# Viewpoints combined classification method in image-based plant identification task

Gábor Szűcs[1], Dávid Papp[2], Dániel Lovas[2]

[1] Inter-University Centre for Telecommunications and Informatics, Kassai str. 26., H-4028, Debrecen, Hungary
[2] Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Magyar Tudósok krt. 2., H-1117, Budapest, Hungary,

szucs@tmit.bme.hu, pappdavid27@gmail.com, lovas.daniel@simonyi.bme.hu

**Abstract.** The image-based plant identification challenge was focused on tree, herbs and ferns species identification based on different types of images. The aim of the task was to produce relevant species for each observation of a plant of the test dataset. We have elaborated a viewpoints combined classification method for this challenge. We have applied dense SIFT for feature detection and description; and Gaussian Mixture Model based Fisher vector was calculated to represent an image with high-level descriptor. The chosen classifier was the C-support vector classification algorithm with RBF (Radial Basis Function) kernel, and we have optimized two hyperparameters (C from C-SVC and $\gamma$ from RBF kernel) by a grid search with two-dimensional grid. We have constructed a combined classifier using the weighted average of reliability values of classifier at each viewpoint. The results show that our combined method exceeds our best classifier among the list of classifiers constructed for different viewpoints.

**Keywords:** GMM based Fisher vector, C-support vector classification, viewpoint combination

## 1 Introduction

Accurate knowledge of the identity, statistics and uses of plants is essential in the agricultural development. Identifying plant species is usually a very difficult task, even for professionals (such as farmers or wood exploiters) or for the botanists themselves. Using image retrieval technologies is nowadays considered by botanists as a promising direction in this problem, and in order to solve it a challenge is announced in the LifeCLEF campaign [3].

The image-based plant identification task [7] was focused on tree, herbs and ferns species identification based on different types of images. There are 7 viewpoints at the images: branch, leaf, scan (scan or scan-like pictures of leaf, briefly "LeafScan"),

flower, fruit, stem, and entire views. The number of species was about 500, which is an important step towards covering the entire flora of a given region.

The aim of the task was to produce a list of relevant species for each observation of a plant of the test dataset, i.e. one or a set of several pictures related to a same event: one same person photographing several detailed views on various organs the same day with the same device with the same lightening conditions observing one same plant. So the task was observation-centered (not image-centered).

The task was based on the Pl@ntView dataset focusing plants on France (some plants observations came from neighbouring countries). It contains more than 60000 pictures belonging each to one of the 7 types of view reported into the meta-data, in an xml file (one per image) with explicit tags, like ObservationId, species names, date, etc.

The task was evaluated as a plant species retrieval task based on multi-image plant observations queries. The goal was to retrieve the correct plant species among the top results of a ranked list of species returned by the evaluated system. An observation may contain 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, and author). Each participating group was allowed to submit up to 4 runs built from different methods.

User rating information (pictures with the average of the user ratings on image quality) was also available, but we have not used this additional information.

## 2 Image-based plant classification

### 2.1 Elaboration of image descriptors

The first part of the classification is the accomplishment of representation of each image based on the visual content. This consists of three steps: (i) feature detection, (ii) feature description, (iii) image description as usual phases in computer vision.

Feature detection: Lots of different feature types can be detected in an image, e.g. corners, edges, ridges, as "interesting" part of an image. Furthermore many possible feature extraction methods are available for images, but we have chosen SIFT (Scale-Invariant Feature Transform) algorithm [11][12], because this is a widely used method in practice and in theoretical works (as well) with some possible further development of this method.

Feature description: In our solution we have used dense sampling method with SIFT (briefly dense SIFT). This sampling method can be considered as a two-dimensional grid upon the image, where SIFT descriptors were calculated at each grid point. After that we have used PCA (Principal Component Analysis) [1][9] to reduce the dimensions of the descriptor vectors from 128 to 80. This descriptor vector belongs to only one "interesting" point of an image, but an image possesses many feature descriptor vectors, which should be aggregated into an image descriptor.

