Vantage Feature Frames For Fine-Grained Categorization

Asma Rejeb Sfar
INRIA Saclay
Palaiseau, France
asma.rejeb_sfar@inria.fr

Nozha Boujemaa
INRIA Saclay
Palaiseau, France
nozha.boujemaa@inria.fr

Donald Geman
Johns Hopkins University
Baltimore, MD, USA
geman@jhu.edu

Abstract

We study fine-grained categorization, the task of distinguishing among (sub)categories of the same generic object class (e.g., birds), focusing on determining botanical species (leaves and orchids) from scanned images. The strategy is to focus attention around several vantage points, which is the approach taken by botanists, but using features dedicated to the individual categories. Our implementation of the strategy is based on vantage feature frames, a novel object representation consisting of two components: a set of coordinate systems centered at the most discriminating local viewpoints for the generic object class and a set of category-dependent features computed in these frames. The features are pooled over frames to build the classifier. Categorization then proceeds from coarse-grained (finding the frames) to fine-grained (finding the category), and hence the vantage feature frames must be both detectable and discriminating. The proposed method outperforms state-of-the art algorithms, in particular those using more distributed representations, on standard databases of leaves.

1. Introduction

Research in automated object recognition is currently very active, driven by applications as well as the intellectual challenge, and there have been notable recent advances using both discriminative learning and object modeling for detecting and localizing instances of generic object classes such as cars, cats and people appearing in digital images [2, 5, 7, 10, 13, 23, 26]. More recently, motivated by applications in areas such as botany, agriculture, medicine and forestry, there has also been considerable interest in more fine-grained discrimination, for example identifying specific species of birds, flowers, leaves and insects; some of this work is summarized in §2. Relative to fine-grained categories, instances from generic object classes tend to be rather distinct from one another, displaying gross differences that are easy for humans to identify [21]. As a result, methods designed for generic categories are usually not well-adapted to isolating and representing the specific information necessary for discriminating among fine-grained visual categories such as species of leaves.

Indeed, due to large intra-class variability and inter-class similarity, fine-grained categorization can be extremely challenging even for experts, especially when the number of relevant categories is very large, as in botany. Even a single taxonomic family (e.g., orchids) may contain many species, each highly varied, and taxonomic categories (e.g., species or varieties) are often determined by subtle differences in shape and texture. In fact, there can even be less variation in appearance between two images from two different (sub)categories than within a single one, as illustrated in Figure 1. Plants may exhibit different shape characteristics due to local context, such as location, climatic conditions and age; for example, (a),(b) and (c) in Figure 1 come from the same species. And whereas the overall shapes may be sufficiently different to distinguish between some species (see e.g., Figure 1 (a) and (g)), other species may display only subtle differences: see the instances of two different species of orchids in Figure 1 (e) and (f) and two instances of leaves in Figure 1 (h) and (i) from two different genera (and hence species).

Our approach to categorizing botanical species is motivated by the strategy used by botanists, in which attention is focused on visual properties of the object in the vicinity of a small number of distinguished landmarks. Whereas these landmarks are the same for each species, it is the local features which permit disambiguation. But both aspects are important: *where to look* and *what to compute*. The vehicle for translating this into a computer vision algorithm, and our main contribution, is the notion of a *vantage feature frame*; we provide algorithms for learning discriminating ones, detecting them online and pooling the features computed in these frames to identify the categories. We refer to the origins of the frames as *vantage points* - special locations from which observing the leaf or flower and in a particular direction can provide discriminating information about the species. Besides location, scale and orientation, each *vantage feature frame* is also equipped with a possi-

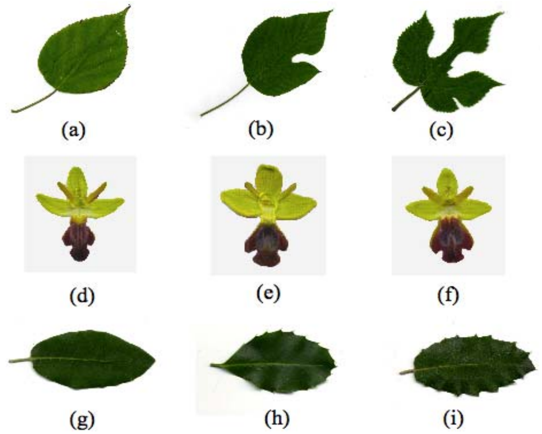

Figure 1. *Intra-class similarity and inter-class differences for leaves and orchids.* Top row: (a), (b) and (c) are *Broussonetia papyrifera*. Middle row: (d) and (f) are *ophrys funerea* while (e) is an *ophrys iricolor*. Bottom row: (g) and (i) are *quercus ilex* while (h) is an *ilex aquifolium*.

bly category-dependent set of features that help to discriminate that category from all others. We will demonstrate that this type of non-distributed representation can be highly effective in distinguishing between closely-related categories, and in particular improves upon the accuracy of existing methods for simple leaves on standard databases.

