# Iterative Search with Local Visual Features for Computer Assisted Plant Identification

Wajih Ouertani, Pierre Bonnet, Michel Crucianu, Nozha Boujemaa, Daniel Barthélémy

**Abstract** — To support computer assisted plant species identification in realistic, uncontrolled picture-taking condition, we put forward an approach relying on local image features. It combines query by example and relevance feedback to support both the localization of potentially interesting image regions and the classification of these regions as representing the target species or not. We show that this approach is successful and makes prior segmentation unnecessary.

**Index Terms** — Assisted identification, biodiversity informatics, local features, local query, object localization, relevance feedback.

## 1 Introduction

Given the large volume and increasing accessibility of biodiversity data—e.g. Encyclopaedia of life [1], Atlas of living Australia [2], or ZipcodeZoo—gathered from all over the world, it is even more important to explore, master and capitalize this type of knowledge [3]. Joint efforts of biologists, information science and data-mining communities are required for solving significant common problems. As biological image databases are increasing rapidly [4], automated species identification based on digital data is of great interest for accelerating biodiversity assessment, researches and monitoring [5]. We put forward here an interactive identification approach in which a botanist having a partially annotated large image database is assisted by a Relevance Feedback search mechanism to identify a plant's specie. The botanist can then easily select the relevant unlabeled images (without having to go through the entire database) and label them at once with the name of the specie.

- *W. Ouertani, M. Crucianu and N. Boujemaa are with INRIA, BP 105, 78153 Le Chesnay cedex, France. E-mail: {Wajih.Ouertani, Michell.Crucianu, Nozha.Boujemaa}@ inria.fr.*
- *P. Bonnet, D. Barthélémy are with INRA, Amap Joint research unit, CIRAD,TA A-51/PS2, 34398 Montpellier cedex 5, France. E-mail: {Pierre.Bonnet, Daniel.Barthelemy}@cirad.fr.*

## 2 CONTEXT AND RESULTING CHALLENGES

### 2.1 Content-based image retrieval and interactive identification

In a query by visual example (QBVE), an example image is first provided to the search engine as a visual query. The engine returns images that are visually similar to the query image, using a metric on the space of the low-level features that represent the images. Motivated by the "semantic gap" issue, *i.e.* the fact that such features seldom reflect user's intention, a Relevance Feedback (RF) [6] mechanism includes the user in the retrieval process. In an RF session the search result is iteratively refined. For a given query, the system first retrieves a set of images ranked according to the predefined similarity measure between the query vector and feature vectors of images in the database. Then, the user provides feedback regarding this result, by qualifying the returned images as either "relevant" or "irrelevant". From this feedback, the engine iteratively learns the visual features of the images and returns improved results to the user. A good RF mechanism should find the user intention with minimal interaction [7].

This retrieval refinement technique was applied to botanical databases with pictures taken in controlled conditions [8], but it has important limitations resulting from the global image description. To remove such restrictions on picture-taking conditions, we extend here RF to the use of *local* features (LF). This is a more adequate representation of image regions and allows users to provide a precise feedback by freely selecting relevant and irrelevant *regions of interest* in images.

### 2.2 Challenges

We address here learning and recognition challenges that come from strong variations in viewpoint, picture-taking conditions, interactivity and generalization requirements. Recent work on plant species identification requires reliable prior segmentation of informative organs such as leaves [9], [10] (with controlled picture-taking conditions) or flowers [11] (less restrictive conditions). With such well-controlled pictures, the shape of a leaf, its margins, or several local and region-based features of flowers are employed for recognition. In general, due to variations in natural environment, plant accessibility, picture-taking system and intention, an object of interest (plant or plant's part) may appear on different backgrounds and cover a potentially small part of the image (see first row in Fig. 1). This supports the use of LF to focus on the target object. Also, in a botanical identification context, some images illustrate global aspects of a plant or of an inflorescence, while others show details having different visual attributes. A same object of interest could thus be represented in various poses and at different scales (see second row in Fig. 1).

Relevance feedback brings in two additional challenges. First, the search engine should respect the interactivity requirement, *i.e.* quickly

respond during each round. Even if joint object segmentation and recognition (*e.g.* [12]) could improve identification, its additional cost makes it inappropriate for interactive retrieval. Second, at each RF round the user only labels *a few* images. For the retrieval session to be successful, the system should generalize well from these few examples. In the next section we propose an approach that addresses these problems using LF.

Fig. 1. Background variations of an inflorescence of *Habenaria* species (1st row), scale and pose variation of an inflorescence of *Cleisomeria lanatum* (Lindl. ex G.Don, 2nd row).

