Leaf Species Classification Based on a Botanical Shape Sub-Classifier Strategy

H. Liu, D. Coquin, L. Valet
LISTIC, Université de Savoie
74944 Annecy le Vieux, France

G. Cerutti
LIRIS, UMR5205, Université de Lyon2
F-69676 Lyon, France

***Abstract*—Within the framework of a smartphone-based application, helping people to identify plant species in the wild, a sub-classifier strategy has been introduced. It aims at recognizing the botanical properties of a leaf, relatively to various global and local shape criteria used in flora books. A decision function is applied on these classified shape categories to produce a final decision on the species of the leaf. In this paper, the fusion strategy and its corresponding Random-Forest-based sub-classifiers are described. The results of these algorithms for botanical leaf shape recognition demonstrate that our classification algorithm can achieve good performance on leaf species identification while providing the user with relevant information for educational purposes.**

***Keywords*—Leaf recognition, Data Classification, Random Forest, Dempster-Shafer theory, Belief functions theory.**

I. INTRODUCTION

The ability to recognize a plant species and to understand its specificities remains a task accessible mostly to specialists. Most flora books promise an arduous time to botanical amateurs who have superficial background knowledge. Recent developed functionalities on smartphone can however help botanical amateurs on plant recognition by introducing botanic knowledge in an interactive way. Nowadays, identifying plant species against a white background using smartphone camera has shown very satisfactory performances [1][2]. Further along that road, we aim at developing an educational smartphone application to help users to recognize a plant species in its natural environment[**]. In order to perform such task, the application first lets the user take a picture of a leaf of the unknown plant with the smartphone camera. Then it extracts high-level morphological features used to predict a list of most corresponding species. This application is also designed to provide an educative and interactive way of transmitting identification skills to amateur botanists.

Plant recognition and leaf image retrieval have been recently been topics of interest for many works in image processing [5][6][7], though most of them limit the problem to leaves on a plain background. Challenges for the community have even been organized such as the ImageCLEF Plant identification Task [3] since 2011. Some authors [1][2] also pursued the goal of conceiving a mobile guide with the LeafSnap application, combining established shape descriptors (Inner-Distance Shape Context) and classification methods to perform remarkably well on white-background images.

Section 2 describes the classification problems related to the leaf data and the inconvenience of the classification strategy presently used for the application. Then, a sub-classifier strategy aiming at giving a more reasonable decision is given in Section 3. It consists of several sub-classifiers based on Random Forest. The two most voted results of each sub-classifier are fused with belief functions theory to give the best decision. Section 4 presented experimentations that show promising general classification accuracy.

II. PROBLEMS OF LEAF DATA CLASSIFICATION

A. Leaf dataset

The leaf classification problem we consider is a multi-class problem, which involves 126 tree and shrub species found in the France territory. We used images present in the Pl@ntLeaves II Dataset [3], that consist of a mix of plain background images (labeled Scan or Pseudoscan) and natural background images (Photograph) of 112 simple leaf species and 14 compound leaf species. Examples of the three different categories are given in **Fig. 1.**

Fig. 1. Examples of scan, pseudoscan and photograph images from the Pl@ntLeaves II Dataset

In accordance with our previous work [4], our system extracts 4 compact sets of attributes to describe the morphological properties of a segmented leaf on a high-level of interpretation. Those sets are designed to capture the information relatively to 4 supposedly independent shape criteria used by botanists to describe leaves, namely the global shape (11 attributes), the basal and apical shapes (5 attributes each), and the margin shape (13 attributes). These descriptors perform a decorrelation of the shape information on the different shape criteria, as illustrates **Fig. 2**.

[**] Folia: free application for iPhone, available on the AppStore : https://itunes.apple.com/app/folia/id547650203

This work has been supported by French National Agency for Research with the reference ANR-10-CORD-005 (ReVeS project).

