Evaluation of Features for Leaf Discrimination

Pedro F.B. Silva[1], André R.S. Marcal[1], and Rubim M. Almeida da Silva[2]

[1] Departamento de Matemática, Faculdade de Ciências da Universidade do Porto
[2] Departamento de Biologia, Faculdade de Ciências da Universidade do Porto

Abstract. A number of shape features for automatic plant recognition based on digital image processing have been proposed by Pauwels et al. in 2009. A database with 15 classes and 171 leaf samples was considered for the evaluation of these measures using linear discriminant analysis and hierarchical clustering. The results obtained match the human visual shape perception with an overall accuracy of 87%.

1 Introduction

The generalization of portable devices with image processing capacity, such as tablets and smartphones, together with the improvement of signal processing techniques, constitutes a framework in which the development of an automatic plant recognition system is possible. Such a system could not only serve the specific academic purposes of biologists and other technicians, but also be of general interest to the wider public.

The development of a plant recognition system requires a set of discriminating variables, as well as a structured database to train statistical models. There are a number of papers in the scientific literature that present variables for leaf classification [1,2,3]. The purpose of this paper is to present an evaluation of the descriptive power of a set of variables proposed in 2009 by Pauwels et al. [1] in the context of the development of an automated image recognition software and to discuss their statistical properties. A database of 15 different plants with a total of 171 leaf samples was used. The paper is structured as follows: in section 2 the features are briefly introduced and the test dataset is presented, section 3 presents the results and discussion, and section 4 presents the conclusions.

2 Materials and Methods

2.1 Geometric Features

A short description of common shape features is initially done following [1]. Let I denote the object of interest, ∂I its border, $D(I)$ the diameter, i.e., the maximum distance between any two points in ∂I and $A(I)$ the area. Let $A(H(I))$ denote the area of the object's convex hull and $L(\partial I)$ the object's contour length. The operator $d(.)$ stands for the Euclidean distance.

1. *Eccentricity* - The eccentricity of the ellipse with identical second moments to I is computed. This value ranges from 0 to 1.

M. Kamel and A. Campilho (Eds.): ICIAR 2013, LNCS 7950, pp. 197–204, 2013.

2. *Aspect Ratio* - Consider any two pixels $X, Y \in \partial I$. Choose X and Y such that $d(X,Y) = D(I)$. Find $Z, W \in \partial I$ maximizing $d(Z,W)$ on the set of all pairs of ∂I that define a segment orthogonal to $[XY]$. Let $D^{\perp} = d(Z,W)$. The aspect ratio is defined as the quotient $D(I)/D^{\perp}$. Values close to 0 indicate an elongated shape.
3. *Elongation* - Compute the maximum escape distance $d_{\max} = \max_{X \in I} d(X, \partial I)$. Elongation is obtained as $1 - 2d_{\max}/D(I)$ and ranges from 0 to 1. The minimum is achieved for a circular region. Note that the ratio $2d_{\max}/D(I)$ is the quotient between the diameter of the largest inscribed circle and the diameter of the smallest circumscribed circle.
4. *Solidity* - The ratio $A(I)/A(H(I))$ is computed, which can be understood as a certain measure of convexity. It measures how well I fits a convex shape.
5. *Stochastic Convexity* - This variable extends the usual notion of convexity in topological sense, using sampling to perform the calculation. The aim is to estimate the probability of a random segment $[XY]$, $X, Y \in I$, to be fully contained in I.
6. *Isoperimetric Factor* - The ratio $4\pi A(I)/L(\partial I)^2$ is calculated. The maximum value of 1 is reached for a circular region. Curvy intertwined contours yield low values.
7. *Maximal Indentation Depth* - Let $C_{H(I)}$ and $L(H(I))$ denote the centroid and arclength of $H(I)$. The distances $d(X, C_{H(I)})$ and $d(Y, C_{H(I)})$ are computed $\forall X \in H(I)$ and $\forall Y \in \partial I$. The indentation function can then be defined as $[d(X, C_{H(I)}) - d(Y, C_{H(I)})]/L(H(I))$, which is sampled at one degree intervals. The maximal indentation depth $\mathfrak{D}$ is the maximum of this function.
8. *Lobedness* - The Fourier Transform of the indentation function above is computed after mean removal. The resulting spectrum is normalized by the total energy. Calculate lobedness as $F \times \mathfrak{D}^2$, where F stands for the smallest frequency at which the cumulated energy exceeds 80%. This feature characterizes how lobed a leaf is.

It is worth noting that the features described were slightly modified in regards to the original definitions in [1].