## 3 IDENTIFICATION APPROACH

We propose to jointly use search by example with local queries and supervised classification (with Support Vector Machines, SVM). Every RF round thus consists of two stages: (1) QBVE using as query the LF that were previously found relevant; (2) result re-ranking by the SVM decision function, applied to the potentially relevant set of features in every returned image. This joint use of QBVE and SVM classification serves two purposes. First, it allows to locate, in the returned images, the potential regions of interest (see Fig. 2, green and red points) that have to be evaluated by the SVM. A region of interest is here the set of LF that were found to be individually similar to some LF in the query. An image can indeed contain objects from multiple classes; our approach will focus on the potentially relevant parts and ignore other, irrelevant parts (blue points in Fig. 2). In this context, the task of the SVM is to solve ambiguity and distinguish sets of LF that belong to the target specie (Fig. 2, middle) from sets composed of LF that are individually similar to relevant LF but, when considered together, do not correspond to the target specie (Fig. 2, right).

Second, QBVE can be very fast with an appropriate index structure—we rely here on *a posteriori* multi-probe locality sensitive hashing [13]—and only images containing hit points (*i.e.* points that are individually similar to relevant LF) have to be evaluated by RF rather than all the images in the database, which significantly improves scalability.

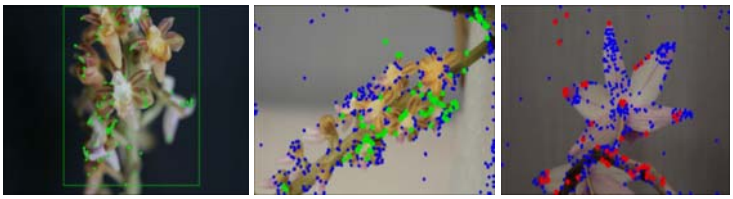

Fig. 2. Region of interest localization: user target (left) and two candidate images with LF belonging to the target (green, middle) or not (red, right). The other LF (blue) are ignored.

We assume that the distribution of LF in the selected sets brings relevant discriminating information with respect to the joint presence of LF, so we employ the pyramidal matching kernel (PMK, [14]) or the kernel based on random histograms (RH, [15]). The SVM has thus to downgrade image regions (sets of LF) whose LF are individually similar to LF of the target specie, but whose distribution does not correspond to this target.

## 4 Experimental Evaluation

We employed two different image databases for the evaluations. The first one was produced by AMAP Joint Unit on Laos orchid's reproductive organs (mainly inflorescences and flowers). It contains 1913 images for 181 orchid species. There are significant variations in scale, pose and lighting (see Fig 1, 2). Botanists manually labelled 2347 regions of interest. The second database is Oxford flowers 17 (www.robots.ox.ac.uk/~vgg/data/flowers/17/), consisting of 17 flower categories with 80 images each. The database includes common UK flowers; there is a significant variation within a same class and close similarity between several classes. There is also a ground truth showing fine flower segmentation for a subset of the images [11].

We compare RF with global image description (`GF_RF`) to RF with local descriptions (`LF_RF_QVE_Harris`, `LF_RF_QVE_SIFT`). The global image description employed (named "joint description" below) concatenates a Laplacian weighted RGB histogram, a Fourier-based histogram and a Hough histogram [2]. Two types of LF were employed: (i) joint description (with coarser histograms) obtained in the neighbourhood of Harris colour points, and (ii) SIFT [16]. The experiments were performed by using the ground truth to emulate user feedback under realistic conditions. Each RF session consists of 8 iterations. At every iteration, the emulated user labels the first 3 relevant and the first 3 irrelevant unlabelled regions. Fig. 3 shows the mean average precision (MAP) of system's responses where recall equals precision (MAP at R=P), for the three RF mechanisms. Only the 10 orchid classes having enough image examples were used for generating RF sessions. Fig. 3 (left) shows that, even with few iterations ($1^{st}$ to $4^{th}$, less than 50% of the available training data), RF with LF outperforms global RF. We also note

that the results obtained with SIFT (features ignoring colour!) are better than those with Harris points whose description includes colour. This is due to the fact that scale and shape variations within a same class are more important than colour differences between classes in this dataset.

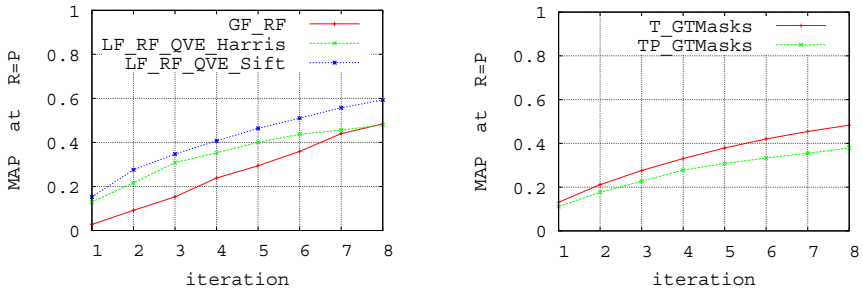


Fig. 3. MAP evolution over RF iterations. Left: on Orchids database. Right: on Oxford flowers 17 database, with and without segmentation masks in prediction stage.

