Int J Comput Vis
DOI 10.1007/s11263-014-0743-3

Confidence Sets for Fine-Grained Categorization and Plant Species Identification

Asma Rejeb Sfar · Nozha Boujemaa · Donald Geman

Received: 1 January 2014 / Accepted: 20 June 2014

Abstract We present a new hierarchical strategy for fine-grained categorization. Standard, fully automated systems report a single estimate of the category, or perhaps a ranked list, but have non-neglible error rates for most realistic scenarios, which limits their utility. Instead, we propose a semi-automated system which outputs a it *confidence set* (CS)—a variable-length list of categories which contains the true one with high probability (e.g., a 99 % CS). Performance is then measured by the expected size of the CS, reflecting the effort required for final identification by the user. The implementation is based on a hierarchical clustering of the full set of categories. This tree of subsets provides a graded family of candidate CS's containing visually similar categories. There is also a learned discriminant score for deciding between each subset and all others combined. Selection of the CS is based on the joint score likelihood under a Bayesian network model. We apply this method to determining the species of a plant from an image of a leaf against either a uniform or natural background. Extensive experiments are reported. We obtain superior results relative to existing methods for point estimates for scanned leaves and report the first useful results for natural images at the expense of asking the user to initialize the process by identifying specific landmarks.

Keywords Fine-grained categorization · Hierarchical representation · Confidence set · Plant identification · Semi-automated

Communicated by Derek Hoiem.

Electronic supplementary material The online version of this article (doi:10.1007/s11263-014-0743-3) contains supplementary material, which is available to authorized users.

A. Rejeb Sfar (✉) · N. Boujemaa
INRIA Saclay, Palaiseau, France
e-mail: asma.rejeb_sfar@inria.fr

N. Boujemaa
e-mail: nozha.boujemaa@inria.fr

D. Geman
Johns Hopkins University, Baltimore, MD, USA
e-mail: geman@jhu.edu

1 Introduction

We study fine-grained categorization, the task of distinguishing among sub-categories of a more basic category, such as an object or shape class, focusing on identifying botanical species from leaf images. Whereas people can usually immediately recognize instances from basic categories (trees, dogs, etc.), fine-grained categories (e.g., species of plants, breeds of dogs) are usually recognized only by experts. The difficulty arises because taxonomic categories often have very fine differences which are hard to notice for the common eye. Figure 1 shows examples from three different leaf species which are evidently visually very similar in appearance. Another source of difficulty is the large variability in shape, color and texture within leaves of the same species, as well as changes due to viewpoint. For instance, leaves may exhibit different appearances due to local context, such as location and climatic conditions, as shown in Fig. 2 (additional examples can be found in Figures S1–S7). They may also vary in form and size even along a single stem as they develop (known as *leaf heteroblasty*); see Figure S8. Finally, there may be a great many biologically distinct fine-grained categories, e.g., about 300 breeds of dogs, and over 10,000 species of birds, 200,000 species of plants and 6,000,000 species of insects are currently known.

Due to these challenges, identifying plant species can be onerous and time-consuming even for skilled taxonomists,

Published online: 05 August 2014

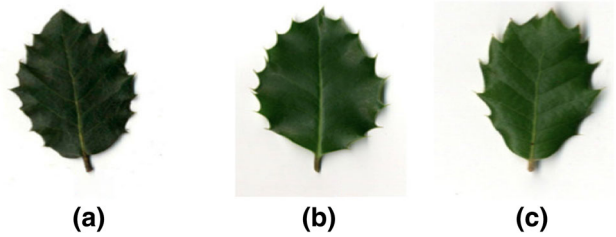

Fig. 1 High visual similarity among species. Three different species are displayed: **a** *Quercus ilex*. **b** *Ilex aquifolium*. **c** *Quercus coccifera*

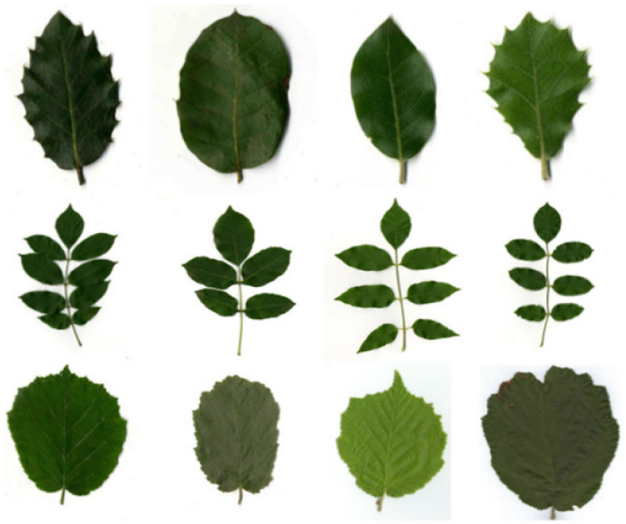

Fig. 2 Large variability within the same species. Displayed are examples from *Quercus ilex* (*top row*), *Fraxinus ornus* (*middle row*) and *Corylus avellana* (*bottom row*)

and nearly impossible for novices. Generally, the situation is the same in other domains of fine-grained categorization, and raises the question of what extent of semi-automation is required to provide useful results. In particular, how can we minimize human intervention while ensuring near-perfect sensitivity, i.e., assuring that the true category is among the ones reported by the system? To this end, we employ the user, with the goal of achieving something sensible between the two extremes of an inaccurate but fully-automated identification and a very accurate but fully-manual identification.

The *baseline* scenario is the standard one with no human intervention: given an image of a leaf, usually scanned against a flat background, the system automatically provides a single estimate of the true species. Even with scanned leaves, the utility of this approach is questionable due to relatively high error rates on large databases which contain very similar species and display high variability within the same species. This motivates the design of semi-automated systems. One natural possibility is to envision human participation at the end of the process in the sense of final disambiguation. That is, given a test image, instead of providing a single estimate,

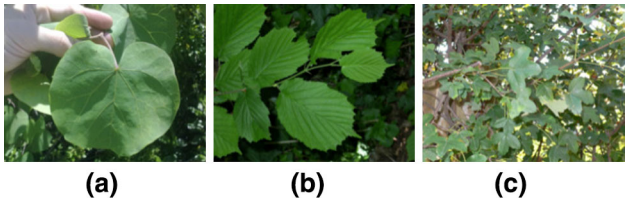

Fig. 3 Examples of unconstrained leaf photographs. One can photograph **a** a picked leaf, **b** a branch or **c** a foliage

the system returns a set of candidate species, *but constrained to contain the true species with high probability*, and the degree of human effort needed is measured by the average size of the set of candidates. In analogy with classical parameter (and Bayesian) estimation in Statistics, we refer to such a pruned list as a *confidence set* or just CS, and as a P% CS when the confidence level is $P/100$.

The degree of the user intervention is then data-dependent, and can range from non-existent (when the CS is a singleton) to significant in ambiguous situations. More generally, for example for natural photos of plants, the degree of interaction between the user and the system may extend to having the user play an important role at both the beginning and the end of the process, with the algorithm in between, in order to still guarantee a near-perfect result. This idea is introduced in Sect. 4.6.2, especially for identifying species of natural, cluttered leaf images (see Fig. 3) without using segmentation algorithms.

The model for generating the CS is generic and has four key ingredients.

- **Hierarchical Representation of Categories:** A distinguished family of subsets of categories of varying sizes indexed by the nodes of a binary tree. Learning is controlled by restricting the confidence sets to those in the hierarchy.
- **Node Discriminant Scores:** A score or test statistic for each node of the hierarchy for the binary classification problem of discriminating categories at the node from all others.
- **Statistical Model:** The joint probability distribution of the true category and the set of scores (see Sect. 4.5). In the model here, the conditional distribution of scores given category is a Gaussian Bayesian network.
- **CS Selection Algorithm:** A procedure for selecting the node of minimal size subject to the constraint of containing the true species with a given confidence level. Performance is measured by the expected size of the CS (Fig. 4).

The main contribution of this work is twofold: (i) the statistical framework above; (ii) an implementation for identifying the species of a plant from an image of an extracted leaf,

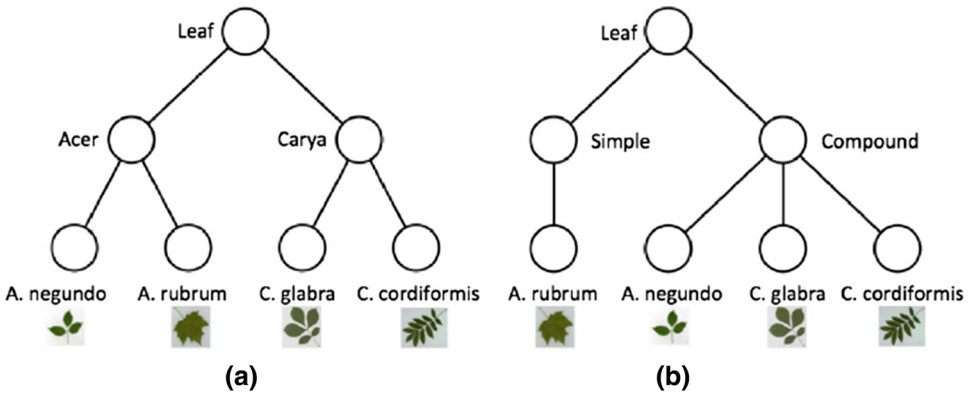

Fig. 4 Simple hierarchical representations of four species. **a** A semantic hierarchy: the second level represents leaf genera and the third level the species. **b** A hierarchy based on morphological leaf characteristics: the second level represents the leaf type (simple or compound) and the third level the species. A thumbnail from each species is displayed

illustrating the various levels of semi-automation. The implementation exploits domain-specific knowledge about landmarks and taxonomy to automatically build the hierarchical representation of species based on purely foliar characteristics. As indicated earlier, we also introduce different identification scenarios and tackle the problem of the cluttered leaf images without using segmentation algorithms. State-of-the-art results are obtained in all cases in which comparisons with previous work are possible. In particular, we achieve more than 90% of accuracy while returning less than two estimates (i.e., $|\hat{C}| = 2$) in average, on different challenging scanned leaf datasets and outperform all previous work on image-CLEF2011[1] cluttered photo category, including those who made use of a manual segmentation; see Sect. 5.2.

The rest of the paper is organized as follows: after describing the related work in Sect. 2, we describe the mathematical framework for defining and learning confidence sets in Sect. 3. The application to plant identification occupies the remainder of the paper: the representation of leaves in Sect. 4.1, the features in Sect. 4.2, the hierarchy of species in Sect. 4.3, the local scores in Sect. 4.4 and the probabilistic model in Sect. 4.5. In Sect. 4.6, we describe the different semi-automated identification scenarios and experiments are reported in Sects. 5 and 6. Finally, we draw some conclusions in Sect 7.

2 Related Work

Our work is related to existing work on fine-grained categorization (Nilsback and Zisserman 2006; Larios et al. 2008; Martínez-Muñoz et al. 2009; Wah et al. 2011; Duan et al. 2012; Yao et al. 2012; Cope et al. 2012; Yang et al. 2012; Zhang et al. 2012; Liu et al. 2012), especially work on plant species identification from leaf images (Belhumeur et al. 2008; Cope et al. 2012; Kumar et al. 2012; Goëau et al. 2011, 2012). But our approach also relates to confidence sets, hierarchical classification and classification with class-selective rejection.

Fine-grained categorization. Different recognition systems and object representations have been introduced or adapted for fine-grained categorization. The main previous scenario was to provide the user with a single estimate; examples include Du et al. (2005); Felzenszwalb and Schwartz (2007); Wu et al. (2007); Larios et al. (2008); Wu and Rehg (2008); Martínez-Muñoz et al. (2009); Branson et al. (2010); Angelova and Zhu (2013). Other studies chose to report the k most similar classes to improve accuracy. Usually, k ranges from five to twenty (Belhumeur et al. (2008); Kumar et al. (2012); Liu et al. (2012); Rejeb Sfar et al. (2013b)).

Several shape-based approaches, including boundary analyzes, have been used, especially for leaves (Ling and Jacobs (2007); Felzenszwalb and Schwartz (2007); Belhumeur et al. (2008); Caballero and Aranda (2010); Kumar et al. (2012); Mouine et al. (2013)). Often, performance is sensitive to the quality of the contour resulting from a segmentation process, which naturally complicates distinguishing between categories with very similar shapes. Other methods adapt systems for detecting instances of generic object classes (Lazebnik et al. 2006; Wang et al. 2010) by encoding an image as a bag of discrete visual codewords and basing classification on histograms of codeword occurrences; examples include Nilsback and Zisserman (2006), Larios et al. (2008). Again, however, the distinctions among fine-grained categories are sometimes too refined to be captured by variations in bags of visual words.

