LifeCLEF Bird Identification Task 2015

Hervé Goëau, Hervé Glotin, Willem-Pier Vellinga, Robert Planqué, Andreas Rauber, Alexis Joly

LifeCLEF Bird Identification Task 2015

Hervé Goëau[1], Hervé Glotin[2], Willem-Pier Vellinga[3], Robert Planqué[3], Andreas Rauber[4], and Alexis Joly[1,5]

[1] Inria ZENITH team, France, name.surname@inria.fr
[2] Aix Marseille Univ., ENSAM, CNRS LSIS, Univ. Toulon, Institut Univ. de France, glotin@univ-tln.fr
[3] Xeno-canto Foundation, The Netherlands, {wp,bob}@xeno-canto.org
[4] Vienna University of Technology, Austria, rauber@ifs.tuwien.ac.at
[5] LIRMM, Montpellier, France

Abstract. The LifeCLEF bird identification task provides a testbed for a system-oriented evaluation of 999 bird species identification. The main originality of this data is that it was specifically built through a citizen science initiative conducted by Xeno-Canto, an international social network of amateur and expert ornithologists. This makes the task closer to the conditions of a real-world application than previous, similar initiatives. This overview presents the resources and the assessments of the task, summarizes the retrieval approaches employed by the participating groups, and provides an analysis of the main evaluation results.

Keywords: LifeCLEF, bird, song, call, species, retrieval, audio, collection, identification, fine-grained classification, evaluation, benchmark, bioacoustics

1 Introduction

Accurate knowledge of the identity, the geographic distribution and the evolution of bird species is essential for a sustainable development of humanity as well as for biodiversity conservation. Unfortunately, such basic information is often only partially available for professional stakeholders, teachers, scientists and citizens. In fact, it is often incomplete for ecosystems that possess the highest diversity, such as tropical regions. A noticeable cause and consequence of this sparse knowledge is that identifying birds is usually impossible for the general public, and often a difficult task for professionals like park rangers, ecology consultants, and of course, the ornithologists themselves. This "taxonomic gap" [22] was actually identified as one of the main ecological challenges to be solved during United Nations Conference in Rio de Janeiro, Brazil, in 1992.

The use of multimedia identification tools is considered to be one of the most promising solutions to help bridging this taxonomic gap [14], [8], [6], [21], [20], [12]. With the recent advances in digital devices, network bandwidth and information storage capacities, the collection of multimedia data has indeed become an easy task. In parallel, the emergence of "citizen science" and social networking tools has fostered the creation of large and structured communities of nature

observers (e.g. eBird[6], Xeno-canto[7], iSpot [8], etc.) that have started to produce outstanding collections of audio and/or visual records. Unfortunately, the performance of the state-of-the-art multimedia analysis techniques on such data is still not well understood and it is far from reaching the real world's requirements in terms of identification tools. Most existing studies or available tools typically identify a few tens of species with moderate accuracy whereas they should be scaled-up to take one, two or three orders of magnitude more, in terms of number of species.

The LifeCLEF Bird task proposes to evaluate one of these challenges [**?**] based on big and real-world data and defined in collaboration with biologists and environmental stakeholders so as to reflect realistic usage scenarios.

Using audio records rather than bird pictures is justified by current practices [6], [21], [20], [5]. Birds are actually not easy to photograph; audio calls and songs have proven to be easier to collect and sufficiently species specific.

Only three notable previous worldwide initiatives on bird species identification based on their songs or calls have taken place, all three in 2013. The first one was the ICML4B bird challenge joint to the International Conference on Machine Learning in Atlanta, June 2013 [2]. It was initiated by the SABIOD MASTODONS CNRS group[9], the University of Toulon and the National Natural History Museum of Paris [9]. It included 35 species, and 76 participants submitted their 400 runs on the Kaggle interface. The second challenge was conducted by F. Brigs at MLSP 2013 workshop, with 15 species, and 79 participants in August 2013. The third challenge, and biggest in 2013, was organised by University of Toulon, SABIOD and Biotope [4], with 80 species from the Provence, France. More than thirty teams participated, reaching 92% of average AUC. Descriptions of the best systems of ICML4B and NIPS4B bird identification challenges are given in the on-line books [2,1] including, in some cases, references to useful scripts.
In collaboration with the organizers of these previous challenges, BirdCLEF 2014 and 2015 go one step further by (i) significantly increasing the species number by almost an order of magnitude (ii) working on real-world data collected by hundreds of recordists (iii) moving to a more usage-driven and system-oriented benchmark by allowing the use of meta-data and defining information retrieval oriented metrics. Overall, the task is expected to be much more difficult than previous benchmarks because of the higher confusion risk between the classes, the higher background noise and the higher diversity in the acquisition conditions (devices, recordists uses, contexts diversity, etc.). It will therefore probably produce substantially lower scores and offer a better progression margin towards building real-world generalist identification tools.

