

Identification of plant species on large botanical image datasets

Naiara Aginako
Vicomtech-IK4
Paseo Mikeletegi 57
20009 Donostia-San Sebastián
+34943309230
naginako@vicomtech.org

Javier Lozano
Vicomtech-IK4
Paseo Mikeletegi 57
20009 Donostia-San Sebastián
+34943309230
jlozano@vicomtech.org

Marco Quartulli
Vicomtech-IK4
Paseo Mikeletegi 57
20009 Donostia-San Sebastián
+34943309230
mquartulli@vicomtech.org

Basilio Sierra
Computer Sciences and
Artificial Intelligence Department
University of the Basque Country
+34943015102
b.sierra@ehu.es

Igor G. Olaizola
Vicomtech-IK4
Paseo Mikeletegi 57
20009 Donostia-San Sebastián
+34943309230
iolaizola@vicomtech.org

ABSTRACT

The continuously growing amount of multimedia content has enabled the application of image content retrieval solutions to different domains. Botanical scientists are working on the classification of plant species in order to infer the relevant knowledge that permits them going forward in their environmental researches. The manual annotation of the existing and newly creation plants datasets is an outsized task that is becoming more and more tedious with the daily incorporation of new images. In this paper we present an automatic system for the identification of plants based on not only the content of images but also on the metadata associated to them. The classification has been defined as a classification plus fusion solution, where the images representing different parts of a plant have been considered independently. The promising results bring to light the chances of the application computer vision solutions to botanical domain.

1. INTRODUCTION

The digital age has brought the development of new technologies that allow making deeper studies about our reality and therefore, winning a more exhaustive knowledge. In addition, the ever increasing use of digital cameras and sensors in several fields, has led to an exponential growth in the amount of multimedia content being generated every day in the world. Nowadays, the whole society is involved in the generation of any kind of content; it's already a fact that digital technologies are introduced in all aspects of our daily lives.

Although the multimedia analysis techniques in their beginning were focused on application sectors directly related with the technology, their penetration in divergent sectors such as medicine, meteorology, environment it's a reality that is bringing huge progress.

Regarding environmental multimedia content, there is an increasing need of techniques for analyzing, interpreting and labelling of the content in order to enrich the actual knowledge. This automatically extracted knowledge leads to the adoption of new strategies that can improve the actual insight of the environment to move forward in the deployment of new directives to help in its protection and care.

Initiatives such as Tela Botanica and projects such as Pl@ntNet foster the development of this kind of technologies. Even more, open competitions as ImageCLEF[1], and more precisely plant identification task [2], where technological researchers focused on multimedia content analysis take part, promote the approach of these two worlds. Newborn mobile applications such as Plantifier, LeafSnap or NatureGate are also examples of the natural synergy tendencies.

The image-based identification of different species of plants that both botanical scientists and expert users have collected has become a key study among plant biology science. On the one hand, one of the peculiarities of plant image analysis is that such images may belong to different plant parts such as leaf, stem or flower. On the other hand, content is also time dependent, thereby increasing the difficulty of the identification task. The latter can be mitigated by using not only image content but also the linked metadata. Thus, the analysis process is enriched and more accurate results can be obtained. This metadata is not only the data that users can add manually but information that nowadays digital cameras impress automatically.

One of the biggest handicaps of multimedia content analysis is to determine the working domain so that afterwards, more domain specific implementations are applied. In the case of ImageCLEF dataset, there is a division of 6 subcategories that identify these domains. Each image has an associated XML which specifies what subcategory belongs to, permitting the abstraction from the domain categorization issue.

In our plant identification approach we used ImageCLEF dataset. This competition was first turned up in 2003. Since then, it has become a benchmark platform for the evaluation of image annotation and retrieval algorithms in several domains such as medical imagery, robot vision imagery or botanical collections. This year, a new lab dedicated to life media LifeCLEF which includes plant identification task has been released. In the past

Copyright © by the paper's authors. Copying permitted only for private and academic purposes.

In: S. Vrochidis, K. Karatzas, A. Karpinnen, A. Joly (eds.): Proceedings of the International Workshop on Environmental Multimedia Retrieval (EMR2014), Glasgow, UK, April 1, 2014, published at <http://ceur-ws.org>

edition, 2013, there were 33 submitted runs. Training data resulted in 20985 images while testing data resulted in 5092.