To account for such distinctions, an increasing number of studies make use of human input in the identification task. In Wah et al. (2011), an interactive system is proposed wherein users click on bird parts and answer questions about attributes (e.g., "white belly", "red-orange beak", "sharp crown"). Farrell et al. (2011b) and Zhang et al. (2012) use annotated data (e.g., key points and object parts) by experts to exploit *poselet* classifiers (Bourdev and Malik 2009) and build fine-grained models. In Rejeb Sfar et al. (2013b), we described scanned leaves and orchid flowers using *vantage points* based

[1] http://www.imageclef.org/2011/Plants

on botanical knowledge, and dedicated features to permit disambiguation between similar species. In other recent work (Deng et al. 2013), an online game *Bubbles* was introduced to reveal discriminative features humans use for bird identification. Our work on leaves is somewhat similar in that we study interactive scenarios using domain knowledge about botanical landmarks. Our feature extraction follows Rejeb Sfar et al. (2013b), but our classification framework is new.

Also, it should be emphasized here that most of previous work on leaves, including those mentioned above, use leaf images with uniform backgrounds (e.g., scans or photographs on white backgrounds). Only few of them addressed the problem of identifying leaves on cluttered backgrounds which is more likely to be the real-world scenario. To tackle this problem, a manual or an interactive segmentation process was generally designed. Obviously, isolating green leaves in an equally green environment seems like an other more difficult issue. Teng et al. (2009) proposed to recover the 3D position of a leaf from different cluttered images with close viewpoints. Then they performed a 2D/3D joint segmentation using 3D distances and color similarity. In Wang et al. (2008), an automatic marker-controlled watershed segmentation method is combined with pre-segmentation and morphological operation to segment leaf images with cluttered background based on the prior shape information. In the case of weed leaves, deformable templates have been used in Manh et al. (2001) to segment one single species *Setaria viridis*, providing promising results. Casanova et al. achieved the best results on natural leaf photo classification either at ImageClef2011 or ImageClef2012 plant identification tasks (Goëau et al. 2011, 2012). At both tasks, they proposed a shape boundary analysis based on a prior leaf segmentation. To this end, a manual segmentation was performed at the first task (Casanova et al. 2011) while a semi-automatic segmentation was performed at the second task (Casanova et al. 2012). In this work, we will show the efficiency of our algorithm on unconstrained photographs of leaves without any segmentation process.

Confidence intervals and sets in statistics. In classical (frequentist) statistics, a confidence interval CI (Neyman 1937) is a data-dependent interval estimate of a single population parameter. For example, the CI provides an indication of the precision of a point estimate such as maximum likelihood. Precision corresponds to the length of the CI and the confidence level can be interpreted as the fraction of times in repeated experiments that this random interval would contain the true parameter. A confidence set (CS) refers to an extension of confidence intervals to a multidimensional parameter (Cook 2005). Bayesian CI's and CS's (Lee 1989) extend these notions to Bayesian statistics wherein a prior distribution over parameters combined with a data model leads to a posterior distribution over parameters given the observations, interpreting the confidence level as a posterior probability. In analogy with these classical tools, we propose a semi-automated system which outputs a CS, a variable-length list of categories which contains the true one with high probability, rather than providing a point estimate (or ranking all candidates).

Hierarchical search. Hierarchy is a powerful organizing principle for both representation and search (Li and Perona 2005; Burl and Perona 1998; Fan and Geman 2004; Jr and Freitas 2011). The idea is to decompose the original problem into more tractable sub-problems sharing more homogeneous properties. One monolithic classifier could be then replaced by a hierarchy of classifiers which gather increasingly detailed information about the object under investigation. Many real-world classification problems, are naturally cast as hierarchical classification problems, where the classes to be predicted are organized into a class hierarchy, typically a tree or a Direct Acyclic Graph (DAG). Many of them utilize semantic class hierarchy, including the sharing of training examples across semantically similar categories (Fergus et al. 2010) or combining information from different levels of the semantic hierarchy (Zweig and Weinshall 2007). Deng et al. (2010) consider exploiting the semantic hierarchy in the context of more than 10, 000 categories.

In the fine-grained field, only few previous work take advantage of the hierarchical structure for identification tasks. To the best of our knowledge, only the natural semantic hierarchy (based on taxonomic groups, e.g., family and genus) were used; examples include those defined in Farrell et al. (2001a) or in Rejeb Sfar et al. (2013b). Using such a hierarchy needs specialized domain knowledge about species and taxonomy. Rather than using pre-defined taxonomic groups, which are defined according to both shared physical and genetic characteristics, we propose to consider purely visual characteristics to automatically build the hierarchy, using an agglomerative clustering on training data.

Class-selective rejection. Class-selective rejection (Ha 1997; Grall-Maës and Beauseroy 2009; Coz et al. 2009) is an extension of basic simple rejection (Yuan and Wegkamp 2010; El-Yaniv and Wiener 2010) in the multi-class case. That is, when an input pattern cannot be reliably assigned to one of the pre-defined classes in a multi-class problem, it is assigned to a subset of classes that are most likely to fit the pattern, instead of simple rejection. Selecting the most promising classes allows to reduce the error rate and to propose a reduced set to another classifier or an expert, which is of great interest in many decision making systems. Examples of class-selective rejection rules include those defined in Gupta 1965; Ha 1997; Horiuchi 1998. The simplest and the most used rule is the *top-n ranking*, in which n takes its values between one and the total number of classes considered. Another popular one is the *constant risk* rule (Gupta (1965)) which consists of selecting, for each pattern, the minimum number of best classes so that the accumulated

Table 1 Notation

I: an image
$\mathcal{Y}$: complete set of categories
$Y = Y(I)$: true category of I
$\mathcal{T}$: a binary tree with nodes $t \in \mathcal{T}$
C_t: categories of $\mathcal{Y}$ associated with t
X_t: discriminant score for testing $Y \in C_t$ vs. $Y \notin C_t$
$\mathbf{X}$: set of scores $\{X_t, t \in \mathcal{T}\}$
$\widehat{C}(\mathbf{X}) \in \{C_t\}$: confidence set
$p = P(Y \in \widehat{C}(\mathbf{X}))$: confidence level

posterior probability exceeds a pre-defined threshold. Ha (1997) defined a new optimality criterion to be the best trade-off between error rate and average number of classes. An *optimum* class-selective rejection rule was then obtained by solving a discrete convex minimization problem. Grall-Maës and Beauseroy (2009) addressed the problem of multi-class decision with class-selective rejection and performance constraints. The problem was defined using three kind of criteria: the label sets, the performance constraints, and the average expected loss. More recently, Deng et al. (2012) connected class-selective rejection with hierarchical classification, to restrict the subset of selected classes to internal nodes of a predefined hierarchy. In our work, we also focus on providing the best subset of classes using a predefined hierarchy but within a novel framework based on the notion of confidence sets in statistics (Table 1).

3 Methodology

Let $\mathcal{Y}$ denote the complete set of (fine-grained) categories and let $Y(I)$ denote the true category of image I. Our task is to predict Y. But rather than provide a single estimate we will provide a *confidence set* $\widehat{C} \subset \mathcal{Y}$ which depends on I and such that $Y \in \widehat{C}$ with high probability, say $P(Y \in \widehat{C}) \geq 1 - \epsilon$. Performance is then measured by the expected size of $\widehat{C}$.

One straightforward way to generate a CS is model-based: delineate a feature vector $Z = Z(I)$ and *model* for the joint distribution $p(z, c)$ of Z and Y. Provided with this, and given an image I (and hence z) the natural recipe for assembling $\widehat{C}(z)$ would be to compute the posterior distribution $p(c|z)$ over categories and aggregate the masses starting from the largest one, say $p(c_1|z) \geq p(c_2|z) \geq \cdots$, until the cumulative probability passes $1 - \epsilon$. That is, $\widehat{C} = \{c_1, ..., c_k\}$ where $p(c_1|z) + \cdots + p(c_{k-1}|z) < 1 - \epsilon$ and $p(c_1|z) + \cdots + p(c_k|z) \geq 1 - \epsilon$. In principle, the CS can then be any subset of categories. We propose a different strategy which is anchored by a hierarchical representation of $\mathcal{Y}$ and which drastically reduces the space of candidate CS's.

3.1 Hierarchical Strategy

Suppose we are provided with a recursive partitioning of $\mathcal{Y}$ indexed by the nodes of a binary tree $\mathcal{T}$. The hierarchy will serve as a platform for defining features and for selecting confidence sets. (The construction for our application to identifying plant species is based on hierarchical clustering of training data and is described in Sect. 4.3.) More specifically, each node $t \in \mathcal{T}$ is identified with a non-empty subset of categories $C_t \subset \mathcal{Y}$. The root node contains all categories and the terminal nodes contain a single category. At each level of $\mathcal{T}$, the subsets partition $\mathcal{Y}$ and the partitioning is recursive in the sense that for every non-terminal node t, $C_t = C_{l(t)} \cup C_{r(t)}$, where $l(t)$ and $r(t)$ are the left and right children of node t. A hierarchy for leaves is shown in Fig. 7.

When visual similarity is the clustering principle which generates the hierarchy, this structure can provide a natural family of (visually) closely-related categories with diverse sizes. This argues for *restricting* $\widehat{C}$ to the subsets $\{C_t, t \in \mathcal{T}\}$. In the standard case of returning a single estimate, the selection is restricted to the terminal nodes which are individual species.

The hierarchy also serves another key role. The data for selecting $\widehat{C}$ is a discriminant function on $\mathcal{T}$, denoted $X_t, t \in \mathcal{T}$; here $X_t \in \mathbb{R}$ represents a "score" for distinguishing between the two hypotheses $Y \in C_t$ versus $Y \notin C_t$. The scores are functions of node-specific feature vectors and are learned from training data using a some machine learning methodology. In our application, we use SVMs trained on features which are of course dedicated to leaves (see Sects. 4.2 and 4.4), but the framework is classifier-independent in that any learning algorithm could be chosen to induce a local discriminant function from the training data; those data points associated with node t serve as positive examples and all others as negative examples. Since the choice of $\widehat{C}$ depends only on the scores $\mathbf{X} = \{X_t, t \in \mathcal{T}\}$, we will sometimes write $\widehat{C}(\mathbf{X})$ to emphasize the dependence on the data.

3.2 Statistical Model

In this framework, the modeling is naturally done at the level of $\mathbf{X}$ and Y, thereby integrating all the evidence from the node scores. Let $p(\mathbf{x}, c)$ be a model for the joint distribution $P(\mathbf{X} = \mathbf{x}, Y = c)$. In order to specify $p(\mathbf{x}, c)$ we fix a prior $p(c)$ over categories (usually uniform); hence the key ingredient is the conditional data distribution $p(\mathbf{x}|c)$, $c \in \mathcal{Y}$. (Note that the score at the root is meaningless since all categories belong to C_{root} and consequently this node can be ignored in what follows.) The components of $\mathbf{x}$ are real-valued and

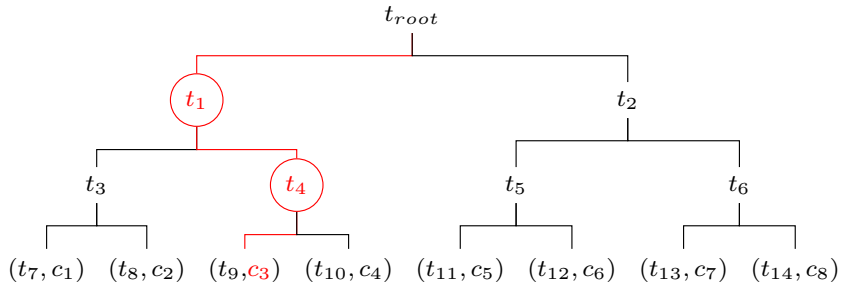

Fig. 5 Example of $\mathcal{T}$ illustrating the key objects for 8 categories $(c_1, ..., c_8)$. $\mathcal{T}$ contains 14 nodes (not counting the root t_{root}), labeled $(t_1, ..., t_{14})$. Associated with each node t is a set of categories C_t, e.g., $C_{t_1} = \{c_1, c_2, c_3, c_4\}$. Here, the true category is $Y = c_3$, $B(\mathbf{x}) = \{t_1, t_4\}$, (*red circles*) which are on the true path (*in red*). So, $T(\mathbf{x}) = t_4$ and $\widehat{C}(\mathbf{x}) = \{c_3, c_4\}$ (Color figure online)

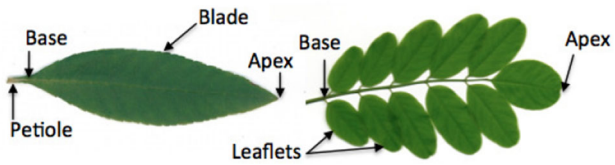

Fig. 6 The different leaf parts, including the leaf base and the leaf apex for both a simple (*on the left*) and a compound (*on the right*) leaf

indexed by the tree $\mathcal{T}$; hence the dimension of $\mathbf{x}$ is basically twice the number of categories. The model we use for $p(\mathbf{x}|c)$ in our application is a Bayesian network (BN) over Gaussian variables and will be described in detail in Sect. 4.5. In brief, the two children t_1 and t_2 of t_{root} serve as roots of the BN, which then has the form:

$$p(\mathbf{x}|c) = f(x_1|c) f(x_2|c) \prod_{t \in \mathcal{T} \setminus \{t_1, t_2\}} f_t(x_t|x_{t-}, c) \tag{1}$$

Here $t-$ denotes the parent of t; $f(x_1|c)$ and $f(x_2|c)$ are the marginal densities of scores X_{t_1}, X_{t_2} given $Y = c$, both assumed univariate normal; and $f_t(x_t|x_{t-}, c)$ is the conditional density of X_t given $\{X_{t-} = x_{t-}, Y = c\}$. Since we are assuming (X_t, X_{t-}) is bivariate Gaussian given $Y = c$, the form of the conditional density follows immediately. Again, the details for our application to plants, including parameter estimation, appear later in Sect. 4.5.