Image description: The final step of the representation creating is the completion of high level representation of each image. We have applied BoW (bag-of-words) model [6][10] for this purpose, where images are treated as documents. According to this, "visual words" (so called "codewords") in images need to be defined from feature descriptors. The whole set of codewords gives the codebook (similarly to dictionary in text tasks). To determine the codebook we used GMM (Gaussian Mixture Model) [15][17]. This is a parametric probability density function represented as a weighted sum of (in our case 256) Gaussian component densities. GMM parameters were estimated based on the training set by using the iterative EM (Expectation Maximization) algorithm [5], but an initial model was needed for EM. In our training procedure the k-means clustering [13] was performed over all the vectors with 256 clusters, which resulted the initial model for EM. As a result of the algorithms described above, a codebook with 256 codewords was available for further calculations, which can be considered as a concise representation of the image set. According to the codebook the next step is to create a descriptor that specifies the distribution of the visual codewords in any image, called high-level descriptor. To represent an image with high-level descriptor, the GMM based Fisher vector [14][15] was calculated. These vectors were the final representation (image descriptor) of the images. The code used to train GMM vocabularies and compute the Fisher vectors is a standalone C++ library, developed by Jorge Sánchez, to support the research of Visual Geometry Group of Oxford University [8].

## 2.2 Training the classifier

For the classification task we have divided the labelled image set into three subsets: training, validation and test set (the last one is used for preliminary testing). The validation image set was used for calibration of the trained model during the validation phase of the training procedure. To train the classifier (classification model) based on training image set, a variation of SVM (Support Vector Machine) was used, the C-SVC (C-support vector classification) [2][4] with RBF (Radial Basis Function) kernel. The SVM is basically a binary linear classifier, thus in order to extend it to a number of classified categories, the one-against-all technique was used. During this method a binary classifier was created for each category in the training set.

The two hyperparameters (C from C-SVC and γ from RBF kernel) were optimized by a grid search with two-dimensional grid. The algorithm was trained with the training image set, and then validated on the validation set, while the hyperparameters were different in each iteration. The parameter pair that gave the best result is selected to train the final classification model (for each category) based on the whole image set.

## 2.3 Preliminary testing

After the training, the codebook was already available and only Fisher vector of each image should be computed. At the preliminary testing we have selected only 50 species (classes) for training and testing as well. RBF based kernel matrix was built from the Fisher vectors of the test and training images. Each C-SVC classifier was parametered with this matrix and the hyperparameters were the same as in the final classification models. Since the classifiers are assigned to species, the generated model for a classifier is responsible to separate the designated class from the other ones. Thus a classifier is able to provide a confidence value showing a certainty of the class in a given image.

We have trained 7 classifiers for each viewpoint and we have evaluated as preliminary testing based on precision and computer run time. The results of the preliminary testing can be seen in Table 1.

**Table 1.** Results of the preliminary testing

| **viewpoint** | **precision** | **testing time (per image) [sec]** |
|---|---|---|
| Branch | 0.341 | 1.82 |
| Leaf | 0.583 | 1.59 |
| LeafScan | 0.965 | 0.95 |
| Stem | 0.492 | 1.39 |
| Flower | 0.512 | 1.61 |
| Entire | 0.314 | 1.44 |
| Fruit | 0.482 | 1.56 |

## 2.4 Viewpoints combination for observation classification

The decision about the observation could be based on majority voting of image decisions, but we have used continuous information instead of discrete one. C-SVC classifier calculates continuous reliability value for each class at each image, and we have constructed a combined classifier using the weighted average of reliability values. Our

combined classifier has applied a formula (as can be seen in Equation 1.) for the aggregated reliability value that an image belongs to class *c* (species *c*).

$$R(c) = \frac{1}{\sum_{i=1}^{NVP} w_i} \sum_{i=1}^{NVP} w_i \cdot R_i(c) = \frac{1}{\sum_{i=1}^{7} w_i} \sum_{i=1}^{7} w_i \frac{1}{N_{i,p}} \sum_{n=1}^{N_{i,p}} r_n(c) \quad (1)$$

- NVP is the number of viewpoints, which equals to seven in this challenge
- $w_i$ is the weight parameter of viewpoint *i*
- $r_n(c)$ is reliability value for class *c* coming from C-SVC classifier
- $N_{i,p}$ is the number of images in viewpoint *i* taken from the p-th plant observed

Based on *R(c)* values the final decision is always the species that possesses the largest *R(c)* value. In the challenge the order of predicted species should have been submitted, and we have constructed the order based on *R(c)* values as well.