The feature *aspect ratio* uses the length of the entire orthogonal axis and not just the maximum distance between a boundary point and $[XY]$.

The original definition of *stochastic convexity*, in [1], is somewhat incomplete. It is claimed that this variable constitutes an accurate estimator of a convexity measure, although no indication is given concerning the required number of points to compute the variable. We tested a variety of cases, and consider the mean value of 3 repetitions with 20 points for the calculation. Some results about this variable's statistical behaviour are presented in section 3.

The feature *lobedness* is computed using a normalized spectrum as mentioned above. Normalization has been applied in order to exclude any scale, translation or rotation effects that could affect erroneously the feature's values.

Furthermore, some non-geometrical features referred in [1] were excluded from this analysis, given the fact that their inclusion worsened the results of linear discriminant analysis over the two first principal components in about 13%.

2.2 Test Dataset

A test dataset was prepared to evaluate the geometric leaf features. A total of 171 images from 15 different classes were collected using an Apple iPAD 2 device. The images, 720 × 960 pixels, 24 bit RGB, were acquired with a contrasting background (e.g. pink, blue, etc.). These were segmented using an automatic process further corrected by human inspection in order to ensure that the leaf pecioles were removed to ensure comparability. Although the segmentation process could be further automated, it is not the focus of this work. Table 1 presents the considered plant species and the number of specimens in each class. An overview of one individual from each class is presented in Figure 1. Inner class variability is illustrated in Figure 2, where the 10 available individuals from class 2 are presented.

Table 1. Test dataset class list and number of individuals

Class	Species	#	Class	Species	#	Class	Species	#
1	*Quercus suber*	12	6	*Crataegus monogyna*	8	11	*Acer palmaturu*	16
2	*Salix atrocinerea*	10	7	*Ilex aquifolium*	10	12	*Celtis sp.*	12
3	*Populus nigra*	10	8	*Nerium oleander*	11	13	*Corylus avellana*	13
4	*Alnus sp.*	8	9	*Betula pubescens*	14	14	*Castanea sativa*	12
5	*Quercus robur*	12	10	*Tilia tomentosa*	13	15	*Populus alba*	10

Fig. 1. Overview of leaves of the plant species considered (from left to right: no. 1-8 on top, no. 9-15 on bottom)

Fig. 2. Example of inner class variability for class 2 - *Salix atrocinerea*

3 Results

The feature dataset was evaluated by various processes, namely: Pearson's correlation, Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Hierarchical Clustering.

3.1 Pearson's Correlation

Pearson's correlation was calculated for the 8 shape features. The results are presented in Table 3.1.

Table 2. Pearson correlation matrix for the 8 shape features

1	2	3	4	5	6	7	8
1	0.67	0.46	0.54	0.46	0.16	-0.57	-0.55
	1	0.64	0.33	0.28	-0.18	-0.29	-0.29
		1	-0.37	-0.4	-0.74	0.36	0.36
			1	0.86	0.79	-0.93	-0.92
				1	0.72	-0.89	-0.91
					1	-0.81	-0.72
						1	0.96

The feature definition (section 2) clearly suggested that there should be high correlation between some features. Redundancy is pointed out by very high values of linear correlation (>0.9) for the following pairs of variables: *maximal indentation depth/lobedness* (0.96), *maximal indentation depth/solidity* (0.93), *solidity/lobedness* (-0.92). These results are quite natural as these features assess closely related shape properties. *Stochastic convexity*, being highly correlated with *lobedness*, is also necessarily very correlated with *maximal indentation depth* and *solidity*. *Isoperimetric factor* exhibits a high level of correlation (0.7-0.8) with *maximal indentation depth*, *solidity*, *elongation* and *isoperimetric factor*.

3.2 Principal Component Analysis (PCA)

Principal component analysis was used to reduce dimensionality and perform exploratory analysis. The correlation matrix was used as features variance have different magnitudes. A summary of the results obtained is presented in Table 3 and graphically in Figure 3.

Table 3. Principal Component Analysis - Weights

	1	2	3	4	5	6	7	8
Comp.1	-0.25	-0.13	0.18	-0.43	-0.42	-0.37	0.44	0.43
Comp.2	0.48	0.58	0.59	0.02	0.00	-0.30	-0.02	-0.03

The first two components together explain 91% of point variability (62%+29 %). The first component works as a "mean" of the shape features contributions, grouping them in two sets accordingly to the geometric properties they measure:

eccentricity, *aspect ratio*, *solidity*, *stochastic convexity* and *isoperimetric factor*, with negative weights and *elongation*, *maximal indentation depth* and *lobedness* with positive weights. Notice as well, that highly correlated shape features have similar weights in the first component.