3.3 Constructing the Confidence Set

The first step in CS selection is to compute the posterior probabilities $P(Y \in C_t|\mathbf{X} = \mathbf{x})$ for each $t \in \mathcal{T}$. This is straightforward given our model:

$$P(Y \in C_t|\mathbf{X} = \mathbf{x}) = \sum_{c \in C_t} P(Y = c|\mathbf{X} = \mathbf{x}) \tag{2}$$

$$= \frac{\sum_{c \in C_t} p(\mathbf{x}|c)}{\sum_{c \in \mathcal{Y}} p(\mathbf{x}|c)} \tag{3}$$

Now define

$$B(\mathbf{x}) = \{t \in \mathcal{T} : P(Y \in C_t|\mathbf{X} = \mathbf{x}) \geq 1 - \epsilon\}.$$

Obviously we can assume $\epsilon < 0.5$; in practice, we take values such as 0.05 and 0.01. It is then easy to see that for every $\mathbf{x}$, the set of nodes $B(\mathbf{x})$ is a non-empty path in $\mathcal{T}$ originating at one of the two roots t_1, t_2 and generally terminating before a terminal node is reached. The natural definition of $\widehat{C}$ is then the *smallest* set C_t in the tree which satisfies the constraint. Specifically,

$$\widehat{C}(\mathbf{x}) \doteq C_{T(\mathbf{x})}, \quad T(\mathbf{x}) = \arg \min_{t \in B(\mathbf{x})} |C_t|.$$

Equivalently, $T(\mathbf{x})$ is the deepest node in $B(\mathbf{x})$. The corresponding *confidence level* for the given data is then

$$p(\mathbf{x}) = P(Y \in \widehat{C}(\mathbf{x})|\mathbf{X} = \mathbf{x})$$

and the average confidence level is

$$Ep(\mathbf{X}) = P(Y \in \widehat{C}(\mathbf{X})).$$

Given the definition of $B(\mathbf{x})$, it follows that $Ep(\mathbf{X}) \geq 1 - \epsilon$.

Figure 5 illustrates the concepts above for a simplified hierarchical structure $\mathcal{T}$ of 8 categories $(c_1, ..., c_8)$. Here $T(\mathbf{x}) = t_4$ is the deepest node in $B(\mathbf{x}) = \{t_1, t_4\}$, and the resulting confidence set is the $\widehat{C} = \{c_3, c_4\}$.

The efficiency of this algorithm will be demonstrated in a variety of experiments in Sect. 5.2, both in terms of comparing with other methods as well as generating high confidence sets.

3.4 Relationship to Non-Bayesian Confidence Sets

In classical (frequentist) statistics, there is no r.v. Y, only a family of probability distributions $\{p(\mathbf{x}|c)\}$ indexed by a parameter $c \in \mathcal{Y}$. A $100(1 - \epsilon)\%$ confidence set for the true parameter c^0 is a random set (i.e., data-dependent) which contains c^0 with probability $1 - \epsilon$. For a continuous real-valued parameter, an interval is often centered at a

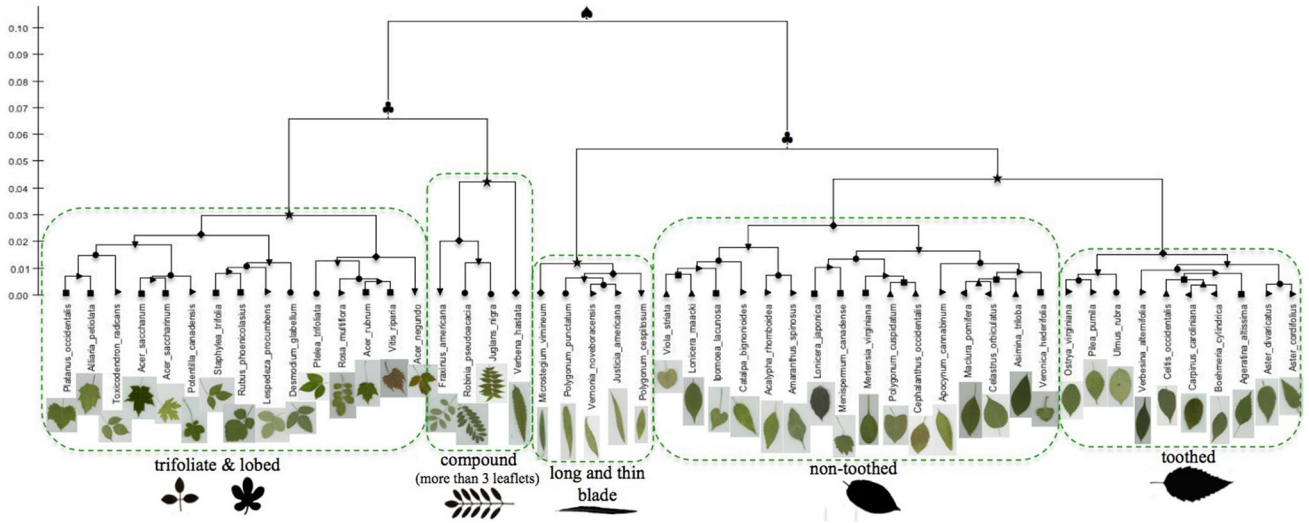

Fig. 7 A dendogram representing a hierarchical clustering of 50 species of Smithsonian leaves. Displayed are the nested groupings of species, similarity levels at which groupings change, and a thumbnail from each species. Many clusters match morphological classes

point estimate $\hat{c}$ such as the maximum likelihood estimator $\hat{c}_{ML} = \arg\sup_c p(\mathbf{x}|c)$.

Following this recipe we would begin with the maximum likelihood estimator $\hat{c}_{ML}$, which coincides with the MAP estimator $\arg\max_c P(Y = c|\mathbf{X} = \mathbf{x})$ in the Bayesian case when the prior is uniform. The tree provides a neighborhood structure: a natural way to "center" the CS at $\hat{c}_{ML}$ is to consider the subsets of categories along the path from $\hat{c}_{ML}$ to the root. However, given such a set $\widehat{C}(\mathbf{x})$ of categories containing $\hat{c}_{ML}(\mathbf{x})$, computing $P(c^0 \in \widehat{C}(\mathbf{X}))$ would require knowing the distribution of the ML estimator under c^0, which appears difficult. The Bayesian argument gives this in an *average* sense. (Note, however, that the CS constructed in the previous section does not necessarily contain the MAP estimator, but nearly always does in practice.)

4 Application to Plant Species

4.1 Leaf Representation

In botany, a leaf is defined as a colored, usually green, expansion growing from the side of a stem, in which the sap for the use of the plant is elaborated under the influence of light. It is one of the parts of a plant which collectively constitute its foliage. Usually, a leaf consists of a blade (i.e., the flat part of a leaf) supported upon a petiole (i.e., the small stalk situated at the lower part of the leaf that joins the blade to the stem), which, continued through the blade as the midrib, gives off woody ribs and veins that support the cellular texture.

A leaf is qualified as being "simple" if its blade is undivided; otherwise it is "compound" (i.e., divided into two or more leaflets); see Fig. 6. According to the leaf architecture manual of Ellis (2009), the internal shape of the blade is characterized by the presence of vascular tissue called veins, while the global shape can be divided into three main parts: (1) The basal part, usually the lower 25 % of the blade; the *base*, which is the point that joins the blade to the petiole, is situated at its center. (2) The apical part, usually the upper 25 % of the blade and centered by a sharp point called the *apex*. (3) The margin, which is the edge of the blade.

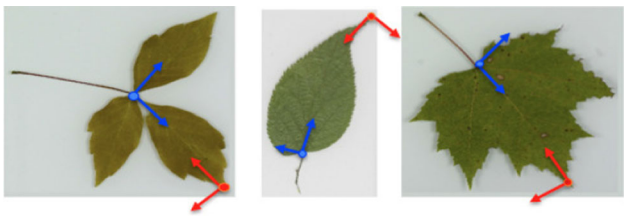

Fig. 8 Two local coordinate systems are used for vantage feature frames, one centered on the leaf apex (*red one*) and one on the leaf base (*blue one*) (Color figure online)

This botanical leaf decomposition gives an interesting alternative for efficient local representation of the leaf and is often used by botanists in the identification tasks. In the manual process, experts generally use the different foliar characters as *identification keys* which are examined sequentially and adaptively (Elpel 2004) to identify the plant species. In essence, one is posing and answering a series of questions about one or more attributes (e.g, shape, color, distinguished landmarks, internal structure) with the aim of narrowing down the set of possible species. Specific leaf parts are often examined by botanists for identification purposes. In computational vision, such a fine-grained

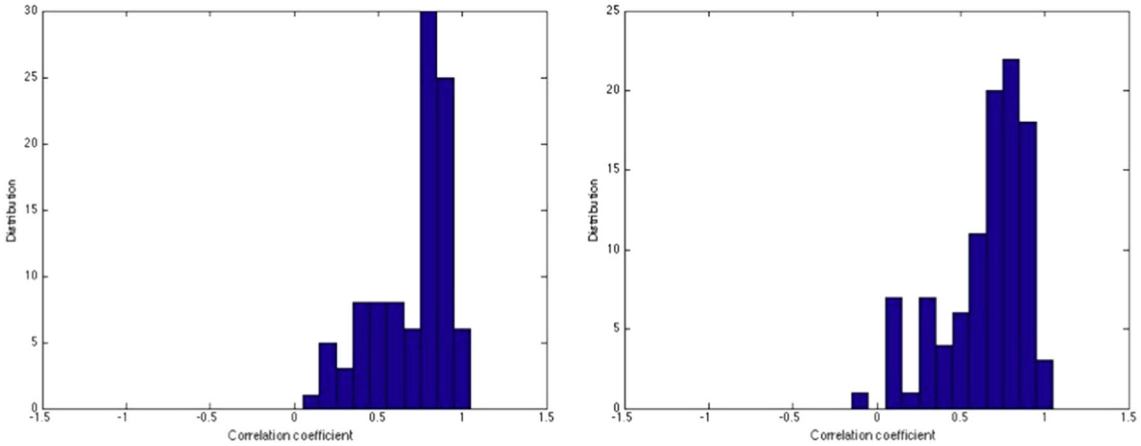

Fig. 9 Histograms of correlation coefficients between a hierarchical node t and its parent $t-$ given a fixed species