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.266

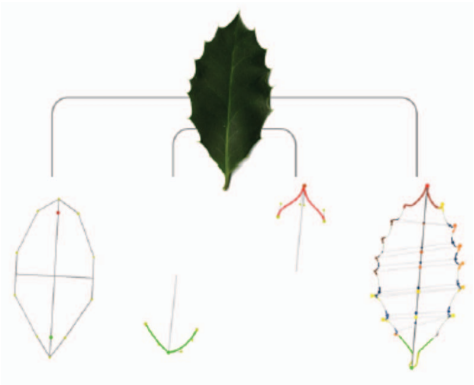

Fig. 2. Leaf shape description decorrelating global, basal, apical and margin shapes

In this paper, we focus on the classification of simple leaf species, though the methods developed in this study can be extended to compound leaf species or to other shape criteria. Among the treated species, some lack enough samples to give sufficient statistical power, and we chose to eliminate classes with less than 5 samples for the study. The resulting problem consists of a multivariate dataset of high dimensionality, consisting of 34 attributes for 7338 individuals distributed on 108 plant species.

The database is potentially noisy, especially in the case of photographs, since their complexity for segmentation may lead to incorrect evaluation of the attributes. In addition, the number of images per class is different: some classes contain more than 200 individuals whereas only 5 individuals represent others. This may affect model training and generate biased results. At last, the species classification problem is characterized by significant intra-class variability, combined with a potentially high inter-class similarity for some species. This induces ambiguity concerning the class of an individual to be predicted, which can be seen on Fig. 3.

Fig. 3. Two leaves of the same species (common holly, *Ilex aquifolium* L.) (a,b) and one of a different species (holly oak, *Quercus ilex* L.) (c)

B. Botanical shape sub-classification

The diversity of shapes for a species is well accounted in botanical reference books, where each shape criterion of a species is generally described by up to three terms. These categorical terms belong to a specific vocabulary, quite obscure at first sight, that covers the set of shapes one can encounter for each criterion. For instance, the global shape of a leaf, can be describes as Triangular, Ovate, Obovate or even Cordate.

The objective of the sub-classification is to automatically associate a leaf with such botanical terms, for each of the 4 considered shape properties. It requires a supervised training, which consists in inferring a classification function from labeled samples, using a training set composed of the corresponding vectors of attributes.

However, the botanical labels of the training samples are unknown *a priori*, and a ground-truth labeling would involve too much subjectivity. To establish a labeling reference nevertheless, we collected botanical information from reference flora books [8][9] and summarized morphological characteristics in a reference table. For example, Tab. 1 shows an extract of the reference table for the sessile oak (*Quercus petraea* Liebl.) and the downy oak (*Quercus pubescens* Willd.) on global shape, base, apex and margin properties. This table is used as a reference to develop our sub-classification models. It is worth noting that some information might be unknown for some species on shape properties.

Table 1. Extract of the reference table showing the shape properties for the sessile and downy oaks.

Property / Species	Global	Base	Apex	Margin
Sessile Oak	Obovate	Cordate	Round	Lobate, Sinuate
Downy Oak	Obovate	Truncate	Round	Lobate, Sinuate

In the case of margin shapes for instance, the information is lacking for some species, while for others several types of margin could be attributed to a same species. In order to have a reliable reference, it was more appropriate to use only the data of plant species with a unique margin shape in the reference table for training the sub-classifier (62 species out of 108).

This allows us to remove the ambiguity and the uncertainty present on the training set. The downside is that some terms had to be left apart since no species presents only this term. However, based on the unambiguous data, we develop a robust model recognizing leaf margin shapes. The model trained on the unique-margin species can later be tested with the other species non-involved in the training set to experiment its efficiency.

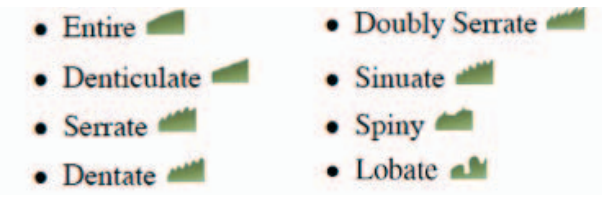

Fig. 4. List of margin shapes considered for sub-classification

The **Fig. 4** gives the list of the retained classes for the margin shape sub-classification. Notice that certain shapes

look very similar, and might be hard to distinguish especially if the segmentation produces an imprecise contour. Technically, this similarity will affect the result of the trained sub-classifier.