The rest of the paper is organized as follows: section 2 describes category dependent image analysis (section 2), divided into two subsections that go in depth in the metadata analysis (section 2.1) and in the image content analysis (section 2.2). Section 3 is focused on the classification algorithms for the plant identification purpose while in section 4 fusion and merging methodologies are described. We conclude with a summarization of the obtained results (section 4), pointing out the challenges ahead for the use of content based retrieval technologies in botanical domain.

2. CATEGORY DEPENDENT IMAGE ANALYSIS

As mentioned in the prior sections, the available dataset for ImageCLEF2013 Plant Identification Task is segmented into 2 main categories, *NaturalBackground* and *SheetAsBackground*, that are also divided into several sub-categories: Scan and Scan-like for *SheetAsBackground* category, that are considered equally in our system, and *Leaf*, *Flower*, *Fruit*, *Stem* and *Entire* for *NaturalBackground* category. Both training and testing images have an associated XML describing their metadata that permits the system to separate the images into groups for the later processing and classification.

This subcategory based groups are the key units of the overall plant identification process until the merging done taking into account the Individual Plant Identification, a metadata parameter that determines images that belong to the same plant. For each of the subcategories or groups is necessary to extract all the relevant knowledge. First, inferring this knowledge from the metadata such as localization and date and second, describing the content of images as in detail as possible and using discriminative factors. Not all the implementations have been considered for all the groups, taken decisions permit obtaining better results.

In the next subsections, more detailed explanations are presented regarding the metadata analysis and the deployed image content description algorithms.

2.1 Image metadata analysis: georeference and seasonal nature

Considering the metadata information attached to each of the images, we determine the inclusion of two metadata parameters: GPS data and the date in order to extract knowledge that can improve the plant identification process. These parameters are included not only for the training dataset but also for the testing dataset.

The schema of categories and subcategories of the image dataset delimits the use of these metadata parameters to the *Natural Background* category. Images included within *SheetAsBackground* category don't belong to natural environments; consequently, their latitude, longitude and date parameters don't represent the plant ecosystem. Including these data in the classification process can insert too much noise in the system preventing good results.

2.1.1 Georeferenced data

Since ancient times, studies to determine the influence of topography on species identification have been done. One of the most important factors is the altitude at which each species grows. Therefore, altitude has been considered one of the key indicators for the classification process. Altitude values have been extracted using the actual digital elevation model (DEM) for Europe as the

vast majority of the images belong to France. The inputs to these models are the latitude and longitude data (GPS data).

In this case, the classification process has been focused in the analysis of the altitude parameter, not taking into account longitude and latitude variables as we judge that it could increase the noise level as all the images pertain to a specific country.

2.1.2 Seasonal nature classification

The plants are species that change throughout the seasons. Although not all plants undergo this change that doesn't affect to different parts of the plants in the same way, this seasonal concept has been considered an important factor that can be determinant in recognizing the plant. As a consequence, date metadata parameter has been added to the classification attribute list.

2.2 Image content analysis

After analyzing the metadata associated to each image, only its content can contribute some meaningful information to improve the classification process. We implemented approaches based on algorithms for the extraction of global characteristics such as DITEC[3]; textural characteristics such as Haralick [4], Zernike [5][6] and Local Binary Patterns (LPB)[7]; and parameters to characterize the principal object of the image calculating its solidity, eccentricity, dominant colour and area-perimeter relationship.

DITEC algorithm is based on the statistical modelling of the Trace Transform for global image description. Its main strength lays on the capacity for the description of an image as it extracts the most robust features for the interpretation of the content. This algorithm provides highly discriminative global descriptors at very low dimensionality.

Textural characteristics are very meaningful when the elements representing the image have texture. This kind of features give information about the spatial arrangement of colour or intensities and are very present in natural scenes. When using these attributes we intend to find repeated spatial patterns in images to make a distinction between them.

The identification of an image principal object brings the capacity to describe the overall image more accurately. Doing a good segmentation of the object is crucial since the parameters that are being extracted are totally related with this object withdrawing all the other elements that form the image. Parameters associated to the description of principal objects have been applied only in the case of *SheetAsBackground* and *Fruit* categories.