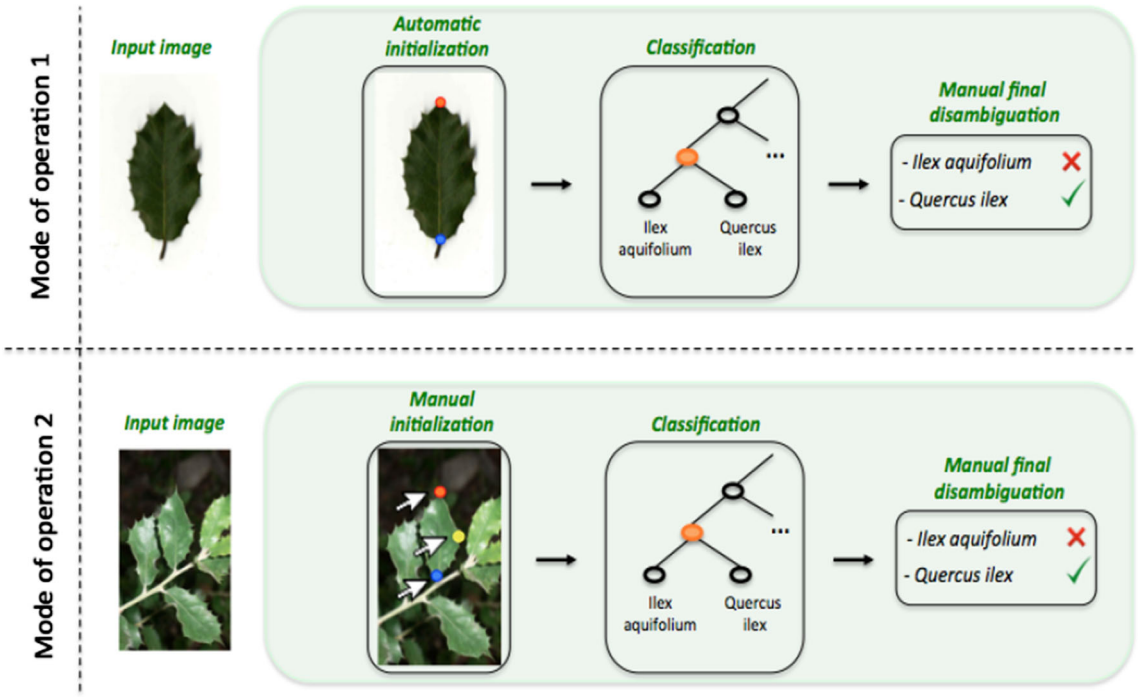

Fig. 10 Given an input leaf image, two modes of operation are proposed. The first mode (*top row*) is semi-automated identification with human intervention only at the end of the process. The second mode (*bottom row*) involves the user at both the beginning, for initialization, and the end, for final disambiguation

recognition conveniently *enables* part-based approaches, because the differences between subcategories are very fine and not noticeable from global, image-level features, and objects within the same basic-level category often have the "same" parts (Tversky and Hemenway 1984), allowing for easier comparison. Note that in basic-level categorization, this approach is more difficult, as there is no natural corresponding parts among instances of, for example, dogs, cars, and plants.

4.2 Feature Frames

In Rejeb Sfar et al. (2013b), we demonstrated that the *leaf apex* and the *leaf base* are very useful to automatically separate one species from another. In fact, such landmarks could be considered more like *vantage points* in that orientation plays a role as well. Naturally, species tend to have certain signature appearance properties and consequently what to look for in the neighborhood of these landmarks may be

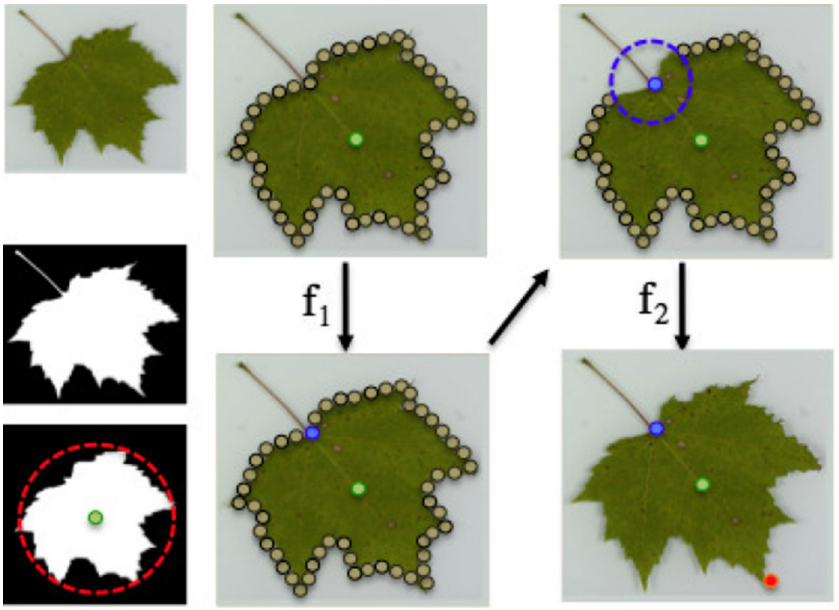

Fig. 11 A test leaf image is first segmented. Then the petiole is removed in order to compute the centroid (*green point*) as well as the approximate bounding circle of the leaf blade (*red dashed circle*). The base (*blue point*) and the apex (red point) are estimated using learned classifiers (f_1, f_2). The proposed locations for both landmarks are restricted to the boundary points. The neighborhood of the first landmark detected is excluded from the list of candidate points for the next detection (*blue dashed circle*) (Color figure online)

species-dependent. Put differently, the conditional distribution over any large family of generic local features may depend strongly on the species or groups of species. This aspect of the identification process was encoded in Rejeb Sfar et al. (2013b) by allowing the set of features associated with each landmark to depend on the species using the notion of *vantage feature frames*. Such frames also ensure that the local appearance properties are largely invariant to the orientation and scale of the object. As will be seen in Sect. 4.3, these frames are defined and used here within the context of a hierarchical representation of the data.

4.3 Hierarchy Construction

Hierarchical representation can open the door to exploring classification algorithms or cost metrics that do not penalize as much misclassification among very similar classes. For example, it may not be as problematic which fir tree it is as long as we do not confuse it with other non-related trees. Botanical species are naturally organized in a hierarchical taxonomy (family-genus-species). However, rather than using pre-defined taxonomic groups, which are defined according to both shared physical and genetic characteristics, we consider purely visual characteristics to automatically build the hierarchy using a hierarchical clustering on training data.

Hierarchical clustering is widely used; a useful review of the standard methods has been given in Jain et al. (1999). In particular, agglomerative procedures produce a series of partitions of the data; the first partition consists of single-member 'clusters'; the last consists of a single group containing all individuals. The variation here is based on domain knowledge about botanical species and landmarks, but the

Fig. 12 Examples of leaf photographs manually marked. For each image, displayed are the leaf base (*blue point*), the leaf apex (*red point*) and a third boundary point (*yellow point*). The approximate width (marked as *s* at each image) of the leaf is defined as the distance between the yellow point and the apex-base line (Color figure online)

principle is quite general: a tree-structured hierarchy is recursively constructed bottom-up by successively merging similar groups. We treat each species as a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all the clusters have been merged into a single cluster that contains all species. We use Ward's criterion (see Ward 1963) and the Euclidean norm. That is, at each step we define the dissimilarity between two clusters

$$dissim(r,s) = \frac{n_r * n_s}{n_r + n_s} \times ||\bar{X}_r - \bar{X}_s||^2$$

where r and s denote two specific clusters with sizes n_r and n_s, $\bar{X}_r$ and $\bar{X}_s$ denote the centers of gravity of the clusters and $||.||$ is the Euclidean norm.

Local features are used to compute the centers of gravity of the clusters. More specifically, texture and shape-based features were used to characterize each leaf image and were defined in two local coordinate systems, one centered on the leaf *base* and the other on the leaf *apex*. The motivation behind using such local coordinate systems is to *focus attention* around each landmark, which is the strategy reported by botanists as explained in Sect. 4.1.

The tree-structured hierarchy provides a useful summary of the data, i.e., an overview of the visual similarities and relationships between species based on both the basal and the apical parts. Figure 7 depicts a dendrogram (Fernández and Gómez 2008) that illustrates the nested grouping of the species produced by a hierarchical clustering on 50 botanical species. Note that many clusters obtained could be matched with morphological classes defined by botanists themselves. In particular, two large, natural clusters are formed at the first level of the hierarchy; one cluster consists of compound and lobed leaves (on the left) and another cluster of simple leaves (on the right). Species with lobed leaves (with 3 lobes) merge with those with trifoliate leaves (with 3 leaflets) while species with compound leaves containing more leaflets are grouped together. Also, toothed leaves are separated from non-toothed leaves. Such a hierarchical clustering could even help botanists to speed up the classification process of large amounts of newly collected leaves by suggesting coarse morphological categories.

4.4 Discriminant Function

The hierarchical representation of species will serve as a platform for the classification algorithm. To this end, we learn local discriminant functions at each node $t \in \mathcal{T}$. As previously mentioned (see Sect. 3.1), the framework is largely classifier-independent in that any learning algorithm could be chosen to induce such local functions from the training data. We have chosen to train an SVM classifier for $Y \in C_t$ versus $Y \notin C_t$ using *vantage feature frames*, with different positive and negative images at each node. The "score" X_t refers here to the SVM score.

Given a tree $\mathcal{T}$, a vantage feature frame $\mathcal{F}$ has two components. One, Θ, is geometric and the other, $\mathcal{Z}$, is appearance-based. The geometric component Θ is category-independent and simply a local coordinate system centered at a specific landmark. The appearance component is a family of pose-indexed features, one element of the family for each category: $\mathcal{Z} = \{\mathcal{Z}_1, \ldots, \mathcal{Z}_N\}$, where $\mathcal{Z}_t$ is the set of local features to compute in frame $\mathcal{F}$ for C_t and N the total number of hierarchical nodes. Obviously, to be useful the frame must be reliably detected and the features must be discriminating. In Rejeb Sfar et al. (2013b) algorithms are given for learning discriminating ones, detecting them online and pooling the features computed in these frames for identification purposes.

We refer to the origins of the frames as *vantage points* - special locations from which observing the leaf and in a particular direction can provide discriminating information about the species. Following the work in Rejeb Sfar et al. (2013b), we consider two *vantage points* for leaves: the *leaf base* and the *leaf apex* as shown in Fig. 8. Also, Hough, EOH, HSV and Fourier histograms are used as base features (more details of these global features can be found in Ferecatu (2005)) to construct the frames. Hence, given the pre-defined hierarchy $\mathcal{T}$, we select specific subset of features $\mathcal{Z}_t$ for each subset of species C_t and only these are used to train local classifiers. The reason for dedicated features is that there is so much variability in the presentation of leaves in the neighborhood of landmarks that some features are far more discriminating than others, and the discriminating ones can depend as well on the vantage point. For example, the discriminating features around the leaf base for estimating a particular group of species might be different from those around the apex for estimating another group.

4.5 Probabilistic Model

As explained in Sect. 3.3, in order to compute $P(Y \in C_t | \mathbf{X} = \mathbf{x})$ we require the joint conditional density $p(\mathbf{x}|c)$ of the scores $\mathbf{x} = (x_t, t \in \mathcal{T})$. (We assume the prior distribution $p(c)$ is uniform over species $c \in \mathcal{Y}$.) We use a Gaussian Bayesian network. The choice of a Gaussian model for individual SVM scores X_t is primarily motivated by simplicity; we have sufficient data to reliably estimate the mean and the variance and this approximation, although rough, works well in practice.

As for the dependency structure among the scores, whereas there are significant (conditional) correlations among many pairs of variables X_t, X_s given the species Y, clearly we must control the complexity of the joint distribution since we do not have sufficient data to reliably estimate all the order $|\mathcal{T}|^2$ parameters involved in a full multivariate Gaussian parameterization. The motivation for the Bayesian network is that the largest of the (absolute) correlations tend to be between parents and children; see Fig. 9. More details are shown in Figure S10. The underlying DAG (directed acyclic graph) is of course the tree $\mathcal{T}$ with arrows from parents to children. With this Gaussian Bayesian network we must estimate three parameters (mean, variance, correlation with parent) for each non-root node and two parameters (mean and variance) for nodes t_1 and t_2. To this end, part of the training data is set aside for parameter estimation in each experiment.

Consequently, the densities $f(x_1|c)$ and $f(x_2|c)$ in Eq. (1) are univariate normal. The densities

$f_t(x_t|x_{t-}, c)$ are obtained by recalling that if U_1, U_2 are jointly normal with means and standard deviations $\mu_1, \mu_2, \sigma_1, \sigma_2$ and correlation coefficient ρ, then $f(u_1|u_2)$ is normal with mean $\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2)$ and variance $(1-\rho^2)\sigma_1^2$. Hence

$$f_t(x_t|x_{t-}, c) = \frac{1}{\sigma_t^c\sqrt{2\pi(1-\rho_t^c)}} \times \exp\left\{-\frac{(x_t - \mu_t^c - \rho_t^c\frac{\sigma_t^c}{\sigma_{t-}^c}(x_{t-} - \mu_{t-}^c))^2}{2(1-\rho_t^{c2})\sigma_t^{c2}}\right\}$$

Fig. 13 Samples from the Swedish dataset. One image from each species is shown

Fig. 14 Samples from the Flavia dataset. One image from each species is shown

where the superscripts indicate the dependence on the species c. Computing $p(\mathbf{x}|c)$ and thus $P(Y \in C_t|\mathbf{X} = \mathbf{x})$ is then straightforward.