Using `LF_RF_QVE_Harris` and the fine segmentation ground truth provided in Oxford flowers 17 database, we performed two experiments in which we use segmented objects as training examples and, for the prediction stage, we either (i) use only hit points (retrieved by QBVE) that fall in pre-segmented objects of interest in a candidate image (`TP_GTMasks`), or (ii) use all the hit points retrieved in a candidate image (`T_GTMasks`). As can be seen in Fig. 3 (right), the object localisation given by the QBVE stage allows to reach a performance that is close to the one obtained with fine prior segmentation. We also find that the inclusion of a small part of object's neighborhood provides a relevant context that increases recognition accuracy.

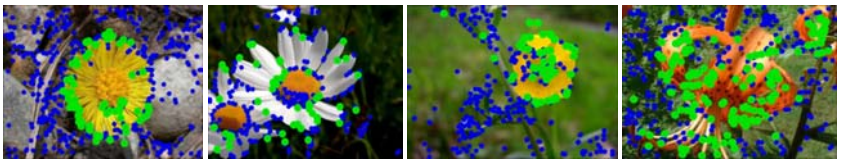

Fig. 4. Object localization examples on Oxford Flower 17 database, green points showing the object of interest. From left to right: Colt's Foot, Daisy Flower, Buttercup, Tiger Lily.

## 5 Conclusion

Content-based image search can have a significant contribution to plant species identification. However, to make it successfully applicable to realistic contexts, we argue that it is necessary to let the user interact with the system on the basis of local image descriptions that allow to focus on the relevant part of an image. We proposed a relevance feedback method

relying on local images features. It also makes use of an LF retrieval stage in order to locate potentially interesting image regions and improve scalability to larger image databases. We have shown that this approach can be successful and that it makes prior segmentation unnecessary. The results also show how important it is to devise local features that are robust to most of the variations that can be expected when pictures are taken in more general, uncontrolled conditions.


### ACKNOWLEDGEMENTS

This work is part of the Pl@ntNet project, http://www.plantnet-project.org, the first flagship project of Agropolis foundation.


### REFERENCES


[1] Wilson E. O., 2003. The encyclopedia of life. Trends in Ecology and Evolution. 18 (2).

[2] Anon., 2008, Atlas of Living Australia – sharing biodiversity knowledge to shape our future, Proc. R. Soc. Western Australia (Nov. 2008).

[3] N. F. Johnson. Biodiversity informatics. Annu. Rev. Entomol. 52, p. 421-438, 2007.

[4] S. J. Baskauf, B. K. Kirchoff. Digital plant images as specimens: toward standards for photographing living plants. Vulpia, Vol. 7, pp. 16–30, 2008.

[5] K. J. Gaston, M. A. O'Neil. Automated species identification: why not? Phil. Trans. R. Soc. B., 359, p. 655-667, 2004.

[6] Xiang Sean Zhou and Thomas S. Huang. Relevance feedback for image retrieval: a comprehensive review. Multimedia Systems, vol. 8, no. 6, p. 536-544, 2003.

[7] M. Ferecatu. Image retrieval with active relevance feedback using both visual and keyword-based descriptors. PhD thesis, Université de Versailles, France, 2005.

[8] M. Coutaud, P. Bonnet, A. Joly, R. Enficiaud, N. Boujemaa and D. Barthélémy. Advances in taxonomic identification by image recognition with the generic content-based image retrieval IKONA. In e-Biosphere 09: Intl. Conf. on Biodiversity Informatics, London, 2009.

[9] I. Yahiaoui, N. Hervé, and N. Boujemaa. Shape-based image retrieval in botanical collections. In 7th Pacific Rim Conf. on Multimedia, LNCS vol. 4261: 357–364, 2006.

[10] P. N. Belhumeur, D. Chen, S. Feiner, D. W. Jacobs, W. J. 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 *European Conf. on Computer Vision*, *LNCS vol.* 5305: 116–129. Springer, 2008.

[11] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In 6th Indian Conf. on Computer Vision, Graphics & Image Proc., p. 722–729, Washington, DC, USA, 2008.IEEE Computer Society.

[12] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. Int. J. Comput. Vision, 81(1):2–23, 2009.

[13] A. Joly and O. Buisson. A posteriori multi-probe locality sensitive hashing. In 16th ACM intl. conf. on Multimedia, pages 209–218, New York, NY, USA, 2008. ACM.

[14] K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. J. Mach. Learn. Res., 8: 725–760, 2007.

[15] W. Dong, Z. Wang, M. Charikar, and K. Li. Efficiently matching sets of features with random histograms. In 16th ACM intl. conf. on Multimedia, p. 179–188, New York, NY, USA, 2008. ACM.

[16] D. G. Lowe. Distinctive image features from scale-invariant keypoints. Intl. J. Comp. Vis. 60(2):91-110, 2004.