Regarding *SheetAsBackground*, is the shape of the object that represents the leaf which best discriminates between different species. Different measure properties of image regions related with shape description have been applied:

- **Eccentricity:** It is the ratio of the distance between the foci of the ellipse and its major axis length.
- **Solidity:** Scalar specifying the proportion of the pixels in the convex hull that are also in the region.
- **Area-perimeter relationship:** The number of pixels that belong to the area of the object divided by the number of pixels belonging to the object perimeter.

Even though, we didn't consider it a very discriminative parameter we also added the dominant colour of the segmented object.

Concerning *Fruit* subcategory, as the segmentation process was not as accurate as in the previous case because the photos had been taken in real scenarios, only dominant colour parameter was

extracted as it's also a factor that can make the difference between different types of plants.

2.2.1 Segmentation

Although usually image segmentation has a crucial significance for content description, as mentioned before, our system only uses it for the *SheetAsBackground* category and *Fruit* subcategory. In the first case, an isolated leaf is represented in the image with uneven illumination and possible shadows. We implemented colour clustering techniques based on Local Relative Entropy Method (LRE) [9] for the subtraction of the background. As this background doesn't represent a real scenario, the results for the segmentation of this uniform area are promising and therefore, valid for the implementation of an automatic segmentation process.

In the case of *Fruit* image segmentation, the assumption about the importance of the flower object itself carries the necessity of isolating it from the forest background. As well as in the previous approximation, colour clustering techniques based on Joint Relative Entropy method (JRE) [10] are used.

Even more, we observe that *Stem* subcategory contains predominantly images with tree trunks both in vertical and horizontal that fill the majority of the image. Hence, in order to minimize the effect of the insertion of noisy backgrounds to the system, four fifths of the images are cropped in a fixed direction. To determine this orientation of the trunk along the image, local gradients are analyzed.

3. PLANT CLASSIFICATION

All the image content retrieval solutions include a classification stage where data mining algorithms are implemented. These algorithms are necessary to infer knowledge from the extracted

features. Five different algorithms have been studied with the aim of determining the best one for each of the subcategories: Bayesian Network [11], Naive Bayes [12], SMO [13], SVM [14] and Kstar [15]. For the comparison between classification algorithms, training dataset is split into two subsets, one for the training and the other one for the validation of the implementation. KNIME [16] is an appropriate framework to carry out this learning approach and for the experimentation with a range of algorithms and parameterization of them. It permits working with several feature spaces at a time, therefore it a very suitable framework to undertake the evaluation of the algorithms with the best performance.

As starting point, we considered the classification as totally independent problem for each of the subcategories. The interdependency between some of the images has not been taking into account till the merging of the results. Most suitable features (see section 5) are extracted from all the images belonging to the same subcategory and they are gathered into five groups when all present. Each of the group is also considered an independent classification approach; therefore, the overall classification process is atomized as a subcategory classifications solution based on feature associations.

For the learning of the classification algorithms the training subset of images has been used and we validate the performance of the five implemented classification algorithms using the validation subset. As a result, we got at most five classification modules per category for each feature group. These modules output is a *ClassID* probability list that represents the probability of each image to belong to a plant species.