4.6 Scenarios for Species Identification

Different levels of interactive identification can be considered depending on whether the background is uniform and or natural (cluttered). The particular scenarios we consider are illustrated in Fig. 10.

4.6.1 Final Disambiguation

Given a leaf image, vantage feature frames are first automatically detected. Detecting the vantage feature frames refers to first estimating vantage points (landmarks) and then the orientation and scale of each frame. The orientation is determined by the centroid, which is directly computed from the raw image data after a segmentation process using the Otsu (1979) algorithm. The scale is taken to be the radius of the bounding circle as illustrated for leaves in Fig. 11.

The landmarks are detected using SVM classifiers. Since we are only using landmarks on the object boundaries (as determined by the segmentation process), we restrict the search to a sample of boundary points to minimize the computation. In addition, after detecting each landmark, we exclude the boundary points in its neighborhood from the list of candidates; see Fig. 11. More details about the detection of vantage feature frames can be found in Rejeb Sfar et al. (2013b).

Once the frames are detected, category-dependent features are extracted as described in Sect. 4.4. Then, given the scores $\mathbf{x}$ for the leaf image being processed, the Bayesian network model and fixing ϵ, we compute the sub-path $B(\mathbf{x})$ of $\mathcal{T}$ and finally provide $\widehat{C}(\mathbf{x})$ to the user. The type of user intervention will then depend on the needs and skills of the user. The novice user may simply accept $\widehat{C}$ as it stands or use reference material to narrow it down. A more skilled user may be able to identify it if it resides in the set or recognize that it does not. Of course, the smaller the confidence set, the more informative and useful it is. Several experiments in Sect. 5.2 will demonstrate the efficiency of such a scenario on leaf images with uniform background (e.g. scanned leaves).

4.6.2 Initialization

For a single leaf image with a cluttered background (Fig. 3a), automatic detection of vantage points requires a very efficient segmentation algorithm (robust to background noise and texture), which is not the case for the algorithm we use (Otsu) or any we are aware of. Also, it could be exceedingly difficult to automatically (and accurately) extract a single leaf boundary from a branch or foliage image; see Fig. 3b, c (additional examples can be found in Figure S9). Moreover, returning a $P\%CS$ is of little value in applications if either $|CS|$ is very large or $P \ll 100$. The minimal intervention we can imagine is asking the user to mark several landmarks; providing a faithful segmentation is another possibility but we are able to obtain good results without this level of intervention.

We ask the user to mark the two terminals of the main vein of the leaf, the base and the apex, as well as a third boundary point which will be used to approximate the width of the leaf (see Fig. 12). The centroid of the leaf is defined

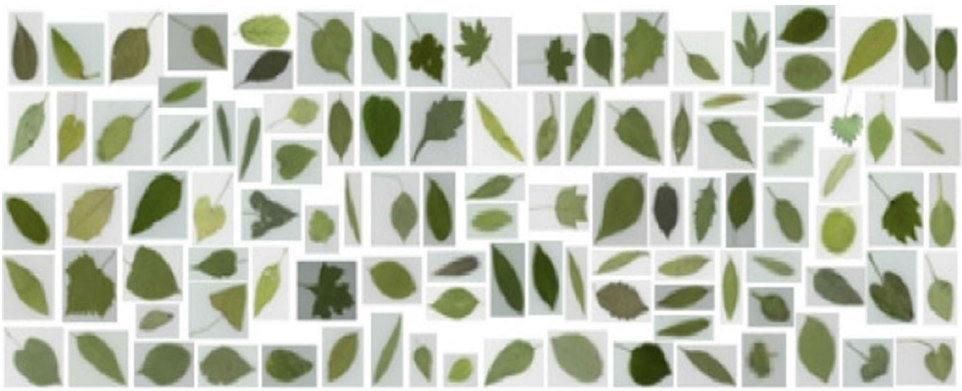

Fig. 15 Samples from the Smithsonian dataset. One image from each species is shown

as the mid-point of the apex-base line. Local features are then extracted in two coordinate systems, one centered on the base and the other on the apex as described in Sect 4.4. The same classification process as in the previous scenario Sect. 4.6.1 is used to provide the user with a CS. A summary of the scenarios is provided in 10.

5 Experiments

5.1 Datasets

We considered four challenging leaf datasets from different geographical areas. Three of them consist of images of single leaves on a white background. The last one consists of unconstrained photographs of leaves.

Swedish: This has 1125 scanned leaf images containing 75 images from each of 15 different Swedish plant species. This dataset was the first publicly available leaf data, introduced by Söderkvist (2001) for research. Although it contains relatively few varieties of species, we chose it in order to be able to compare our work with various approaches, including generic shape classification approaches such as Felzenszwalb and Schwartz (2007); Ling and Jacobs (2007); Wu and Rehg (2008) which were applied on leaves (Fig. 13)

Flavia: The Flavia dataset is composed of 1907 scans of leaves. It consists of 32 species with 50-60 observations in each species; see Fig. 14. It was introduced by Wu et al. (2007) and was used to evaluate retrieval systems but also some leaf classification algorithms (Wang et al. 2005; Du et al. 2005; Gu et al. 2005).

Smithsonian: This dataset has 5466 leaf images containing 148 different species from the Northeastern U.S area. The number of exemplars per species varies from 2 to 63. These images were provided by the Smithsonian botanical institution within the framework of the US National Herbarium. One particularity of these data is that the images present only simple leaves with various poses and orientations of leaves as well as different structures of basal and apical parts as shown in Fig. 15.

ImageCLEF2011: Used in the ImageCLEF2011 plant identification task[2], the complete leaf collection contains three categories of images: scans of leaves acquired using a flat-bed scanner, scan-like leaf images acquired using a digital camera and free natural photos. Here, we focus only on photos; see Fig. 16. This category has 1469 unconstrained photographs of leaves; 930 images for training and 539 test images. It was constructed through a citizen sciences initiative conducted by Telabotanica[3], a French social network of amateur and expert botanists (more details can be found in Goëau et al. (2011)). As a result, the task it represents is quite close to the conditions encountered in a real-world application. Each image can represent either a single leaf, a branch or a foliage as shown in Fig. 3. In particular, the training leaves were collected from 269 plants and those of the test set from 99 other different plants. Not all the species were considered for testing. Only samples from 26 species were available for testing using 40 training species.

5.2 Results and Analysis

For ease of notation, we label three scenarios:

- **CS0:** The confidence set CS is generated by ranking the posterior probabilities and accumulating species until the total mass exceeds $1-\epsilon$; see Sect. 3.
- **CS1:** The process is automatically initialized and the $\widehat{C}$ used is described in Sect. 3.3. Note that the size of $\widehat{C}$ is necessarily at least as large as the CS returned by **CS0**.
- **CS2:** The process is manually initialized and the $\widehat{C}$ used is described in Sect. 3.3.

We will also refer to two *baseline cases* where the confidence set is restricted to a singleton.

[2] http://www.imageclef.org/2011/Plants

[3] http://www.tela-botanica.org

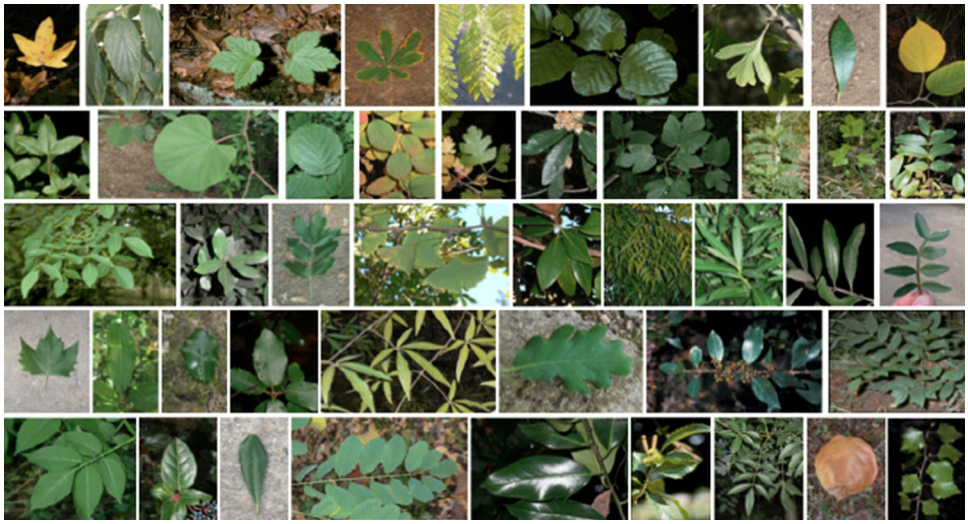

Fig. 16 Samples from the ImageCLEF2011 leaf photos. One image from each species is shown

– **MAP:** Only the species with the highest posterior mass is returned.
– **F-SVM:** Only the species with the highest SVM score is returned, i.e., a "flat classifier" using one-vs-all SVM's.

To evaluate the performance of the proposed framework, we first provide the rate on the holdout test data at which the true species appears among the list of estimates, and second analyze the size of the response.

In order to be able to compare our performance with that of other methods, we will also adopt other evaluation metrics: (1) the accuracy rate among the top k estimates for the Swedish, Flavia, and Smithsonian datasets, (2) the evaluation metric[4] used for the ImageCLEF2011 plant identification task, for the ImageCLEF photo subset, which allows us to compare our performance with that of all the task participants. Such a metric refers to a *normalized classification rate* evaluated on the first species returned for each test image while taking into account the individual plant and the author (more details about the metric definition and the participants can be found in Goëau et al. (2011)). In all the following experiments , we use $\epsilon = 0.01$.

Swedish Data: Following all previous work on this dataset, we randomly select 25 training images from each species and test the remaining images in order to evaluate our performance. One third of the training set was used to estimate the Bayesian network parameters. We use this dataset to evaluate CS1.

As shown in Table 2, the correct species belongs to the CS returned 99.5% of the time while applying the CS1 scenario and 99.2% while applying CS0. Figure 17 illustrates the distribution of the size of the CS in both cases. The average size is less than 1.5; see Table 2. Both CS1 and CS0 do achieve near-perfect results while returning a single estimate at most of the time.

Table 2 Comparison between CS0 and CS1 on the Swedish dataset

Scenario	Accuracy (%)	Average size of the response
CS0	99.2	1.2
CS1	99.5	1.3

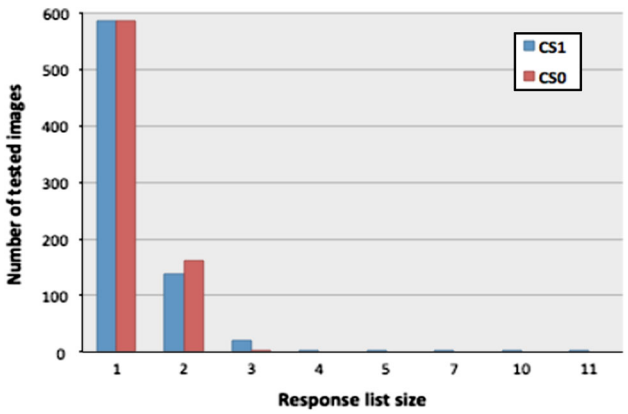

Fig. 17 The distribution of $|\widehat{C}|$, the size of the CS returned, for both methods of constructing the CS when testing on the Swedish leaves

By construction, both strategies are equivalent when only one estimate is returned. Note that one advantage of the proposed approach compared with CS0 is that CS1 provides visually coherent sets for the user; see Fig. 18. An additional advantage will be demonstrated on the Smithsonian and the ImageClef data.

