Maestra at LifeCLEF 2014 Plant Task: Plant Identification using Visual Data

Ivica Dimitrovski[1], Gjorgji Madjarov[1], Petre Lameski[1], and Dragi Kocev[2]

[1] Faculty of Computer Science and Engineering, University of Ss Cyril and Methodius
Rugjer Boshkovikj 16, 1000 Skopje, Macedonia
[2] Department of Knowledge Technologies, Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
ivica.dimitrovski@finki.ukim.mk, gjorgji.madjarov@finki.ukim.mk,
dragi.kocev@ijs.si, petre.lameski@finki.ukim.mk

Abstract. In this paper, we describe an approach to the automatic plant identification task of the LifeCLEF 2014 lab. The image descriptors for all submitted runs were obtained using the bag-of-visual-words method. For the leaf scans, we use multiscale triangular shape descriptor and for the other plant organs Opponent SIFT extracted around points of interest obtained using Harris-Laplace detector. We then use approximate k-means (AKM) algorithm to cluster these descriptors in large number of clusters/visual words (approximately 200K). Each image in the training and test dataset is represented as a sparse high-dimensional histogram of term (visual word) occurrences. The similarity between two images is defined as a L_2 distance over the obtained histograms. We use the standard tf-idf weighting scheme, which reduces the contribution that commonly occurring, and therefore less discriminative, words make to the similarity. To obtain the predictions, we employ a late fusion scheme for combining the similarities/ranks from multiple ranked image lists build for each type of view. Overall the proposed methods performed well, we ranked fifth, out of 10 competing groups.

Keywords: plant identification, opponent SIFT, TSLA, bag-of-visual-words, approximate k-means

1 Introduction

The ImageCLEF plant identication competition is organized every year since 2011 and aims to benchmark the progress in the area of plant identication from images [5]. Similar to the previous years, the task in 2014 is evaluated as a plant species retrieval task based on multi-image plant observations queries. The goal is to retrieve the correct plant species among the top results of a ranked list of species returned by the evaluated system. The number of species in this year task is about 500, which is an important step towards covering the entire flora of a given region.

Contrary to previous plant identification benchmarks, queries are not defined as single images but as plant observations, meaning a set of 1 to 5 images depicting the same individual plant observed by the same person the same day.

Each image of a query observation is associated with a single view type (entire plant, branch, leaf, fruit, flower, stem or leaf scan) and with contextual meta-data (data, location, author). The motivation of the task is to fit better with a real scenario where one user tries to identify a plant by observing its different organs. The details of this competition are described in [6].

In this paper, we describe our approach and runs submitted to the LifeCLEF 2014 Plant Task. The approach is based on bag-of-visual-words representation. We are using Harris-Laplace detector to detect points of interest. From these points, local invariant descriptors are then extracted. We used Opponent SIFT as local descriptors [13]. For the leaf scans we use the multiscale triangular shape descriptor [10]. Approximate k-means (AKM) algorithm is applied to cluster these descriptors in large number of clusters/visual words (approximately 200K) [12]. In AKM, the exact nearest neighbor search is replaced with approximate nearest neighbor search in the assignment step when searching for the nearest cluster center for each point. Each image in the training and testing dataset is represented as a sparse high-dimensional histogram of term (visual word) occurrences. The similarity between each query/test image histogram and each histogram from the training set is measured by using a L_2 distance. We use the standard tf-idf weighting scheme [1], which down-weights the contribution that commonly occurring, and therefore less discriminative, words make to the relevance score.

The remainder of this paper is organized as follows. Section 2 briefly presents the training and test dataset. The image processing and feature extraction algorithms are described in Section 3. Section 4 presents the information fusion and classification algorithms that we used to obtain the predictions. Section 5 presents the results from the experimental evaluation. Finally, the conclusions and a summary are given in Section 6.

2 Training and Test Dataset

The Plant Identification task is based on the Pl@ntView dataset which focuses on 500 herb, tree and fern species centered on France (some plants observations are from neighboring countries) [6]. The complete dataset contains 60961 images belonging each to one of the 7 types of view reported into the meta-data, in a xml file (one per image) with explicit tags. The views are as follows: Scan (scan or scan-like pictures of leaf), photos of Flower, Fruit, Stem, Leaf, Branch and Entire views. On Figure 1 example images from each type of view are shown.

The distribution of training and test data of the Pl@ntView dataset is depicted in Table 1. As can be seen from the presented data, most of the images in the training and the test dataset are from the Flower view.