At the estimation of weight parameters we have taken the goodness of different viewpoint classifiers into the consideration. As can be seen in the results of the preliminary testing (at Table 1), the LeafScan has the best precision. So the LeafScan has got the largest weight parameter, and on an empirical way we have chosen the following weight parameters: LeafScan: 7.5, Leaf: 2.5, Flower: 1.5, Fruit: 1.5, Stem: 1.5, Branch: 1.5, Entire: 1.5.

## 3 Evaluation

### 3.1 Evaluation metrics

In the official evaluation instead of precision (as used in our preliminary testing) a new evaluation metric was defined for measurement of goodness of the observation classification. This metric (S score) is defined as follows.

$$S = \frac{1}{U} \sum_{u=1}^{U} \frac{1}{P_u} \sum_{p=1}^{P_u} S_{u,p} \quad (2)$$

- U : number of users (who have at least one image in the test data)
- Pu : number of individual plants observed by the u-th user
- Nu,p : number of pictures taken from the p-th plant observed by the u-th user
- Su,p : score between 1 and 0 equals to the inverse of the rank of the correct species (for the p-th plant observed by the u-th user)

Although the goal was to classify the observations containing more images, an additional metric was defined for the image classification as can be seen in Equation 3.

$$S_{image} = \frac{1}{U}\sum_{u=1}^{U}\frac{1}{P_u}\sum_{p=1}^{P_u}\frac{1}{N_{u,p}}\sum_{n=1}^{N_{u,p}}S_{u,p,n} \tag{3}$$

- U : number of users (who have at least one image in the test data)
- Pu : number of individual plants observed by the u-th user
- Nu,p : number of pictures taken from the p-th plant observed by the u-th user
- Su,p,n : score between 1 and 0 equals to the inverse of the rank of the correct species (for the n-th picture taken from the p-th plant observed by the u-th user)

### 3.2 Final official results

$S_{image}$ score can be calculated for each viewpoint, and these scores can be compared. Our final official results for each viewpoint and the observation can be seen in Table 2., and it can be shown that S score of observation exceeds the best S score of all viewpoints.

**Table 2.** Our final official results

| **viewpoints and observation** | **S score** |
|---|---|
| Branch | 0.052 |
| Leaf | 0.019 |
| LeafScan | 0.119 |
| Stem | 0.072 |
| Flower | 0.115 |
| Entire | 0.06 |
| Fruit | 0.07 |
| **Observation** | **0.255** |

Our final official observation results (BME TMIT) compared with other participants can be seen in Fig. 1.

