Floristic participation at LifeCLEF 2016 Plant Identification Task

Julien Champ[1,2], Hervé Goëau[3], and Alexis Joly[1,2]

[1] Inria ZENITH team, France, name.surname@inria.fr
[2] LIRMM, Montpellier, France
[3] IRD, UMR AMAP, Montpellier, France

Abstract. This paper describes the participation of the Floristic consortium to the LifeCLEF 2016 plant identification challenge[18]. The aim of the task was to produce a list of relevant species for a large set of plant images related to 1000 species of trees, herbs and ferns living in Western Europe, knowing that some of these images belonged to unseen categories in the training set like plant species from other areas, horticultural plants or even off topic images (people, keyboards, animals, etc). To address this challenge, we first experimented as a baseline, without any rejection procedure, a Convolutional Neural Network (CNN) approach based on a slightly modified GoogLeNet model. In a second run, we applied a simple rejection criteria based on probability threshold estimation on the output of the CNN, one for each species, for removing automatically species propositions judged irrelevant. In the third run, rather than definitely eliminating some species predictions with the risk to remove false negative propositions, we applied various attenuation factors in order to revise the probability distributions given by the CNN as confident score expressing how much a query was related or not to the known species. More precisely, for this last run we used the geographical information and several cohesion measures in terms of observation, "organ" tags and taxonomy (genus and family levels) based on a knn similarity search results within the training set.

Keywords: LifeCLEF, plant, leaves, leaf, flower, fruit, bark, stem, branch, species, retrieval, images, collection, species identification, citizen-science, fine-grained classification, evaluation, benchmark

1 Introduction

Content-based image retrieval and computer vision approaches are considered as one of the most promising solutions to help bridging the taxonomic gap, as discussed in [5,2,26,24,17]. We therefore see an increasing interest in this transdisciplinary challenge in the multimedia community (e.g. in [23,10,3,22,19,12]. Beyond the raw identification performances achievable by state-of-the-art computer vision algorithms, recent visual search paradigms actually offer much more efficient and interactive ways of browsing large flora than standard field guides or online web catalogs ([4]). Smartphone applications relying on such image-based identification services are particularly promising for setting-up massive ecological monitoring systems, involving thousands of contributors at a very low cost. A first step in this way has been achieved by the US consortium behind LeafSnap[4], an i-phone application allowing the identification of 184 common American plant species based on pictures of cut leaves on an uniform background (see [21] for more details). Then, the French consortium supporting Pl@ntNet ([17]) went one step beyond by building an interactive image-based plant identification application that is continuously enriched by the members of a social network specialized in botany. Inspired by the principles of citizen sciences and participatory sensing, this project quickly met a large public with more than 1M downloads of the mobile applications ([8,7]). A related initiative is the plant identification evaluation task organized since 2011 in the context of the international evaluation forum CLEF[5] and that is based on the data collected within Pl@ntNet.

Since few years, *deep convolutional neural networks* repeatedly demonstrated record breaking results in generic object recognition problems such as ImageNet [20] and do attract more and more interest in the computer and multimedia vision communities. The promising effectiveness of this kind of approaches on more specific and fine grained classification problems like plant identification was confirmed last year [9] with impressive results regarding the fineness of the classes (at species level) and the unbalanced data in terms of available images per species. Rather than extracting the features according to hand-tuned or psycho-vision oriented filters, such methods directly work on the image signal. The weights learned by the first convolutional layers allows to automatically build relevant image filters whereas the intermediate layers are in charge of pooling these raw responses into high-level visual patterns. The last fully connected layers work more traditionally as any discriminative classifier on the image representation resulting from the previous layers.

A known drawback of Deep Convolutional Neural Networks is that they require a lot of training data mainly because of the huge number of parameters to be learned. This is particularly true here where the training set is highly unbalanced and includes many classes with few instances. The possibility to efficiently fine tune an already learned model, to adapt the architecture and resume training from the already learned model weight, is one a the main strength of CNN. This is one key explaining such results obtained last year on the plant identification task.

However, this year the task introduce an additional challenge by considering an open set classification problem, i.e. where some of the queries of the test set do not belong to the known species[6]. More precisely, according to the description of the task, these unseen images came from the plantnet mobile application and reflect the diversity of the visual content which the users produce despite of

[4] http://leafsnap.com/

[5] http://www.clef-initiative.eu/

the plantnet application is dedicated to wild plants from Western Europe. More precisely these pictures can be:

- off topic pictures like peoples, keyboards, landscapes, etc,
- horticultural plants (house & garden plants, vegetables & fruits),
- and wild plants but observed from all around the world and outside from the list of known species in the training set.

