Are Species Identification Tools Biodiversity-friendly ?

A cross-tasks analysis of LifeCLEF 2014 results

Alexis Joly[1,*], Hervé Goëau[1], Pierre Bonnet[2], Concetto Spampinato[3], Hervé Glotin[4], Andreas Rauber[5], Robert B. Fisher[6], Henning Müller[7]

1 INRIA, LIRMM, France
2 CIRAD, France
3 University of Catania, Italy
4 IUF & Univ. de Toulon, France
5 Vienna Univ. of Tech., Austria
6 Edinburgh Univ., UK
7 HES-SO, Switzerland
∗ E-mail: alexis.joly@inria.fr

Abstract

This paper discusses the results of the LifeCLEF 2014 multimedia identification challenges with regards to the requirements of real-world ecological surveillance systems. In particular, we study the identification performances of the evaluated systems as a function of the ordinariness or rarity of the species in the dataset. This allows us to assess the ability of the underlying methods to be robust to heavily tailed distributions such as the ones encountered in real-world collections of life observations. Results show that all methods are more or less affected by the long-tail curse but that the best methods making use of classifiers with good discrimination capacities do resist the phenomenon pretty well.

1 Introduction

LifeCLEF [14][1] is one of the labs of the CLEF[2] evaluation forum dedicated to the evaluation of multimedia-oriented life species identification systems. Using multimedia identification tools is considered as one of the most promising solutions to help bridge the taxonomic gap and build accurate knowledge of the identity, the geographic distribution and the evolution of living species [16, 3, 21, 18, 1, 20, 12]. Unfortunately, the performance of the state-of-the-art multimedia analysis techniques on such data is still not well understood and we are far from reaching the real world's requirements in terms of identification tools. The LifeCLEF lab evaluates these challenges around 3 tasks related to multimedia information retrieval and fine-grained classification problems in 3 subdomains. Each task is based on large and real-world data and the measured challenges are defined in collaboration with biologists and environmental stakeholders in order to reflect realistic usage scenarios. As the purpose of this paper

[1] www.lifeclef.org
[2] www.clef-initiative.eu

is to focus on a deeper analysis of the raw results of the 2014 campaign, we refer reader to the complete overview of the lab [14] for the details about each of the tasks including the data description, the metrics, the run formats, etc. We here only synthetise the main tasks:

PlantCLEF[3]: an image-based plant identification task continues the three previous plant identification challenges of ImageCLEF in 2011 [7], 2012 [8] and 2013 [13]. The 2014 PlantCLEF dataset was composed of 60,962 pictures belonging to 19,504 observations of 500 species of trees, herbs and ferns living in a European region centered around France. This data was collected by 1,608 members of TelaBotanica[4], a French-speaking social network of 23,000 amateur and expert botanists. Each picture belongs to one of the 7 types of view reported in the meta-data (entire plant, fruit, leaf, flower, stem, branch, leaf scan) and is associated with a single plant observation identifier allowing to link it with the other pictures of the same individual plant (observed the same day by the same person).

BirdCLEF[5]: an audio-based bird identification task based on the audio recordings collected by Xeno-canto[6], a web-based community of bird sound recordists worldwide with about 1,500 active contributors that have already collected more than 180,000 recordings of about 9,000 species. Nearly 500 species from Brazilian forests are used in the BirdCLEF dataset totalling about 14,000 recordings produced by hundreds of users. As a comparison, the previous largest bird species bioacoustic classification was the NIPS4B 2013 challenge with (only) 80 species from French Provence [11].

FishCLEF: a video-based fish identification task based on the Fish4Knowledge[7] underwater video repository, which contains about 700k 10-minutes video clips that were taken in the past five years to monitor Taiwan's coral reefs. More specifically, the FishCLEF dataset consists of about 3,000 videos with several tens of thousands of detected fish instances that were identified for the 10 most common species.