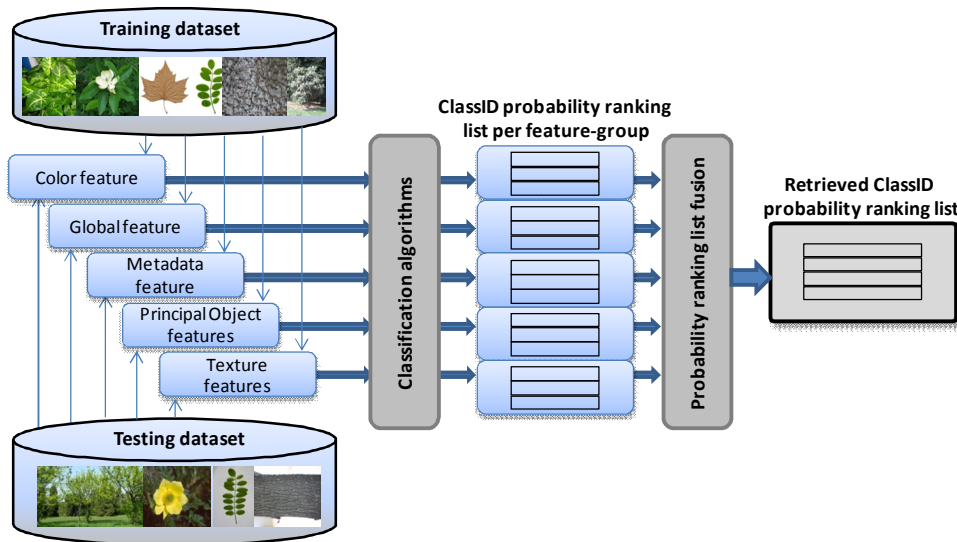


Figure 1 Computation of plant identification based on feature space level fusion

4. FUSION AND MERGING OF CLASSIFICATION RESULTS

We grouped the extracted features in five different groups to analyze their relevance in the identification task results. In general, most of the Content Based Image Retrieval (CBIR)

systems employ a unique probability output to determine the belonging class of a new query image. Multiple feature fusion is a classical technique used in CBIR and pattern recognition to improve the efficiency and robustness of results but this fusion is usually done to feature level. As an alternative to this, we propose an approach that computes the fusion of the classification results at feature space level. Probability scores lists for each of the

feature group are fused using a Leave Out algorithm (LO) [8]. Despite the algorithm was defined for its application using similarity scores, the adoption to probability lists is direct.

For the plant identification of a new query image, its features are extracted taking into account the aforementioned five feature spaces: colour, principal object, texture, global (DITEC) and metadata (see Figure 1). Classification modules have been already trained at feature space level so each feature group vector is classified by the corresponding classifier. As the output of this classification stage, we get a *ClassID* probability ranking list that denotes the probability for that image to belong to each of the plant classes regarding a concrete feature space.

In order to get a unique output, these probability lists are fused by setting the probability of an image belonging to a class to the maximum of the probabilities in each list. The resulting probability list represents the ranking for the plant identification *ClassID*.

$$M_{ij} = \text{Prob.}(\text{img}_{\text{feature space } i} \in \text{ID}_{W_j})$$

where; $\text{ID}_{W_j} = \text{sort}\{\text{Prob}(\text{img} \in \overline{\text{ID}})\}$ $\overline{\text{ID}} = \{\text{ClassID}\}$

M_{ij} matrix is composed of cells representing a tupla that contains the *ClassID* and the probability value of pertaining to that class. Each of the columns represents the probability ranking list for each of the feature spaces.

$$\vec{R}_{jDC} = \text{sort}\{\vec{M}_j\} \quad j = 1, \dots, J$$

\vec{R}_{jDC} vector represent the retrieved *ClassID* probability ranking list (see Figure 1).

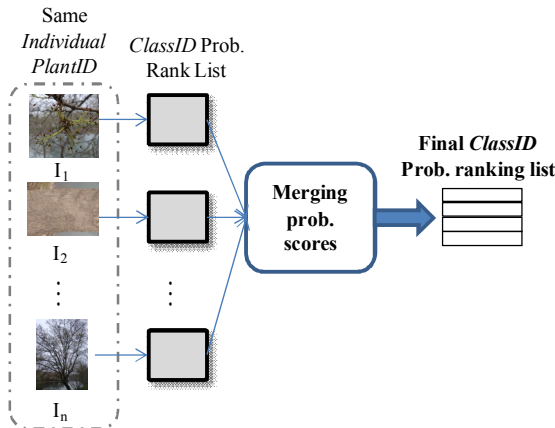


Figure 2 Merging of *ClassID* identification results for images belonging to the same plant (same *IndividualPlantID*)

But there is another fact that must be taken into consideration when estimating classification results: ImageCLEF dataset includes a metadata that must be considered during the plant identification; it is the *IndividualPlantID* which represents an exclusive number identifying images taken from the same plant. Therefore, there is a need of merging results coming from the same plant (see Figure 2). The *ClassID* probability lists belonging to the same plant are merged by means of empirically obtained weights for each of the subcategories.

$$\vec{W}_{SC} = \text{mean}(\overline{\text{Accuracy}}|_{\text{ClassID}})|_{\{T_{SC}\}}$$

where $\{T_{SC}\}$ is the group of images selected for the validation of the classification modules and the definition of the weights.