Considering the off topic pictures, one can guess that it must be rather easy to build a system predicting low or scattered probabilities on the 1000 known species, since visual content should be very different from the training dataset. Indeed, strong lines and corners from a manufactured object will produce certainly visual features very different from textured and mostly green visual contents learned from the training dataset. The difficulty of the task is most probably concentrated on the queries related to horticultural and wild plants with images sharing with the training set more visual similarities.

That said, CNN, like a vast majority of machine learning tools and recognition systems are designed for a static closed world, where the primary assumption is that all categories are known. We can admit that is a classification problem not so much explored in computer vision while it is a frequent usage case in the real world, even if some previous works are yet done in this direction with the CNNs [1].

2 Related work

2.1 Floristic Run 1

To address this challenge, we used a CNN model without any rejection procedure in order to obtain a first run considered here as a baseline, with the expectation that a query related to a known species will obtain a probability distribution concentrated on one or few relevant species (for instance species related to a same genus). The opposite expectation is that a query related to a unseen class will obtain a probability distribution spread over many classes.

We have used Caffe [14], a Deep Learning Framework, allowing us to use CNN architectures and models from the literature. We have chosen and slightly modified the "GoogLeNet GPU implementation" model in the Caffe model Zoo, based on the Google winning architecture in the ImageNet 2014 Challenge [25]. The GoogLeNet architecture consists of a 22 layers deep network with a softmax loss as the top classifier. It is composed of three "inception modules" stacked on top of each other. Each intermediate inception module is connected to an auxiliary classifier during training, so as to encourage discrimination in the lower stages of the classifier, increase the gradient signal that gets propagated back, and provide additional regularization. These auxiliary classifiers are only used during the training part, and then discarded.

We modified this model network by adding a batch normalisation at each level between the pooling and the Local Response Normalization layers in order to accelerate the learning phase [13]. As it is mentioned in this paper, we also removed the dropout layers. Combined with Parametric Rectified Linear Unit (PReLU) instead of ReLU layers, this model finally prevent the risk of overfitting [11]. Since we didn't find a such GoogleNet implementation, we learn this model on the ImageNet 2014 dataset (one week, 1,100,000 iterations with a batch size of 32, reaching a final train loss cost around 0.12).

Finally, we fine-tuned this model on the LifeCLEF Plant Task 2016 training dataset. For each image in the training and test sets, we therefore cropped the largest square in the center, and re-sized it to 256x256 pixels. As it was implemented within Caffe library, it makes also use of a simple data augmentation technique, consisting in cropping randomly a 224x224 pixels image, and mirroring it horizontally.

As a reminder, here are the most important parameters for Caffe to obtain our first submitted run "Floristic Run 1". The base learning rate parameter was set to 0.0075 which is rather high compared to usual learning rates applied to models without batch normalisation. The learning rate is divided by 10 every 42451 iterations with a batch size of 16 involving that each training images will pass 6 times during a step (113204 images x 6 / 16 gives the step size). We used only 2 steps and finally stopped the training after 90k iterations. For information, this fine-tuned model stopped with a top-1 loss accuracy on the training set itself of 0.9378.

To obtain the first run "Floristic Run 1", we directly used this fine-tuned model on the 8000 test images and limited the responses to the first 50 predicted species for each (when necessary).

2.2 Floristic Run 2: run 1 + rejection procedure

In a second run we added to the first approach a simple rejection procedure based on the estimation of probability thresholds, one for each species. The main idea was here to detect and remove some species predictions judged irrelevant. If the thresholds are correctly estimated, a query related to an unseen category should not be associated to any predictions and finally should not occur in the run file.

Given a species, for estimating its probability threshold, we computed the probability for each of its training image, and then selected the lowest value as a threshold. This threshold represents in a way the limit of the visual knowledge of the species according to the model and its available training images.

Finally, we produced the run file "Floristic Run 2" by applying the estimated thresholds on the predictions given by the run 1. This approach divided by two on average the number of species predictions of each query, with numerous queries associated to only one species prediction (1470 queries among 8000, while the run 1 contained only 83 queries with a response size of 1). But finally, none of the queries where entirely rejected.