127 research groups worldwide registered to at least one task of the lab, of which 22 crossed the finish line by submitting runs (27 runs for the plant task, 29 runs

[3]supported by Agropolis fundation through the Pl@ntNet project (http://www.plantnet-project.org/)
[4]http://www.tela-botanica.org/
[5]supported by CNRS MASTODONS SABIOD project (http://sabiod.org)
[6]http://www.xeno-canto.org/
[7]www.fish4knowledge.eu

for the bird task, 6 runs for the fish task). Details on the methods used in the runs and the results achieved by all teams are synthesised in the overview working notes of each task [10, 9, 5]. Overall, quite impressive identification rates are achieved by the best runs of each task with mean average precision scores close to 0.5 for the hundreds of species of the bird or the plant task, and up to 0.95 for the 10 species of the fish task.

2 The big problem with data

Building effective multimedia analysis and machine learning techniques is unfortunately not the only side of the taxonomic gap problem. Whatever the used algorithms, the availability of rich and appropriate training data is actually equally challenging towards setting up powerful identification tools at large scales. If we look at the popular ImageNet dataset [6], widely used for the evaluation of large-scale image classification methods, it is essential to notice that the average number of training images per category is in the range 600-1200. And this is actually several orders of magnitude richer than most existing collections of multimedia life observations. Even the *Encyclopedia Of Life*[8], which is the world's largest data centralization effort concerning multimedia data for life on Earth, does not have more than few images per class of interest for the vast majority of species. Thanks to the integration of hundreds of large expert collections built in the past, the global plant index is for instance now approximating the 700K images, which is an outstanding number. But from a machine learning point of view, the problem is that these images are scattered across tens of thousands distinct taxa and across tens of distinct types of views or organs in a given taxon.

Overall, as discussed in [12], most existing multimedia collections suffer from one or several problems preventing their easy use as training data. The *long-tail problem* is one of the most common ones, particularly in the context of collaborative data. The symptom is that the distribution of the number of samples per species generally follows a long-tailed distribution, with very few species well populated, and the vast majority of species with one or few images. This more generally reflects the heterogeneous knowledge that we have of plants and animals, with a huge volume of information on widespread and useful species for human beings, and very little information (in term of geographical distribution, morphological description, etc.) on most of the plant species of a given area. A rather good average number of multimedia documents per taxon can therefore be misleading regarding the real coverage of all species. Figure 1 illustrates the distribution of the Bugwood[9]-ForestryImages[10] dataset which includes 187K images of about 18K plants species of economic concern. Similarly, Figures 3 and 4 present the long-tailed distribution of the number of observations per species

[8] http://eol.org/
[9] http://www.bugwood.org/
[10] http://www.forestryimages.org/