First, retrieved *ClassID* probability lists (\vec{R}_{jDC}) with the same *IndividualPlantID* are gathered. Taking into consideration the subcategory that the images belong to, probabilities are multiplied by a factor that has been deduced from the performance of the system for each of the subcategories (\vec{W}_{SC}). More precisely, the weight represents the mean accuracy value of the two best classification methods for each of the subcategories. In order to infer this value training dataset has been split into two sets, one for the training and the other for the validation of the classification system. *SheetAsBackground* and *Flower* subcategories are the ones with the highest weight while *Stem* and *Entire* have rather lower values.

Table 1 Weight values for each of the subcategories

SaB	Flower	Fruit	Leaf	Stem	Entire
0,6	0,562	0,409	0,534	0,175	0,07

Second, weighted probability lists are merged by means of the highest probability score that will determine the *ClassID* of the images with the same *IndividualPlantID*.

$$\vec{V}_{kj} = W_{Type} \cdot R_{1kj}|_{\{k\}}; \quad k = \forall \text{img} \cap \text{PlantID}_z$$

$$\text{Prob}(\text{img} \in \text{ClassID}) = \max(\vec{V}_{kj}|_{\{\text{ClassID}\}})$$

5. RESULTS

In order to validate the influence of each of the extracted features in the overall process of plant identification we considered to analyze the results of the classification process for each of the subcategories. The results presented in this section are the rate of correct predictions for each of the subcategories. These prediction results have been computed using only the training dataset, splitting this dataset into two sets, 90% of the images for the training of the classification and fusion modules and the other 10% for the validation.

As summarized in **Table 2**, not all the features have been contemplated for all the subcategories, as an example aforementioned associated metadata has not been included in the classification of images pertaining to *SheetAsBackground* category. In addition, all the extracted attributes concerning the identification of the principal object of the image such as the solidity, eccentricity or area-perimeter relationship has only been rated for the *SheetAsBackground* category. By contrast, principal object dominant colour attribute is extracted from both *Flower* and *SheetAsBackground* categories.

Concerning *Leaf* and *Stem* subcategories, metadata, textural and DITEC attributes have been included as the most representative features. As there is no a clear principal object in the image and the colour is not something characteristic other attributes were not considered.

In the case of *Entire* subcategory, images contain the entire natural scene where the plants grow, so the elements of the image are very diverse. This fact introduces lots of noise in the system and the classification of this subcategory is considered the most ambitious. In this case, metadata features and DITEC have been selected for the description.

Fruit and *Flower* are the subcategories where image colouring is a leading figure. Hence, for both subcategories metadata and colour attributes are extracted. Even at first it was considered to add the dominant colour attribute to both cases, due to the weak

results of the segmentation algorithms for *Flower* images we dismiss that possibility and it was only included for *Fruit*. The

opposite of textural features, that are more descriptive in the case of *Flower* subcategory.

Table 2 Subcategory dependent image description features extraction

Subcategory	Texture features			Metadata features		Principal Object features				Global feature	Color
	LBP	Zern.	Haral.	Data	Geo.	Dom. Co.	Sol.	AP rel.	Ecc.	DITEC	HSV
SaB	x	x	x	x	x	x	✓	✓	✓	✓	✓
Flower	✓	✓	✓	✓	✓	x	x	x	x	x	✓
Fruit	x	x	x	✓	✓	✓	x	x	x	x	✓
Leaf	✓	✓	✓	✓	✓	x	x	x	x	✓	x
Stem	✓	✓	✓	✓	✓	x	x	x	x	✓	x
Entire	x	x	x	✓	✓	x	x	x	x	✓	x