3 Feature Extraction and Image Description

For image description, we used the bag-of-visual-words approach [14], [4], [3]. It consists of three phases: creation of visual vocabulary, image description and

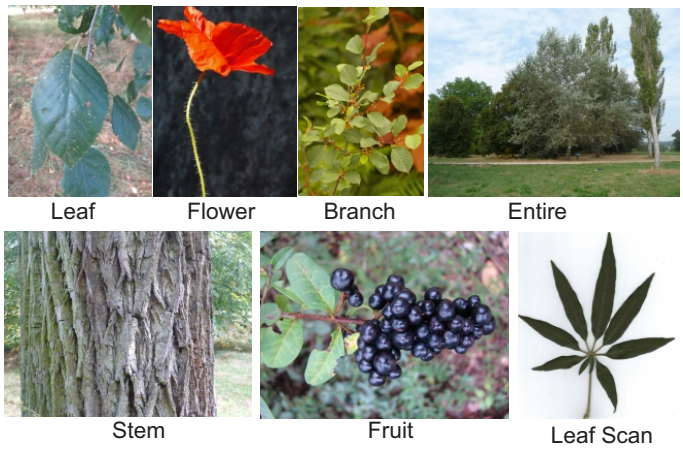

Fig. 1. Example images from each plant view/organ: Scan (scan or scan-like pictures of leaf), photos of Flower, Fruit, Stem, Leaf, Branch and Entire.

Table 1. Distribution of the images in the Pl@ntView dataset.

	Scan	Flower	Fruit	Stem	Leaf	Branch	Entire	
Training dataset	11335	13164	3753	3466	7754	1987	6356	47815
Test dataset	696	4559	1184	935	2058	731	2983	13146
Complete dataset	12031	17723	4937	4401	9812	2718	9339	60961

similarity definition. The creation of the visual vocabulary starts with detection of interesting points in the images, and then proceeds with extracting local invariant descriptors from them. Finally, the visual codebook is obtained by clustering the large set of descriptors obtained from all of the images. The resulting clusters represent the visual words, while all the visual words comprise the visual codebook. The image description phase assigns all of the local image descriptors to the visual words from the visual codebook. Each image is then described with a high-dimensional histogram and each component from the histogram is the number of descriptors that are assigned to a given visual word. Finally, the images are ranked using term frequency inverse document frequency (tf-idf) scores which reduce the influence of visual words which occur in many images. In the reminder of this section, we explain the phases in more details.

3.1 Image Processing and Feature Extraction

The images can be categorized in two groups. The first group is represented by scan and scan-like images, and in the second group are images from plants organs in natural surroundings, like branch, leaf, fruit, stem, flower and images from the entire plant. Having this in mind, we used two different feature extraction algorithms for the given images.

For the first group of images (scans and scan-like images of leaf) we used the triangle side lengths and angle (TSLA) descriptor from [10], [2]. TSLA is a multiscale triangular shape descriptor where the triangles are described by their lengths and an angle. Similar as in [2], the leaf contour in our experiments is described by 400 sample points, each point is represented by 10 triangles, with a distance d=5 between the triangle points at two successive scales. The TSLA descriptors require a preliminary leaf boundary extraction/segmentation of the image. In our experiments, we performed the boundary detection with the Otsu thresholding method [11]. The resulting descriptor for each image is a set of 400 points, each point represented with 30 float values.

For the second group of images, we used Opponent SIFTs as local descriptors extracted over the area around points of interest [13]. First, we extracted points of interest in the images using a Harris-Laplace interest point detector [9]. The Harris-Laplace detector uses the Harris corner detector to find scale-invariant interesting points. It then selects a subset of these points for which the Laplacian-of-Gaussians reaches a maximum over scale [9]. For the given set of images, especially (for the flowers and branch) more than 20000 points were sampled per image. In addition, a rhomboid-shaped mask was applied to the input image to minimize the effect of the cluttered background, and to reduce the number of points as in [2]. We kept only the points that were inside the applied mask. This assumption is justified because in most of the images the observed plant organ is placed in the center.