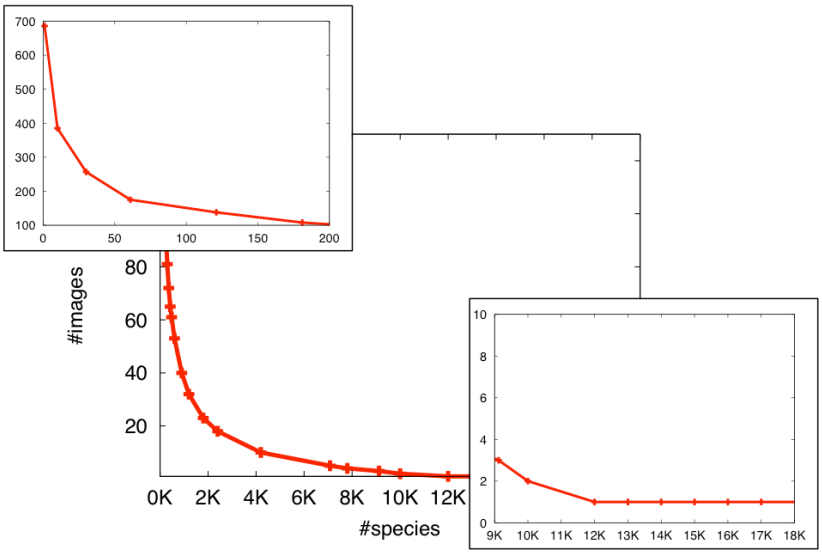

Figure 1: Long tail distribution of Bugwood-Forestry Images dataset

within the plant and the bird datasets used for LifeCLEF 2014 challenges. The histograms notably show that the plant dataset is more heavily tailed than the bird dataset. This is partially due to the natural abundance distribution of the species but also to the different characteristics of the social networks that collected the data (less contributors to the bird dataset but with a more homogeneous and wider expertise). But still for the plant dataset, the set of the kept species in LifeCLEF challenge is only the tip of the iceberg. This is illustrated by Figure 2, which presents the distribution of the whole Pl@ntNet dataset [12] which in itself is the largest existing dataset for the French flora. As shown on the graph, the 500 hundred species of the PlantCLEF dataset (coloured in red) are mainly concentrated at the head of the distribution and therefore rather focus on the most common species, regardless the rich biodiversity existing in the country (estimated to be about 6000 plant species). If we now recall that the vast majority of the hundreds of thousands of plant species on Earth are even more incomplete, it gives a nice picture of the enormously challenging problem we are facing in order to build well-balanced and well-populated training data. The ecosystems that possess the highest plant diversity are actually also the least studied and understood (particularly tropical and Mediterranean regions). It is consequently very difficult to collect in these regions as much data as in well covered areas such as France.

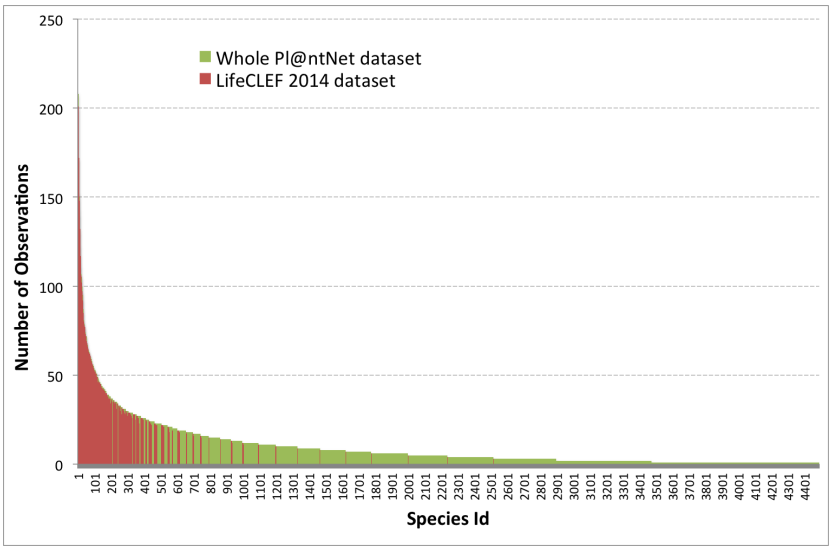

Figure 2: Long tail distribution of the whole Pl@ntNet dataset (with PlantCLEF 2014 subset in red)

3 Abundance-aware evaluation of LifeCLEF runs

In order to cover sufficient biodiversity, it is therefore crucial that multimedia identification tools also work for the species in the long tail and not only on the most populated and popular species. In this paper, we thus propose to analyse the results of the LifeCLEF 2014 challenge with regard to the ability of the systems to deal with the less populated classes. For each of the three tasks, we therefore split the species into 3 categories according to the number of observations populating these species in the datasets. The 3 resulting splits are illustrated in Figures 3, 4 and 5. For the three challenges, the category A of species (blue) corresponds to the most populated species (i.e. the tip of the distribution), the category B (red) corresponds to intermediate species with a relatively high number of observations, the category C (green) corresponds to the less populated species in the long tail. Note that for the fish task, as the total number of species is very low (restricted to the 10 most common species), we only included one species in the long tail category C. Therefore, the results will be statistically less relevant than for the bird and the plant task. For the bird and the plant datasets, we used the same thresholds on the cumulative distributions to define the categories (cat. A is represented by the first 20 percent of the observations belonging to the most populated species, cat. B by the next 50 percent of the observations, cat. C by the least common 30 percent).