In order to be able to compare CS1 with previous work using the same evaluation framework, we use the two baseline strategies for providing a single estimate. As seen in Table 3, we achieve the best performance (98.7% accuracy) while considering the species in $\widehat{C}$ with the highest posterior

[4] http://www.imageclef.org/2011/Plants

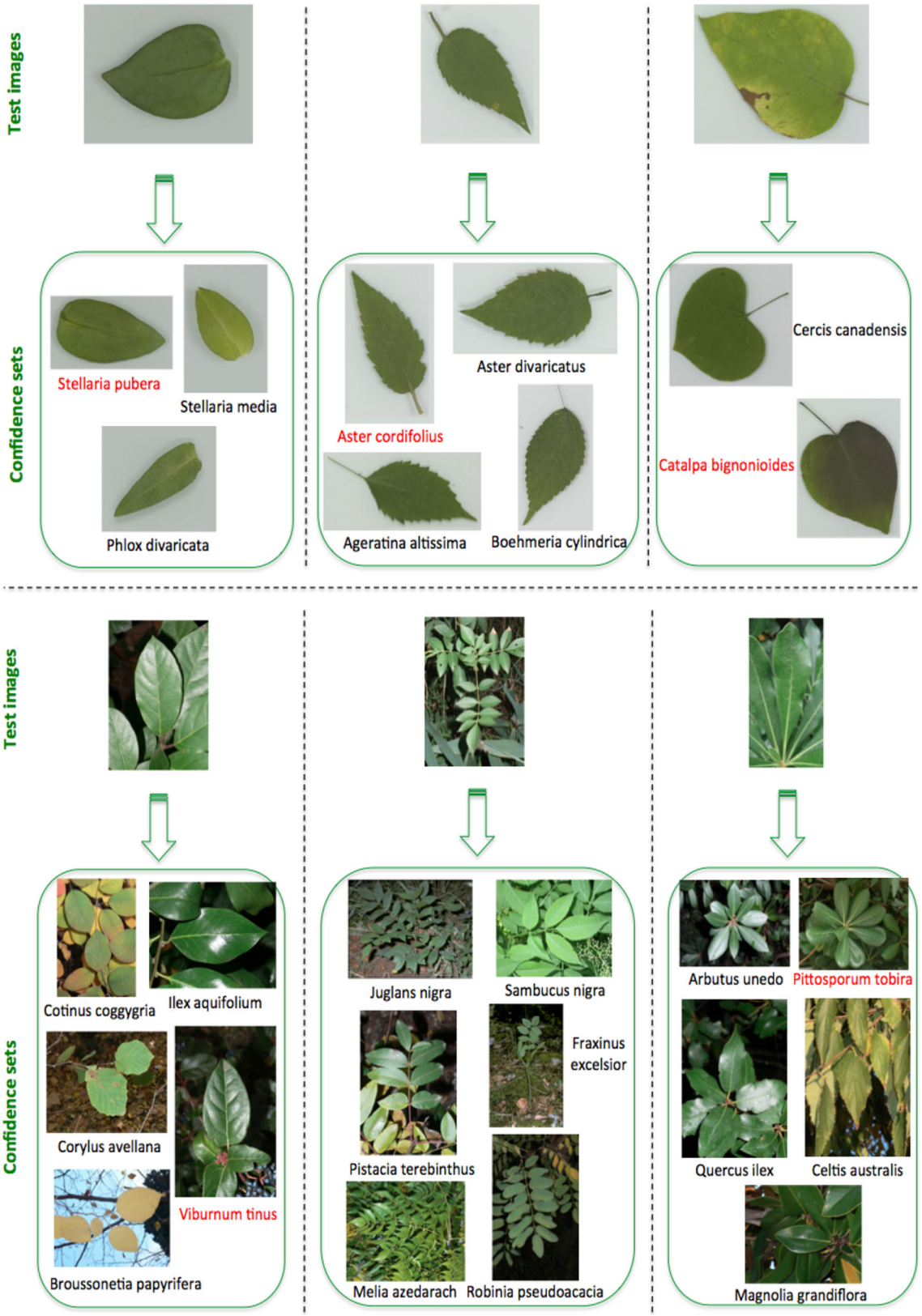

Fig. 18 A sample of test leaf images (scans and photos) with non-singleton confidence sets (CS) of species. For each CS, a training image from each species is displayed. For each test image, the red species is the true one. The CS are generally visually coherent

Table 3 Different results on the Swedish data while considering a single estimate (top-1)

Methods	Accuracy (%)
CS1 - *MAP*	**98.7**
IdKeys [(Rejeb Sfar et al. (2013a))	98.4
sPACT (Wu and Rehg (2008))	97.9
TSLA (Mouine et al. (2013))	96.5
Shape-Tree (Felzenszwalb and Schwartz (2007))	96.3
SPTC+DP (Ling and Jacobs (2007))	95.3
IDSC+DP (Ling and Jacobs (2007))	94.1
F-SVM	93.3
SC+DP (Ling and Jacobs (2007))	88.1
Söderkvist (Söderkvist (2001))	82.4

The number in bold indicates the best performance

Table 4 Comparison between CS0 and CS1 on the Flavia dataset

Scenario	Accuracy (%)	Average size of the response
CS0	97.1	1.6
CS1	98.1	2.2

mass. However, the one-vs-all SVM classifier yields only 93.3% accuracy; see Table 3. Using a hierarchical model is clearly of value for this dataset.

Flavia Data: Following Wu et al. (2007), we used 10 leaves from each of 32 species to evaluate the performance of our approach, so that a total of 320 leaves are used for testing the algorithms and the remaining leaves for training and estimating the Gaussian Bayesian network. As with the Swedish leaves, we achieve near-perfect results: using CS1, the average size of the CS is 2.2 for an accuracy rate of 98.1% as shown in Table 4. Figure 19 shows the distribution of $\widehat{C}$ while applying both CS0 and CS1. We have $|\widehat{C}| = 1$ 87.5% of the time. From Table 4, we see that CS1 is slightly more accurate than CS0, but at the expense of providing a slightly larger CS on average.

Finally, we use the same evaluation framework as in Wu et al. (2007) to enable a direct comparison with some previous methods. We consider only a single estimate. As shown in Table 5, we outperform other methods, including the the SVM-based flat classifier, by returning the species in $\widehat{C}$ with the highest posterior mass.

Smithsonian Data: As in Rejeb Sfar et al. (2013b), we also use two-thirds of the images for training (one third for learning discriminant functions and one third for estimating Gaussian parameters) and the remaining images for testing. With CS1, we achieve 92.4% accuracy while returning about 4 estimates on average; the accuracy with CS0 is 91.8%. As shown in Fig. 20, we do return a single estimate about 94% of the time. In order to compare CS1 with Rejeb Sfar et al. (2013b) on such a dataset, we rank the list of species

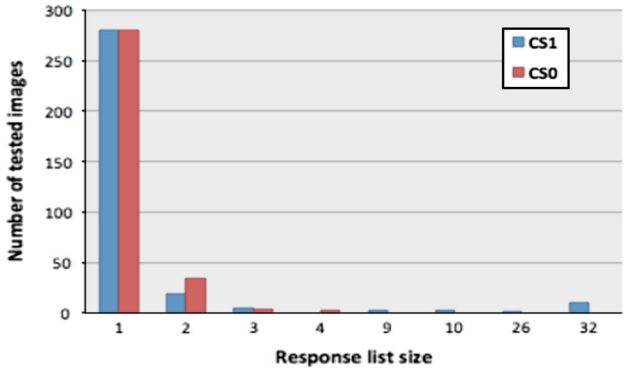

Fig. 19 The distribution of $|\widehat{C}|$, the size of the CS returned, while testing on the Flavia leaves

Table 5 Different results on the Flavia data while considering a single estimate (top-1)

Methods	Accuracy (%)
CS1 - *MAP*	**97**
F-SVM	94
RBFNN (Du et al. (2005))	94
MLNN (Du et al. (2005))	94
1-NN (Gu et al. (2005))	93
MMC (Wang et al. (2005))	92
BPNN (Wang et al. (2005))	92
RBPNN (Gu et al. (2005))	91
PNN (Wu et al. (2007))	90

The number in bold indicates the best performance

in $\widehat{C}$ for each test image, using their posterior masses. In this case, we also achieve about 90% accuracy for the top response compared with 79% for Rejeb Sfar et al. (2013b), where achieving 90% accuracy required using the top four responses.

Whereas using both strategies, the CS is estimated by the model to capture the true species with very high probability, this of course does not necessarily occur in practice due to errors in estimating the true posterior distribution. This is why CS1 out-performs CS0 in accuracy even though, *in principle*, CS0 should be better. To illustrate this, Fig. 21 shows the distribution of the posterior masses of the *true species* on the Smithsonian leaves. Note the high value (at least 0.9) for the majority of the tested images; in this special case, CS0 is equivalent to CS1 as both achieve perfect results. However, CS1 is more efficient when the true species has low mass under the model; the CS1 strategy can recover from such a catastrophic error in estimation due to the way the CS is constructed as long as there are species with non-trivial posterior masses which are visually similar to the true one. In Fig. 21, 4.2% of the images for which the posterior probability of the true species is less than 0.1 were missed by CS0 but not by CS1, but never vice-versa.

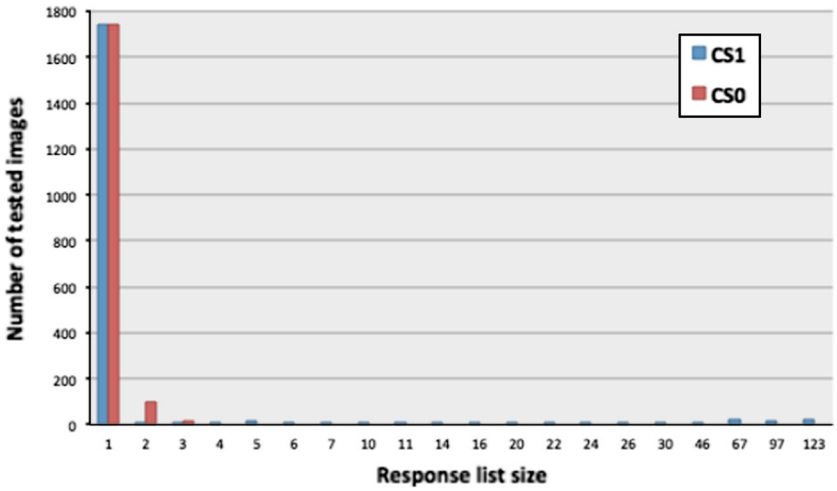

Fig. 20 The distribution of $|\widehat{C}|$, the size of the CS returned, while testing on the Smithsonian leaves

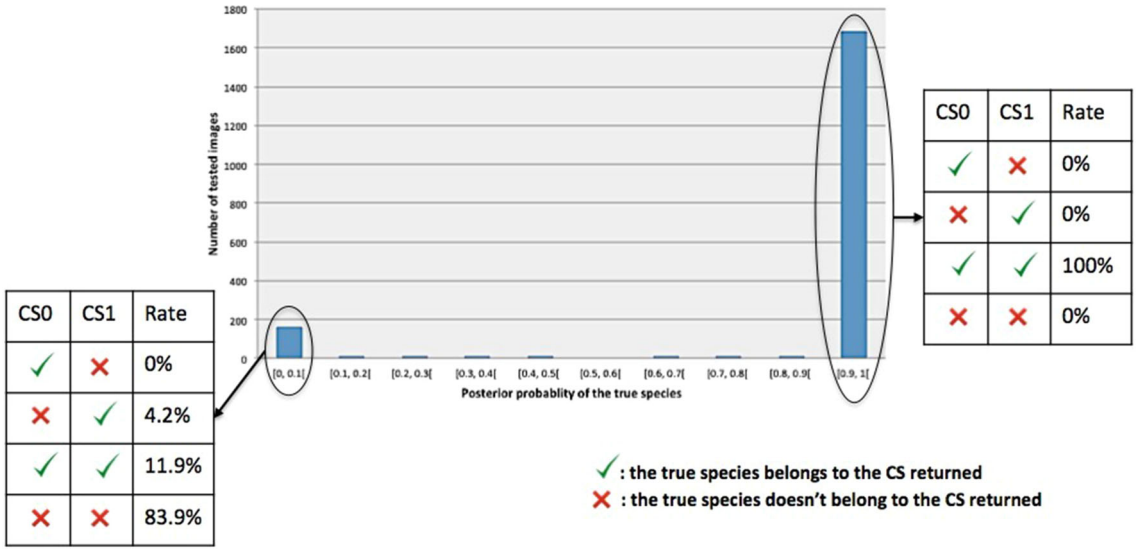

Fig. 21 The histogram (*in blue*) of the posterior masses on the true species for the leaves in the Smithsonian dataset. The two tables compare the performance of CS1 and CS0 at the two extremes, i.e., when the posterior mass on the true species is very low and very high. In the former case (drastic estimation error), the CS1 strategy is able to recover (generate a CS with the true species) but CS0 does not for 4.2% of the images, but there are no images for which the opposite occurs, i.e., CS0 succeeds but CS1 does not (Color figure online)

ImageCLEF Data: Finally, we apply our approach in a real-world context using unconstrained photographs. For this subset, we focus on CS2, using human input to mark some landmarks at the beginning of the process as explained in Sect. 4.6.2. First, we compare our method with the entries to the ImageCLEF2011 plant identification task on the photo category using the MAP baseline. In this task, each entry was assigned a normalized classification score s[5] as explained in Sect. 5.1. Figure 22 shows the scores of all the submitted runs of the eight participants; details about the participants can be found in Goëau et al. (2011). We achieve the best score: s = 0.525. More specifically, two groups can be formed among the participants, the methods which use segmentation process (in red) and those which do not use segmentation (in blue). One can notice a relatively big gap between these two groups in terms of performance, i.e., there is a difference of about ± 0.3 between the best scores of the two groups; see Fig. 22. We outperform all the previous work on such data, including segmentation-based methods. Note that the best score (s=0.523 for "IFSC UPS run2") among the participants was obtained using a *manual* segmentation which is not feasible in real-world application.