C. Previous classification strategy

Addressing to the aforementioned particularity of the data set, Random Forest was used in the previous work for the leaf classification. Random Forest (**RF**), introduced by Breiman [11], is an ensemble classifier that consists of a number of decision trees voting for the classes. The forest outputs the most frequent result all trees, the mode of classes output by individual trees. To build a decision tree, the base component of a Random Forest, a bootstrap sample of the data is used and at each node a set of random variable is selected to split on. This strategy of random sampling increases the error of each tree but reduces the correlation between trees. As a consequence, the ensemble achieves both low variance and low bias [11]. First in [10], the Random Forest was applied on the whole 74 attributes for the simple leaves classification. The parameters of the Random Forest (200 decision trees, 9 variables selected for each split) are chosen empirically and allow obtaining satisfactory classification results. With Pl@ntLeaves II Dataset, the average classification score is 42%. This average is computed for the classification rate for scan, pseudo-scan and photos images.

However, as mentioned above, the ambiguity caused by inter-class similarity and the intra-class variability makes a sample possibly classified into several plant species with high and balanced probabilities. This can affect the output of a Random Forest, as the real class of a leaf may not always be the first choice of the ensemble decision tree, but the second one. So, instead of giving a crisp decision on plant species, we prefer to compute a fuzzy one using for instance the five species having the most of the votes. Thus a user can make a more reasonable decision.

Moreover, one of our objectives is to offer an instructive way for smartphone users to learn how to recognize leaf species. The previous classification strategy gives no information about how the classification has been done. Hence, we propose an approach based on sub-classifiers strategy that gives a better explanation of the obtained results.

III. Shape Sub-Classification Strategy

We propose a sub-classifier strategy in order to solve the classification problems induced by the dataset. The method consists in using several sub-classifiers to characterize a leaf, according to various aspects such as global shape, base, apex, and margin. Then, the combination of all the property information given by each sub-classifier will produce the final prediction.

A. Sub-classification step

On the one hand, dividing the original classification problem into several sub-problems can simplify the task, as the classification is implemented on a sub-set possessing less number of variables; and on the other hand, this is somewhat similar to the intuitive way of a real botanist in recognizing plant species. Providing the users elementary information, regarding the vocabulary associated with the visible leaf properties, and how they have been used to make a decision is also a way of transmitting a knowledge concerning the particularities of the species.

As shown in **Fig. 5**, the attributes extracted by image processing methods [7] are split into four subsets A_{Bi}, A_{Ai}, A_{Mi} and A_{Si} in order to create four datasets. Each of these databases consists of some attributes describing one leaf property (B=base, A=apex, M=margin and S=global shape). Then, each dataset is split up into training and test set, following a proportion of 75% of samples devoted to training of the sub-classifiers RF_i. The samples are randomly selected.

The dataset is labeled using the reference data compiling the possible shapes of the species (manually filled from botanical information sources) and each example considered for training is associated with the corresponding shape category K^i of its species. The four random forest sub-classifiers RF_B, RF_A, RF_M and RF_S are trained using this way.

Then, given a vector of attributes for a new coming example, the sub-classifiers associate the leaf with shape categories and output four qualitative values K^B, K^A, K^M and K^S indicating how a leaf appears in terms of its base, apex, margin and global shape. In general, plant species correspond to a combination of the values $\{K^B, K^A, K^M, K^S\}$ that describe its typical shape. Then, a decision function $D(K^B, K^A, K^M, K^S)$ is used to combine the information provided by each sub-classifier to predict the plant species S_i $(i=1...108)$ of a leaf, by giving probability on several prospective species.

