Interactive Learning of Heterogeneous Visual Concepts with Local Features

Wajih Ouertani
INRIA – IMEDIA project
and INRA, France
Wajih.Ouertani@inria.fr

Michel Crucianu
INRIA – IMEDIA project
and CEDRIC – CNAM, France
Michel.Crucianu@cnam.fr

Nozha Boujemaa
INRIA – IMEDIA project
78153 Le Chesnay, France
Nozha.Boujemaa@inria.fr

ABSTRACT

In the context of computer-assisted plant identification we are facing challenging information retrieval problems because of the very high within-class variability and of the limited number of training examples. To address these problems, we suggest a new interactive learning approach that combines similarity-based retrieval and re-ranking by SVM using local feature distributions. This approach leads to improved sample selection, allowing to obtain better results.

Categories and Subject Descriptors: H.3.3 [Information Systems]: Multimedia Information Search and Retrieval; H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing methods.

General Terms: Algorithms.

Keywords: Relevance feedback, interactive retrieval, object localization.

1. INTRODUCTION

The interactive construction of a concept class of images, based on visual criteria, can be very useful in several applications such as identification and recognition or mass annotation in various domains including botany, medicine and earth observation. For example, to identify the plant shown in a picture, a user can perform a query by visual similarity and then refine the query through several relevance feedback interactions to find the most similar images; if the images were labeled with the appropriate names of the species they represent, the user has a good set of candidate species for the unknown plant. Besides, a botanist having a partially annotated large image database can interactively retrieve, through the relevance feedback mechanism, a set of images containing plants sharing an attribute that could help labeling the species, the plant's organ, etc.; the botanist can then easily select the most relevant images returned (without having to go through the entire large database) and label them at once with the name of the species.

There are commonalities between these different application domains. First, usually only a part of an image is relevant for the target class of an interactive retrieval session; the rest of the image represents noisy background or items belonging to other classes. We introduce a new interactive retrieval method dedicated to image descriptions with local features (LF), where for each round of relevance feedback (RF) we proceed in two stages. During the first stage, an adequate index allows to quickly retrieve images containing LF that are similar to the features found relevant in the previous round. In the second stage, the potentially relevant sets of features are evaluated by an SVM (with cost-effective kernels for sets of features) and the images are re-ranked according to this result. Then, the user provides feedback by selecting regions (sets of LF) as positive or negative examples. Since only the images returned by the first stage are evaluated by the SVM, the response time is compatible with the time constraints of interactive retrieval. Another important characteristic is that the relevance information obtained from the user during a retrieval session is limited and can be very noisy. The two-stage interactive retrieval mechanism also addresses this problem.

The next section further describes the difficulties we are facing in the context of plant image databases and briefly analyzes the capabilities of state of the art methods in this context. Our approach, sketched above, is presented in detail in section 3. Section 4 shows a comparison between this approach and some alternative proposals on an existing ground truth database.

2. CHALLENGES AND RELATED WORK

In the Fig. 1, we are showing two typical examples representing relevant visual characteristics for assisted plant identification or mass annotation in botany. An object of interest only covers a (potentially small) part of an image and several different objects from various classes can be simultaneously present. Recent work on plant identification requires reliable prior segmentation of a leaf [19, 2] (with pictures taken in controlled conditions) or of a flower [14] (less restrictive picture taking conditions). In such controlled image capture conditions, the shape of the leaf contour or several local and region-based features of the flower are then employed for recognition. The problems we are addressing in this study come from the uncontrolled viewpoint and picture-taking conditions that make the solution harder to find. In this regard, the object is often a sparse inflorescence and can even be some other organ having a characteristic visual attribute, on a potentially complex background, so prior segmentation cannot reliably delimit the region of interest. Joint object

MM'10, October 25–29, 2010, Firenze, Italy.