2.3 Floristic Run 3: run 1 + mitigating factors

The risk of the approach using rejection procedure like in the run 2 is to definitely remove false negative species propositions, notably if the thresholds are too high while the probability of a correct species on a query is too loo. Therefore, in the third run, we preferred to keep all the species predictions produced in run 1 and apply on it some various attenuation factors. Indeed, the metric of the task is the classification MAP, i.e. the Mean of the Average Precision of each class taken individually, so, given a class, all the queries are sorted by their probability associated to this class, and the Average Precision depends directly on the ranks of the queries which really belong to this class. The main idea here was to reorder the list of the queries by downscaling their initial probability value by several factors between $[\alpha, 1.0]$ (α fixed arbitrarily to 0.9) with the expectation that irrelevant queries will be finally pushed on the tail of the list while relevant queries will maintain their rank.

For each query, six distinct factors were applied, mixing some information available in the metadata provided in the dataset with some consistency measures computed on the response given by a visual similarity search approach. The similarity search is produced by a fast nearest neighbors indexing and search method applied to the 1024 dimensional high level feature vector extracted with the CNN model from the second to last layer "pool5/7x7 s1". Each image feature is compressed with RMMH[15] (Random Maximum Margin Hashing) and its approximate k-nearest neighbors are searched by probing multiple neighboring buckets in the consulted hash table (according to the a posteriori multi-probe algorithm described in [16]). In that way, the knn search gives a complementary views of the training dataset from which we re-examine the species predictions given by the softmax output in the CNN model. More precisely, we can compare the metadata of the knns returned by the system with the metadata of a query for computing several factors (five here) reporting various contextual information:

- a factor $S_{classes}$ based on the classes returned by the knns,
- a factor S_{organs} based on the "organ" tags,
- two "taxonomic" factors S_{genus} and S_{family} at the genus and family levels,
- and a geolocalisation factor S_{geoloc}.

Factors are estimated individually for each query i, with values belonging to $[0.9, 1.0]$ and are directly applied to the probability distribution P_i in order to obtain some confident scores C_i:

$$C_i = P_i * S_{classes} * S_{organs} * S_{genus} * S_{family} * S_{geoloc}$$

For computing these factors, we choose to select arbitrarily the most visually similar images belonging to distinct observations. We didn't take directly the 5 most similar images because it can potentially be only near duplicate images belonging to a same observation and thus report poor contextual information.

Class distribution factor $S_{classes}$: this factor represents the convergence of the knns to a same class or not: if the knns belong to the same class, the factor will be neutral ($S_classes = 1$), while more the returned classes are distinct, more the factor will tend to $\alpha = 0.9$. Based on the occurrences of the classes appearing in the knns, we can compute a probability distribution P and then compute the entropy H_c defined by:

$$H_c = -\sum_{i=1}^{k} P_i \; log_2 \; P_i$$

with $k = 5$ observations. The entropy H_c will be equal to 0 when all the knns belong to a same same class, while it will be equal to its maximal value $H_{cmax} = log_2(k) = log_2(5)$ when each knn belong to a different class. Then an affine function gives directly the factor:

$$S_{classes} = 1 - 0.1 * \frac{H_c}{log_2(5)}$$

Organ factor S_{organs}: following the same previous approach, we count here the number of distinct tags *organ* reported by the knns among the available tags (flower, fruit, leaf, scan, stem, entire, branch), extract a probability distribution on these organs, compute the entropy H_o and finally compute the factor:

$$S_{organs} = 1 - 0.1 * \frac{H_o}{log_2(5)}$$

Taxonomic factors S_{genus} and S_{family}: following the same previous formulas, from the occurrences of the distinct genera (families) reported by the nns, we can extract a probability distributions on the genera (family), compute the entropy H_g (and H_f) and compute finally the factors:

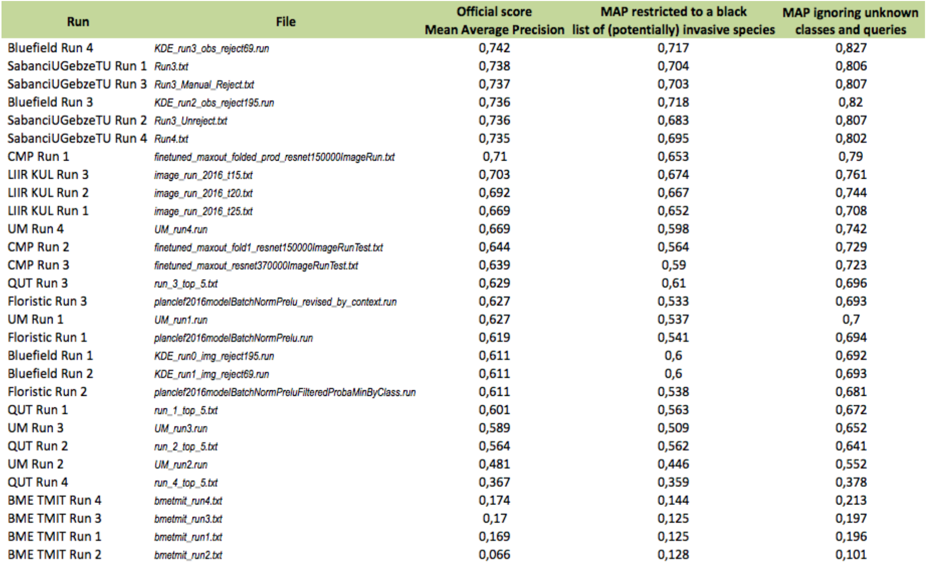
$$S_{genus} = 1 - 0.1 * \frac{H_g}{log_2(5)}$$

$$S_{family} = 1 - 0.1 * \frac{H_f}{log_2(5)}$$

Geolocalisation factor S_{geoloc}: here we didn't use a visual similarity knn search, but computed directly a factor based on the great circle distance *dist* between the GPS coordinates given by the metadata of a query and the coordinates representing more or less the center of France (latitude = 46.3, longitude = 2.3): $S_{geoloc} = 1 - \frac{dist}{distancemax}$ where $distancemax = 20000$ kms is more or less the farthest distance on earth from the center of France. By default $dist = 500$ kms if the metadata of a query doesn't contain some GPS coordinates, which is giving a factor of $S_{geoloc} = 0.975$

3 Official Results

Table 1 reports the scores of the 29 submitted runs, and figure 2 gives a complementary graphical overview of all results obtained by the participants.



Run	File	Official score Mean Average Precision	MAP restricted to a black list of (potentially) invasive species	MAP ignoring unknown classes and queries
Bluefield Run 4	*KDE_run3_obs_reject69.run*	0,742	0,717	0,827
SabanciUGebzeTU Run 1	*Run3.txt*	0,738	0,704	0,806
SabanciUGebzeTU Run 3	*Run3_Manual_Reject.txt*	0,737	0,703	0,807
Bluefield Run 3	*KDE_run2_obs_reject195.run*	0,736	0,718	0,82
SabanciUGebzeTU Run 2	*Run3_Unreject.txt*	0,736	0,683	0,807
SabanciUGebzeTU Run 4	*Run4.txt*	0,735	0,695	0,802
CMP Run 1	*finetuned_maxout_folded_prod_resnet150000ImageRun.txt*	0,71	0,653	0,79
LIIR KUL Run 3	*image_run_2016_t15.txt*	0,703	0,674	0,761
LIIR KUL Run 2	*image_run_2016_t20.txt*	0,692	0,667	0,744
LIIR KUL Run 1	*image_run_2016_t25.txt*	0,669	0,652	0,708
UM Run 4	*UM_run4.run*	0,669	0,598	0,742
CMP Run 2	*finetuned_maxout_fold1_resnet150000ImageRunTest.txt*	0,644	0,564	0,729
CMP Run 3	*finetuned_maxout_resnet370000ImageRunTest.txt*	0,639	0,59	0,723
QUT Run 3	*run_3_top_5.txt*	0,629	0,61	0,696
Floristic Run 3	*planclef2016modelBatchNormPrelu_revised_by_context.run*	0,627	0,533	0,693
UM Run 1	*UM_run1.run*	0,627	0,537	0,7
Floristic Run 1	*planclef2016modelBatchNormPrelu.run*	0,619	0,541	0,694
Bluefield Run 1	*KDE_run0_img_reject195.run*	0,611	0,6	0,692
Bluefield Run 2	*KDE_run1_img_reject69.run*	0,611	0,6	0,693
Floristic Run 2	*planclef2016modelBatchNormPreluFilteredProbaMinByClass.run*	0,611	0,538	0,681
QUT Run 1	*run_1_top_5.txt*	0,601	0,563	0,672
UM Run 3	*UM_run3.run*	0,589	0,509	0,652
QUT Run 2	*run_2_top_5.txt*	0,564	0,562	0,641
UM Run 2	*UM_run2.run*	0,481	0,446	0,552
QUT Run 4	*run_4_top_5.txt*	0,367	0,359	0,378
BME TMIT Run 4	*bmetmit_run4.txt*	0,174	0,144	0,213
BME TMIT Run 3	*bmetmit_run3.txt*	0,17	0,125	0,197
BME TMIT Run 1	*bmetmit_run1.txt*	0,169	0,125	0,196
BME TMIT Run 2	*bmetmit_run2.txt*	0,066	0,128	0,101