2. Related work

There is a growing body of work investigating fine-grained image classification of birds [8, 24, 27], insects [15, 18], flowers [6, 19] and leaves [1, 6, 14].

Several shape-based approaches, including boundary analyses, have been adapted for fine-grained categorization, especially for leaves [1, 4, 17]. Often, performance is sensitive to the quality of the contour resulting from a segmentation process, which naturally complicates distinguising between categories with very similar shapes. Other methods adapt systems for detecting instances of generic object classes [16, 25] by encoding an image as a bag of discrete visual codewords and basing classification on histograms of codeword occurrences; examples include [19, 27]. Again, however, the distinctions among fine-grained categories are sometimes too refined (see Figure 1) to be captured by variations in bags of visual words.

To account for such distinctions, an increasing number of studies utlize information from experts. In [24], an interactive system is proposed wherein humans click on bird parts and answer questions about attributes (e.g., ”white belly”, ”red-orange beak”, ”sharp crown”). In other recent work [9, 28] annotated training data (e.g., key points and objects parts) are obtained from experts. In [9], classifiers based on *poselets* (parts of the object from a given viewpoint) [3] are employed to extract part and shape information for building fine-grained models. However, this approach requires 3-D pose annotation, which is based on volumetric primitives that are costly to obtain manually and present other difficulties (see Figure 1 of [28]); instead, the authors of [28] advocate a 2-D rather than 3-D representation in order to reduce the level of annotation required to generate the *poselets*. Our work on leaves and orchids is somewhat similar in that the detection of the vantage frames primes the identification of the botanical species; however, unlike the work in [9, 28], our representation is based on frames not *poselets*, i.e., on coordinate systems and corresponding local features which are, by construction, invariant to variations in pose, thereby avoiding any need for global image transforms, e.g., geometric normalization. Of course birds and leaves present different kinds of challenges; the former exhibit higher intra-species variation (e.g., birds may be flying, swimming or perched), whereas the latter exhibit more inter-species similarity (e.g., in color, overall shape and internal structure).

3. Vantage Feature Frames

Let $\{C_1, ..., C_N\}$ denote N categories. In the botanical applications which motivate this work, such as assigning a species to a scanned image of a leaf, there is often useful domain knowledge, for instance named landmarks $\mathcal{L} = \{l_1, ..., l_K\}$ around which botanists focus in order to separate one species from another. (See Figure 2.) Usually, N is of order tens or hundreds and K ranges from two to four. In fact, landmarks are more like “vantage points” in that orientation plays a role as well, in other words, where the landmarks are in relation to one another. Naturally, species tend to have certain signature appearance properties and consequently what to look for in the neighborhood of the landmarks may be species-dependent. Put differently, the conditional distribution over any large family of generic local features may depend strongly on the species. This aspect of the identification process will be encoded by allowing the set of features associated with each landmark to depend on the category. We also want to ensure that the local appearance properties are largely invariant to the orientation and scale of the object. Finally, in order to identify an unknown specimen, botanists proceed hierarchically (or at least so describe their reasoning process) from coarse-grained categories (i.e., higher taxa than species such as family or genus) to the species level. Consequently, information about an underlying taxonomy can be useful in organizing the search.

With these considerations in mind, a *vantage feature frame* $\mathcal{F}$ has two components. One, Θ, is geometric and the other, $\mathcal{X}$, is appearance-based. The geometric component Θ is category-independent and simply a local coordinate system centered at one of the landmarks l; the scale and orientation are discussed below. The appearance component is a family of pose-indexed features, one element of the family for each category: $\mathcal{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_N\}$, where $\mathcal{X}_t$ is the set of local features to compute in frame $\mathcal{F}$ for category C_t. Obviously, to be useful the frame must be reliably detected and the features must be discriminating.

4. Construction

As with the representation above, we will first describe the general process in abstract terms, providing more specific examples of frames and features in ensuing sections.

