LifeCLEF Bird Identification Task 2016: The arrival of Deep learning

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

LifeCLEF Bird Identification Task 2016

The arrival of Deep learning

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

[1] IRD, UMR AMAP, Montpellier, France herve.goeau@cirad.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] Inria ZENITH team, Montpellier, France alexis.joly@inria.fr
[5] LIRMM, Montpellier, France

Abstract. The LifeCLEF bird identification challenge provides a large-scale testbed for the system-oriented evaluation of bird species identification based on audio recordings. One of its main strength is that the data used for the evaluation is collected through Xeno-Canto, the largest network of bird sound recordists in the world. This makes the task closer to the conditions of a real-world application than previous, similar initiatives. The main novelty of the 2016-th edition of the challenge was the inclusion of *soundscape recordings* in addition to the usual xeno-canto recordings that focus on a single foreground species. This paper reports the methodology of the conducted evaluation, the overview of the systems experimented by the 6 participating research groups and a synthetic analysis of the obtained results.

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

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. The general public as well as professionals like park rangers, ecological consultants and of course the ornithologists themselves are potential users of an automated bird identifying system, typically in the context of wider initiatives related to ecological surveillance or biodiversity conservation. The LifeCLEF Bird challenge proposes to evaluate the state-of-the-art of audio-based bird identification systems at a very large scale. Before LifeCLEF started in 2014, three previous initiatives on the evaluation of acoustic bird species identification took place, including two from the SABIOD[6] group

[6] Scaled Acoustic Biodiversity http://sabiod.univ-tln.fr

[4,3,1]. In collaboration with the organizers of these previous challenges, BirdCLEF 2014, 2015 and 2016 challenges went one step further by (i) significantly increasing the species number by an order of magnitude, (ii) working on real-world social data built from thousands of recordists, and (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 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 (different recording devices, contexts diversity, etc.). It therefore produces substantially lower scores and offers a better progression margin towards building real-world generalist identification tools.
The main novelty of the 2016-th edition of the challenge with respect to the two previous years was the inclusion of *soundscape recordings* in addition to the usual xeno-canto recordings that focus on a single foreground species (usually thanks to mono-directional recording devices). Soundscapes, on the other hand, are generally based on omnidirectional recording devices that continuously monitor a specific environment over a long period. This new kind of recording fits better to the (possibly crowdsourced) passive acoustic monitoring scenario that could augment the number of collected records by several orders of magnitude. In this paper, we report the methodology of the conducted evaluation as well as the synthetic analysis of the results achieved by the 6 participating groups.

2 Dataset

The training and test data of the challenge consists of audio recordings collected by Xeno-canto (XC)[7]. Xeno-canto is a web-based community of bird sound recordists worldwide with about 3,000 active contributors that have already collected more than 300,000 recordings of about 9550 species (numbers for June 2016). Nearly 1000 (in fact 999) species were used in the BirdCLEF dataset, representing the 999 species with the highest number of recordings in October 2014 (14 or more) from the combined area of Brazil, French Guiana, Surinam, Guyana, Venezuela and Colombia, totalling 33,203 recordings produced by thousands of users. This dataset includes the entire dataset from the 2015 BirdCLEF challenge [5], which contained about 33,000 recordings. The newly introduced test data in 2016, contains 925 soundscapes provided by 7 xeno-canto members, sometimes working in pairs. Most of the soundscapes have a length of (more or less) 10 minutes, each coming often from a set of 10-12 successive recording made at one location. The total duration of new testing data to process and analyse is thus equivalent to approximately 6 days of continuous sound recording. The number of known species (i.e. belonging to the 999 species in the training dataset) varies from 1 to 25 species, with an average of 10.1 species per soundscape.

To avoid any bias related to the used audio devices in the evaluation , each audio file was normalized to a constant bandwidth of 44.1 kHz and coded with 16 bits in wav mono format (the right channel is selected by default). The conversion

[7] `http://www.xeno-canto.org/`

from the original Xeno-canto data set was done using ffmpeg, sox and matlab scripts. The optimized 16 Mel Filter Cepstrum Coefficients for bird identification (according to an extended benchmark [2]) were 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. However, due to some technical limitations, the soundscapes were not normalized and directly provided to the participants in mp3 format (shared on the xeno-canto website, the original raw files being not available).

All audio records are associated with various meta-data including the species name 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 may be derived), some textual comments by the authors, multilingual 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 all the active singing birds species in each query file. It was forbidden to correlate the test set of the challenge with the original annotated Xeno-canto database (or with any external content as many of them are circulating on the web). The whole data was split in two parts, one for training (and/or indexing) and one for testing. The test set was composed of (i) all the newly introduced soundscapes recordings and (ii), the entire test set used in 2015 (equal to about 1/3 of the observations in the whole 2015 dataset). The training set was exactly the same as the one used in 2015 (i.e. the remaining 2/3 of the observations). Note that recordings of the same species made by the same person on 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 taxon name (the ground truth), the vernacular name (common name of the bird) 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 were kept unaltered.

