysis for plant identification. This paper reports the experiment results of the IFSC/USP team in participating of this task. The main goal is investigate the performance of Complex Network method for feature extraction and classification of plant species. The achieved results are promising and can help the botanists in the future.

Keywords: Complex Network, FDA, Taxonomy, Plant identification, Leaves.

1 Introduction

In the world is estimated that there are 400,000 species, of which only 270,000 have been named and identified by botanists. The Plant Taxonomy is the responsible science for survey of the fauna and your consequent classification. Although we have many researches in this field, the taxonomy of species is still a hard task. One of reasons for this is the fact of the flowers and fruits (the main sources used for diagnostic of characteristics) are not available for studies throughout the year, but only at certain times. Although available throughout most of the year, the leaves do not have sufficient visible characteristics to differentiate between many species. Methods of computer vision can help in this point. The main idea is extract more good characteristics of the leaves, using computer vision methods, than traditional manual inspection.

The ImageCLEF series use this idea and propose an ongoing campaign's on plant identification task. The main goal of this task is provide a forum for researchers that work on image analysis and artificial intelligence methods, share ideas and compare their systems in order to help the taxonomic process with leaves information.

* Dalcimar Casanova gratefully acknowledges the financial support FAPESP (São Paulo Research Foundation, Brazil) (2008/57313-2) for his PhD grant.

** João Batista Florindo gratefully acknowledges the financial support CNPq (National Council for Scientific and Technological Development) (870336/1997-5) for his PhD grant.

Our group, which has already been working on plant identification in recent years, accepts this challenge. In this year we use a new method of shape analysis, based on Complex Network theory [1], to characterize the contour of leaves. This method is based on simple measurements of Complex Networks. Although very simple, the use of these features has shown good results in other works of shape analysis.

The following Section 2 describes the materials and methods utilized. In Section 3, we explain the experiments and obtained results. Finally, conclusions are presented in Section 4.

2 Material and Methods

2.1 Database

The experiments are performed over Pl@antLeaves dataset [3]. This database is maintained by the French project Pl@ntNet (INRIA, CIRAD, Telabotanica). The full database contains 5436 images of 71 tree species of real-world. The images are taken under 3 different practical conditions:

1. Scan: contains 3070 scans of leaves collected using flatbed scanners. These images are oriented vertically along the main natural axis and with the petiole visible.
2. Scan-like photos: contains 897 photos which look similar to the scans images. Those images have uniform background but with some luminance variations, optical distortions, shadows and color derivations.
3. Free natural photos: contains 1469 photos taken directly on the trees. No acquisition protocol is used, which results in a non-uniform background, rotated and bad-scaled images, among others problems.

Each image has an xml file associated that contain the date, type (single leaf, single dead leaf or foliage), name of the author and GPS coordinates of the observation among others information. But we do not use any of this information in classification process. Just characteristics of the images are used to make our recognition system.

The full database is split in training and test dataset. The training dataset have 4004 images (2329 scans, 686 scan-like photos, 989 natural photos) and de test dataset have 1432 images (741 scans, 211 scan-like photos, 480 natural photos).

2.2 Pre-processing

All images of both test and training dataset are firstly segmented. For Scan and Scan-like photos a simple Otsu [4] method was employed. For Free natural photos a manual segmentation is performed where Otsu method do not have good results.

In sequence, we apply a contour detection method to extract only contour of leaves. We do not bother to treat open or imperfect contours, because the method of shape analysis that we will use is robust to such problems.