Based on these relative abundance categories A, B and C, we computed a per-category score for all the runs submitted to LifeCLEF. This was done by

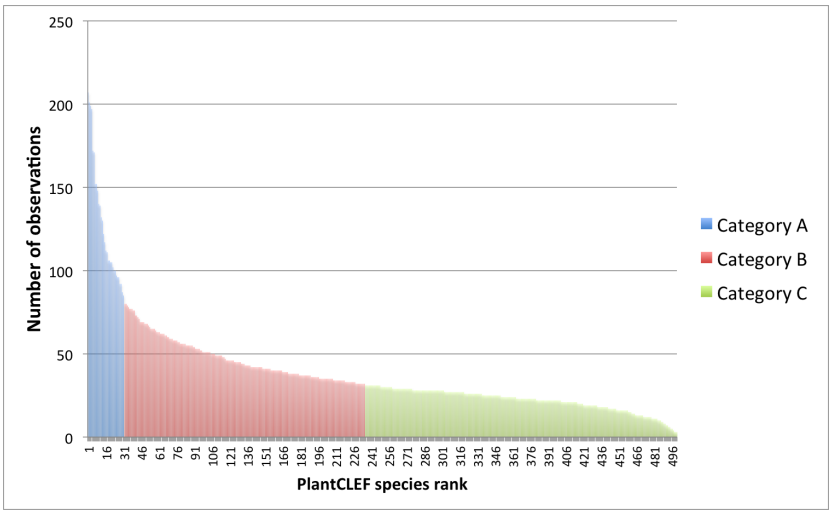

Figure 3: Distribution of PlantCLEF 2014 dataset - split in 3 ordinariness categories

first computing the per-species official score of each run and then averaging the scores of the species that belong to the same category (details of the scores used for each task can be found in their respective CLEF working notes [10, 9, 5]). Results are displayed within Figures 7, 6 and 8 (for the fish video challenge we only considered subtask 3). Note that the initial ranking of the runs has been preserved in accordance to the overall official score of each run. This allows us to analyse whether the per-category rankings of the methods differ from the global one (and/or between each other). Furthermore, the graphs allow us to check whether the overall performance of a given run is achieved to the detriment of the less populated species. To further quantify this *biodiversity-friendliness*, we also displayed on the graph the *coefficient of variation* of each run (i.e. the standard deviation of the 3 categorical scores divided by their mean).

A first overall conclusion is that the performances of all systems degrade with the ordinariness of the targeted species (i.e. none of the evaluated systems has a lower score on category A than on category B and this is also the case for category A vs. C and B vs. C). This is clearly not surprising and simply demonstrates that all participants use common statistical machine learning strategies whose performances are correlated with the class statistics. As it was not an objective of the measured challenges, none of the participants specifically tried to emphasize the less populated species or to balance the classes. This raises the question of whether it would be meaningful to foster rare species in the evaluation protocol of future LifeCLEF challenges (e.g. through a biodiversity-friendly evaluation metric or through a balanced distribution of the queries across the classes). On one hand, this would bias the evaluation because the natural dis-

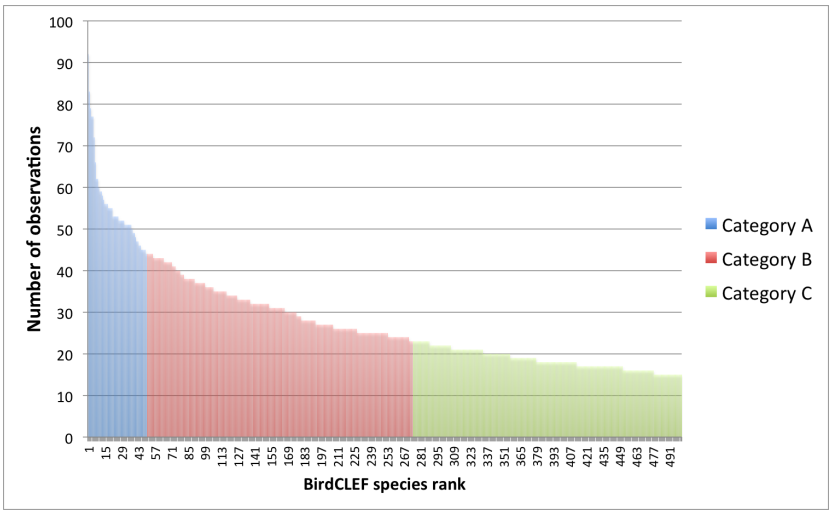

Figure 4: Distribution of BirdCLEF 2014 dataset - split in 3 ordinariness categories