[5] http://www.imageclef.org/2011/Plants

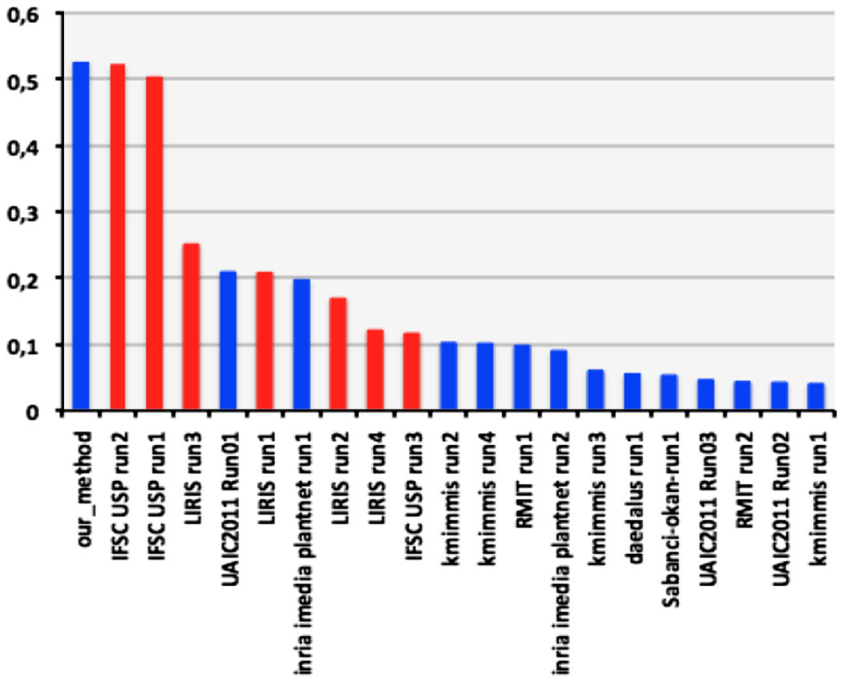

Fig. 22 Classification scores on the leaf photos of the ImageCLEF2011 dataset. In *red* are the scores of the methods which use segmentation and in *blue* are the scores of those which do not use segmentation (Color figure online)

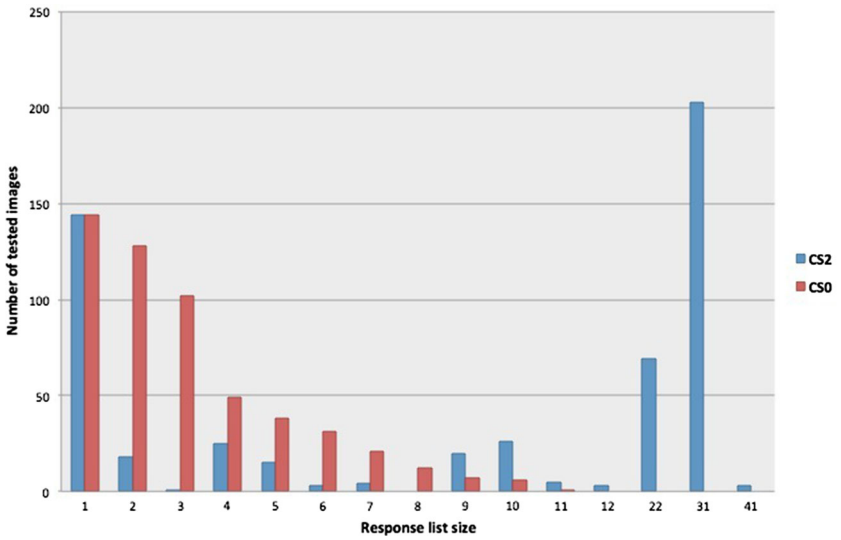

Fig. 23 The distribution of $|\widehat{C}|$, the size of the CS returned, while testing on Image CLEF2011 leaf photos. The *blue histogram* is CS2 and the *red* is CS0, both with manual landmark identification (Color figure online)

Figure 23 illustrates the distribution of $|\widehat{C}]$ while applying CS2 to the ImageCLEF2011 photos. About 50% of the time we have $|\widehat{C}| \leq 10$. However, we only achieve 58.4% accuracy due to the difficulty of this task compared with identifying leaves on a uniform background; evidently, the posterior probabilities are poorly estimated. Figure 24 shows the distribution of the posterior masses of the true species on the ImageClef2011 photos. In contrast with Fig. 21, note the low value (less than 0.1) of this mass for the majority of the tested images, which accounts for the even lower accuracy of CS0 strategy, namely 38.4%. The superior performance of CS2 occurs because for 32.5% of the images for which the posterior mass on the true species is less than 0.1, the CS generated by CS0 does not contain the true species but the one generated by CS2 does.

Moreover, additional issues are revealed from a more detailed analysis and which would explain the relatively low accuracy rate (comparing to other data). Figure 26 illustrates the different accuracies obtained per species. We completely fail to recognize those which have only few training samples (between zero and six); see the red boxes in Fig. 26. Note that four tested species do not appear among the training species and these represent about 12% of the test images. Also, using different image types (i.e., leaf, branch and foliage photos) has made the task more challenging, especially since the number of samples per image type is not balanced. For

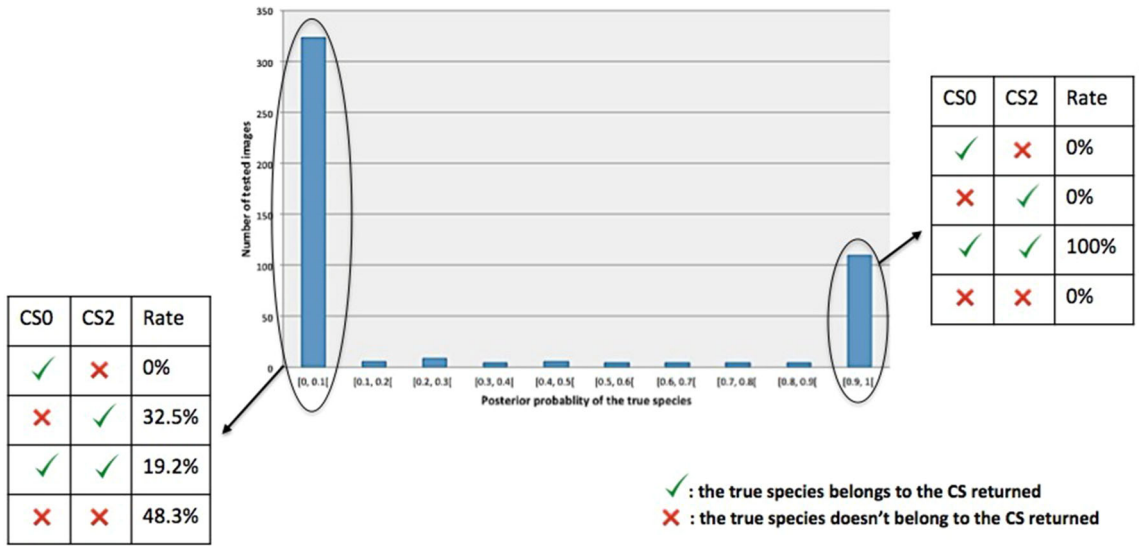

Fig. 24 The histogram (*in blue*) of the posterior masses on the true species for the ImageClef2011 photos. The two tables compare the performance of CS2 and CS0 at the two extremes, i.e., when the posterior mass on the true species is very low and very high. Among the low ones, the the CS2 strategy succeeds and the CS0 strategy does not in 32.5% of the cases, but never the opposite (Color figure online)

Fig. 25 Random sample of incorrectly identified imageCLEF011 leaf photos

example, one has only very few foliage images to predict a picked leaf image from the same species. However, we manage to recognize species from different image types with approximatively "equivalent" performances, especially for branch and foliage photos as shown in Table 6.

More generally, the quality of the photographs affects the performance. Of course, a well-photographed leaf would be easier to identify and also well-photographed training samples would lead to a better learning algorithm. For example, a close-up photo where the leaf covers a large part of the picture, is sharp whereas the background is optically blurred due to a short deep-of-field, would provide more useful visual content than a picture which is globally blurred or in which the leaf is out of focus, too damaged (e.g., dry leaves), too

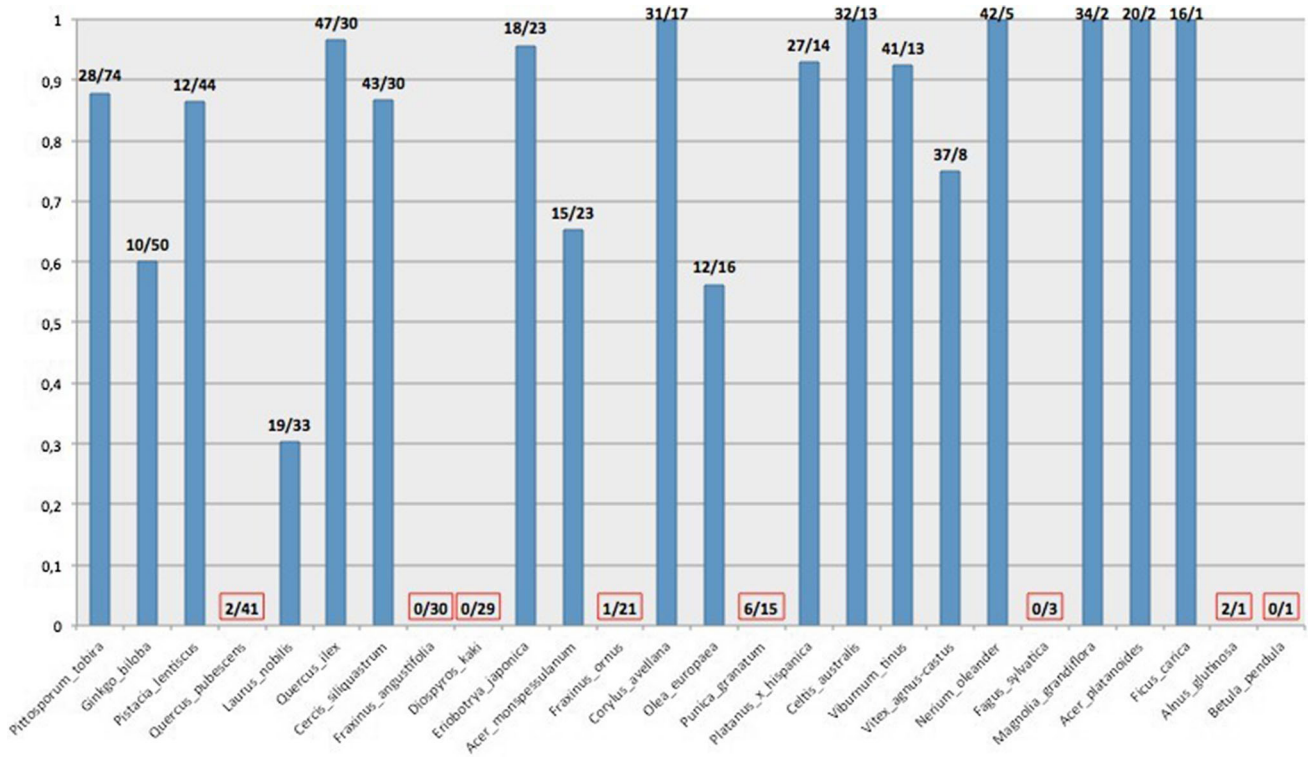

Fig. 26 Illustration of the performance per species on the ImageCLEF photo subset. Each bin is labeled by two numbers separated by a slash. The first one refers to the number of training samples in the species considered and the second one refers to the number of testing samples

Table 6 Performance of CS2 on different image types of ImageCLEF2011 photos

Image type	Accuracy (%)
Leaf	53.4
Branch	60.3
Foliage	61.5

small or/and the background is predominant with a sharp visual content like grass or foliage of other plants, etc. Figure 25 illustrates some cases of failure.

6 Multiple Leaf Images

In a botanical field scenario where the basic unit of observation is a plant, botanists can examine different samples of leaves from the same plant in order to determine the species. In fact, one sample alone might not capture sufficient information for accurate identification.

