Improving a leaves automatic recognition process using PCA

[1]Jordi Solé-Casals, [2]Carlos M. Travieso, [2]Jesús B. Alonso, [2]Miguel A. Ferrer

[1]Signal Processing Group
University of Vic
c/ de la Laura, 13, E-08500, Vic, Barcelona, Spain.
[2]Department of Signals and Communications. CeTIC.
University of Las Palmas de Gran Canaria
Campus de Universitário Tafira s/n, E-35017, Las Palmas de Gran Canaria, Spain
jordi.sole@uvic.cat ,{ctravieso, jalonso, mferrer}@dsc.ulpgc.es

Abstract In this work we present a simulation of a recognition process with perimeter characterization of a simple plant leaves as a unique discriminating parameter. Data coding allowing for independence of leaves size and orientation may penalize performance recognition for some varieties. Border description sequences are then used, and Principal Component Analysis (PCA) is applied in order to study which is the best number of components for the classification task, implemented by means of a Support Vector Machine (SVM) System. Obtained results are satisfactory, and compared with [4] our system improves the recognition success, diminishing the variance at the same time.

Keywords Principal Component Analysis, Pattern Recognition, Leaves Recognition, Parameterization, Characteristics selection.

1 Introduction

Recognition of tree varieties using samples of leaves, in spite of its biological accuracy limitations, is s simple and effective method of taxonomy [1]. Laurisilva Canariensis is a relatively isolated tree species, in the Canary Islands, biologically well studied and characterized. Twenty-two varieties are present in the archipelago and have simple and composed regular leaves. Our study takes into account sixteen of the twenty simple leaf varieties, with totals of seventy-five individuals per each one. They have been picked over different islands, pressed (for conservation purposes) and scanned in gray tonalities.

From a biological perspective, attention has to be brought to the fact that emphasis on structural characteristics, which are consistent among individuals of a species, instead of quality parameterization (as color, size or tonality), improves recognition performance. Quality parameterization lack of accuracy is due to the fact of leaves individual variability on the same variety as well on leaf variability on a single plant. Plant age, light, humidity, context behavior or distribution of soil characteristics, among other things, contributes for such anomaly.

In spite of the fact that we may consider several biological parameters, as we have done previously [2], in order to generalize such study, in this paper we have just considered a border parameterization. This system was classified by Hidden Markov Model (HMM) [3] achieving a success of 78.33% [4].

In this present work, we have improved that previous study using the transformation and reduction of border parameterization using Principal Component Analysis (PCA) [5], and classifying its result with Support Vector Machines (SVM) [6][7]. The rest of this paper presents our proposal.

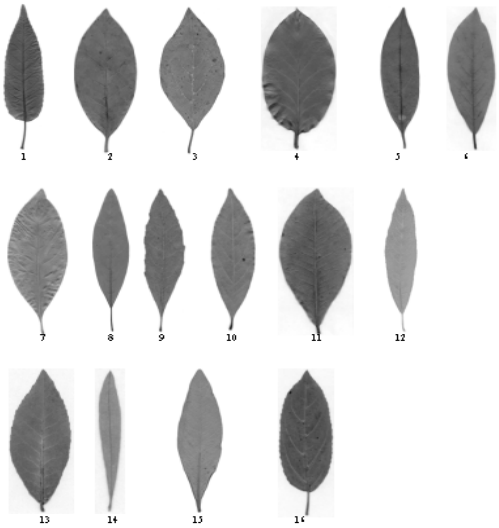

Fig.1 Images of the 16 varieties of canariensis laurisilva considered for the present study. Images are presented regardless of size.

2 Leaves Database

In order to create a recognition system of different vegetable species it is necessary to build a database. This database should contain the samples of the different species of study. The number of samples will be large enough to, first train the classifier with guarantees and second, test this classifier to assess the results obtained. On top of this, the amount of chosen samples, for each vegetable species must cover the largest amount of shapes and structures that this unique specie can take. In this way, a robust study of the different vegetable species is ensured.

Attending to this reasoning, the sample collection was made at different times of the year, trying in this way to cover all the colors and shapes that the leaves take throughout the four seasons. Besides, a special attention was made to reject those samples that were degraded so that the selected samples were in good condition.