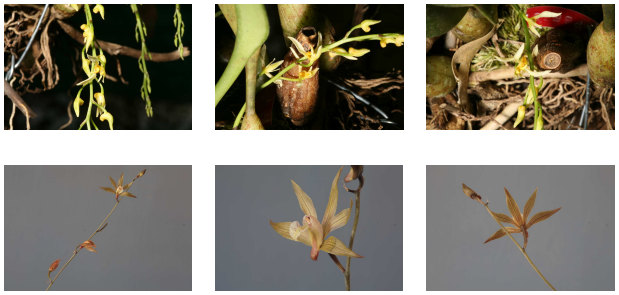

Figure 1: Concept heterogeneity due to viewpoint variations and to diverse natural environment.

segmentation and recognition (e.g. [16]) could result in better segmentation, but is not appropriate for interactive retrieval because it requires a large training set and is too time consuming. It is nevertheless important to locate potential regions of interest in order to apply the class recognition tool to those regions. The typical solution is to use windows moving over the entire image. Such a method can also provide good results with LF (see e.g. [13]) but, in spite of improvements that accelerate the identification of the best matching window, remains too slow for interactive retrieval from a large set of images. When LF are employed, another approach is possible: perform a similarity-based retrieval using as query a set of relevant features and take into account, in the returned images, only the regions where the similar LF are found. This type of retrieval can be very fast if an adequate index structure is employed (e.g. [9]).

For general object class recognition, discriminant LF (see [17, 3]) can be selected and provide good results (e.g. [15]). When different classes share many LF, the *distribution* of the features in these classes can be more discriminant, and results can be improved with SVMs and kernels for sets of features like the Pyramid Matching Kernel (PMK) [7] or Random Histograms (RH) [5]. When dealing with rather rigid objects and not too large viewpoint variations, the (local) geometric configuration of the features can provide very relevant discrimination information [18, 8, 12]. Specific kernels taking configuration information into account were also developed, see e.g. [1]. This additional information can even allow to find some classes of objects in an unsupervised way (e.g. [4]). For the plant image databases we focus on, viewpoint variation is often significant, so geometric configurations do not provide reliable information. We only employ here distributions of LF, obtained from positive and negative examples corresponding to user-selected image regions.

The selections often concern concave or "sparse" objects and users tend to be generous in delimiting the relevant regions. So the sets of LF provided as positive or negative examples include a large share of features that should not be taken into account for the the definition of the target class of objects. It appears necessary then to perform feature selection, in a difficult context because during an interactive session very little labeled data is provided. One could think of an alternative approach, consisting in directly performing object class recognition rather than interactively defining a class of interest. But this alternative approach can only be used for predefined classes and requires a large annotated database for training the class recognizers. While for general object classes the contribution of a broad community can support large scale annotation initiatives like LabelMe or ImageNet, for specialized databases (like plant image collections) expert knowledge is needed and such a solution cannot be employed. On the other hand, the interactive construction of a concept class of images can make expert annotation easier.

3. PROPOSED APPROACH

A relevance feedback (RF) session is divided into several consecutive rounds. Starting from an initial query (here, a set of LF), at every round the user provides feedback regarding the retrieval results, by selecting regions in the returned images and qualifying each region as either a positive or a negative example. From this feedback, the search engine learns the features describing the truly relevant image regions and returns improved results to the user. For the RF method we suggest, every round consists of two stages (see the **for** loop in Algorithm 1): (i) query by example (QBE) using as query the LF that were previously found relevant, (ii) result re-ranking according to the SVM decision function applied to the potentially relevant set of features in every returned image.