Interpretation of the second principal component is not so obvious. Notice though that the variables exhibiting high correlation among each other (*maximal indentation depth*, *lobedness*) *solidity* and *stochastic convexity*, have low weights in this component. On the other hand, variables exhibiting lower values of linear correlation with other variables (*eccentricity*, *aspect ratio* and *elongation*) have higher weights in this component. Thus, variables with less redundancy, are contributing more to explain point variance in this component.

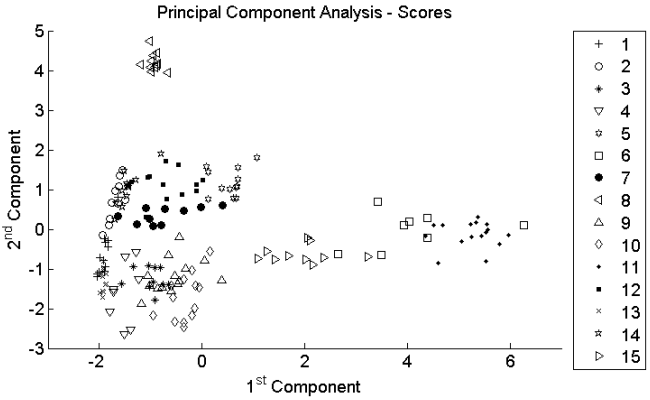

Fig. 3. Principal Component Analysis of the 8 shape features

It has been shown that principal components are the continuous solution of the cluster membership indicators in the K-means clustering method, i.e., the PCA dimension reduction automatically performs data clustering according to the K-means objective function [4]. The point cloud in Figure 3 is therefore an indication of possible satisfactory classification analysis with these shape features, due to the reasonable point dispersion and reduced overlap between classes.

3.3 Linear Discriminant Analysis (LDA)

Linear discriminant analysis was performed, using the entire dataset for further descriptive purposes. The predictive capacity of the analysed shape features was tested using a cross validation scheme, splitting the dataset in training (70%) and testing (30%). Linear discriminant analysis was performed on the randomized training and testing sets and the mean misclassification rate along 1000 executions was 12.7%.

Considering the entire database the lowest misclassification rate achieved was 13% using all shape features except *eccentricity*. Table 4 shows the accuracy rate for the exercise classes. The accuracy for classes 5, 8, 10, 11 and 15 is 100%, as all images were correctly identified. These appear, from a human perception perspective, to be the most easily distinguishable in the dataset (cf. Figure 1).

Table 4. Contingency Table for Linear Discriminant Analysis

Class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	∑	Accuracy
1	8	1											3			12	67%
2	1	7												2		10	70%
3			8						1				1			10	80%
4			1	6									1			8	75%
5					12											12	100%
6						7									1	8	87%
7		1					8					1				10	80%
8								11								11	100%
9			2	1			1		10							14	71%
10										13						13	100%
11											16					16	100%
12												10		2		12	83%
13	1												12			13	92%
14		3										1		8		12	67%
15															10	10	100%

3.4 Hierarchical Clustering

The centroid of each class was calculated in the space spanned by the first and the second principal components. The distances between centroids were computed using the Euclidean distance. Hierarchical clustering was afterwards performed considering the Ward's linkage. The result of this analysis is presented as a dendrogram in Figure 4.

The result of this analysis is quite satisfactory. Leaves were first separated in two groups: simple (1,2,3,4,5,7,8,9,10,12,13,14) and lobed (6,11,15). Considering the first group a further division into two groups can be found: leaves with an elongated shape (2,5,8,7,12,14) and leaves with a roundish shape (1,3,4,9,10,13).

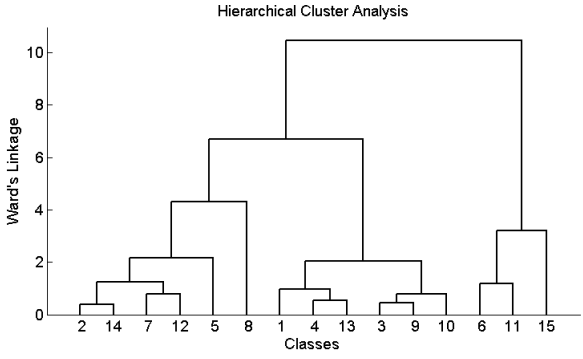

Fig. 4. Dendrogram of the Hierarchical Clustering Analysis

In the group of the elongated leaves a further division takes place, setting class 8 apart from the other classes, due to its extremal aspect ratio. Further association, namely between the classes 2 - 14 and 7 - 12, match in a reasonable way the human visual perception.