Therefore, this database is composed of 16 classes (see Fig. 1), with 75 samples each one. The images that form the database has been stored in a grey scale using a "jpeg" format (Joint Photographic Experts Group) with Huffman compression. The images have been digitalized to 300 dpi, with 8 bit accuracy.

3 Parameterization System

We have considered just the leaf perimeter. This image is considered without its petiole that has been extracted automatically from the shadow image. Leaves are scanned fixed on white paper sheets, placed more or less on the center, upward (petiole down) and reverse side to scan.

Border determination as (x,y) positioning perimeter pixels of black intensity, has been achieved by processes of shadowing (black shape over white background), filtering of isolated points, and perimeter point to point continuous follow.

3.1 Perimeter interpolation.

As shown in table 1, perimeter size variability induces us to consider a convenient perimeter point interpolation, in order to standardize perimeter vector description. For an interpolating process, in order to achieve reconstruction of the original shape, we may use any of the well known algorithms as mentioned in [8], [9], [10], but a simple control point’s choice criterion in 1-D analysis allows for an appropriate performance ratio on uniform control point’s number and approximation error for all individuals of all varieties studied.

Class	Mean size	Mean Error	
		Uniform	Monotonic
01	2665.6	9.1474	2.0226
02	1885.1067	3.5651	0.43655
03	2657.68	11.0432	5.3732
04	2845.8133	31.6506	2.8447
05	1994.68	1.8569	0.42231
06	2483.04	0.4425	0.71093
07	2365.2667	9.711	0.68609
08	3265.48	0.4753	0.49015
09	2033.2267	19.7583	3.4516
10	2258.2533	3.9345	2.4034
11	1158.9867	5.4739	1.0286
12	1934	1.3393	0.40771
13	1183.4	1.2064	0.39012
14	981.4	0.2752	0.23671
15	3159.08	11.575	8.8491
16	1973.3733	47.4766	6.6833

Table 1. A comparative table of mean error, obtained from a uniform criterion of control point selection and the monotonic way

The general idea, for such choice, is to consider (x,y) positional perimeter points as (x,F(x)) graph points of a 1-D relation F.

Consideration of y coordinate as y = F(x) is done, because of the way, leaves images are presented in our study: leaves have been scanned with maximum size placed over x ordinate.

For a relation G to be considered as a one-dimensional function, there is need to preserver a correct sequencing definition (monotonic behavior).

That is: A graph,

$$G = \{ i = 1..n, (x_i, y_i) / \; y_i = f(x_i) \} \tag{1}$$

It is the description of a function f if ordinate points $x_i, \; i = 1..n$ must be such that: $x_i < x_{i+1}, \; i = 1..n-1$.

We consider then the border relation F as a union of piece like curves (graphs) preserving the monotonic behavior criterion, i.e.

$$F = \bigcup_{j \in J} G_j \tag{2}$$

where: $G_j \subseteq F, \forall\, j \in J$ and $G_j = \{\alpha_j \in J_j, (x_{\alpha_j}, y_{\alpha_j}) / \, y_{\alpha_j} = f_j\}$,

For convenient sets of index J, J_j and restriction functions $f_j = f_{|\{x_{\alpha_j} | \alpha_i \in J_j\}}$, such that the next point following the last of G_j is the first one of G_{j+1}. Gj graphs are correct f_j functions descriptions.

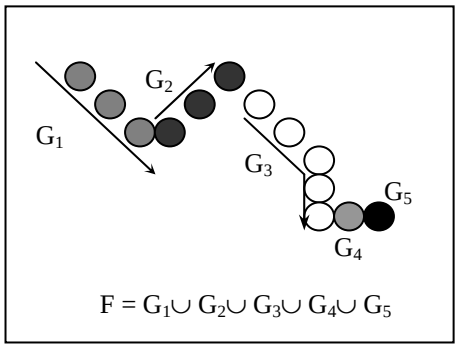

Fig.2 Example of an F relation decomposed in graphs with a correct function description..

Building the Gj sets is a very straightforward operation:

- Beginning with a first point we include the next one of F.
- As soon as this point doesn't preserve monotonic behavior we begin with a new $G_{j+}1$.
- Processes stop when all F points are assigned.