Secondly, over the area around points of interest, Opponent SIFT descriptors were extracted. Opponent SIFT describes all the channels in the opponent color space (eq. 1) using SIFT descriptors [8]. The information in the O_3 channel is equal to the intensity information, while the other channels (O_1 and O_2) describe the color information in the image. These other channels do contain some intensity information, but due to the normalization of the SIFT descriptor they are invariant to changes in light intensity. The R, G and B values in eq. 1 represent the channels of the RGB color space. The resulting descriptor for each image in this case is a set of 1000 points, each point represented with 384 integer values.

$$\begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R+G+B}{\sqrt{3}} \end{pmatrix} \tag{1}$$

3.2 Approximate k-means (AKM)

The construction of a visual codebook is an essential part of the bag-of-visual-words approach to image representation. For example, in our case, we are clustering more than 10M local descriptors into more than 200K clusters. Generating clusters from such a large quantity of data presents challenges to traditionally used algorithms like k-means [12]. As a alternative, we use approximate k-means.

In typical k-means, the vast majority of computation time is spent on calculating nearest neighbours between the points and cluster centers. We replace

this exact computation by an approximate nearest neighbor method, and use a forest of 8 randomized k-d trees built over the cluster centers at the beginning of each iteration to increase speed. We use the implementation from Philbin et al. [12]. This implementation uses randomized k-d tree code, optimized for matching SIFT descriptors [8]. Usually in a k-d tree, each node splits the dataset using the dimension with the highest variance for all the data points falling into that node and the splitting value is found by taking the median value along that dimension (although the mean can also be used). In the randomized version, the splitting dimension is chosen at random from among a set of the dimensions with highest variance and the split value is randomly chosen using a point close to the median.

The conjunction of these trees creates an overlapping partition of the feature space and helps to mitigate quantization effects, where features which fall close to a partition boundary are assigned to an incorrect nearest neighbour. This robustness is especially important in high-dimensions, where due to the "curse of dimensionality" [12], points will be more likely to lie close to a boundary. A new data point is assigned to the (approximately) closest cluster center as follows. Initially, each tree is traversed to a leaf and the distances to the discriminating boundaries are recorded in a single priority queue for all trees. Then, iteratively the most promising branch from all trees is chosen and keep adding unseen nodes into the priority queue. The stop criteria is the exploration of a fixed number of tree paths. This way, more trees can be used without significantly increasing the search time.

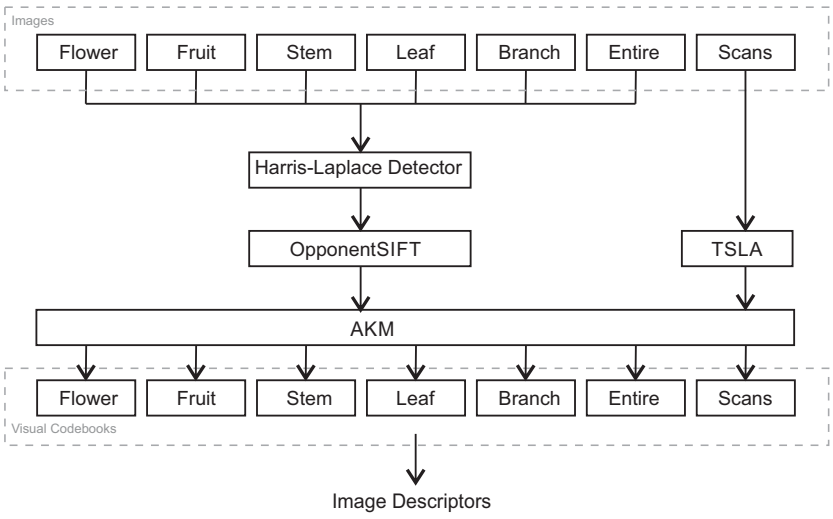

Fig. 2. The pipeline used for obtaining the image descriptors.

3.3 Image Description

The complete pipeline for extracting the visual descriptors, creating the visual codebook and obtaining the image descriptors is presented in Figure 2. The proposed pipeline for obtaining the image descriptors is as follows. First, we apply Harris-Laplace detector on the images (training and test) that belong to the second group of images (leaf, flower, fruit, stem, entire and branch) and generate opponent SIFT local descriptors around the detected points. For the first group of images (scans of leaves), we generate TSLA descriptors. Next, we use the generated descriptors to construct the visual codebooks. Note that, different visual codebooks were created for the different views of the plants (seven in total, one for each view). We randomly select a subset of the local descriptors (TSLA and opponent SIFT) from the training images for each view separately. The number of the local descriptors is varying from 6M for the scans to 10M for the flowers. We use these descriptors as a input to the approximate k-means algorithm to obtain the clusters/visual words that will constitute the visual codebooks. Finally, each image in the training and test dataset is represented as a sparse high-dimensional histogram of term (visual word) occurrences.