2.3 Complex Network Features

To apply the Complex Network method [1] an graph $G = (V, E)$ should be built using the contour of the leaf. To this, each pixel of the contour $S = \{s_1, s_2, ..., s_N\}$ is represented as a vertex in the network (i.e. $|S| = |V|$) and for each pair of vertices an edge w_{ij} is added as their Euclidean distance:

$$w_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{1}$$

Therefore, the network G is represented by the $N \times N$ weight matrix W normalized into the interval $[0, 1]$,

$$W = \frac{W}{\max_{w_{ij} \in W}} \tag{2}$$

A complete graph is obtained from this. So, relevant properties are extracted from the posterior transformation of this network using a set of thresholds $T = \{t_1, t_2, \ldots, t_L\}$:

$$A_{T_l} \forall w \in W \begin{cases} a_{ij} = 0, & \text{if } w_{ij} \geq t_l \\ a_{ij} = 1, & \text{if } w_{ij} < t_l \end{cases} \tag{3}$$

In this experiments we use $T = \{0.025, 0.050, 0.075, \ldots, 0.925\}$, totaling 13 thresholds ($|T| = 13$). We measure, for each threshold, the mean degree and maximum degree, given respectively by:

$$k_\mu = \frac{1}{N} \sum_{i=1}^{N} k_i \tag{4}$$

$$k_\kappa = \max_i k_i \tag{5}$$

where the degree k_i of a node i is the number of edges directly connected to node, and it is defined in terms of the adjacency matrix A as:

$$k_i = \sum_{j=1}^{N} a_{ij} \tag{6}$$

An normalization of the vertices degree by the number of vertices in the network is necessary before computing these measurements. This normalization is performed in order to reduce the influence of the network size in the computed descriptors, and it is performed as follows:

$$\forall k_i = \frac{k_i}{N} \tag{7}$$

Thus, a feature vector $\mathbf{x}$ for each leaf image is composed by 26 features (13 of k_μ and 13 of k_κ). The Fig. 1 shows all process to computing these features vectors.

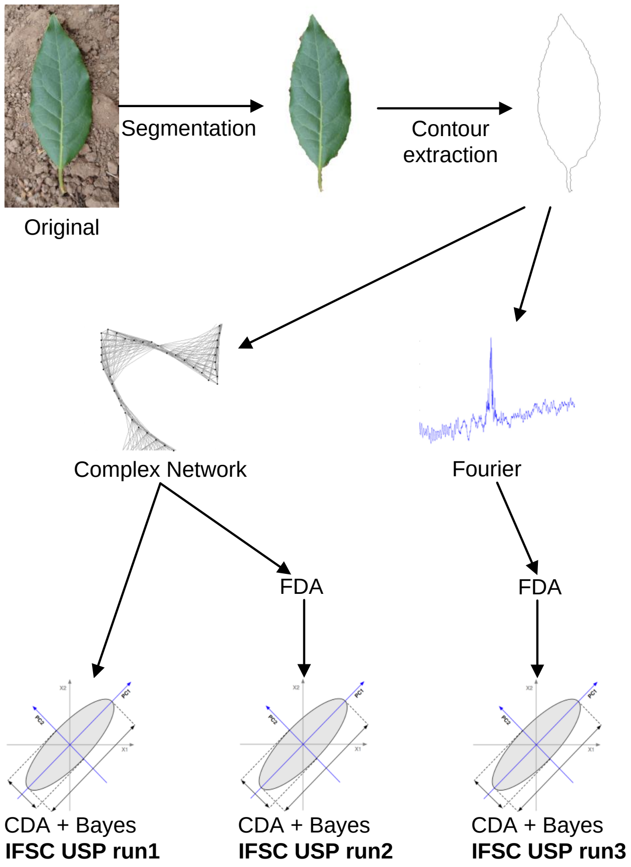

Fig. 1. Overview of runs IFSC USP

2.4 Fourier Features

The Fourier descriptors consist of the sum of the main components of the normalized power spectrum. It was used here 40 frequencies. These frequencies are then called Fourier descriptors and make up the feature vector $\mathbf{x}$.

2.5 Functional Data Analysis

Functional Data Analysis (FDA) [6, 2] is a statistical approach alternative to multivariate analysis. In FDA, a set of variables is handled as a unique entity, more exactly, an analytical function. Such function may be obtained through any sort of interpolation method. Thus, the function f may be calculated by:

$$f = \sum_{j=1}^{q} \alpha_j(f)\phi_j, \tag{8}$$

where ϕ are the basis functions and α are the basis coefficients.