In order to avoid building G_j reduced to singletons, as show in figure 2 (G_4 and G_5) the original F relation may be simplified to preserve only the first point of constant x ordinate series.

Afterwards, spreading of a constant number of points is done proportional to the length of the G_j and always setting in it is first one.

The point's choice criterion mentioned before allows, in two-dimensional interpolation, for taking account on points where reverse direction changes take place. Irregularity, of the surface curve, is taken into account with a sufficient number of interpolating points, as done in the uniform spreading way. Results on table 1 allows for comparison between choice of control points with the criterion motioned before and the uniform one. Such results show the benefit of choosing control points with the monotonic criterion instead of the uniform one.

The 1-D interpolation has been perform using 359 control points, with spline, lineal or closest interpolated point neighborhood, depending on the number of control points present in the decomposed curve. As a reference at 300 dpi a crayon free hand trace is about 5 to 6 points wide.

Table 1 also shows size variability of the different varieties ranging in mean, between 981 pixels for class 14 to 3255 for class 8. With 359 points chosen with

the monotonic criterion, all perimeter point vectors have a standard size and errors representation is negligible.

Due to perimeter size variability inside a class, for example in class 15 ranging between 2115 points to 4276 with a standard deviation of about 521, coding of (*x*,*y*) control perimeter points have been transformed taking account for size independence.

Considering the following definitions:

Γ the set of n, a fixed number, of control points, $\Gamma = \{X_{i=1..n} / X_i = (x_i, y_i)\}$

Where (x_i,y_i) are point coordinates of control perimeter points.

C_0 the central point of the Γ set: $C_0 = (1/n)(\sum_{i=1..n} x_i, \sum_{i=1..n} y_i)$, $/(x_i, y_i)_{i=1..n} \in \Gamma, \beta_i = angle(C_0 X_i X_{i+1}), \alpha_i = angle(X_i C_0 X_{i+1})$ angles defined for each interpolating points of Γ.

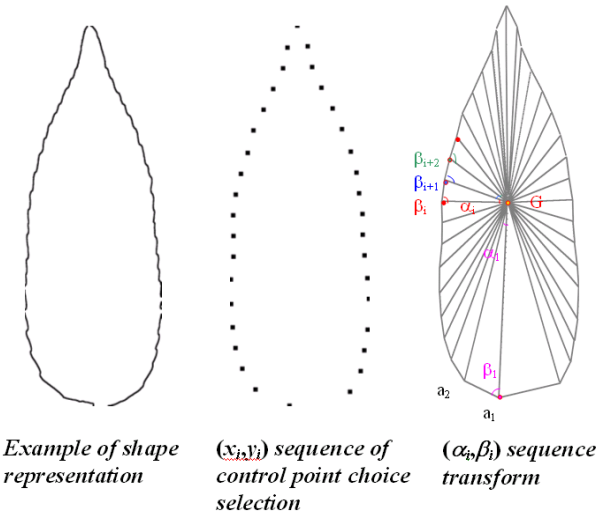

Fig. 3. Example of an angular coding for a 30 control points selection.

An example is shown in figure 3. Sequences of (x_i,y_i) positional points are then transformed in sequence of (φ_i, β_i) angular points.

The choice of a starting and a central point accounts for scale and leaf orientation. Placement of both points sets the scale: its distance separation. Relative point positioning sets the orientation of the interpolating shape. Given a sequence of such angles α_i and β_i, it's then possible to reconstruct the interpolating shape of a leaf. Geometrical properties of triangle similarities make such sequence size and orientation free.

4 Reduction parameters

Principal Components Analysis (PCA) is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences [5]. Since patterns in data can be hard to find in data of high dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analyzing data. The other main advantage of PCA is that once you have found these patterns in the data, and you compress the data, i.e. by reducing the number of dimensions, without much loss of information.

PCA is an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. PCA is theoretically the optimum transform for a given data in least square terms.