tribution of the data somehow reflects the usage of a real-world identification system in which the most common species attract the most user requests [12]. Maximising the average score across all the observations consequently also meets the objective of maximizing the average user satisfaction. But on the other hand, boosting the visibility of the less populated species in a real-world application might help compensate for the long-tail curse in the long term. This might degrade the brute-force performances on the most common queries. On the other hand, when a user meets a rare plant or animal, this would give him a better chance to identify it and consequently enrich the system with this useful observation. In other words, this would stabilize the positive feedback loop typically observed in crowd-sourced information systems that tends to accentuate the inequalities and put too much emphasis on the most popular items (e.g. the distribution of user ratings in a social network tends to be more and more heavily tailed when their number increases [17]).

Now, the most important question is whether the methods used by the different participants are equally biodiversity-friendly or not. Let us first start with the bird task results (cf. Figure 6). For this challenge, we can observe that the overall ranking of the runs remains roughly the same whatever the category (A,B or C). This means that none of the methods is critically more affected by the long-tail issue than the other ones. Some variations can however be observed. The per-category scores of MNB TSA runs are notably more homogeneous than the ones of QMUL (with a variation coefficient of 0.19 vs. 0.28 for the best run of each team). That means that in addition to be better on average, the runs of MNB TSA are also more biodiversity-friendly. This confirms the good skills of the segment-probabilities features used by this group [15] as well as the good

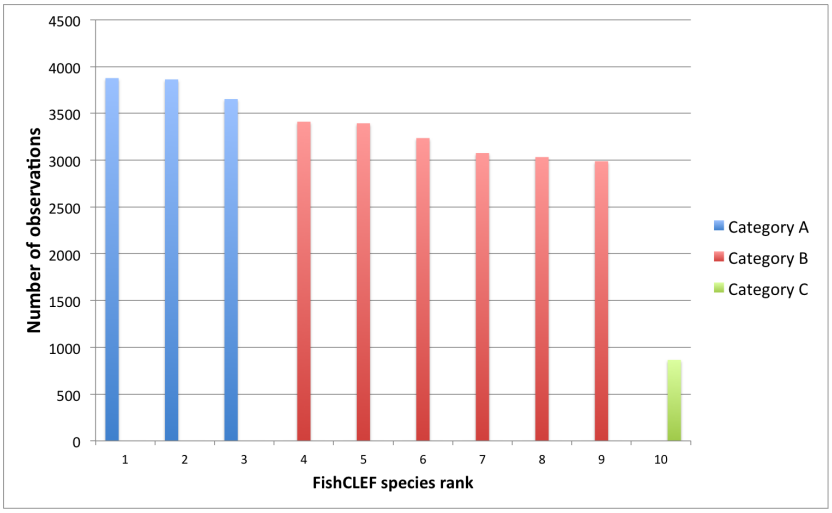

Figure 5: Distribution of FishCLEF 2014 dataset - split in 3 ordinariness categories

capacity of the ensemble of randomized decision trees they are using as classifier. On the contrary, the per-category scores of INRIA Zenith runs are a little bit more scattered than the others, as illustrated by the emerging blue peak of INRIA Zenith Run 2 and the higher values of the variation coefficient ($0,38$ for Run 1). As discussed again later for the plant task, this might be due to the use of a K-NN majority voting classifier on top of their discriminant features selection and matching scheme. Finally, the Run 1 of HLT is particularly compact across the three categories (with a very low variation coefficient of 0.11). This is presumably due to the local temporal pooling strategy they used exclusively in that run. But as the overall performance of that run is rather low, its ability to identify well the less populated species remains much lower than the best runs of the challenge.

If we now look at the fish task results (cf. Figure 8), we can observe that here again the ranking of the runs is preserved for the three categories but that the variations in the homogeneity of the scores are much more accentuated. Whereas all runs have achieved comparable performances on the category A, the performances of the I3S runs clearly crashed on the less populated species leading to very bad biodiversity-friendliness values (variation coefficient greater than 1.0). As discussed in the working note of the fish task [5], the lower performances of the I3S runs are mainly due to their fish detection algorithm that has a much lower recall than the ViBe [2] background modeling approach used in the baseline. Interestingly, we can see here that the deficit of recall is mainly concentrated on the less populated species in the dataset that are probably also the less visible ones in the video contents (smaller fishes and/or less numerous ones).