[6] http://ebird.org/
[7] http://www.xeno-canto.org/
[8] http://www.ispotnature.org/communities/global
[9] `http://sabiod.univ-tln.fr`

2 Dataset

The training and test data of the bird task is composed by audio recordings hosted on xeno-canto.org (XC). Xeno-canto is a web-based community of bird sound recordists worldwide with more than 2300 active contributors that have already collected more than 240,000 recordings of about 9330 species (may 2015). 999 species from Brazil are used in the BirdCLEF dataset. They represent the species of that country with the highest number of recordings on XC, totalling 33,862 recordings contributed by hundreds of users. The dataset has between 13 and 234 recordings per species, recorded by between 1 and 72 recordists. This dataset also contains the entire dataset from the 2014 BirdCLEF challenge [10], which contained about 14,000 recordings from 501 species.

To avoid any bias in the evaluation related to the audio devices used, each audio file has been normalized to a constant bandwidth of 44.1 kHz and coded over 16 bits in .wav mono format (the right channel was selected by default). The conversion from the original Xeno-canto data set was done using ffmpeg, sox and matlab scripts. An optimized 16 Mel Filter Cepstrum Coefficients for bird identification (according to an extended benchmark [7]) have been computed with their first and second temporal derivatives on the whole set. They were used in the best systems run in ICML4B and NIPS4B challenges [2], [1],[4], [9].

Audio records are associated with various meta-data including the species of the most active singing bird, the species of the other birds audible in the background, the type of sound (call, song, alarm, flight, etc.), the date and location of the observations (from which rich statistics on species distribution can be derived), common names and collaborative quality ratings. All of them were produced collaboratively by the Xeno-canto community.

3 Task Description

Participants were asked to determine the species of the most active singing birds in each query file. The background noise can be used as any other meta-data, but it is forbidden to correlate the test set of the challenge with the original annotated Xeno-canto data base (or with any external content as many of them are circulating on the web). More precisely, the whole BirdCLEF dataset has been split in two parts, one for training (and/or indexing) and one for testing. The test set was built by randomly choosing 1/3 of the observations of each species whereas the remaining observations were kept in the reference training set. Recordings of the same species done by the same person the same day are considered as being part of the same observation and cannot be split across the test and training set. The xml files containing the meta-data of the *query* recordings were purged so as to erase the foreground and background species names (the ground truth), the vernacular names (common names of the birds) and the collaborative quality ratings (that would not be available at query stage in a real-world mobile application). Meta-data of the recordings in the training set are kept unaltered.

The groups participating to the task were asked to produce up to 4 runs containing a ranked list of the most probable species for each record of the test set. Each species had to be associated with a normalized score in the range $[0, 1]$ reflecting the likelihood that this species was singing in the sample. For each submitted run, participants had to say if the run was performed fully automatically or with a human assistance in the processing of the queries, and if they used a method based on only audio analysis or with the use of the metadata. The metric used to compare the runs was the Mean Average Precision averaged across all queries. Since the audio records contain a main species and often some background species belonging to the set of 501 species in the training, we decided to use two metrics, one focusing on all species (MAP1) and a second one focusing only on the main species (MAP2).

4 Participants and methods

137 research groups worldwide registered for the task and downloaded the data (from a total of 189 groups that registered for at least one of the three LifeCLEF tasks). This shows the high attractiveness of the challenge in both the multimedia community (presumably interested in several tasks) and in the audio and bioacoustics community (presumably registered only to the bird songs task). Finally, 6 of the registrants crossed the finish line by submitting runs and 5 of them submitted working notes explaining their runs in details. We list them hereafter in alphabetical order and give a brief overview of the techniques they used in their runs. We would like to point out that the LifeCLEF benchmark is a system-oriented evaluation and not a deep or fine evaluation of the underlying algorithms. Readers interested in the scientific and technical details of the implemented methods should refer to the LifeCLEF 2015 working notes or to the research papers of each participant (referenced below):

CHIN. AC. SC., China, 3 runs: This participant attempted to experiment a baseline audio classification system based on the classification of Mel-bands representations and their scattering refinements [3] using a Gaussian Mixture Model. The first run used only MFCC features with 128 Gaussian mixtures, the second run used the scattering refinements with 32 Gaussian mixtures, the third run used the scattering refinements with 128 Gaussian mixtures.