This joint use of QBE (retrieval by similarity) and SVM classification serves two purposes. First, it allows to locate, in the returned images, the potential regions of interest 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. In this context, the task of the SVM is to distinguish sets of LF that belong to the target class from sets composed of LF that are individually similar to relevant LF but, when considered together, do not correspond to the target class. Second, QBE can be very fast with an appropriate index structure, so the SVM only has to be applied to relatively few sets of LF for the selection of an unlabeled sample for the next round. QBE relies here on *a posteriori* multi-probe locality sensitive hashing [9]. In Algorithm 1 the following notations are employed: $\mathcal{P}(I)$ is the set of LF in image I; $\mathcal{Q}$ is a set of LF used as a query; $\mathcal{R}^+$ and $\mathcal{R}^-$ are sets of LF defining a relevant and, respectively, an irrelevant region; $\mathcal{I}_k$ is a set of images; $\mathcal{P}_s(\mathcal{I}_k)$ is the set of LF in the images $\mathcal{I}_k$ that contributed to their retrieval as the k most similar images to the corresponding query $\mathcal{Q}$; f is the decision function of an SVM.

Figure 2 shows examples of images retrieved by QBE (lines 3 and 11 of Algorithm 1); the LF found similar to the query, called below "candidate LF", are shown in green (when belonging to a target object) or red (when belonging to another object), while the other LF in the image are in blue. The decision function of the SVM is only computed for the candidate LF (line 12), the other LF of the image are ignored.

We assume that the *distribution* of LF in the selected sets brings relevant discrimination information with respect to the joint presence of LF, so we employ PMK [7] or the kernel based on RH [5]. We noticed that for the plant image databases we have to deal with LF configurations are too diverse, within a same class, to provide reliable information, so we do not employ here configuration-based kernels. In line 12 of Algorithm 1, the SVM has to downgrade image regions (sets of LF) whose LF that are individually similar to LF of the target object, but whose distribution does not correspond to a target object.

Algorithm 1 Proposed relevance feedback method

1: User provides one relevant region $\mathcal{R}_0^+$ in one image;
2: $\mathcal{Q}_0 = \mathcal{R}_0^+$;
3: QBE with $\mathcal{Q}_0 \rightarrow$ set $\mathcal{I}_{0k}$ of k images;
4: Rank $I_j \in \mathcal{I}_{0k}$, $0 \leq j \leq k$, by decreasing similarity between $\mathcal{P}(I_j)$ and $\mathcal{Q}_0$;
5: Evaluate the ranking with respect to ground truth;
6: **for** $i \in 1, \ldots, T$ **do**
7: User marks set of relevant regions $\mathcal{E}_i^+ = \{\mathcal{R}_{il}^+\}_{l \in \{1,\ldots,L_i\}}$ and set of irrelevant regions $\mathcal{E}_i^- = \{\mathcal{R}_{im}^-\}_{m \in \{1,\ldots,M_i\}}$;
8: Define $\mathcal{E}_{is}^+ = \{\mathcal{R}_{il}^+ \cap \mathcal{P}_s(\mathcal{I}_{i-1,k})\}_{l \in \{1,\ldots,L_i\}}$ and $\mathcal{E}_{is}^- = \{\mathcal{R}_{im}^- \cap \mathcal{P}_s(\mathcal{I}_{i-1,k})\}_{m \in \{1,\ldots,M_i\}}$;
9: Train SVMi (to obtain f_i) on $\bigcup_{j=0}^{i} \mathcal{E}_{js}^+$ and $\bigcup_{j=0}^{i} \mathcal{E}_{js}^-$;
10: $\mathcal{Q}_i = \mathcal{Q}_{i-1} \cup (\bigcup_{l \in \{1,\ldots,L_i\}} \mathcal{R}_{il}^+)$;
11: QBE with $\mathcal{Q}_i \rightarrow$ set $\mathcal{I}_{ik}$ of k images;
12: Rank $I_j \in \mathcal{I}_{ik}$, $j \in \{1, \ldots, k\}$, by decreasing $f_i(\mathcal{P}(I_j) \cap \mathcal{P}_s(\mathcal{I}_{ik}))$;
13: Evaluate the ranking with respect to ground truth;
14: **end for**

Figure 2: Candidate LF belonging to the target (green) or not (red), and the other LF (blue).