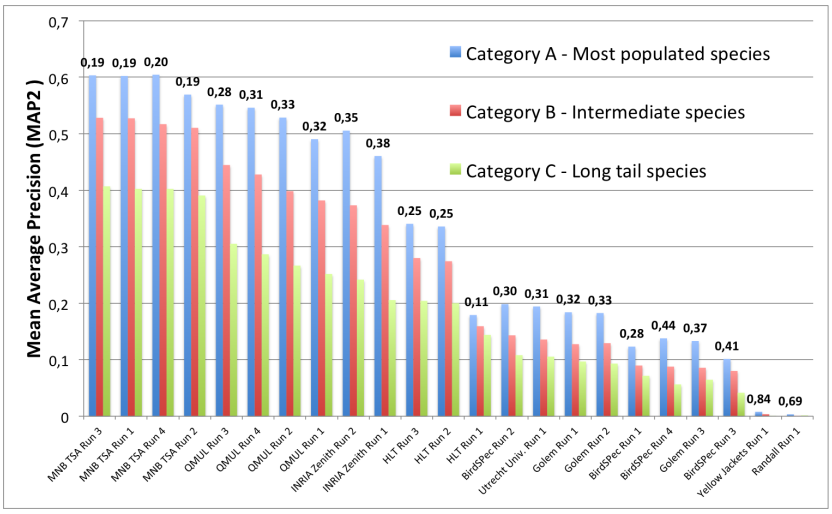

Figure 6: BirdCLEF results detailed by ordinariness categories

The results of the plant task (Figure 7) are probably the most informative ones concerning our biodiversity-friendliness analysis. The more heavily-tailed distribution of the plant dataset actually accentuates its impact on the methods that are the most sensitive to multi-class imbalanced problems. This is typically the case for the runs of SZTE, FINKI and PlantNet that result in high values of the variation coefficient and low classification scores for the species belonging to the long tail. The common point of all these runs is that they rely on instance-based classifiers that are more directly dependent on the feature density and thus more sensitive to imbalanced problems. The most affected runs are the ones of SZTE and FINKI that directly use a K-NN classifier on global visual features. The two best runs of PlantNet (2 and 3) resist better because of the use of a Borda count to fuse the different features and organ modalities instead of the weighted majority voting strategy used in run 1 and 4.
On the contrary, using large margin classifiers appears to be a rather good strategy to limit the inequalities between the different categories of species. This mainly concerns the runs of IBM AU, BME TMIT and Sanbanci Okan whose variation coefficients values are all lower than $0,38$. But there are still some important variations between them, meaning that other factors enter the equation. The three first runs of IBM [4] clearly outperform all other runs in terms of the coefficient of variation. Similarly to the bird task, this shows that in addition to being the best runs on average, the runs of IBM AU are also the most biodiversity-friendly. The classification score of the best of their runs on category C is impressively better than the classification score of any other teams on all categories. This confirms that using linear support vectors on top of high-dimensional Fisher vectors is definitely a good strategy to reach state-of-the-art performances on this benchmark. But as the runs of BME TMIT are based on