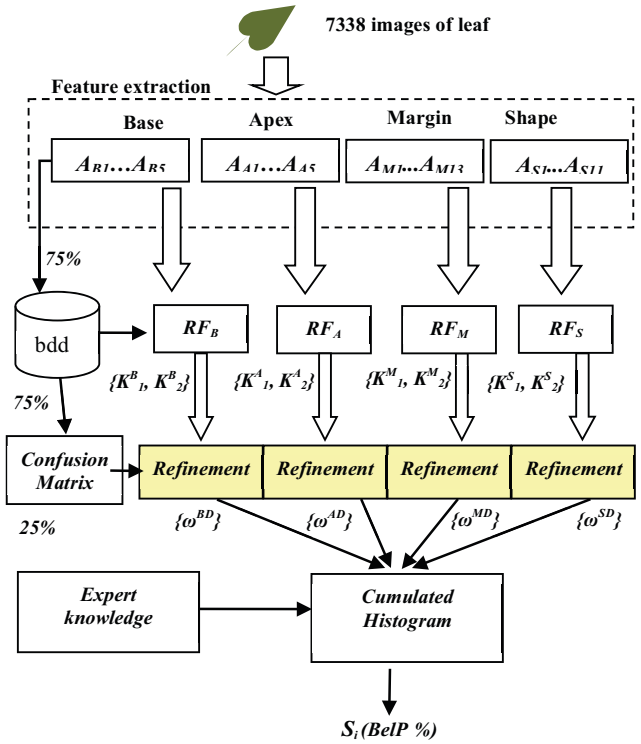

Fig. 5. Sub-classifier strategy based on leaf shape features

B. Refinement step

As we have discussed and shown in **Fig. 5**, there is high risk of confusion between species and the distinction is sometimes very difficult to make, especially for a novice in botany. The botanical terms themselves are also a source of confusion, given their unclear limits and their subjectivity. Due to this morphological confusion, the first output values of the sub-classifiers $\boldsymbol{RF_i}$ not necessarily reflect the actual properties of the leaf. In this case, it is advantageous to consider also the second output to identify the leaf. To do this, a refinement step based on the Dempster-Shafer theory has been implemented to determine which output value we must take into account. This refinement step is shown in **Fig. 6** and will be detailed in the following, applied to the **margin shape**.

As frame of discernment $\Omega_c = \left\{K_1^M, K_2^M\right\}$, we have the two first answers K_1^M and K_2^M as possible margin leaf, provided by the margin sub-classifier $\boldsymbol{RF_M}$. These responses are obtained by the two most important votes from the ***RF*** random forest. Dividing by the number of trees in the forest (*here 200*), the occurrence frequencies of these responses are obtained. These frequencies can be considered as degrees of belief that the random forest attributed to these two types of margin. Thus, we obtain masses of belief $m_{RF_M}\left(\left\{K_1^M\right\}\right)$ and $m_{RF_M}\left(\left\{K_2^M\right\}\right)$ given by sub-classifier $\boldsymbol{RF_M}$ (*the first information source*).

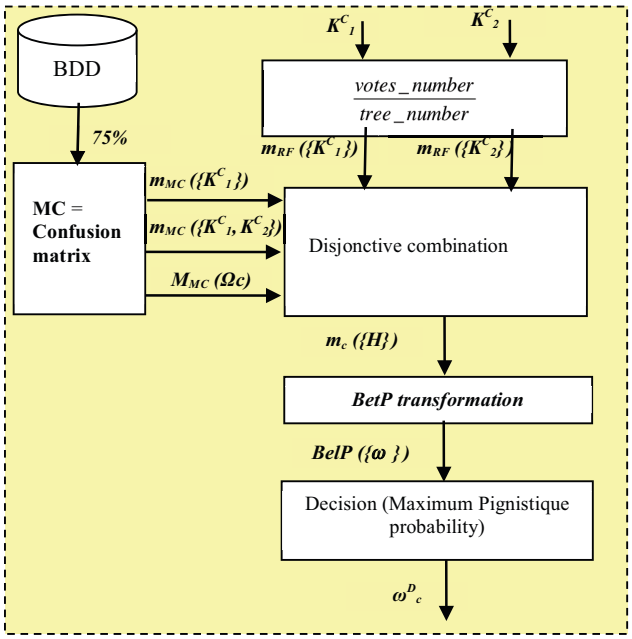

Fig. 6. Refinement step using Dempster-Shafer theory

On the other hand, it should be noted that given the similarity existing account and variability on different margins, our model might be wrong. To view the overall effectiveness of decision trees to characterize the margins of leaves, we built, from 75% of the data in the database BDD, a confusion matrix denoted ***MC***, one line $\boldsymbol{MC_i}$ (i = 1 ... n, n = number of margins) represents the number of occurrences of an outline $\boldsymbol{C_i}$ estimated by random forest trees on a set of individuals of the training set, and a column $\boldsymbol{MC_j}$ (j = 1 ... n) is the number individuals having a margin $\boldsymbol{C_j}$. Thus, after the normalization of the values relative to the sum of rows or columns, we can get information about the good margin recognition rate $\boldsymbol{MC_{ii}}$ (i = j) and the level of confusion between different margins $\boldsymbol{MC_{ij}}$ (i ≠ j). These values are used as another information source, denoted ***MC***, for refining the results of sub-classifiers.