Golem, Mexico, 3 runs [15]: This participant experimented a simple yet highly scalable system based on the classification of Mel-bands representations using a random forest. The extracted Mel bands per recording were actually pooled through simple statistics (i.e. mean, standard deviation, median and skewness), resulting in time- and space-efficient 320-dimensional features to be trained by the classifier.

Inria Zenith, France, 3 runs [11]: Inspired by recent works on fine-grained image classification, this group introduced a new match kernel based on the shared nearest neighbors of the low level audio features extracted at the frame level. To make such strategy scalable to the tens of millions of MFCC features extracted from the training set, they make use of high-dimensional hashing techniques coupled with an efficient approximate nearest neighbors search algorithm with controlled quality. Further improvements are obtained by (i) using a sliding window for the temporal pooling of the raw matches (ii) weighting each low level feature according to the semantic coherence of its nearest neighbors. The final classification was then completed thanks to a support vector machine trained on top of the resulting matching-based representations.

MARF, Canada, 4 runs [17]: These participants mainly attempted to transpose a speech processing method they developed earlier to the birds case (Modular Audio Recognition Framework (MARF)'s API, [16]). The first run was using only 20 LPC coefficients as features and the Chebyshev distance. The second run was using only the meta-data features using the MARFCAT approach [16] to represent the XML meta-data as a wave form without pre-processing, and using 512-window FFT features and cosine similarity measure. The third run was a concatenation of Run 1 and Run 2. The fourth run used the same set up as Run 1 but split the training data by quality ratings attributes.

MNB TSA, Germany, 4 runs [13]: This participant combined two main categories of features for the classification: parametric acoustic features (see openSMILE Audio Statistics) and probabilities of species-specific spectrogram segments (see Segment-Probabilities). This second source of information, which performs the best, consists in extracting for each species, a set of representative segments from spectrogram images. These segments are then used to extract Segment-Probabilities for each file by calculating the maxima of the normalized cross-correlation between all segments and the target spectrogram image via template matching. Due to the very large amount of audio data not all files belonging to a certain species were used as a source for segmentation (i.e. only good quality files without background species were used). Additionally, to further reduce the computation time, the spectrogram images were downsmapled before computing the template matching. The classification problem was then formulated as a multi-label regression task completed by training ensembles of randomized decision trees with probabilistic outputs. The training was performed in two passes, one selecting a small subset of the most discriminant features, and one training the final classifiers on the selected features (Run 1). To further improve classification results a bagging approach was used consisting in calculating further Segment-Probabilities from additional segments and to combine them either by averaging (Run 2) or by blending (Run 3 and Run 4 with more blends).

QMUL, UK, 1 run [18]: This group focused on unsupervised feature learning in order to learn regularities in spectro-temporal content without reference to

the training labels and further help the classifier to generalise to further content of the same type. MFCC features and several temporal variants are first extracted from the audio signal after a median-based thresholding pre-processing. Extracted low level features were then reduced through PCA whitening and clustered via spherical k-means (and a two-layer variant of it) to build the vocabulary. During classification, MFCC features are pooled by projecting them on the vocabulary with different temporal pooling strategies. Final supervised classification is achieved thanks to a random forest classifier. This method is the subject of a full-length article which can be read at [19]. Details of the different parameters settings used in each run are detailed in the working note [**?**].

5 Results

Figure 1 and table 1 show the scores obtained by all the runs for the two distinct measured Mean Average Precision (MAP) evaluation measures: MAP 1 when considering only the foreground species of each test recording and MAP 2 when considering additionally the species listed in the *Background species* field of the metadata.

Table 1: Raw results of the LifeCLEF 2014 Bird Identification Task

Run name	Type	MAP 1 (without Bg. Sp.)	MAP 2 (with Bg Sp.)
MNB TSA Run 4	AUDIO	0.454	0.414
MNB TSA Run 3	AUDIO	0.442	0.411
MNB TSA Run 2	AUDIO	0.442	0.405
MNB TSA Run 1	AUDIO	0.424	0.388
INRIA ZENITH Run 2	AUDIO	0.334	0.291
QMUL Run 1	AUDIO	0.302	0.262
INRIA ZENITH Run 3	AUDIO	0.292	0.259
INRIA ZENITH Run 1	AUDIO	0.265	0.240
GOLEM Run 2	AUDIO	0.171	0.149
GOLEM Run 1	AUDIO	0.161	0.139
CHIN. AC. SC. Run 1	AUDIO	0.01	0.009
CHIN. AC. SC. Run 3	AUDIO	0.009	0.01
CHIN. AC. SC. Run 2	AUDIO	0.007	0.008
MARF Run 1	AUDIO	0.006	0.005
MARF Run 2	METADATA	0.003	0.002
MARF Run 3	AUDIO & METADATA	0.005	0.005
MARF Run 4	AUDIO	0.000	0.000