4. EXPERIMENTAL EVALUATION

The evaluation aims to provide answers to several questions: (i) is the contribution of each component of the proposed method (QBE and SVM learning) valuable, (ii) does the *distribution* of LF in the selected sets bring relevant discrimination information with respect to the joint presence of LF, (iii) knowing that users' selections are typically broader than the truly relevant sets of LF, what improvement can we expect from a method that would remove those selected LF that do not belong to the target object?

To answer the first two questions, we compare the method described above to a basic query expansion solution (denoted BasicQE), a "naive" feedback (denoted NaiveRF) and to an existing RF solution based on boosting [11]. BasicQE simply adds to the query, at every feedback round, the sets of LF that were marked "relevant" by the user. NaiveRF consists in using *all* the LF in an image to compare against the discrimination frontier of the SVM (there is no preliminary QBE to find only the potentially relevant set of LF).

Since we do not have a large enough ground truth for our plant image database, we looked for existing benchmarks that could raise similar difficulties. The best candidate we found is Graz-02[1] also used in [15, 11]. This database (see figure 2 for examples) consists of images of bikes (365), cars (420) and people (311), plus 380 images not containing any of these objects. The objects appear in a natural scene, the bikes are "sparse" and there may be more than one relevant object in an image. These characteristics are shared with the plant image database mentioned in Section 2; however, the plant image database has many more classes and some of them are quite similar to each other. For Graz-02, the basic ground truth is provided as rectangles (this is also what can be expected from real users) surrounding the objects belonging to the target class.

The comparisons were (classically) performed in batch mode, using the ground truth to emulate user feedback. Each RF session consists of 8 iterations; at each iteration, the emulated user labels the first 3 relevant and the first 3 irrelevant unlabeled regions, the ranking being obtained from the SVM decision function (see also Algorithm 1). For the results are reported here, the kernel is the normalized dot product on Random Histograms (RH) [5] embedding space. We employ a 5 times folded concatenation of 20 elementary 2^{10}-dimensional histograms (B=10, M=5, N=20 with the notations in [5]). Even with such fine histograms, the system is sufficiently fast for real time interaction. Retrieval can be accelerated by trading speed against accuracy in producing the RH embedding and by using more localised interest regions (having fewer LF). Two types of LF were employed: (i) a concatenation of histograms (Laplacian weighted RGB histogram, Fourier-based histogram and Hough histogram, see [6]) obtained in the neighborhood of Harris color points, and (ii) SIFT. Figure 3 shows the evolution of the mean average precision (MAP) obtained for the point where Recall = Precision, and averaged over 450 RF sessions (150 sessions for each class: bike, car, person).

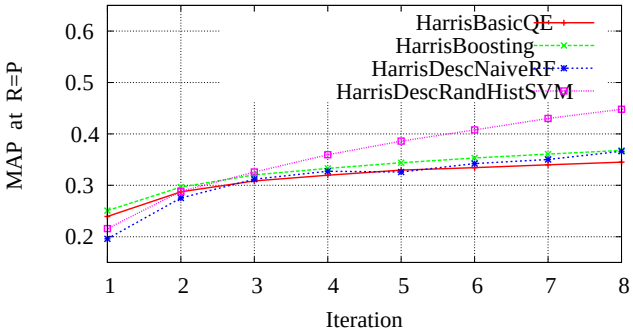

Figure 3: Comparison between BasicQE, NaiveRF, boosting and the proposed approach.

As seen in Figure 3, after a few RF iterations, the proposed method significantly outperforms BasicQE, NaiveRF and boosting. This shows that both QBE and SVM learning have a valuable contribution to the overall performance (comparison with BasicQE and NaiveRF). Also, the distribution of LF brings relevant discrimination information with respect to the joint presence of LF (comparison with boost-

[1] http://www.emt.tugraz.at/~pinz/data/GRAZ_02/

ing [11]). The results obtained with SIFT are not as good; to avoid overload, they were not shown in Figure 3.