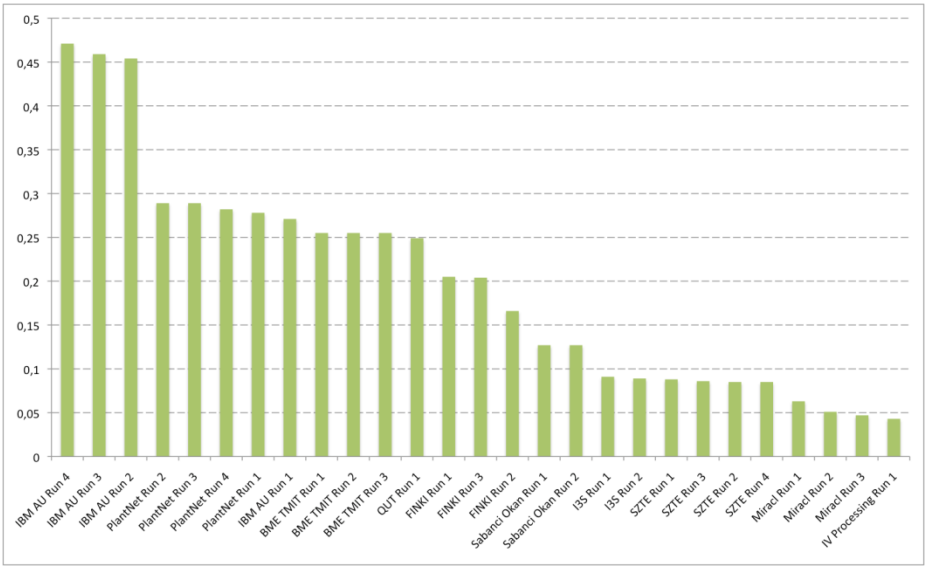


**Fig. 1.** Final official observation results of participants

## 4 Conclusion

We have elaborated a viewpoints combined classification method for image-based plant identification task. We have applied dense SIFT for feature detection and description; and Gaussian Mixture Model based Fisher vector was calculated to represent an image with high-level descriptor. The chosen classifier was the C-support vector classification algorithm with RBF (Radial Basis Function) kernel, and we have optimized two hyperparameters (C from C-SVC and $\gamma$ from RBF kernel) by a grid search with two-dimensional grid. We have constructed a combined classifier using the weighted average of reliability values of classifier at each viewpoint. The weight parameters of the combined classifier were based on our preliminary testing results. Our observation result of the combined method exceeds our best score of all viewpoints. At the official evaluation our solution has reached 0.255 score value.

## Acknowledgement

The publication was supported by the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project. The project has been supported by the European Union, co-financed by the European Social Fund.

## References


1. Abdi H., Williams L. J.: Principal Component Analysis, *Wiley Interdisciplinary Reviews: Computational Statistics*, Vol 2. No. 4, pp. 433-459 (2010)
2. Boser, B., Guyon, I., Vapnik, V.: *A Training Algorithm for Optimal Margin Classifier, Proc. of the 5th Annual ACM Workshop on Computational Learning Theory*, pp. 144-152 (1992)
3. Joly, A., Müller, H., Goëau, H., Glotin, H., Spampinato, C., Rauber, A., Bonnet, P., Vellinga, W.P., Fisher, B.: Lifeclef 2014: multimedia life species identification challenges. In: *Proceedings of CLEF 2014* (2014)
4. Cortes, C., Vapnik, V.: *Support-vector networks, Machine Learning*, Vol. 20, No. 3, pp. 273-297 (1995)
5. Dempster A., Laird N., Rubin D.: Maximum likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society*, Vol. 39, No. 1, pp. 1-38 (1977)
6. Fei-Fei, L., Fergus, R., & A. Torralba, A.: *Recognizing and Learning Object Categories, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, (2007)
7. Goeau, H., Joly, A., Bonnet, P., Selmi, S., Molino, J.F., Barthélémy, D., Boujemaa, N.: Lifeclef plant identifcation task 2014. In: *CLEF working notes 2014* (2014)
8. K. Chatfield, V. Lempitsky, A. Vedaldi and A. Zisserman.: The devil is in the details: an evaluation of recent feature encoding methods, *British Machine Vision Conference*, pp. 76.1-76.12, (2011)
9. Ke, Y., & Sukthankar, R.: PCA-SIFT: A more distinctive representation for local image descriptors, In *Computer Vision and Pattern Recognition, CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, Vol. 2, pp. II-506. (2004)
10. Lazebnik, S., Schmid, C. and Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, New York, Vol. 2, pp. 2169-2178 (2006)
11. Lowe, D. G.: *Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision*, Vol. 60, No 2., pp. 91-110 (2004)
12. Lowe, D. G.: Object Recognition from local scale-invariant features, In *International Conference on Computer Vision*, Corfu, Greece, pp. 1150-1157 (1999)
13. MacQueen, J.: Some methods for classification and analysis of multivariate observations, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 281-297 (1967)
14. Perronnin, F., Dance, C.: Fisher kernel on visual vocabularies for image categorization, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, (2007)
15. Reynolds D. A.: Gaussian Mixture Models, *Encyclopedia of Biometric Recognition*, Springer, February, pp. 659-663 (2009)
16. Sánchez, J. Perronnin, F., Mensink, T.: *Improved Fisher Vector for Large Scale Image Classification, In Proc. of the 11th European Conference on Computer Vision (ECCV): Part IV*, September 05-11, pp. 143-156 (2010)
17. Tomasi C.: Estimating gaussian mixture densities with EM: A tutorial, *(Tech. rep., Duke University); Chinese Journal of Electron Devices*, pp, 15-18 (2004)