In PCA, the basis vectors are obtained by solving the algebraic eigenvalue problem $\mathbf{R}^{\mathbf{T}}\left(\mathbf{X}\mathbf{X}^{\mathbf{T}}\right)\mathbf{R} = \Lambda$ where $\mathbf{X}$ is a data matrix whose columns are centered samples, $\mathbf{R}$ is a matrix of eigenvectors, and Λ is the corresponding diagonal matrix of eigenvalues. The projection of data $\mathbf{C}_n = \mathbf{R}_n^{\mathbf{T}}\mathbf{X}$ **,** from the original p dimensional space to a subspace spanned by n principal eigenvectors is optimal in the mean squared error sense.

Another possibility, not presented here, is to use Independent Component Analysis (ICA) [11] [12] instead of PCA. In this case, we obtain independent coordinates and not only orthogonal as in previous case. ICA has been used for dimensional reduction and classification improvement with success [13].

In our problem we have 16 different classes of leaves, and for each class we have 75 different samples, where each one is a two column matrix of 359 points. First column corresponds to interior angles α_i and second column to exterior angles β_i , as explained before (see Fig. 3).

The procedure for applying PCA can be summarized as follows:

1. Subtract the mean from each of the data dimensions. This produces a data set whose mean is zero (**X**).
2. Calculate de covariance matrix ($\mathbf{Cov}_x$)
3. Calculate the eigenvectors e_i and eigenvalues λ_i of $\mathbf{Cov}_x$
4. Order the eigenvectors e_i by eigenvalue λ_i, highest to lowest. This gives us the components in order of significance.
5. Form a feature vector by taking the eigenvectors that we want to keep from the list of eigenvectors, and forming a matrix (**R**) with these eigenvectors in the columns.
6. Project the data to a subspace spanned by these n principal components.

5 Classification

For the classification system based on the SVM [6], [7], in order to establishing efficiency, we have calculated error, success and rejected rates on recognition.

Particularly, we have used an implementation of Vapnik`s Support Vector Machine known as SVM light [6], [7] which is a fast optimization algorithm for pattern recognition, regression problem, and learning retrieval functions from unobtrusive feedback to propose a ranking function. The algorithm has scalable memory requirements and can handle problems with many thousands of support vectors efficiently.

In the next figure, we can see the detection of support vectors and the creation of a boundary, one per each class, because it is a bi-class classifier (see figure 4). In our implementation, we have built a multi-classes classification module, from this SVM light.

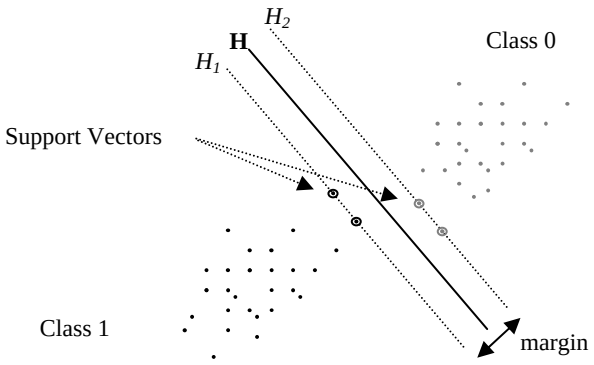

Fig.4 Separate lineal Hyperplane in SVM.

6 Experiments and results

We did several experiments in order to find the best dimension reduction for leaves automatic recognition. In our experiments we observed that the first column of each sample, that corresponds to the interior angles α_i are not useful for the classification purpose. Hence, we use only exterior angles β_i.

To apply PCA to these values we construct a global matrix with 37 of 75 different samples that we have for each class, arranged in rows, resulting in a 592x359 global matrix. Procedure detailed in Section 4 is applied to this matrix in

order to obtain the projected data by using the subspace spanned by 1 to 15 principal components (only odd numbers in our experiments).