To answer the third question mentioned above, we took advantage of the more refined segmentation that is also available for Graz-02. Figure 4 shows the results of a comparison between the use of rectangular selections as above (RandHistSVM) and of refined segmentations (RandHistSVMWithMasks). Surprisingly, the use of refined segmentations brings almost no improvement to accuracy. This shows that feature selection for removing those selected LF that do not belong to the target object may not be essential.

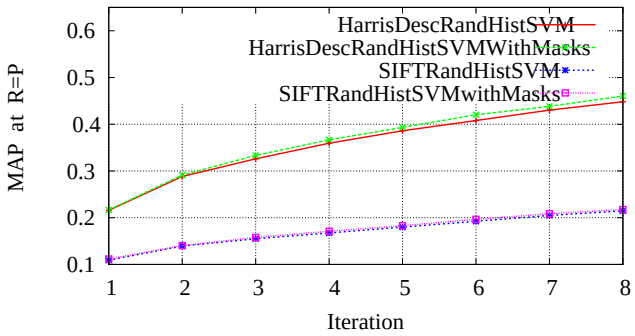

Figure 4: Rectangular selections vs. segmentation

5. CONCLUSION

To address the challenging problem of learning heterogeneous visual concepts with local features, we present a new interactive learning method that performs both the identification of potentially relevant image regions and the subsequent classification of these regions based on distributions of local features. We evaluate this method on the Graz-02 database that represents, to some extent, the difficulties raised by real plant image databases. The evaluation results shows that the distribution of local features brings relevant discrimination information in this challenging learning context, and that the combined use of retrieval by similarity and SVM-based re-ranking allows to improve retrieval results.

Acknowledgments

This work is part of the Pl@ntNet project of the Agropolis foundation. Plant images are courtesy of AMAP.

6. REFERENCES

[1] F. R. Bach. Graph kernels between point clouds. In *25th Intl. Conf. on Machine Learning*, pages 25–32, New York, NY, USA, 2008. ACM.

[2] 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.

[3] G. J. Burghouts and J.-M. Geusebroek. Performance evaluation of local colour invariants. *Comput. Vis. Image Underst.*, 113(1):48–62, 2009.

[4] O. Chum, M. Perd'och, and J. Matas. Geometric min-hashing: Finding a (thick) needle in a haystack. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 17–24, June 20–26 2009.

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

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

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

[8] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *10th European Conf. on Computer Vision*, pages 304–317, Berlin, Heidelberg, 2008. Springer-Verlag.

[9] 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.

[10] H. Ling and S. Soatto. Proximity distribution kernels for geometric context in category recognition. In *11th IEEE Intl. Conf. on Computer Vision*, pages 1–8. IEEE, 2007.

[11] S. Litayem, A. Joly, and N. Boujemaa. Interactive objects retrieval with efficient boosting. In *17th ACM intl. conf. on Multimedia*, pages 545–548, New York, NY, USA, 2009. ACM.

[12] D. Liu, G. Hua, P. Viola, and T. Chen. Integrated Feature Selection and Higher-order Spatial Feature Extraction for Object Categorization. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[13] M. H. Nguyen, L. Torresani, F. de la Torre, and C. Rother. Weakly supervised discriminative localization and classification: a joint learning process. In *IEEE Intl. Conf. on Computer Vision*, 2009.

[14] 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.*, pages 722–729, Washington, DC, USA, 2008. IEEE Computer Society.

[15] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer. Generic object recognition with boosting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(3):416–431, 2006.

[16] 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.

[17] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluation of color descriptors for object and scene recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[18] Z. Wu, Q. F. Ke, M. Isard, and J. Sun. Bundling features for large-scale partial-duplicate web image search. In *IEEE Conf. on Computer Vision and Pattern Recognition*, June 20–26 2009.

[19] 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.