4.1. Learning the frames

Learning the most discriminating frames from scratch would evidently be a major challenge, and we do not attempt this. As indicated above, by leveraging domain knowledge, we begin with a list of candidate origins $l_1, ..., l_K$. There will be frames associated with a subset of these. Moreover, since we are dealing with images of single objects (e.g., scanned images of leaves) we declare the orientation of the frame to be determined by the centroid of the object, that is, the landmark points to the centroid, and the unit distance to be the approximate scale of the object. The choice of landmarks or vantage points is performance-based. Assume we are given a classifier for each set of vantage feature frames; our particular choice is described in §6. Given $|\mathcal{L}| = K$ candidate landmarks, there are then 2^{K-1} possible set of coordinate systems, evaluating them one-by-one might be infeasible, in which case one might adopt a greedy strategy: the efficiency of each candidate could be measured by the improvement in the overall classification rate obtained by adding the corresponding frame to the existing list of frames.

However, for simple leaves and orchids only three "universal" landmarks $\mathcal{L} = \{l_1, l_2, l_3\}$ have been suggested by botanists; they are described in §6.1 and illustrated in Figure 2. For each of the $2^3 - 1 = 7$ combinations of frames, we estimated the classification accuracy using cross-validation. Feature extraction and classification are described in §4.3 and §5 respectively. It should be noted that for this learning process the locations of the landmarks were determined by manually annotating the training data. As a result, the errors that are inevitably made in automatically detecting the landmarks (see Section §6.3) are not taken into account in choosing the best set of frames. One might expect that the more frames the better the performance, and hence using all three would be optimal. However, this was not the case; Table 1 shows the recognition rates for the seven possible combinations of frames used for simple leaves. The best performance is obtained with two frames corresponding to apex and base of the leaf.

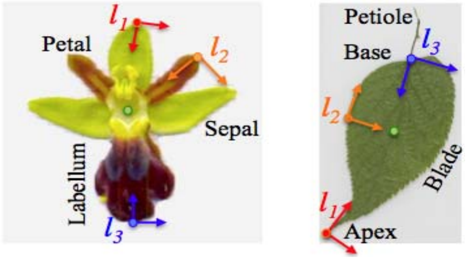

Figure 2. Candidate frames for orchids and leaves.

4.2. Detecting the frames

The first step in classifying an image is to estimate the location, orientation and scale of each frame. As indicated above, the orientation is determined by the centroid, which is directly computed from the raw image data after a segmentation process using the Otsu algorithm [20]. The scale is taken to be the radius of the bounding circle as illustrated for leaves in Figure 5. The landmarks are detected by dedicated classifiers trained on manually annotated images. 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.

In order to detect each vantage point, a classifier (see §5) based on SVM scores is built from positive and negative training examples. Positive images are annotated by the landmark considered and negative images are randomly annotated. The features for SVM learning are defined in the local coordinate system centered on the candidate landmarks (i.e., the x-axis is directed towards the centroid as described above). Invariant focusing of this nature is enabled by the type of "pose-indexed" (or "frame-indexed") features X introduced in [13] for detecting cats. Basically, given a frame consisting of two distinguished points and a distinguished scale, there is a candidate feature $X = X(w, j)$ for each (local) window w in frame coordinates and for each local image property j: the feature X is just the property histogram in w. We refer to [13] for details. We use color, shape and texture as properties; specifically, we used HSV, Hough, EOH and Fourier histograms as base features (more details can be found in [12]).

4.3. Learning the features

The appearance-based component is category-dependent. Whereas we use the same class of features to learn landmark detectors, we construct a separate binary classifier for each category C_t for distinguishing that category from all others and which employs a learned subset of features $\mathcal{X}_t$. The reason for dedicated features is that there is so much variability in the presentation

Set of coordinate systems	l_1	l_2	l_3	l_1, l_2	l_1, l_3	l_2, l_3	l_1, l_2, l_3
Recognition rate	0.75	0.72	0.73	0.76	**0.8**	0.77	0.78

Table 1. Cross-validated recognition rates for leaves (from the Smithsonian database) for each of seven possible sets of frames sets with centers l_1, l_2, l_3. The best result (in bold) is obtained with two frames centered at the base l_1 and apex l_3.

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 the genera might be different from those around the apex for estimating either the genera or the species; and the best features in any given frame may be genus- and species-dependent. Hence, we select a category-dependent subset of features $\mathcal{X}_t$ and only these are used to train SVM classifiers.