Fig. 1. LifeCLEF 2016 Plant Task Official results

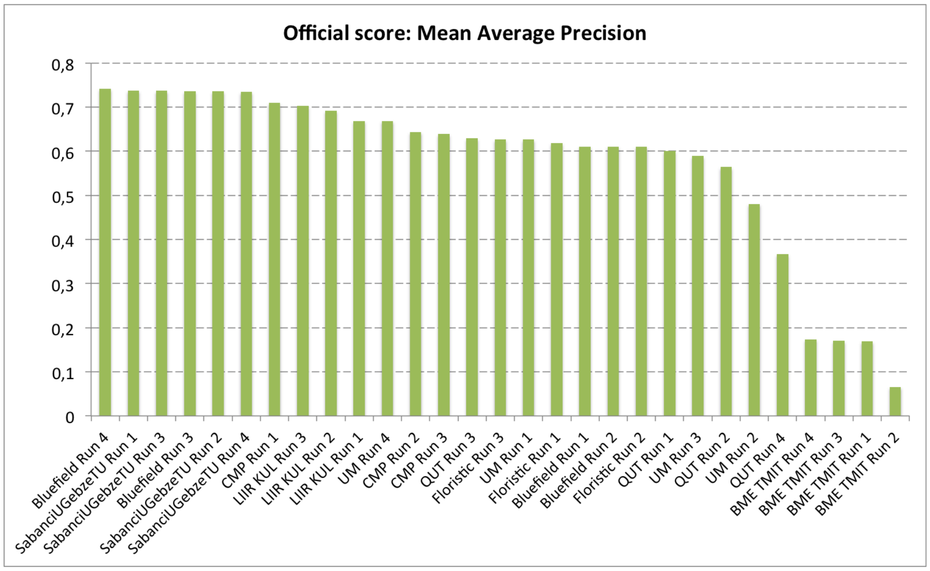


Fig. 2. LifeCLEF 2016 Plant Task Official results

4 Conclusion

Floristic team submitted 3 runs: the first run *Floristic Run 1* was based on the well-known GoogLeNet CNN architecture, but slightly modified with the use of Parametric Rectified Linear Units and the use of batch normalisation layers in order to accelerate and prevent from overfitting the learned model. This first approach obtained an intermediate MAP of 0.619 while the best system obtained a MAP of 0.742. Unfortunately, by adding the rejection criteria, we degraded slightly the MAP (down to 6.111 obtained by "Floristic Run 2"). This rejection criteria was certainly too strong, with estimated probability thresholds too high, and have probably removed too much correct species predictions. On another side, contextual information exploited in "Floristic Run 3" for revising the species predictions slightly improved the MAP (up to 0.627), but not enough for reaching the performances of the best systems.