In this work, we used B-splines basis. Then, we extract features from the Complex Network or Fourier descriptors by applying a transform to the coefficients α ($alpha = \mathbf{x}$) [8, 2]. The features are represented by $\beta(f)$ and provided by:

$$\beta(f) = S\alpha(f), \tag{9}$$

where S is the Choleski decomposition of $\varPhi$ matrix, whose elements are:

$$\varPhi(k,l) =< \phi_k, \phi_l > . \tag{10}$$

The above transform turns possible the expression of the original data on the basis functions algebraic space. In this way, it becomes a richer analysis tool emphasizing the global relevant aspect of the original data.

3 Canonical Discriminant Analysis + Naive Bayes

With a single feature vector for each leaf we have chosen to use the Naive Bayes as classifier. In addition, we have used a 10-fold cross validation. For g groups, the Bayes rule assigns an object to the group i when:

$$P(i|\mathbf{x}) > P(j|\mathbf{x}),\ for\ \forall j \neq i \tag{11}$$

In this case, assuming the hypothesis of independence, we have for the random variables:

$$P(i|\mathbf{x}) = \frac{P(i)\prod_{k=1}^{n} P(x_k|i)}{\prod_{k=1}^{n} P(x_k)} \tag{12}$$

where:

$$P(x_k|i) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} e^{\frac{(x_i-\mu_{ik})^2}{2\sigma_{ik}^2}} \tag{13}$$

being $P(\mathbf{x}|i)$ the probability of obtaining a particular set of features $\mathbf{x}$, given that the object belongs to the group i and $P(i)$ is the probability *a priori*, that is the probability of choosing the group i without any feature of the known object.

In addition to Naive Bayes we use the Canonical Discriminant Analysis [7]. This method aims maximize the separation between classes. Given the matrix S, indicating the total dispersion among the feature vectors, defined as:

$$S = \sum_{i=1}^{N} (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)' \tag{14}$$

and the matrix S_i indicating the dispersion of objects of C_i:

$$S_i = \sum_{i \in C_i} (\mathbf{x}_i - \mu_i)(\mathbf{x}_i - \mu_i)' \tag{15}$$

we can define the intra-class variability S_{intra} (indicating the combined dispersion within each class) and interclass variability S_{inter} (indicating the dispersion of the classes in terms of their centroids) as:

$$S_{intra} = \sum_{i=1}^{K} S_i \tag{16}$$

$$S_{inter} = \sum_{i=1}^{K} N_i(\mu_i - \mu)(\mu_i - \mu)' \tag{17}$$

where K is the number of classes, N, the number of samples, N_i, the number of objects in class i, C_i, the set of samples of class i, μ, the global average, and μ_i, the average of objects in class i. For these measures of dispersion we have necessarily:

$$S = S_{intra} + S_{inter} \tag{18}$$

Thus, the i-th canonical discriminant function is given by:

$$Z_i = a_{i1}\mathbf{X}_1 + a_{i2}\mathbf{X}_2 + \cdots + a_{ip}\mathbf{X}_p \tag{19}$$

where p is the number of features of the model and a_ij are the elements of the eigenvector $a_i = (a_{i1}, a_{i2}, \ldots, a_{ip})$ of matrix C given by:

$$C = S_{inter} * S_{intra}^{-1} \tag{20}$$

In general a reduction in the number of features is desired. Thus, the system of random variability of the original vector with p-original variables is approximated by the variability of the random vector containing the k-principal components.

4 Experiments and Results

We submitted 3 different runs to the plant identification task. The main differences between those 3 runs can be seen in Fig. 1.

In the first run we apply the Complex Network method in shapes of the leaves. In sequence the methods of Canonical Discriminant Analysis followed by

Naive Bayes classifier are employed. For that, only 10 canonical variables are used in the Naive Bayes classifier. These 10 main components represent 99.99% of total variance.