Let K_1^M and K_2^M be the two highest output votes of sub-classifier $\boldsymbol{RF_M}$, corresponding to classes $\boldsymbol{C_1}$ and $\boldsymbol{C_2}$. If we consider the confusion matrix as another source, we can calculate beliefs about these classes, $m_{MC}\left\{C_1\right\} = MC_{11}$ either or $m_{MC}\left\{C_2\right\} = MC_{22}$, but also on the whole $m_{MC}\left\{C_1, C_2\right\} = MC_{12}$ or $m_{MC}\left\{C_2, C_1\right\} = MC_{21}$. These can be interpreted as beliefs that we have on the set $\left\{C_1, C_2\right\}$, knowing that the sub-classifier gives a $\boldsymbol{C_1}$ or $\boldsymbol{C_2}$ but plausibly confused $\boldsymbol{C_2}$ or $\boldsymbol{C_1}$ result. In addition, to meet a total belief equal to 1, we distribute the remainder of the mass of the entire outline, so that:

$$m_{MC}\left(\Omega_c\right) = 1 - m_{MC}\left(\left\{C_1\right\}\right) - m_{MC}\left(\left\{C_1\right\}, \left\{C_2\right\}\right) \qquad (1)$$

Since we have no information on the reliability of each source (***RF*** and ***MC***), we applied the disjunctive rule, to maintain a more cautious reasoning on a combination of sources [12]. The rule stated in [12] is used to revise the belief mass of each assumption $H \in \Omega_c$:

$$m_c\left(H\right) = m_{RF} \cup m_{MC}\left(H\right) = \sum_{A \cup B = H} m_{RF}\left(\left\{A\right\}\right) m_{MC}\left(\left\{B\right\}\right) \qquad (2)$$

where is and A and B are the focal elements of each source whose mass is greater than 0.

Once the masses of the revised assumptions, we convert into a pignistic probability by the following equation:

$$BelP\left(\omega\right) = \sum_{\left\{H \subseteq \Omega_c, \omega \in H\right\}} \frac{m_c\left(H\right)}{\left(1 - m_c\left(\varnothing\right)\right) card(H)} \qquad (3)$$

$card(H)$ is the cardinality of the set ***H*** and $\boldsymbol{m_c(\varnothing)}$ is the mass of the empty set is equal to 0, because our problem is expressed in a closed world. Then, the criterion of maximum probability pignistic is used to determine the value of the margin ω_M^D to be taken into account for the further identification of the leaf:

$$\omega_M^D = \max_{\omega} BelP\left(\omega\right) \qquad (4)$$

Weight ω_M^D is assigned to all related cash refined property. For example, if the sub-classifier Margin after refinement said that the the most probable margin shape category is $\boldsymbol{C_1}$ with a probability $BelP(\omega_M^D)$. Based on the botanical reference of species, we identify all the species $\boldsymbol{S_i}$ that present the margin $\boldsymbol{C_1}$ in the reference table, and assign them a weight of

$BelP(\omega_M^D)$. These weights are cumulated for all 4 shape properties by simply summing the pignistic probabilities obtained for each recognized shape category. This results in a histogram representing a distribution of probabilities over all the considered species. The most likely species will then simply be the one with the maximal cumulated probability.

Following this decision scheme, once the shape sub-classifiers are trained, the identification of the species relies only on an expert knowledge coming from botanical references. Each species is associated with a set of possible shapes for each criterion forming the reference table that is used to compute the species probabilities.

In particular, this means that new species can be added to the scope of identification without the need of any training image data. The description of a species in terms of shape properties, from a trustable source of knowledge, would be sufficient to make it possible for the system to recognize it, provided the ability of the sub-classifiers to correctly generalize the shape categorization they have learned.