Specifically, we first estimate the probability distribution of each feature X under both hypotheses $I \in C_t$ and $I \notin C_t$ (where I is the image be classified) from the positive and negative examples. For each distinct (taxonomic) category, images belonging to that category are positive and all others negative. For feature $X(w, j)$, denote the two distributions by $p^+_{w,j}$ and $p^-_{w,j}$ and let $d_{w,j} = |p^+_{w,j} - p^-_{w,j}|$ be the difference in the L1 norm. Then $\mathcal{X}_t$ consists of the features with the M largest differences, and M is chosen by cross-validation. Figure 3 illustrates the recognition rate for leaf genera for various M. Selecting category-dependent features increases recognition performance and decreases computation. For instance, we achieve over 75% recognition rate of leaf genus while considering only the first genus returned and using between about 500 and 2500 category-dependent features against only 67% without any selection i.e, $M = 5808$ (see Figure 3).

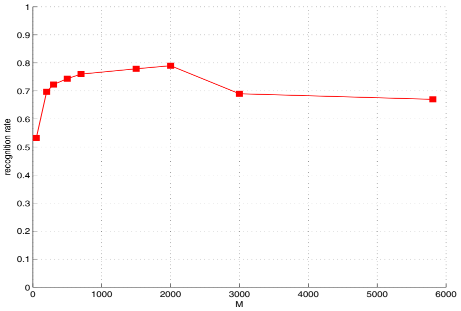

Figure 3. Recognition rates for leaf genera from the Smithsonian data (see §6.2) while considering only the first genus returned and using M selected features.

5. Fine-grained categorization

As indicated above, the category identification is also hierarchical, coarse-grained to fine-grained, which is another way of exploiting domain knowledge. In our experiments we consider two-levels, the first for genera and second for the species, the ultimate target.

Let T denote the full tree-structured graph. Associated with every node $t \in T$, a one-vs-all classifier $f_t = f_t(I)$ is designed to separate images in C_t from images in the complement of C_t (i.e., all other taxonomic categories). The framework is largely classifier-independent in that any learning algorithm could be chosen to induce f_t from the training data at node t. We have chosen to use SVM scores for test statistics and a likelihood framework. The SVM score F_t is trained using the feature set $\mathcal{X}_t$ defined in §4.3, and the corresponding classifier f_t is based on the likelihood ratio:

$$L_t(I) = \frac{P(F_t = F_t(I)|I \in C_t)}{P(F_t = F_t(I)|I \notin C_t)}$$

Several detected genera may be considered for species identification. If C_t corresponds to a genus, we define

$$f_t(I) = \begin{cases} 1 & \text{if } \log(L_t(I)) > \rho \\ 0 & \text{else} \end{cases}$$

Here, ρ is a threshold used to control the false negative genus rate, that is to allow only a very small number of instances in which $I \in C_t$ but $f_t(I) = 0$ (missed detections). This can be accomplished at the expense of (temporary) low specificity (i.e., a high false positive rate), but this is a favorable tradeoff in our context.

The hierarchy is processed breadth-first coarse-to-fine: at each level, all the children of a *positive* node t (i.e., one for which $f_t(I) = 1$) are retained and tested at the next level. Whereas false positives can be successively pruned, if the true hypothesis is rejected at a node containing it then it cannot be recovered. Hence only the classifiers for species which belong to the retained genera are performed. Finally, those species for which $f_t(I) = 1$ (where t is the node for the genus of the species) are then sorted according to their likelihood ratios.

The advantage of mapping the SVM score to a likelihood ratio is that it takes into account the distribution under both hypotheses. In particular, this mapping is *not* monotone, i.e, does not preserve the ordering of SVM scores across a level, which might naturally occur on different scales. This is illustrated in Figure 4, which shows two pairs of distributions for two classes of leaf species C_1 and C_2. The dashed (respectively, solid) red and blue lines correspond respectively

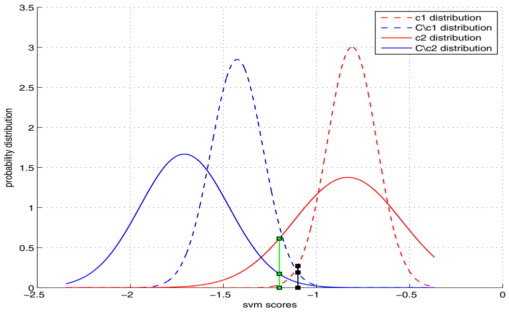

Figure 4. Comparison between the SVM score distributions of two different genera from the Smithsonian database (see §6.2). Both distributions are approximated by Gaussian densitites with the estimated means and variances.

to the SVM score distribution of images in C_1 (resp., C_2) and in the complement of C_1 (resp., C_2). Also shown are the scores achieved by an image which would be classified as C_1 if we only considered the raw SVM scores (black for C_1 and green for C_2) but in fact is classified as C_2 based on likelihoods.