The similarity between two images is defined as a L_2 distance over the obtained histograms. We use the standard tf-idf weighting scheme [1], which down-weights the contribution that commonly occurring, and therefore less discriminative, words make to the similarity.

4 Information Fusion and Classification

For each run, we used the fact that images in the test dataset are associated with plant observations to perform multiple image queries for all image organs and scans having the same ObservationID value [2]. The overall process is presented in Figure 3. More precisely, for each descriptor:

- We first grouped all the images I_1, ..., I_k coming from the same plant observation using the ObservationID in metadata.
- Then, we computed similarity ranking lists of the retrieved images L_1, ..., L_k corresponding to the query images I_1, ..., I_k.
- Finally, the 300 first image results were kept for each list and were merged into a final list L using a late fusion scheme.

We used three different late fusion schemes to obtain the final predictions.

1. **Min. rank fusion**: For this fusion scheme, we used the Leave Out algorithm (LO) [7]. Lists L_1, ..., L_k are merged by setting the rank of an image to the minimum of the ranks in each list. Thus, the best position of an image among the returned lists is kept. The minimal ranks of the classes associated to the corresponding images are considered as a final predictions of the observations.

2. **Probability fusion**: For this fusion scheme, first the classes associated to the images from the lists L_1, ..., L_k are ranked per organ (i.e. scans), according to the average L_2 distance between the corresponding query images and the images from their ranked lists L_1, ..., L_k. We took into account only the best two ranked images of one observation. The final predictions (per observation) are obtained by calculating the minimal ranks of the classes.
3. **Mixed fusion**: This fusion scheme is a combination of the previous two schemes. In particular, for this setup we used min. rank fusion for the scans images and probability fusion for the organ images.

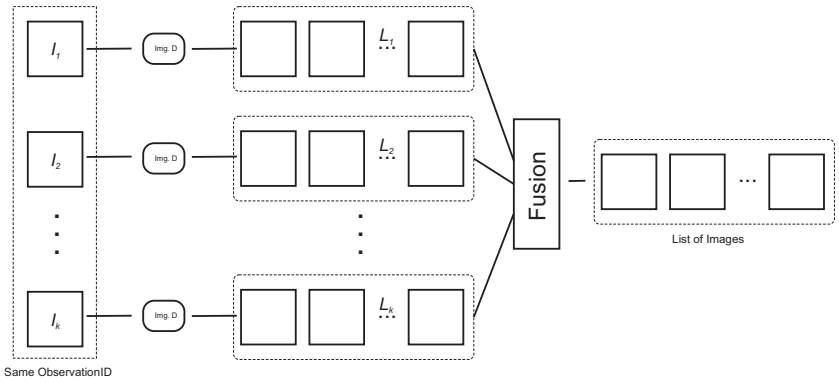

Fig. 3. Multiple image queries. I_1, ...I_k are leaf images associated to the same ObservationID and Img. D is the image descriptor either TSLA or Opponent SIFT.

5 Experimental Results

We submitted three runs for the LifeCLEF 2014 Plant Task. As we stated previously, the three runs rely on the same visual descriptors but we used different fusion schemes to obtain the final predictions. The results from the runs are presented in Tabale 2. The table contains the scores by image and observation. In our submitted runs, these two values for each run are the same. First we obtained the predictions for the observations and later on, we just apply these predictions for the images that are part of the corresponding observation.

The best performing run is the run named FINKI Run 1. This run is a combination of the other two runs. Namely, the predictions for the test images denoted with Leaf Scans were taken from the FINKI Run 2 and the predictions for the other images were taken from the run with name FINKI Run 3. We made this combination because in the validation phase, when we apply the algorithm on the ImageCLEF 2013 Plant Task, we obtained better results for the images denoted with Leaf Scan using the technique implemented in FINKI Run 1.

By comparing the second and third run in Table 2 we can conclude that taking into consideration the distribution of the images across the different species does help in boosting the predictive performance. The run named FINKI Run 3 has better score compared to the run with name FINKI Run 2.

Table 2. Scores per image and observation of the 3 runs submitted to the LifeCLEF 2014 Plant Task.