Number of Components	Success Rates	Type of kernel	Γ
1	41.53% ± 3.72	Lineal	---
3	44.58% ± 0.33		
5	49.94% ± 7.49		
7	58.62% ± 0.90		
9	62.33% ± 3.40		
11	63.70% ± 0.11		
13	63.43% ± 0.12		
15	69.25% ± 0.33		
1	58.27% ± 0.98	RBF	8×10^{-2}
3	62.23% ± 0.70		0.9
5	78.91% ± 0.33		7×10^{-2}
7	83.69% ± 0.80		0.7
9	83.69% ± 2.01		0.6
11	84.58% ± 0.62		1×10^{-1}
13	86.25% ± 0.45		6×10^{-1}
15	87.14% ± 0.86		7×10^{-2}

Table 2. Results with SVM classifier

Data processed with PCA is used then with the SVM classificatory and results are shown in Table 2. As the methodology of our experiments was a cross-validation method repeating each experiments 10 times, Table 2 shows the obtained average ± typical deviation success rate for different number of PCA components considered in the experiments.

We compare our results with the results obtained in [4], where a HMM of 40 stages was used in the best case, giving a success rate of 78.33% ± 6.06. As can be seen in Table 2, RBF kernel for a SVM system give much better results than lineal kernel whatever the number of components is used. In all of the cases of SVM with RBF kernel and 5 components or more, we outperforms the HMM results in success rate and we diminish the variance as well. The best case is obtained with RBF kernel and a PCA of 15 components, that significantly improves the HMM results.

7 Conclusions

In this present work, we have presented an improvement of an automatic leaves recognition system using Principal Component Analysis and classifying with Support Vector Machines. The transformation and reduction of data contribute to increase its discrimination, from 78.33% using contour parameterization + HMM

to 87.14% using contour parameterization + PCA + SVM. The advantage of using PCA is twofold: first, we increase the classification results, and second we diminish the features dimension, giving as a result a less complex classifier. Future work will be done using ICA as an alternative method to PCA for improving results, and other kind of classifiers will be explored.

Acknowledgments The first author acknowledges support from the Ministerio de Educación y Ciencia of Spain under the grant TEC2007-61535/TCM, and from the Universitat de Vic under the grant R0912

References

1. Lu, F., Milios, E.E.: Optimal Spline Fitting to Planar Shape. Elsevier Signal Processing, No. 37, pp 129-140. (1994)
2. Loncaric, S.: A Survey of Shape Analysis Techniques. Pattern Recognition, Vol. 31 Nº. 8, pp 983-1001. (1998)
3. Rabiner, L., Juang, BH.: *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, New Jersey. (1993)
4. Briceño, J.C., Travieso, C.M., Ferrer, M.A.: Automatic Recognition of Simple Laurisilva Canariensis Leaves, by Perimeter Characterization. IASTED International Conference on Signal Processing, Pattern Recognition and its Applications, pp. 249 – 254. (2002)
5. Jolliffe I.T. *Principal Component Analysis*, Series: Springer Series in Statistics, 2nd ed., Springer, NY, 2002, ISBN 978-0-387-95442-4
6. Cui, G., Feng G., Shan S.: Face Recognition Based on Support Vector Method. Proceedings of 5th Asian Conference on Computer Vision, pp 23-28. (2002)
7. Guo, G., Li, S. Z., Kapluk, C.: Face recognition by support vector machines. Image and Vision Computing, Vol. 19 Nº9- 10, pp. 631–638. (2001)
8. Lu, F., Milios, E.E.: Optimal Spline Fitting to Planar Shape. Elsevier Signal Processing Nº. 37- pp 129-140. (1994)
9. Loncaric, S.: A Survey of Shape Analysis Techniques. Pattern Recognition. Vol 31 Nº. 8, pp. 983-1001. (1998)
10. Huang, Z., Cohen, F.: Affine-Invariant B-Spline Moments for Curve Matching. IEEE Transactions on Image Processing, Vol. 5. No. 10. pp 824-836. (1996).
11. C.Jutten, J. Herault, " Blind separation of sources, Part 1: an adaptive algorithm based on neuromimetic architecture", Signal Processing (Elsevier), Vol. 24 , Issue 1 (July 1991), ISSN:0165-1684
12. A. Hyvärinen, J. Karhunen, E. Oja, *Independent Component Analysis*, New York, USA: John Wiley & Sons, 2001
13. V.Sanchez Poblador, E. Monte Moreno, J.Solé-Casals, "ICA as a preprocessing technique for Classification", ICA 2004, Granada, Spain,. Lecture Notes in Computer Science, Springer-Verlag Volume 3195/2004