Using multiple-image queries rather than a single leaf image can then improve the identification accuracy by taking advantage of the added information. We applied the same algorithm used in Rejeb Sfar et al. (2013a) to collate the individual results into a single set of estimates for an unknown plant. Experiments on the ImageCLEF subset demonstrate the efficiency of the proposed scenarios using multiple leaf images of an unknown plant. Note that we use here only the imageCLEF data since they are the only leaf images for which we know the plant identity thanks to the additional annotation provided with this dataset; see Sect. 5.1. We improved to 74.5% accuracy (a gain of 16.1%) and the normalized ImageCLEF2011 score reaches $s = 0.626$ (a gain of about 0.01).

7 Conclusion

We have introduced a new approach to fine-grained categorization. In analogy with confidence sets in classical statistics, we output a set of categories rather than a single estimate. Our approach is model-based and Bayesian. The expected size of the confidence set plays the role of the width of the confidence interval in standard statistics and the posterior probability that the true category belongs to the confidence set plays the role of the confidence level.

We have applied this approach to identifying species of plants from images of leaves, considering images with both uniform and cluttered backgrounds. The confidence sets are restricted to the elements of a hierarchical representation

of leaf species based on visual similarity. We have shown the superior performance of this hierarchical model-based method relative to a baseline of flat classification using one-vs-all SVM as well as the straightforward way to generate a CS by aggregating ranked posterior masses, which is shown to be less robust against estimation errors. In fact, the hierarchical methods achieved better accuracy on all the datasets, including both flat and cluttered backgrounds.

We have also considered various levels of human intervention, and demonstrated how an interactive, semi-automated system can be utilized to obtain practical results. Our recognition rates outperform the state-of-art on several challenging datasets. Still, further improvements are necessary to determine the plant species from unconstrained photographs of leaves with high accuracy. Using multiple leaf images per plant or images of different organs as well as leaves (e.g., flowers and fruits) could potentially improve the recognition rates and render fine-grained categorization of plants of further interest to amateurs and botanists alike.

References

Angelova, A. & Zhu, S. (2013). Efficient object detection and segmentation for fine-grained recognition. In: CVPR.

Belhumeur, P.N., Chen, D., Feiner, S., Jacobs, D.W., Kress, W.J., Ling, H., Lopez, I.C., Ramamoorthi, R., Sheorey, S., White, S., Zhang, L. (2008). Searching the world's herbaria: A system for visual identification of plant species. In: ECCV (4), pp 116–129.

Bourdev, L.D. & Malik, J. (2009). Poselets: Body part detectors trained using 3d human pose annotations. In: ICCV, pp 1365–1372.

Branson, S., Wah, C., Schroff, F., Babenko, B., Welinder, P., Perona, P. & Belongie, S. (2010). Visual recognition with humans in the loop. In: ECCV (4), pp 438–451.

Burl, M.C. & Perona, P. (1998). Using hierarchical shape models to spot keywords in cursive handwriting data. In: CVPR, pp 535–540.

Caballero, C. & Aranda, M.C. (2010). Plant species identification using leaf image retrieval. In: CIVR, pp 327–334.

Casanova, D., Florindo, J.B. & Bruno, O.M. (2011). Ifsc/usp at imageclef 2011: Plant identication task. In: CLEF (Notebook Papers/Labs/Workshop).

Casanova, D., Florindo, J.B., Gonçalves, W.N. & Bruno, O.M. (2012) Ifsc/usp at imageclef 2012: Plant identification task. In: CLEF (Online Working Notes/Labs/Workshop).

Cook, N. R. (2005). *Confidence Intervals and Sets*. : John Wiley and Sons Ltd.

Cope, J. S., Corney, D. P. A., Clark, J. Y., Remagnino, P., & Wilkin, P. (2012). Plant species identification using digital morphometrics: A review. *Expert Syst Appl*, *39*(8), 7562–7573.

del Coz, J. J., Díez, J., & Bahamonde, A. (2009). Learning nondeterministic classifiers. *Journal of Machine Learning Research*, *10*, 2273–2293.

Deng, J., Berg, A.C., Li, K., Li, F.F. (2010). What does classifying more than 10, 000 image categories tell us? In: ECCV (5), pp 71–84.

Deng, J., Krause, J., Berg, A.C. & Li, F.F. (2012). Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In: CVPR, pp 3450–3457.

Deng, J., Krause, J. & Li, F.F. (2013). Fine-grained crowdsourcing for fine-grained recognition. In: CVPR, pp 580–587.

Du, J.X., Huang, D., Wang, X. & Gu, X. (2005). Shape recognition based on radial basis probabilistic neural network and application to plant species identification. In: ISNN (2), pp 281–285.

Duan, K., Parikh, D., Crandall, D.J. & Grauman, K. (2012). Discovering localized attributes for fine-grained recognition. In: CVPR, pp 3474–3481.

El-Yaniv, R., & Wiener, Y. (2010). On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, *11*, 1605–1641.

Ellis, B. (2009). Manual of leaf architecture. Cornell paperbacks, Published in association with the New York Botanical Garden.

Elpel, T. (2004). *Botany in a Day: The Patterns Method of Plant Identification. Thomas J. Elpel's herbal field guide to plant families of North America.* : Hops Press.

Fan, X. & Geman, D. (2004). Hierarchical object indexing and sequential learning. In: ICPR (3), pp 65–68.

Farrell, R., Oza, O., Zhang, N., Morariu, V.I., Darrell, T., Davis, L.S. (2011a). Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. In: ICCV, pp 161–168.

Farrell, R., Oza, O., Zhang, Z., Morariu, V., Darrell, T. & Davis, L. (2011b). Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. In: ICCV, pp 161–168.

Felzenszwalb, P.F. & Schwartz, J.D. (2007). Hierarchical matching of deformable shapes. In: CVPR.

Ferecatu, M. (2005). Image retrieval with active relevance feedback using both visual and keyword-based descriptors. PhD thesis, Université de Versailles SaintQuentin-en-Yvelines.

Fergus, R., Bernal, H., Weiss, Y. & Torralba, A. (2010). Semantic label sharing for learning with many categories. In: ECCV (1), pp 762–775.

Fernández, A., & Gómez, S. (2008). Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms. *J Classification*, *25*(1), 43–65.

Goëau, H., Bonnet, P., Joly, A., Boujemaa, N., Barthelemy, D., Molino, J.F., Birnbaum, P., Mouysset, E. & Picard, M. (2011). The clef 2011 plant images classification task. In: CLEF (Notebook Papers/Labs/Workshop).

Goëau, H., Bonnet, P., Joly, A., Yahiaoui, I., Barthelemy, D., Boujemaa, N. & Molino, J.F. (2012). The imageclef 2012 plant identification task. In: CLEF (Online Working Notes/Labs/Workshop).

Grall-Maës, E., & Beauseroy, P. (2009). Optimal decision rule with class-selective rejection and performance constraints. *IEEE Trans Pattern Anal Mach Intell*, *31*(11), 2073–2082.

Gu, X., Du, J.X. & Wang, X. (2005). Leaf recognition based on the combination of wavelet transform and gaussian interpolation. In: ICIC (1), pp 253–262.

Gupta, S. S. (1965). On some multiple decision (selection and ranking) rules. *Technometrics*, *7*(2), 225–245.

Ha, T. M. (1997). The optimum class-selective rejection rule. *IEEE Trans Pattern Anal Mach Intell*, *19*(6), 608–615.

Horiuchi, T. (1998). Class-selective rejection rule to minimize the maximum distance between selected classes. *Pattern Recognition*, *31*(10), 1579–1588.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Comput Surv*, *31*(3), 264–323.

Jr, C. N. S., & Freitas, A. A. (2011). A survey of hierarchical classification across different application domains. *Data Min Knowl Discov*, *22*(1–2), 31–72.

Kumar, N., Belhumeur, P.N., Biswas, A., Jacobs, D.W., Kress, W.J., Lopez, I.C. & Soares, J.V.B. (2012). Leafsnap: A computer vision system for automatic plant species identification. In: ECCV (2), pp 502–516.

Larios, N., Deng, H., Zhang, W., Sarpola, M., Yuen, J., Paasch, R., et al. (2008). Automated insect identification through concatenated histograms of local appearance features: feature vector generation

and region detection for deformable objects. *Mach Vis Appl*, *19*(2), 105–123.
Lazebnik, S., Schmid, C. & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR (2), pp 2169–2178.
Lee, P. (1989). Bayesian statistics: an introduction. No. v. 2 in A Charles Griffin Book, Oxford University Press, http://books.google.fr/books?id=_hXvAAAAMAAJ
Li, F.F. & Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In: CVPR (2), pp 524–531.
Ling, H., & Jacobs, D. W. (2007). Shape classification using the inner-distance. *IEEE Trans Pattern Anal Mach Intell*, *29*(2), 286–299.
Liu, J., Kanazawa, A., Jacobs, D.W., Belhumeur, P.N. (2012). Dog breed classification using part localization. In: ECCV (1), pp 172–185.
Manh, A. G., Rabatel, G., Assemat, L., & Aldon, M. J. (2001). Weed leaf image segmentation by deformable templates. *Journal of agricultural engineering research*, *80*(2), 139–146.
Martínez-Muñoz, G., Delgado, N.L., Mortensen, E.N., Zhang, W., Yamamuro, A., Paasch, R., Payet, N., Lytle, D.A., Shapiro, L.G., Todorovic, S., Moldenke, A. & Dietterich, T.G. (2009). Dictionary-free categorization of very similar objects via stacked evidence trees. In: CVPR, pp 549–556.
Mouine, S., Yahiaoui, I., Verroust-Blondet, A. (2013). A shape-based approach for leaf classification using multiscaletriangular representation. In: ICMR, pp 127–134.
Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. Philosophical Transactions of the Royal Society of London Series A, Mathematical and Physical Sciences 236(767):pp. 333–380, http://www.jstor.org/stable/91337
Nilsback, M.E. & Zisserman, A. (2006). A visual vocabulary for flower classification. In: CVPR (2), pp 1447–1454.
Otsu, N. (1979). *A Threshold Selection Method from Gray-level Histograms*. Man and Cybernetics: IEEE Transactions on Systems.
Rejeb Sfar, A., Boujemaa, N. & Geman, D. (2013a). Identification of plants from multiple images and botanical idkeys. In: ICMR, pp 191–198.
Rejeb Sfar, A., Boujemaa, N., Geman, D. (2013b). Vantage feature frames for fine-grained categorization. In: CVPR, pp 835–842.
Söderkvist, O. (2001). Computer vision classification of leaves from swedish trees. Master's thesis, Linköping University, SE-581 83 Linköping, Sweden, liTH-ISY-EX-3132.
Teng, C.H., Kuo, Y.T. & Chen, Y.S. (2009). Leaf segmentation, its 3d position estimation and leaf classification from a few images with very close viewpoints. In: ICIAR, pp 937–946.
Tversky, B. & Hemenway, K. (1984). Objects, parts, and categories. Experimental Psychology: General.
Wah, C., Branson, S., Perona, P. & Belongie, S. (2011). Multiclass recognition and part localization with humans in the loop. In: ICCV, pp 2524–2531.
Wang, J., Yang, J., Yu, K., Lv, F., Huang, T. & Gong, Y. (2010). Locality-constrained linear coding for image classification. In: CVPR, pp 3360–3367.
Wang, X., Du, J.X. & Zhang, G.J. (2005). Recognition of leaf images based on shape features using a hypersphere classifier. In: ICIC (1), pp 87–96.
Wang, X. F., Huang, D. S., Du, J. X., Xu, H., & Heutte, L. (2008). Classification of plant leaf images with complicated background. *Applied Mathematics and Computation*, *205*(2), 916–926.
Ward, J, Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*(301), 236–244.
Wu, J. & Rehg, J.M. (2008). Where am i: Place instance and category recognition using spatial pact. In: CVPR.
Wu, S.G., Bao, F.S., Xu, E.Y., Wang, Y., Chang, Y.F. & Xiang, Q.L. (2007). A leaf recognition algorithm for plant classification using probabilistic neural network. CoRR abs/0707.4289.
Yang, S., Bo, L., Wang, J. & Shapiro, L.G. (2012). Unsupervised template learning for fine-grained object recognition. In: NIPS, pp 3131–3139.
Yao, B., Bradski, G.R. & Li, F.F. (2012). A codebook-free and annotation-free approach for fine-grained image categorization. In: CVPR, pp 3466–3473.
Yuan, M., & Wegkamp, M. H. (2010). Classification methods with reject option based on convex risk minimization. *Journal of Machine Learning Research*, *11*, 111–130.
Zhang, N., Farrell, R. & Darrell, T. (2012). Pose pooling kernels for sub-category recognition. In: CVPR, pp 3665–3672.
Zweig, A. & Weinshall, D. (2007). Exploiting object hierarchy: Combining models from different category levels. In: ICCV, pp 1–8.