Run name	Run filename	Score Image	Score Observation
FINKI Run 1	run_maestra_per_image_mixed	0.205	0.205
FINKI Run 3	run_maestra_per_image_prob	0.204	0.204
FINKI Run 2	run_maestra_per_image_min_rank	0.166	0.166

In Table 3, we present the detailed scores obtained for each type of plant organs. The best results are obtained for the Leaf Scan images. This is to be expected because these images contain only leaves and are taken in very controlled environment, in most of the cases on a white sheet as a background. The second best score is obtained for the images with flowers. The lowest score is obtained for the images with branches and images that contain the entire plant. These images are most challenging in respect to the variant background and lightening conditions under which these images are taken.

Table 3. Distribution of the images in Pl@ntView dataset.

Run name	Branch	Entire	Flower	Fruit	Leaf	Leaf Scan	Stem
FINKI Run 1	0.088	0.117	0.255	0.177	0.160	0.400	0.157
FINKI Run 2	0.108	0.099	0.187	0.160	0.140	0.399	0.180
FINKI Run 3	0.088	0.117	0.255	0.177	0.162	0.360	0.159

Our best performing run was ranked fifth from 10 different participants/research group.

6 Summary and Discussion

We submitted three runs on LifeCLEF 2014 Plant Task. The image descriptors for all three runs are obtained using the bag-of-visual-words approach. For the leaf scans we are using multiscale triangular shape descriptor and for the other plant organs we are using Opponent SIFT extracted around points of interest obtained using Harris-Laplace detector. We are using approximate k-means (AKM) algorithm to cluster these descriptors in large number of clusters/visual words (approximately 200K). Each image in the training and test dataset is represented as a sparse high-dimensional histogram of term (visual word) occurrences. The similarity between two images is defined as a L_2 distance over

the obtained histograms. We use the standard tf-idf weighting scheme, which reduces the contribution that commonly occurring, and therefore less discriminative, words make to the similarity.

Applied on the LifeCLEF 2014 Plant Task our approach was ranked fifth, out of 10 competing groups. The approach we presented is general. We are planning to extend it with different image descriptors in order to tackle the different aspects of each plant organ/view. The inclusion of more image descriptors requires development of different and more complex weighting/fusion schemes.

Acknowledgment

We would like to acknowledge the support of the European Commission through the project MAESTRA - Learning from Massive, Incompletely annotated, and Structured Data (Grant number ICT-2013-612944).

References

1. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. ACM Press (1999)
2. Bakic, V., Mouine, S., Ouertani-Litayem, S., Verroust-Blondet, A., Yahiaoui, I., Goëau, H., Joly, A.: Inria's participation at ImageCLEF 2013 Plant Identification Task. In: CLEF (Online Working Notes/Labs/Workshop) 2013 (2013)
3. Dimitrovski, I., Kocev, D., Loskovska, S., Dzeroski, S.: Fast and scalable image retrieval using predictive clustering trees. In: Discovery Science. pp. 33–48 (2013)
4. Dimitrovski, I., Kocev, D., Loskovska, S., Dzeroski, S.: Fast and efficient visual codebook construction for multi-label annotation using predictive clustering trees. Pattern Recognition Letters 38, 38–45 (2014)
5. Goëau, H., Bonnet, P., Joly, A., Yahiaoui, I., Barthelemy, D., Boujemaa, N., Molino, J.F.: The imageclef 2012 plant identification task. In: CLEF (Online Working Notes/Labs/Workshop) (2012)
6. Goëau, H., Joly, A., Bonnet, P., Molino, J.F., Barthélémy, D., Boujemaa, N.: Lifeclef plant identification task 2014. In: CLEF working notes 2014 (2014)
7. Jovi, M., Hatakeyama, Y., Dong, F., Hirota, K.: Image retrieval based on similarity score fusion from feature similarity ranking lists. In: Wang, L., Jiao, L., Shi, G., Li, X., Liu, J. (eds.) Fuzzy Systems and Knowledge Discovery, Lecture Notes in Computer Science, vol. 4223, pp. 461–470. Springer Berlin Heidelberg (2006)
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
9. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V.: A comparison of affine region detectors. International Journal of Computer Vision 65(1-2), 43–72 (2005)
10. Mouine, S., Yahiaoui, I., Verroust-Blondet, A.: A shape-based approach for leaf classification using multiscaletriangular representation. In: Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval. pp. 127–134. ICMR '13 (2013)
11. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man and Cybernetics 9(1), 62–66 (1979)

12. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8 (2007)
13. van de Sande, K., Gevers, T., Snoek, C.: Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9), 1582–1596 (2010)
14. Sivic, J., Zisserman, A.: Video google: a text retrieval approach to object matching in videos. In: IEEE Conference on Computer Vision. pp. 1470–1477 (2003)