The group of roundish leaves was segmented in two subgroups: a group of leaves with elliptic-like shape (1,4,13) and a group of leaves with a heart-like shape (3,9,10). Posterior subdivision also matches the human visual perception.

3.5 Stochastic Convexity Feature Evaluation

The original definition of *stochastic convexity* in [1] does not mention how the feature computation is actually performed. The authors suggest that a random samples is made but without indicating the number of points used. The index is computationally demanding as all points are used.

Some tests on the stability of this feature were conducted for 10, 20, 30, 50 and 100 points using one leaf from class 11 and for 1, 3, 5, 10 and 15 repetitions. In Table 5 the mean value of stochastic convexity and the necessary time for computation are presented. For this last case (100 points) a total of 4950 pairs are tested, which results in very long execution times.

Table 5. Stochastic convexity and execution time

	Stochastic Convexity					Execution Time (s)				
	10	20	30	50	100	10	20	30	50	100
1	0.71	0.58	0.66	0.75	0.72	0	2	4	12	48
3	0.75	0.57	0.68	0.66	0.68	1	6	13	36	144
5	0.72	0.66	0.70	0.70	0.66	2	9	21	59	240
10	0.61	0.72	0.68	0.66	0.67	4	18	42	119	480
15	0.63	0.65	0.68	0.69	0.68	7	28	64	177	722

It is obvious from the definition that this shape is very stable for a potentially convex shape, because of the high probability that the selected points span segments fully contained in the set. Considering the results from Table 5 it can be observed that the stability of the measure seems to be more dependent on the number of repetitions used for the calculation than on the number of points used. Controlling the results for stability is computationally expensive as can be observed in Table 5. In this paper an average of 3 repetitions using 20 points was used, as this seemed a good compromise between some stability and computational burden.

Although the *stochastic convexity* feature, proposed in [1], is unstable, it cannot be said that this feature damages the classification results. In this dataset this feature was highly correlated with *solidity*. The result of a linear discriminant analysis over principal component analysis considering and excluding *stochastic convexity* as input feature yielded the same misclassification rate, approximately 16%.

4 Conclusions

The analysis carried out suggest that the shape features proposed by [1] can be adequate for the development of an automatic leaf classification system based on digital images. Such a system can be implemented on mobile devices, such as smartphones, as the computational effort for the calculation of the presented measures is acceptable, with exception of the feature *stochastic convexity*, which can only be used in its lighter form.

The result of an overall classification scheme using one of the simplest classifiers available (LDA) exhibits a high accuracy rate (87%). Better results could possibly be achieved considering other aspects than shape, such as texture and color properties, or using other shape features like elliptic Fourier descriptors, as well as other classifiers. This approach was not pursuit here, as its computational complexity would likely inhibit development on a mobile platform.

As pointed out by the performed hierarchical analysis, the results of automatic classification match natural human pattern recognition perception. This could allow the user to select the ultimate result by being given a list of alternatives sorted out by matching likelihood. This strategy was refered in [1] using a K-Nearest Neighbours classifier. Hierarchical clustering can be a better approach given the natural interpretation of the results and the ability to automatically generate a leaf shape taxonomy.

A possible drawback of the set of features used, as indicated in [1], is the somewhat high correlation values between some of the features. However, in our experimental test, there was no evidence that high correlation hampered the classification results.

References

1. Pauwels, E.J., de Zeeuw, P.M., Ranguelova, E.: Computer-assisted tree taxonomy by automated image recognition. Eng. Appl. of AI 22(1), 26–31 (2009)
2. Du, J.X., Wang, X., Zhang, G.J.: Leaf shape based plant species recognition. Applied Mathematics and Computation 185(2), 883–893 (2007)
3. Cope, J.S., Corney, D.P.A., Clark, J.Y., Remagnino, P., Wilkin, P.: Plant species identification using digital morphometrics: A review. Expert Syst. Appl. 39(8), 7562–7573 (2012)
4. Ding, C., He, X.: K-means clustering via principal component analysis. In: Proceedings of the Twenty-First International Conference on Machine Learning, ICML 2004, 29– ACM, New York (2004)
5. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley Interscience (2000)
6. Hu, M.K.: Visual pattern recognition by moment invariants. IEEE Transactions on Information Theory 8, 179–187 (1962)
7. Oonincx, P.J., de Zeeuw, P.M.: Adaptive Lifting For Shape-Based Image Retrieval. Pattern Recognition 36, 2663–2672 (2003)