Note that the same framework was used to learn the vantage point detectors i.e., likelihood framework based on SVM scores. However, for those points only a single estimate is retained, namely the one corresponding to the candidate at which the likelihood ratio is maximized and thus a single classifier (and thus a single SVM) is learned to detect each vantage point.

6. Experiments

In this section we describe the landmarks for leaves and orchids, the datasets we have used to evaluate our approach, and compare our results with those previously obtained.

6.1. Botanical landmarks

To analyze leaves, experts usually focus on the apex, the base and the leaf margin, whereas an orchid specialist focuses on the sepals, petals and the labellum. These are illustrated in Figure 2. Let l_1 denote the leaf apex (respectively, the central sepal for orchids), l_2 the first intersection point between the perpendicular to the apex-base line throughout the centroid of the blade and the leaf boundary (resp., the petal on the right of l_1 for orchids) and l_3 the leaf base (resp., the bottom of the orchid labellum) as shown in Figure 2. Note that for leaves, the centroid corresponds to the center of mass of the blade; the leaf petiole is removed before computing the centroid (see Figure 5). As for the segmentation process, it is important to note that we are not concerned by imperfect contours or incomplete petiole removal since our method is robust to such problems. Figure 5 illustrates the vantage point detection process for a leaf image, namely the leaf base and the leaf apex detection.

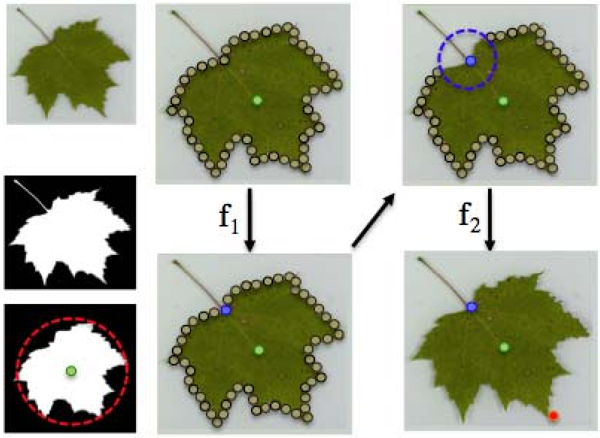

Figure 5. A test leaf image is first segmented. Then the petiole is removed in order to compute the centroid (green point) as well as the approximate bounding circle of the leaf blade (red dashed circle). The base (blue point) and the apex (red point) are estimated using learned classifiers (f_1, f_2). 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).

6.2. Datasets

We considered three challenging simple-leaf datasets from different geographical areas as well as a dataset of Mediterranean rare orchids. Each image represents a scanned object on a white background.

Smithsonian: This dataset has 5466 simple-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 various poses and orientations of leaves as well as different structures of basal and apical parts as shown in Figure 6. Thus, good performance on such a dataset suggests robust and effective landmark detection.

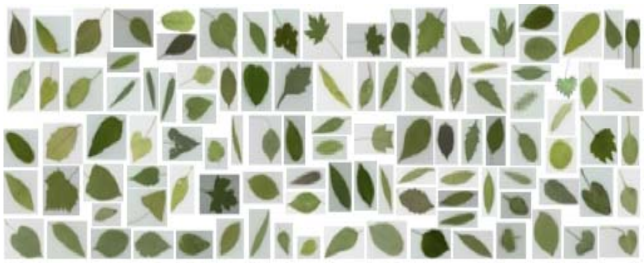

Figure 6. Random sample from the SmithSonian dataset.

Swedish: This is the subset of simple leaf images of the first publicly available leaf data for research, introduced by the authors of [22]. It has 975 images containing 75 images from each of 13 different Swedish simple species (after removing the two compound species from the original dataset[?]

ImageCLEF2011: This dataset is the subset of the ImageCLEF2011[1] data containing all the scanned simple ImageCLEF leaves. It is composed of 46 species from the French Mediterranean and was constructed through a citizen science initiative conducted by Telabotanica[2], a French social network of amateur and expert botanists. As a result, the task it represents is quite close to the conditions encountered in a real-world application. We refer to [14] for details.