IV. Experimentation

We tested our sub-classifier strategy on the Pl@ntLeaves II dataset, and tried to measure its performance both in terms of shape category sub-classification and species identification, and evaluate the effect of the refinement step on the results.

A. Sub-classifier Results (margin shape)

To tackle the identification problem, we used a classification scheme that takes into account several most voted shape categories given by Random Forest. As seen in **Fig. 6**, the classification accuracy can be enhanced significantly if we take into account the 2nd and 3rd most voted margins for classification, with increasing rate about 20%. Through this way, a much more complete information on the shape category can be used for making a decision, and therefore reduce the miss-classified cases.

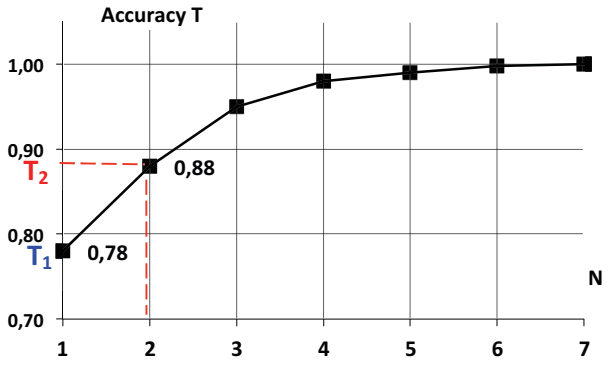

Fig. 6. Variation of classification accuracy for leaf margin, according to the number *N* of margins type taken into account for classification (for all images of Pl@ntLeaves II dataset).

The implemented classification models output a property combination describing plant species. For example, if we have a property combination such as Heart-shaped (base), Acuminate (apex), Serrate (margin) and Cordate (shape), we can say that the underlying leaf would probably to be a lime (*Tilia cordata* Mill.) for which the expert description is a perfect match.

Presenting such information to the user would be extremely beneficial from an educational point of view. It is first a way of transmitting a shape vocabulary which might otherwise be rather obscure. The performance reached by the sub-classifiers make them clear candidates to fulfill this role. And of course, providing the link between this now concrete shape description and the identified species constitutes the heart of the botanical identification skill.

B. Classification Results

The results of our method applied to the same dataset Pl@ntLeaves II, are presented table 2 to table 4. Surprisingly, it shows better performances with the pseudoscan images than with the generally less noisy scan images. This behavior has been observed by quite important number of team during the challenge ImageCLEF 2012. Thus, we suppose it is mostly due to species presence in each type of image, as some class are inherently more difficult to distinguish than others.

Table 2: The enhancement induced by the sub-classification and the refinement step on the classification of **scan images.**

Sub-classifier	**Base**		**Apex**		**Margin**		**Shape**	
Number of output for RF	N=1	N=2	N=1	N=2	N=1	N=2	N=1	N=2
Recognition rate for the properties	78%	92%	61%	81%	75%	86%	74%	88%
Refinement with Dempster Shafer		92%		**86%**		**89%**		**82%**
Recognition rate of species with N=1	49%							
Recognition rate of species with N=2	**64%**							

Table 3: The enhancement induced by the sub-classification and the refinement step on the classification of **pseudoscan images**.

Sub-classifier	**Base**		**Apex**		**Margin**		**Shape**	
Number of output for RF	N=1	N=2	N=1	N=2	N=1	N=2	N=1	N=2
Recognition rate for the properties	70%	88%	56%	85%	74%	84%	68%	83%
Refinement with Dempster Shafer		92%		**86%**		**89%**		**75%**
Recognition rate of species with N=1	41%							
Recognition rate of species with N=2	**65%**							

Table 4: The enhancement induced by the sub-classification and the refinement step on the classification of **photo images**.

Sub-classifier	**Base**		**Apex**		**Margin**		**Shape**	
Number of output for RF	N=1	N=2	N=1	N=2	N=1	N=2	N=1	N=2
Recognition rate for the properties	65%	83%	43%	67%	65%	87%	67%	86%
Refinement with Dempster Shafer		87%		**87%**		**93%**		**95%**
Recognition rate of species with N=1	45%							
Recognition rate of species with N=2	**55%**							

With the use of sub-classifiers applied to shape categories, the classification results appear to be increased compared with the direct Random Forest classification of the whole attributes [10]. Using the second output of the random forest and a method based on the evidential theory to avoid confusion, it further improves the results.