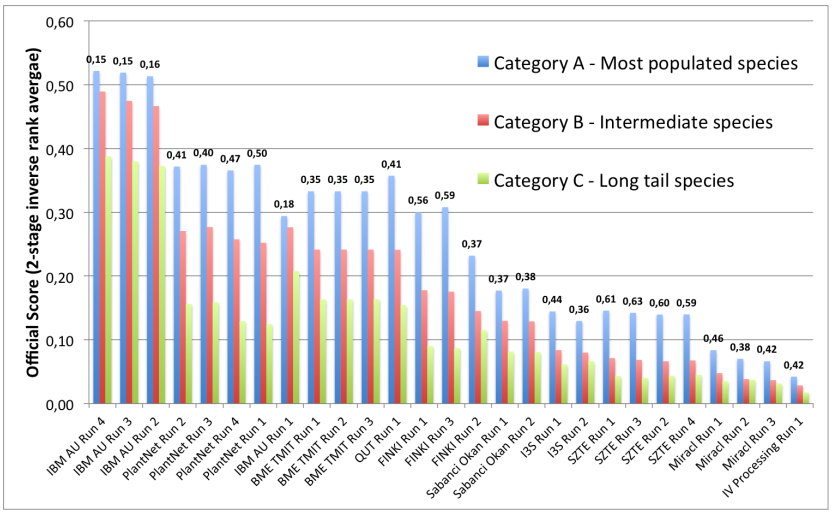

Figure 7: PlantCLEF results detailed by ordinariness categories

the same strategy it also shows that the devil is in the details. The main differences inferred from the working notes of both participants [4, 19] concern the use of (i) color moment features in addition to SIFT, (ii) power normalization of the Fisher vectors, (iii) a 512 components GMM model instead of 256, (iv) the use of a linear support vector machine rather than a RBF kernel in the BME MIT runs, and (v) an observation-oriented split of the data for cross-validation. A last interesting insight we can derive from the plant task results concerns the last run of IBM AU (*IBM AU Run 4*) which is the only one purely based on a deep convolutional neural network. Because of the relatively low average number of training samples per class it actually failed in learning as good visual features as the hand-tuned features of the PlantNet runs. But from the biodiversity-friendliness point of view, it clearly outperformed them, presumably thanks to the much better generalization capacity of the last fully connected layers of the network compared to the instance-based classification scheme employed in the PlantNet runs.

4 Conclusion and perspectives

This paper reported a complementary analysis of the raw results of the LifeCLEF 2014 challenge with regard to the biodiversity-friendliness of the evaluated methods, i.e., their capacity to classify well both the most and the least populated species. The good news is that the best performing methods of each task are also the most biodiversity-friendly ones. The question of whether we should introduce a specific biodiversity-friendly evaluation metric for future LifeCLEF campaigns is consequently less important. It appears that well-designed discriminant classification schemes are naturally more robust to the long-tail

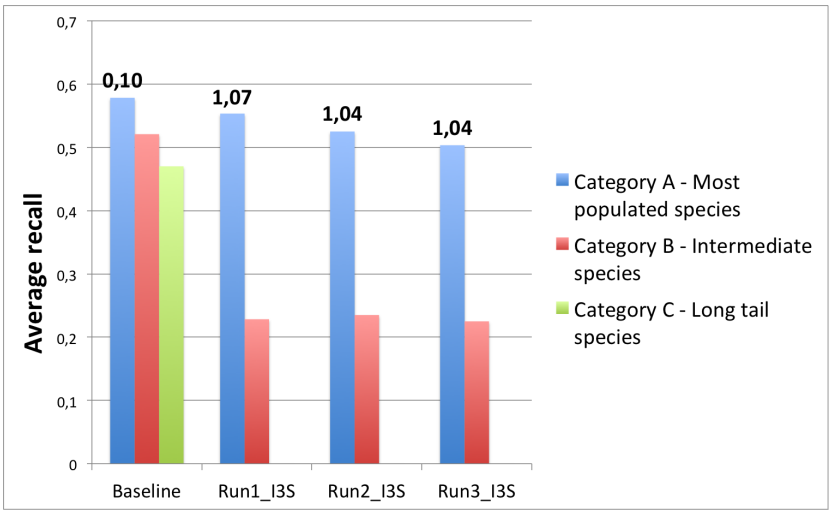

Figure 8: FishCLEF results detailed by ordinariness categories

curse and finally provides relatively good performances even on the less-populated classes of the long tail. This, however, has to be mitigated by the fact that the species used in the evaluations are still the tip of the iceberg and that the real long tail still has to come out. For the upcoming LifeCLEF evaluations, we will study the feasibility of providing more species as well as the feasibility of authorizing any other external training data to further increase the biodiversity cover of our challenges. One might also consider adding a none-of-the-above categories that would help filter new species and focus future ground-truthing efforts.

References

[1] *MAED '12: Proceedings of the 1st ACM International Workshop on Multimedia Analysis for Ecological Data*, New York, NY, USA, 2012. ACM. 433127.