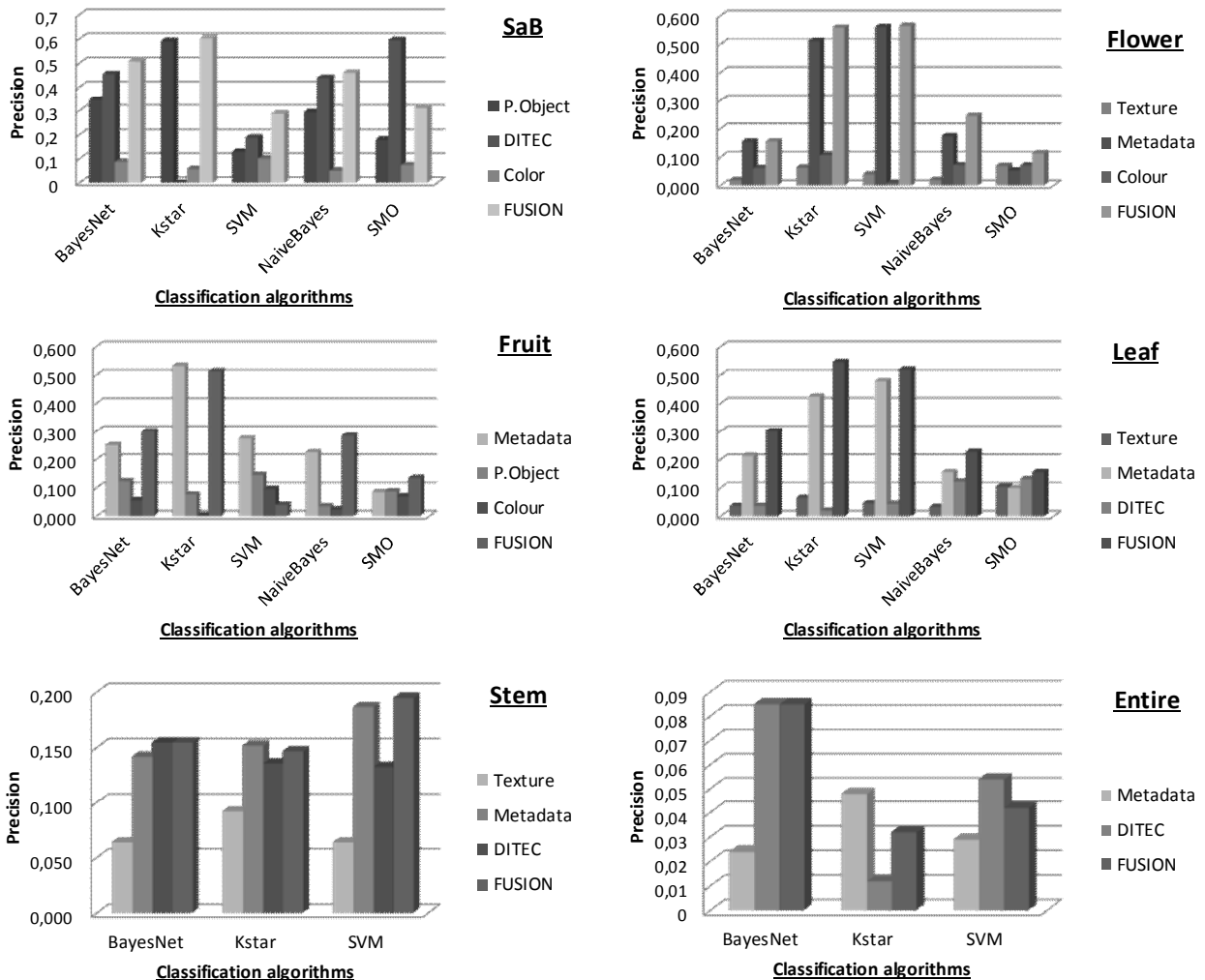


Figure 3 Precision results for each of the subcategories and using different classification algorithms. FUSION label represent the precision results after the application of probability list fusion.

In Figure 3 we resume the results obtained for the classification process visualized separately for each subcategory. For *Flower*, *Fruit* and *Leaf* categories metadata attributes are the ones with the best precision rates. The results for the *SaB*, *Flower*, *Fruit* and *Leaf* categories are quite promising while *Stem* and *Entire* classification doesn't give very good results. In the case of *Entire* category, the inclusion of very diverse elements in the images can distort the general perception of the plant itself and therefore identification task becomes quite difficult. However, if we consider the *Stem* category, we conclude that the extracted features are not feasible for the identification of this type of images.