The main outcome of the evaluation is that the use of matching-based scores as high-dimensional features to be classified by supervised classifiers (as done

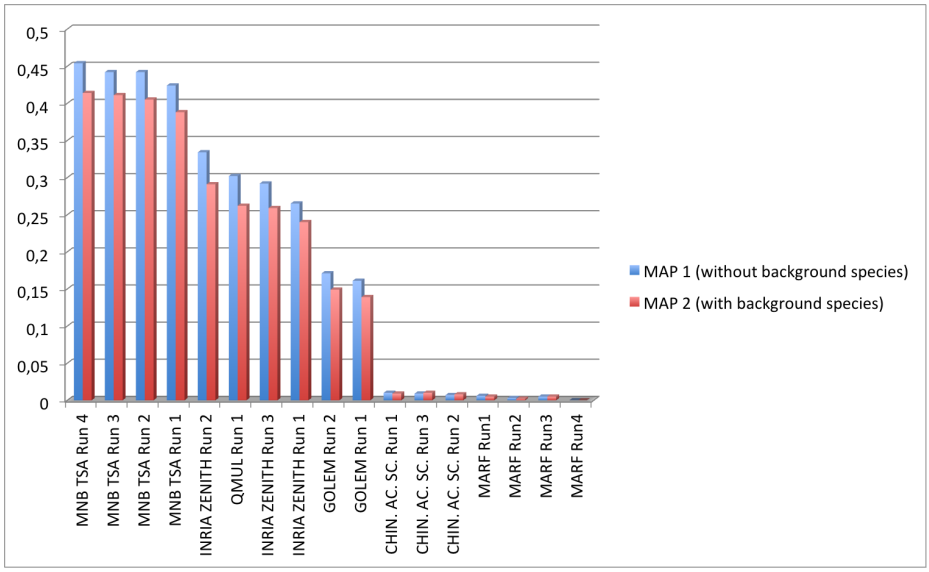

Fig. 1. Official scores of the LifeCLEF Bird Identification challenge 2015. MAP 2 is the Mean Average Precision averaged across all queries taking into account the Background species (while MAP 1 considers only the foreground species).

by MNB TSA and INRIA ZENITH) provides the best results, with a Mean Average Precision up to 0.454 for the fourth run of the MNB TSA group. These approaches notably outperform the unsupervised feature learning framework of the QMUL group as well as the baseline method of the Golem group. The matching of all the audio recordings however remains a very time-consuming process that had to be carefully designed in order to process a large-scale dataset such as the one deployed within the challenge. The MNB TSA group notably reduced as much as possible the number of audio segments to be matched thanks to an effective audio pre-processing and segmentation framework. They also restricted the extraction of these segments to the files having the best quality according to the user ratings and that do not have background species. On the other side, the INRIA ZENITH group did not use any segmentation but attempted to speed-up the matching though the use of a hash-based approximate k-nearest neighbors search scheme (on top of MFCC features). The better performance of the MNB TSA runs shows that cleaning the audio segments vocabulary before applying the matching is clearly beneficial. But using a scalable knn-based matching as the one of the INRIA ZENITH runs could be a complementary way to speed up the matching phase.

It is interesting to notice that the first run of the MNB TSA group is roughly the same method than the one they used within the BirdCLEF challenge of the previous year [10] and which achieved the best results (with a MAP1 equals to

0.511 vs. 0.424 this year). This shows that the impact of the increasing difficulty of the challenge (with twice the number of species) is far from negligible. The performance loss is notably not compensated by the bagging extension of the method which resulted in a MAP1 equals to 0.454 for MNB TSA run 4.
As a final comment on this evaluation study, it is worth noting that none of the participants attempted to evaluate deep learning approaches such as using deep convolutional neural networks (CNN) that have been recently shown to achieve excellent classification performance on both image and audio contents. The most likely reason is that the use of external training data was not allowed. It was consequently not possible to employ transfer learning mechanisms such as specializing a CNN previously trained on a large generalist training set. Without using such strategy, the provided training data might be insufficiently large to train the millions of parameters of the deep networks.