Orchids[3] **:** There are 1610 images representing 23 species of a relatively rare orchid flower family provided by the "Mediterranean Orchid Society" (*Société Méditerranéenne d'Orchidologie*).

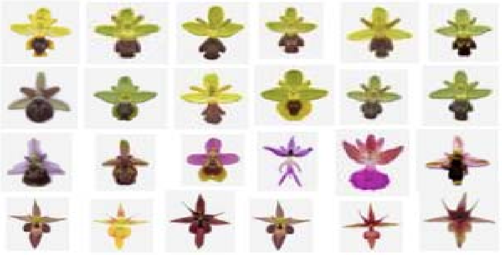

Figure 7. Samples from the Orchids dataset. One image from each species is shown. Note that the color is not a discriminative feature; many differently colored orchids could belong to the same genus or species.

We organized each dataset into its proper taxonomic hierarchy (genus, species) and annotated it with landmarks. These annotations, together with taxonomic labels, will be made available to other researchers. To evaluate the performance of our approach, we used two-thirds of the images for training and one-third for testing.

6.3. Detection results

First, we present the results of the vantage point detection for all the data introduced in the previous section in Table 2, achieving over 90% accuracy in each case and thereby confirming reasonable invariance to shape and structure. Figure 8 shows vantage point detection results for orchids and different type of leaves (e.g., toothed, lobed, concave, convex, symmetric, asymmetric).

6.4. Identification results

To evaluate the performance of species identification, we provide the rate on the holdout test data at which the true species appears among our top n estimates for $n = 1, ..., 5$ for the Smithsonian, Swedish and orchid subsets. However, for ImageCLEF2011 data we adopt the evaluation metric[1] used for the ImageCLEF2011 plant identification task, which allows us to compare our performance with that of all the task participants. We used cross-validation on the training images in order to fix the number of the selected features ($|\mathcal{X}_t|$) for each $t \in T$. We typically select about 1000 features for estimating the genera and about 1500 features for the species. We also fixed the threshold $\rho = -4$; the negative value promotes a low missed detection rate.

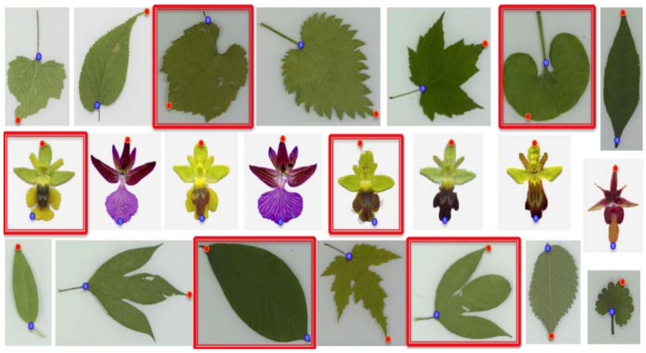

Figure 8. Random sample of test images with the estimated vantage points for both Smithsonian leaves and orchids. False detections are framed with a red box. Note that the entire detection process is considered erroneous if any vantage point is not accurately detected.

Dataset	**Detection Rate**
Smithsonian leaves	92%
Swedish leaves	96%
ImageCLEF leaves	93%
Orchids	95%

Table 2. Rate of well detected vantage points

n	**1**	**2**	**3**	**4**	**5**
Smithsonian data	79%	86%	89%	90%	91%
Orchid data	81%	92%	94%	96%	97%

Table 3. Recognition rates using *vantage feature frames* on both Smithsonian leaves and Orchids

Smithsonian Data: The first row of Table 3 reports the recognition rates for different values of n. We achieve 79% accuracy for the top-ranked species ($n = 1$) and 91% for $n = 5$. The results on a random sample of test images is shown in Figure 9. Of particular note is the similarity between the appearance of the true and estimated species in the misclassified cases and the impact of poor vantage point estimation; for example, note the errors in estimating both the base and the apex for the third test leaf in Figure 9.