In general, fusion algorithms increase precision results so a deeper analysis of the consequences of the utilization of these approaches is recommended for plant identification solutions.

5.1 Comparison with ImageCLEF official results

In this subsection some comparative indicators about the results obtained with the method presented in this paper and the overall results of ImageCLEF participants are presented. ImageCLEF results are divided into two different blocks: one of them including only image from *SheetAsBackground* category and the other one for the rest of the dataset images considered as *NaturalBackground* category. All the values for the final validation have been computed only for the testing dataset.

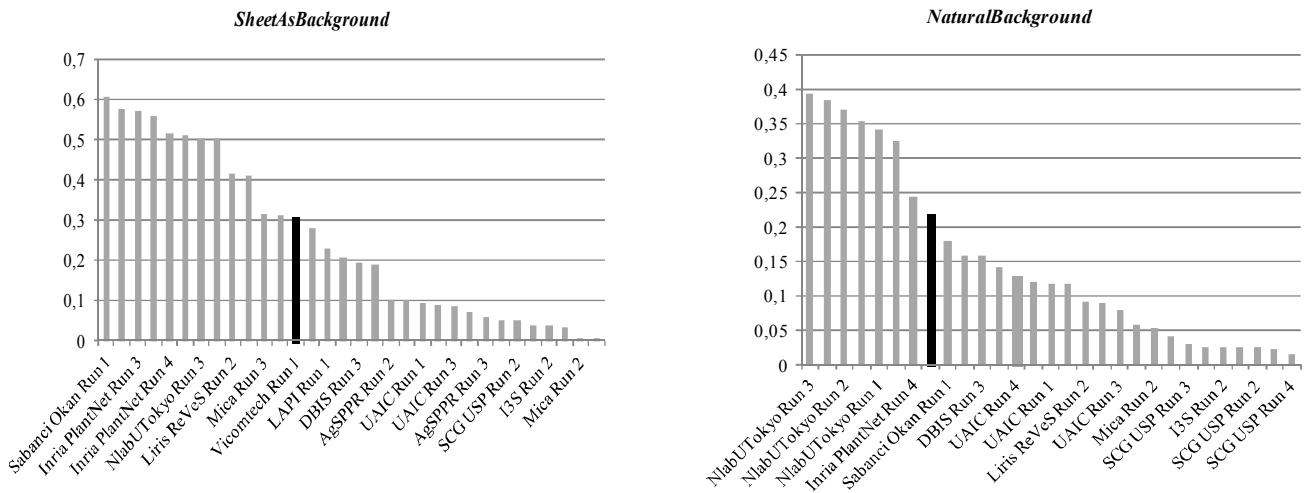


Figure 5 Scores obtained for *SheetAsBackground* and *NaturalBackground* categories

As appreciated in the figures, the results obtained with the described method are among the first half of the participants. In the case of *SheetAsBackground* category, more emphasis must be done in the segmentation process in order to have a better defined content for the analysis.

The bad results obtained for *Stem* and *Entire* subcategories have a direct influence in the scores of the *NaturalBackground* category so better approaches for the classification of these two subcategories are going to be implemented in the near future.

6. CONCLUSION

This paper presents a system for the identification of several plant species based on the analysis of metadata associated to an image and the content of the image. The inclusion of metadata

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

Figure 4 Primary metric used to evaluate the submitted runs in plant identification task of ImageCLEF 2013

As shown in Figure 4 the metric is a score related to the rank of the correct species in the list of retrieved species, where,

- **U** : number of users (who have at least one image in the test data)
- **P_u** : number of individual plants observed by the u-th user
- **N_{u,p}** : number of pictures taken from the p-th plant observed by the u-th user
- **S_{u,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)

In the following figures, highlighted in the graphics, the results of the described method for both categories compared with the results of all the participants of ImageCLEF 2013.

parameters reveals an opportunity to refine the results of the image content analysis. Even the described system has been proved for ImageCLEF dataset, the approaches defined in this paper are applicable to collections that contain plant images, only the categorization of plant parts should be kept in mind.

Concerning technical aspect of the system, remark the need of the inclusion of new algorithms that overcome the actual results especially for *Entire* and *Stem* categories. Additionally, merging strategies should consider the insertion of unique image instance identifiers previously in the classification process.