For the second run, the FDA method is employed over Complex Network descriptors. The new obtained descriptors by FDA method are then used as input to the CDA method.

The third run is exactly equal of the second run, except that, in this we use Fourier descriptors in replace of Complex Network descriptors.

The results are showed in Table 1. We see that Complex Network descriptors obtain best results than Fourier descriptors. You can also observe that the FDA method make a small improving on the success rate.

Is important emphasize here that, though both methods aims improve the quality of descriptors, each one acts in a different way on these. On Canonical Discriminant Analysis the objective is maximize the separation of classes, while the Functional Data Analysis aims highlight some features of the original feature vector.

Run name	No. of descriptors	scan	scan-like	photos	mean Sucess rate
IFSC USP_run1	26	0.411	0.430	0.503	0.448
IFSC USP_run2	**26**	**0.562**	**0.402**	**0.523**	**0.496**
IFSC USP_run3	40	0.356	0.187	0.116	0.220

Table 1. Results for all runs in plant database.

It is important to emphasize that the method of Complex Networks do not need a closed contour, since the method is invariant to rotation and scale, problems that we have in the dataset. We also have a good robustness against noise and spurious contour points [1]. Perhaps due to these characteristics, the CN method fared better than Fourier.

Is not surprising the good success rate achieved by the photo images (if compared with scan and scan-like images). This is probably due to the manual segmentation performed on these images.

5 Conclusion

Although we have obtained good results, they are still far from ideal. We perceived as the main problem to lack of standardization of the images, especially images of free natural photos. For these images we do not have a good generic method to make the correct segmentation of all the images.

It is important to remember that other relevant information contained in the associated XML was not used. Such information can help achieve better success rates.

Is important to note that there are other diagnostic keys that can be used to leaf identification, texture [7] and venation [5] are just some examples. These attributes appear to contain richer information than the leaf contour. However, in order to use of these attributes, images with higher resolution and a standard procedure for capturing images need be used.

Thus, there are good prospects to achieve a good system of leaf identification with the use of these variables. Such a system would be very helpful to botanists and other professionals.

Bibliography

[1] A. R. Backes, D. Casanova, and O. M. Bruno. A complex network-based approach for boundary shape analysis. *Pattern Recognition*, 42(1):54–67, 2009.

[2] J. B. Florindo, M. D. Castro, and O. M. Bruno. Enhancing multiscale fractal descriptors using functional data analysis. *International Journal of Bifurcation and Chaos*, 20(11):3443–3460, 2010.

[3] H. Goëau, P. Bonnet, A. Joly, N. Boujemaa, D. Barthelemy, J.-F. Molino, P. Birnbaum, E. Mouysset, and M. Picard. The clef 2011 plant images classification task. In *CLEF 2011 working notes*, Amsterdam, The Netherlands, 2011.

[4] N. Otsu. A threshold selection method from grey-level histograms. *IEEE Trans. Systems, Man and Cybernetics*, 9(1):62–66, 1979.

[5] R. O. Plotze, M. Falvo, J. G. Pádua, L. C. Bernacci, M. L. C. Vieira, G. C. X. Oliveira, and O. M. Bruno. Leaf shape analysis using the multiscale minkowski fractal dimension, a new morphometric method: a study with passiflora (passifloraceae). *Canadian Journal of Botany*, 83:287–301, 2005.

[6] J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis: Methods and Case Studies*. Springer-Verlag, New York, 2002.

[7] D. R. Rossatto, D. Casanova, R. M. Kolb, and O. M. Bruno. Fractal analysis of leaf-texture properties as a tool for taxonomic and identification purposes: a case study with species from neotropical melastomataceae (miconieae tribe). *Plant Systematics and Evolution*, 291(1):103–116, 2010.

[8] F. Rossi, N. Delannay, B. Conan-Guez, and M. Verleysen. Representation of functional data in neural networks. *Neurocomputing*, 64(1):183–210, 2005.