[2] O. Barnich and M. Van Droogenbroeck. Vibe: A universal background subtraction algorithm for video sequences. *Image Processing, IEEE Transactions on*, 20(6):1709–1724, 2011.

[3] J. Cai, D. Ee, B. Pham, P. Roe, and J. Zhang. Sensor network for the monitoring of ecosystem: Bird species recognition. In *Intelligent Sensors, Sensor Networks and Information, 2007. ISSNIP 2007. 3rd International Conference on*, pages 293–298, Dec 2007.

[4] Q. Chen, M. Abedini, R. Garnavi, and X. Liang. Ibm research australia at lifeclef2014: Plant identification task. In *Working notes of CLEF 2014 conference*, 2014.

[5] S. Concetto, R. Fisher, and B. Boom. Lifeclef fish identification task 2014. In *CLEF working notes 2014*, 2014.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE conference on Computer Vision and Patter Recognition*, pages 248–245, 2009.

[7] H. Goëau, P. Bonnet, A. Joly, N. Boujemaa, D. Barthélémy, J.-F. Molino, P. Birnbaum, E. Mouysset, and M. Picard. The ImageCLEF 2011 plant images classification task. In *CLEF working notes*, 2011.

[8] H. Goëau, P. Bonnet, A. Joly, I. Yahiaoui, D. Barthelemy, N. Boujemaa, and J.-F. Molino. The ImageCLEF 2012 Plant Identification Task. In *CLEF working notes*, 2012.

[9] H. Goëau, H. Glotin, W.-P. Vellinga, A. Rauber, and R. Planqué. LifeCLEF Bird Identification Task 2014. In *CLEF working notes 2014*, 2014.

[10] H. Goëau, A. Joly, P. Bonnet, J.-F. Molino, D. Barthélémy, and N. Boujemaa. LifeCLEF Plant Identification Task 2014. In *CLEF working notes 2014*, 2014.

[11] *Proc. Neural Inform. Proc. Scaled for Bioacoustics, from Neurons to Big Data*, 2013. http://sabiod.org/nips4b.

[12] A. Joly, H. Goeau, P. Bonnet, V. Bakić, J. Barbe, S. Selmi, I. Yahiaoui, J. Carré, E. Mouysset, J.-F. Molino, N. Boujemaa, and D. Barthélémy. Interactive plant identification based on social image data. *Ecological Informatics*, 2013.

[13] A. Joly, H. Goëau, P. Bonnet, V. Bakic, J.-F. Molino, D. Barthélémy, N. Boujemaa, et al. The imageclef plant identification task 2013. In *International workshop on Multimedia analysis for ecological data*, 2013.

[14] A. Joly, H. Müller, H. Goëau, H. Glotin, C. Spampinato, A. Rauber, P. Bonnet, W.-P. Vellinga, R. Fisher, and R. Planqué. Lifeclef 2014: multimedia life species identification challenges.

[15] M. Lasseck. Large-scale identification of birds in audio recordings. In *Working notes of CLEF 2014 conference*, 2014.

[16] D.-J. Lee, R. B. Schoenberger, D. Shiozawa, X. Xu, and P. Zhan. Contour matching for a fish recognition and migration-monitoring system. In *Optics East*, pages 37–48. International Society for Optics and Photonics, 2004.

[17] M. J. Salganik, P. S. Dodds, and D. J. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *science*, 311(5762):854–856, 2006.

[18] C. Spampinato, Y.-H. Chen-Burger, G. Nadarajan, and R. B. Fisher. Detecting, tracking and counting fish in low quality unconstrained underwater videos. In *VISAPP (2)*, pages 514–519. Citeseer, 2008.

[19] G. Szúcs, P. Dávid, and D. Lovas. Viewpoints combined classification method in image-based plant identification task. In *Working notes of CLEF 2014 conference*, 2014.

[20] M. Towsey, B. Planitz, A. Nantes, J. Wimmer, and P. Roe. A toolbox for animal call recognition. *Bioacoustics*, 21(2):107–125, 2012.

[21] V. M. Trifa, A. N. Kirschel, C. E. Taylor, and E. E. Vallejo. Automated species recognition of antbirds in a mexican rainforest using hidden markov models. *The Journal of the Acoustical Society of America*, 123:2424, 2008.