The growing botanical collections ease the inclusion of image retrieval solutions which are considered as very promising by experimented scientists. Competitions such as ImageCLEF are key factors on the approach between image analysis research groups and botanists which permits faster scientific discovery.

Having an accurate knowledge about the identity of plant species is essential for our biodiversity conservation.

7. ACKNOWLEDGMENTS

Our thanks to ImageCLEF organizers and all members of Pl@ntNet project and Tela Botanica initiative who brought us the possibility of researching on the application of multimedia analysis techniques applied to environmental data.

8. REFERENCES

- [1] Caputo, Barbara and Müller, Henning and Thomee, Bart and Villegas, Mauricio and Paredes, Roberto and Zellhofer, David and Gobeau, Herve and Joly, Alexis and Bonnet, Pierre and Martínez-Gómez, Jesus and García-Varea, Ismael and Cazorla, Miguel. ImageCLEF 2013: *The Vision, the Data and the Open Challenges*. Springer, *Lecture Notes in Computer Science*, p. 250-268, 2013
- [2] Gobeau, Herve and Joly, Alexis and Bonnet, Pierre and Bakic, Vera and Bartheemy, Daniel and Boujemaa, Nozha and Molino, Jean-Francois. The imageCLEF 2013 Plant Identification Task, *CLEF 2013 Evaluation Labs and Workshop, Online Working Note*, 2013.
- [3] Olaizola, I.G; Quartulli, M.; Florez, J.; Sierra, B. Trace Transform Based Method for Color Image Domain Identification. 2014. *Multimedia, IEEE Transactions on*
- [4] Haralick, R.M. and Shanmugam, K. and Dinstein, Its'Hak, 1973. Textural Features for Image Classification. *Systems, Man and Cybernetics, IEEE Transactions on*, SMC-3 610-621.
- [5] Hu, Ming-Kuei. Visual pattern recognition by moment invariants. *Information Theory, IRE Transactions on*, 8.2 (1962): 179-187.
- [6] Teague, Michael Reed. Image analysis via the general theory of moments*. *JOSA* 70.8 (1980): 920-930.
- [7] DC. He and L. Wang (1990), Texture Unit, Texture Spectrum, And Texture Analysis, *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 28, pp. 509 - 512.
- [8] Jović, Mladen and Hatakeyama, Yutaka and Dong, Fangyan and Hirota, Kaoru. Image Retrieval Based on Similarity Score Fusion from Feature Similarity Ranking Lists, *Springer Berlin Heidelberg*, 4223, 2006, pp. 461-470, DOI: http://dx.doi.org/10.1007/11881599_54
- [9] Chein-I Chang, Kebo Chen, Jianwei Wang and Mark L. G. Althouse, A relative entropy-based approach to image thresholding, 1994, Pergamon Pattern Recognition. Vol. 27, No. 9. pp. 1275 1289.
- [10] Wang, J. and Eliza Yingzi Du and Chein-I Chang and Thouin, P.D. Relative entropy-based methods for image thresholding, *Circuits and Systems, 2002. ISCAS 2002. IEEE International Symposium on*, II-265-II-268 vol.2.
- [11] Pearl, J. (1985), Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning, *Proceedings of the 7th Conference of the Cognitive Science Society*, University of California, Irvine, CA. pp. 329-334.
- [12] George H. John, Pat Langley., Estimating Continuous Distributions in Bayesian Classifiers, *Eleventh Conference on Uncertainty in Artificial Intelligence*, San Mateo, 338-345, 1995.
- [13] J. Platt, Fast Training of Support Vector Machines using Sequential Minimal Optimization, *B. Schoelkopf and C. Burges and A. Smola, editors, Advances in Kernel Methods - Support Vector Learning*, 1998.
- [14] Cristianini, Nello; Shawe-Taylor, John; An Introduction to Support Vector Machines and other kernel-based learning methods, *Cambridge University Press*, 2000
- [15] John G. Cleary, Leonard E. Trigg: K*: An Instance-based Learner Using an Entropic Distance Measure. *12th International Conference on Machine Learning*, 108-114, 1995.
- [16] <http://www.knime.org/>