6 Conclusion

This paper presented the overview and the results of the first LifeCLEF bird identification challenge 2015. With a number of registrant exceeding hundred, it showed a high interest of the multimedia and the bio-acoustic communities in applying their technologies to real-world environmental data such as the ones collected by Xeno-canto. The main outcome of this evaluation is a snapshot of the performances of state-of-the-art techniques that will hopefully serve as a guideline for developers interested in building end-user applications. One important conclusion of the campaign is that the two best performing methods were based on matching approaches attempting to construct high-dimensional representations of the audio recordings based on their matching scores in a large vocabulary of audio segments. The results of the evaluation clearly show the superiority of these approaches in terms of effectiveness but also point out the underlying scalability issues in terms of efficiency. The increasing complexity of the challenge over the previous year in terms of the number species and items, notably conducted to a consistent loss of the raw identification performance despite the progress of the underlying methods. Considering that the number of bird species on earth is more than 10,000 and that the number of singing insects is even much larger, we believe it is important to continue working on such large-scale identification issues in the next years.

References

1. Proc. of Neural Information Processing Scaled for Bioacoustics: from Neurons to Big Data, joint to NIPS (2013), `http://sabiod.univ-tln.fr/NIPS4B2013_book.pdf`
2. Proc. of the first workshop on Machine Learning for Bioacoustics, joint to ICML (2013), `http://sabiod.univ-tln.fr/ICML4B2013_book.pdf`
3. Andén, J., Mallat, S.: Multiscale scattering for audio classification. In: ISMIR. pp. 657–662 (2011)

4. Bas, Y., Dufour, O., Glotin, H.: Overview of the nips4b bird classification. In: Proc. of Neural Information Processing Scaled for Bioacoustics: from Neurons to Big Data, joint to NIPS. pp. 12–16 (2013), http://sabiod.univ-tln.fr/NIPS4B2013_book.pdf
5. Briggs, F., Lakshminarayanan, B., Neal, L., Fern, X.Z., Raich, R., Hadley, S.J., Hadley, A.S., Betts, M.G.: Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. The Journal of the Acoustical Society of America 131, 4640 (2012)
6. 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)
7. Dufour, O., Artieres, T., Glotin, H., Giraudet, P.: Clusterized mel filter cepstral coefficients and support vector machines for bird song idenfication. In: Soundscape Semiotics - Localization and Categorization, Glotin (Ed.) (2014)
8. Gaston, K.J., O'Neill, M.A.: Automated species identification: why not? Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 359(1444), 655–667 (2004), http://rstb.royalsocietypublishing.org/content/359/1444/655.abstract
9. Glotin, H., Sueur, J.: Overview of the 1st int'l challenge on bird classification. In: Proc. of the first workshop on Machine Learning for Bioacoustics, joint to ICML. pp. 17–21 (2013), http://sabiod.univ-tln.fr/ICML4B2013_book.pdf
10. Goëau, H., Glotin, H., Vellinga, W.P., Rauber, A.: Lifeclef bird identification task 2014
11. Joly, A., Champ, J., Buisson, O.: Shared nearest neighbors match kernel for bird songs identification - lifeclef 2015 challenge. In: Working notes of CLEF 201 conference (2015)
12. 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)
13. Lasseck, M.: Improved automatic bird identification through decision tree based feature selection and bagging. In: Working notes of CLEF 2015 conference (2015)
14. Lee, D.J., Schoenberger, R.B., Shiozawa, D., Xu, X., Zhan, P.: Contour matching for a fish recognition and migration-monitoring system. In: Optics East. pp. 37–48. International Society for Optics and Photonics (2004)
15. Meza, I., Espino-Gamez, A., Solano, F., Villarreal, E.:
16. Mokhov, S.A.: Study of best algorithm combinations for speech processing tasks in machine learning using median vs. mean clusters in marf. In: Proceedings of the 2008 C 3 S 2 E conference. pp. 29–43. ACM (2008)
17. Mokhov, S.A.: A marfclef approach to lifeclef 2015 tasks. In: Working notes of CLEF 2015 conference (2015)
18. Stowell, D.: Birdclef 2015 submission: Unsupervised feature learning from audio. In: Working notes of CLEF 2015 conference (2015)
19. Stowell, D., Plumbley, M.D.: Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. arXiv preprint arXiv:1405.6524 (2014)
20. Towsey, M., Planitz, B., Nantes, A., Wimmer, J., Roe, P.: A toolbox for animal call recognition. Bioacoustics 21(2), 107–125 (2012)
21. Trifa, V.M., Kirschel, A.N., Taylor, C.E., Vallejo, E.E.: Automated species recognition of antbirds in a mexican rainforest using hidden markov models. The Journal of the Acoustical Society of America 123, 2424 (2008)

22. Wheeler, Q.D., Raven, P.H., Wilson, E.O.: Taxonomy: Impediment or expedient? Science 303(5656), 285 (2004), http://www.sciencemag.org/content/303/5656/285.short