In [1], the IDSC was used with a KNN classifier to identify the species within two subsets of the Smithsonian database, achieving a recognition rate of $60\% - 70\%$

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

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

[3] Courtesy of Roland Martin and Errol Vela

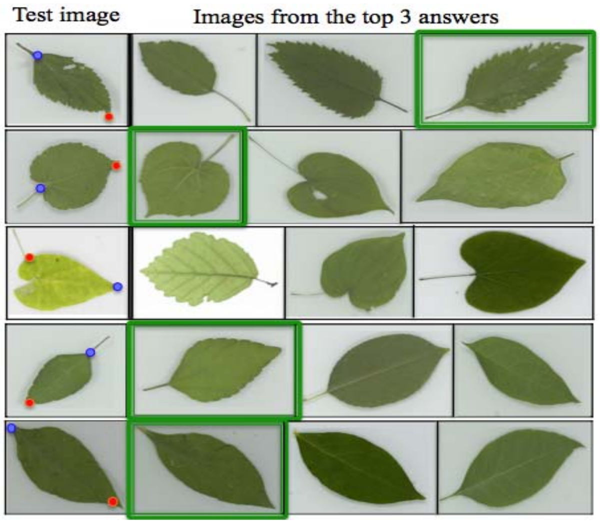

Figure 9. A sample of test leaf images with the estimated vantage points and the top three species returned by our algorithm. For each test image, the red point refers to the estimated leaf apex and the blue point to the estimated leaf base. The examples framed in green come from the same species as the test image.

n	**1**	**2**	**3**	**4**	**5**
VFF	95%	98%	99%	99%	99%
IDSC [17]	94%	97%	98%	98%	98%

Table 4. Recognition rates on the Swedish data

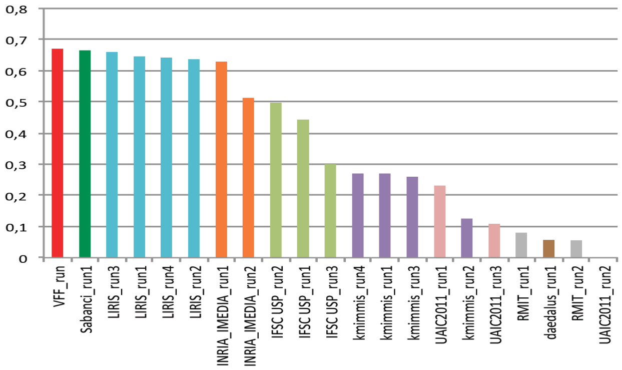

Figure 10. Classification scores on the scanned simple leaves of the ImageCLEF2011 dataset.

($n = 1$). We tried to get those subsets but they were not available. Consequently, we applied the IDSC method to our subset of simple leaves from the Smithsonian dataset using the same parameters as in [1, 17]. However, due to the sensitivity of the IDSC method to boundary resolution and noise (see §4 of [1]), the result (around 20%) is not representative. In particular, our binary images are obtained from Otsu algorithm [20] rather than the finer method of segmentation used in [1], which avoids variations in lighting across the image and shadows cast by other leaves.

Swedish Data: We also compare our results with the IDSC on the Swedish leaves [22] since the IDSC achieved better results than other methods on this dataset according to [17]. In this case, the binarized images were provided by the authors of [17], which substantially improved the performance of the IDSC method. Table 4 reports both our results (”VFF”) and the IDSC results. Both methods reach over 90% classification rate. We achieve the best result with 95% accuracy for $n = 1$. We were not able to perform a direct comparison with other methods, e.g., [11], which report good results on the whole Swedish dataset (which contains both simple and compound leaves) since the algorithms were not publicly available. However, it should be noted that compound leaves have very different characteristics than simple leaves. In particular, they exhibit greater inter-species variation, and thus identifying the species of compound leaves is easier. For this reason we focus here on only simple leaves.

ImageCLEF2011 Data: Finally, we compare our method with the entries to the ImageCLEF2011 plant identification task on the scanned simple leaves (46 species). All the scores, including ours, are provided by the administrators of the competition. In this task, each entry was assigned a normalized classification score s [1]. Figure 10 shows the scores of all the submitted runs of the eight participants; details about the participants can be found in [14]. We achieve the best score: $s = 0.67$.

Orchid Data: To the best of our knowledge, there is no previous work on this family of flowers. We applied the vantage feature frame approach on this data to demonstrate how it could be readily applied to a different type of closely-related botanical species. We achieve 81% accuracy for the top-ranked species ($n = 1$) and 97% for $n = 5$ as shown in the second row of Table 3.

7. Conclusion

We have introduced a novel approach for fine-grained categorization using the concept of *vantage feature frames*. The different characteristics of these frames, namely, the geometric and the appearance-based components, combine to provide the cues needed to distinguish between closely-related categories such as botanical species. Our recognition rates outperform the state-of-art on several challenging datasets. Future work is aimed at applications involving cluttered backgrounds and at automatically determining candidate landmarks for constructing the vantage feature frames.

Acknowledgements

We would particularly like to thank Hervé Goëau and Vera Bakic for their help in conducting our experiments.