References

1. Bendale, A., Boult, T.: Towards open set deep networks. arXiv preprint arXiv:1511.06233 (2015)
2. Cai, J., Ee, D., Pham, B., Roe, P., Zhang, J.: Sensor network for the monitoring of ecosystem: Bird species recognition. In: Intelligent Sensors, Sensor Networks and Information, 2007. ISSNIP 2007. 3rd International Conference on. pp. 293–298 (Dec 2007)
3. Cerutti, G., Tougne, L., Vacavant, A., Coquin, D.: A Parametric Active Polygon for Leaf Segmentation and Shape Estimation. In: 7th International Symposium on Visual Computing. p. 1. Las Vegas, United States (Sep 2011), https://hal.archives-ouvertes.fr/hal-00622269
4. Ellison, A.M., Farnsworth, E.J., Chu, M., Kress, W.J., Neill, A.K., Best, J.H., Pickering, J., Stevenson, R.D., Courtney, G.W., VanDyk, J.K.: Next-generation field guides. BioScience (2013)
5. Gaston, K.J., O'Neill, M.A.: Automated species identification: why not? Philosophical Transactions of the Royal Society of London B: Biological Sciences 359(1444), 655–667 (2004)
6. Goëau, H., Bonnet, P., Joly, A.: Plant identification in an open-world (lifeclef 2016). In: CLEF working notes 2016 (2016)
7. Goëau, H., Bonnet, P., Joly, A., Affouard, A., Bakic, V., Barbe, J., Dufour, S., Selmi, S., Yahiaoui, I., Vignau, C., et al.: Pl@ ntnet mobile 2014: Android port and new features. In: Proceedings of International Conference on Multimedia Retrieval. p. 527. ACM (2014)
8. Goëau, H., Bonnet, P., Joly, A., Bakić, V., Barbe, J., Yahiaoui, I., Selmi, S., Carré, J., Barthélémy, D., Boujemaa, N., et al.: Pl@ ntnet mobile app. In: Proceedings of the 21st ACM international conference on Multimedia. pp. 423–424. ACM (2013)
9. Goëau, H., Joly, A., Bonnet, P.: Lifeclef plant identification task 2015. In: CLEF working notes 2015 (2015)
10. Goëau, H., Joly, A., Selmi, S., Bonnet, P., Mouysset, E., Joyeux, L.: Visual-based plant species identification from crowdsourced data. In: MM'11 - ACM Multimedia 2011. pp. 0–0. ACM, Scottsdale, United States (Nov 2011), https://hal.inria.fr/hal-00642236

11. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. CoRR abs/1502.01852 (2015), http://arxiv.org/abs/1502.01852
12. Hsu, T.H., Lee, C.H., Chen, L.H.: An interactive flower image recognition system. Multimedia Tools Appl. 53(1), 53–73 (May 2011), http://dx.doi.org/10.1007/s11042-010-0490-6
13. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. CoRR abs/1502.03167 (2015), http://arxiv.org/abs/1502.03167
14. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
15. Joly, A., Buisson, O.: Random maximum margin hashing. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. pp. 873–880 (June 2011)
16. Joly, A., Buisson, O.: A posteriori multi-probe locality sensitive hashing. In: Proceedings of the 16th ACM International Conference on Multimedia. pp. 209–218. MM '08, ACM, New York, NY, USA (2008), http://doi.acm.org/10.1145/1459359.1459388
17. Joly, A., Goëau, H., Bonnet, P., Bakić, V., Barbe, J., Selmi, S., Yahiaoui, I., Carré, J., Mouysset, E., Molino, J.F., et al.: Interactive plant identification based on social image data. Ecological Informatics 23, 22–34 (2014)
18. Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W.P., Champ, J., Planqué, R., Palazzo, S., Müller, H.: Lifeclef 2016: multimedia life species identification challenges. In: Proceedings of CLEF 2016 (2016)
19. Kebapci, H., Yanikoglu, B., Unal, G.: Plant image retrieval using color, shape and texture features. Comput. J. 54(9), 1475–1490 (Sep 2011), http://dx.doi.org/10.1093/comjnl/bxq037
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
21. Kumar, N., Belhumeur, P.N., Biswas, A., Jacobs, D.W., Kress, W.J., Lopez, I.C., Soares, J.V.: Leafsnap: A computer vision system for automatic plant species identification. In: Computer Vision–ECCV 2012, pp. 502–516. Springer (2012)
22. Mouine, S., Yahiaoui, I., Verroust-Blondet, A.: Advanced shape context for plant species identification using leaf image retrieval. In: Ip, H.H.S., Rui, Y. (eds.) ICMR '12 - 2nd ACM International Conference on Multimedia Retrieval. ACM, Hong Kong, China (Jun 2012), https://hal.inria.fr/hal-00726785
23. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Computer Vision, Graphics Image Processing, 2008. ICVGIP '08. Sixth Indian Conference on. pp. 722–729 (Dec 2008)
24. Spampinato, C., Mezaris, V., van Ossenbruggen, J.: Multimedia analysis for ecological data. In: Proceedings of the 20th ACM international conference on Multimedia. pp. 1507–1508. ACM (2012)
25. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. CoRR abs/1409.4842 (2014), http://arxiv.org/abs/1409.4842
26. Trifa, V.M., Kirschel, A.N.G., Taylor, C.E., Vallejo, E.E.: Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models. Journal of The Acoustical Society of America 123 (2008)