The performance reached by our system has to be put into perspective with the complexity of the problem we are addressing (high number of classes with a critical inter-class similarity). It is however very satisfying as it constitutes an improvement compared to a method that was competitive enough to reach state-of-the-art performance [10]. The added value provided by the explicit shape description constitute an important extension, that less-dedicated identification methods cannot share. It makes it possible to consider interactive educational applications, enriching the primary goal of a good identification performance.

V. CONCLUSION

Following previous work on leaf shape description, a new sub-classifier strategy is proposed on leaf shape vocabulary to classify leaf species from the different shape criteria used in botany. The developed sub-classifiers exploit the second most voted of Random Forest in order to get rid of the ambiguity on leaf feature and enhance the classification accuracy. The outputs of the sub-classifiers are then used as inputs of a decision function that enables to predict the leaf species according to the shape properties, based on the floristic references describing species. The use of the evidential theory in this fusion process provides a significant enhancement in terms of species classification accuracy.

Some issues remain to perform sub-classification on the whole set of existing shapes, due to a lack of unambiguous data. Furthermore, the morphological similarity between species that may share the same combination of properties in botanical references is a source of confusion for the classification of a leaf. We consider using external information such as geographical data to help reducing this ambiguity between species.

The proposed system is however of great interest in the context of an educational mobile application **Folia**, which could be downloaded on AppStore. Its ability to transmit a vocabulary as well as an identification skill is a way to help an aspiring botanist user to make his way into the fascinating, and otherwise difficult to reach, world of plants.

REFERENCES

[1] P. Belhumeur, D. Chen, S. Feiner, D. Jacobs, W. Kress, H. Ling, I. Lopez, R. Ra-mamoorthi, S. Sheorey, S. White, and L. Zhang. “Searching the world's herbaria: A system for visual identication of plant species”, In *European Conference on Computer Vision*, 2008

[2] N. Kumar, P. Belhumeur, A. Biswas, D. Jacobs, I. Kress, W. Lopez, and J. Soares. “Leafsnap : A computer vision system for automatic plant species identification”, In *European Conference on Computer Vision*, pages 502–516, 2012.

[3] H. Goëau, P. Bonnet, A. Joly, D. Barthelemy, N. Boujemaa, and J.-F. Molino, “The imageclef 2012 plant image identification task,” in *ImageCLEF 2012Working Notes*, 2012.

[4] G. Cerutti, L. Tougne, J. Mille, A. Vacavant, and D. Coquin. "Understanding leaves in natural images - A model-based approach for tree species identification", In *Computer Vision and Image Understanding*, 117 (10): 1482-1501, 2013.

[5] X.F. Wang, D.S. Huang, J.X. Du, X. Huan, and L. Heutte. “Classification of plant leaf images with complicated background”, *Applied Mathematics and Computation*, 205(2):916-926, 2008.

[6] F. Mokhtarian and S. Abbasi. “Matching shapes with self-intersections : Application to leaf classification”, *IEEE Transactions on Image Processing*, 13(5): 653–661, 2004

[7] C. Caballero and M. Carmen Aranda. “Plant species identification using leaf image retrieval”, In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR 2010, pages 327-334, 2010.

[8] H. Coste. “Flore descriptive et illustrée de la France de la Corse et des contrées limitrophes”. Librairie de l’Académie impériale de Médecine, Paris, 1906.

[9] J.C. Rameau, D. Mansion, G. Dumé, J. Timbal, A. Lecointe, P. Dupont, and R. Keller. “Flore forestière française : Guide écologique illustré”, Institut pour le Développement Forestier, 1989.

[10] G. Cerutti, V. Antoine, L. Tougne, J. Mille, L. Valet, D. Coquin, A. Vacavant. ReVeS Participation “Tree Species Classification using Random Forests and Botanical Features”, *Conference and Labs of the Evaluation Forum, Rome*, 2012.

[11] L. Breiman. “Random forests”, *Machine learning*, 45(1):5-32, 2001.

[12] T. Denoeux. “The cautious rule of combination for belief functions and some extensions”, *Proceeding of the 9th International Conference on Information Fusion*, Florence, Italy, 2006.