References

[1] P. Belhumeur, D. Chen, S. Feiner, D. Jacobs, W. Kress, H. Ling, I. Lopez, R. Ramamoorthi, S. Sheorey, S. White, and L. Zhang. Searching the world's herbaria: A system for visual identification of plant species. In *ECCV*, pages 116–129, 2008. 2, 6, 7

[2] A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In *ICCV*, 2007. 1

[3] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009. 2

[4] C. Caballero and M. C. Aranda. Plant species identification using leaf image retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '10, pages 327–334, New York, NY, USA, 2010. ACM. 2

[5] L.-B. Chang, Y. Jin, W. Zhang, E. Borenstein, and S. Geman. Context, computation, and optimal roc performance in hierarchical models. *International Journal of Computer Vision*, 93(2):117–140, 2011. 1

[6] J. S. Cope, D. Corney, J. Y. Clark, P. Remagnino, and P. Wilkin. Plant species identification using digital morphometrics: A review. *Expert Systems with Applications*, 39(8):7562 – 7573, 2012. 2

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR (1)*, pages 886–893, 2005. 1

[8] K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *CVPR*, pages 3474–3481, 2012. 2

[9] R. Farrell, O. Oza, Z. Zhang, V. Morariu, T. Darrell, and L. Davis. Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. In *ICCV*, pages 161–168, 2011. 2

[10] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, 2010. 1

[11] P. Felzenszwalb and J. Schwartz. Hierarchical matching of deformable shapes. In *CVPR*, pages 1–8, june 2007. 7

[12] M. Ferecatu. *Image retrieval with active relevance feedback using both visual and keyword-based descriptors*. PhD thesis, Université de Versailles SaintQuentin-en-Yvelines, 2005. 3

[13] F. Fleuret and D. Geman. Stationary features and cat detection. *Journal of Machine Learning Research (JMLR)*, 9:2549–2578, 2008. 1, 3

[14] H. Goëau, P. Bonnet, A. Joly, N. Boujemaa, D. Barthelemy, J.-F. Molino, P. Birnbaum, E. Mouysset, and M. Picard. The clef 2011 plant images classification task. In *CLEF (Notebook Papers/Labs/Workshop)*, 2011. 2, 6, 7

[15] N. Larios, H. Deng, W. Zhang, J. Sarpola, M.and Yuen, R. Paasch, A. Moldenke, D. Lytle, S. Ruiz-Correa, E. Mortensen, L. Shapiro, and T. Dietterich. 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, 2008. 2

[16] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society. 2

[17] H. Ling and D. Jacobs. Shape classification using the inner-distance. *IEEE Trans. Pattern Anal. Mach. Intell*, 29:286–299, 2007. 2, 7

[18] G. Martnez-Muoz, N. Larios Delgado, E. Mortensen, W. Zhang, A. Yamamuro, R. Paasch, N. Payet, D. Lytle, L. Shapiro, S. Todorovic, A. Moldenke, and T. Dietterich. Dictionary-free categorization of very similar objects via stacked evidence trees. In *CVPR*, pages 549–556, 2009. 2

[19] M.-E. Nilsback and A. Zisserman. A visual vocabulary for flower classification. In *CVPR*, volume 2, pages 1447–1454, 2006. 2

[20] N. Otsu. A Threshold Selection Method from Gray-level Histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66, 1979. 3, 7

[21] E. Rosch, C. Mervis, W. Gray, D. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 1976. 1

[22] O. Söderkvist. Computer vision classification of leaves from swedish trees. Master's thesis, Linköping University, SE-581 83 Linköping, Sweden, September 2001. LiTH-ISY-EX-3132. 5, 7

[23] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *ICCV*, pages 606–613, 2009. 1

[24] C. Wah, S. Branson, P. Perona, and S. Belongie. Multiclass recognition and part localization with humans in the loop. In *ICCV*, Barcelona, 2011. 2

[25] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, pages 3360–3367, 2010. 2

[26] T. Wu and S. Zhu. A numerical study of the bottom-up and top-down inference processes in and-or graphs. *International Journal of Computer Vision*, 93(2):226–252, 2011. 1

[27] G. Yao, G. Bradski, and L. Fei-Fei. A codebook-free and annotation-free approach for fine-grained image categorization. In *CVPR*, Providence, RI, USA, June 2012. 2

[28] N. Zhang, R. Farrell, and T. Darrell. Pose pooling kernels for sub-category recognition. In *CVPR*, pages 3665–3672, 2012. 2