The groups participating in the task were asked to produce up to 4 runs containing a ranked list of the most probable species for each query records of the test set. Each species was associated with a normalized score in the range $[0, 1]$ reflecting the likelihood that this species is singing in the sample. For each submitted run, participants had to say if the run was performed fully automatically or with human assistance in the processing of the queries, and if they used a method based only on audio analysis or with the use of the metadata.

The primary metric used was the mean Average Precision (mAP) averaged across all queries, considering each audio file of the test set as a query and

computed as:

$$mAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q},$$

where Q is the number of test audio files and $AveP(q)$ for a given test file q is computed as

$$AveP(q) = \frac{\sum_{k=1}^{n}(P(k) \times rel(k))}{\text{number of relevant documents}}.$$

Here k is the rank in the sequence of returned species, n is the total number of returned species, $P(k)$ is the precision at cut-off k in the list and $rel(k)$ is an indicator function equaling 1 if the item at rank k is a relevant species (i.e. one of the species in the ground truth).

4 Participants and methods

84 research groups worldwide registered for the task and downloaded the data (from a total of 130 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 detail. 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 2016 working notes or to the research papers of each participant (referenced below):

BME TMIT, Hungary, 4 runs [11]: BME TMIT is one of the three teams who used a Convolutional Neural Network with CUBE and WUT teams. As pre-processing, they first downsampled each audio file to 16 kHz frequency and applied a low-pass filter with cutoff frequency of 6250 Hz in order to reduce the size of the training data. Then they subdivided the spectograms into cells of 0.5 seconds x 10 bands of frequency, and removed the cells with few information (according to the mean and variance). After these preprocessing steps, they assembled and re-split the remaining parts of the spectrograms to five second long pieces, and obtained arrays of 200310 (where 310 samples corresponds to five seconds), used as input of the CNN. They used two distincts CNN architectures: the well-know AlexNet [6] with the addition of a batch normalisation (run 1 & 2), and a CNN more inspired by audio recognition systems based on 4 convolutional layers, one full connected layer, ReLU activation functions and batch normalisation (run 3 & 4).

CUBE, Switzerland, 4 runs: This system is based on a CNN architecture

of 5 convolutional layers combined with the use of a rectify activation function followed by a max-pooling layer. Based on spectrogram analysis and morphological operations, silent and noisy parts were first detected and separated from the call and song parts. Spectrograms were then split into chunks of 3 seconds that were used as inputs of the CNN after several data augmentation techniques. Each chunk identified as a singing bird was first concatenated with 3 randomly selected chunks of background noise. Time shift, pitch sift and mixes of audio files from the same species were then used as complementary data augmentation techniques. Considering one test record, all predictions from its distinct chunks are finally averaged. Run 1 was an intermediate result obtained after only one day of training. Run 2 differs from run 3 by using 50% smaller spectrograms in (pixel) size for doubling the batch size and thus allowing to have more iterations for the same training time (4 days). Run 4 is the average of predictions from run 2 and 3 and reaches the best performance, showing the benefit of bagging.

DYNI LSIS, France, 1 runs [10]: The algorithm presented here is quite standard and was initially used on smaller datasets to improve, in a late fusion scheme, a classifier based on pairs of spectrogram peaks, described in the context of audio fingerprinting. The method is based on the bag-of-words approach: first the 44.1 kHz audio files were split in 0.2s segments with 50% overlap, and only the segments having energy values higher than a relative (to the whole audio file) value and spectral flatness values smaller than an absolute thresh-old were kept for Mel Frequency Cepstral Coefficient computation (MFCC). A k-means clustering was performed on all the MFCC and their derivatives with k=500, in order to extract for every files the normalized histogram of MFCC-based words (i.e. the 500 clusters), using only segments kept in step 2. The resulting feature vectors were then fed to a random forest classifier.

MNB TSA, Germany, 4 runs [8]: As in 2014 and 2015, this participant used two hand-crafted parametric acoustic features and probabilities of species-specific spectrogram segments in a template matching approach. Long segments extracted during BirdCLEF2015 were re-segmented with a more sensitive algorithm. The segments were 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 were used as a source for segmentation (i.e. only good quality files without background species were used). The classification problem was then formulated as a multi-label regression task solved by training ensembles of randomized decision trees with probabilistic outputs. The training was performed in 2 passes, one selecting a small subset of the most discriminant features by optimizing the internal mAP score on the training set, and one training the final classifiers on the selected features. Run 1 used one single model on a small but highly optimized selection of Segment-Probabilities. A bagging approach was used consisting in calculating further Segment-Probabilities from additional segments and to combine them either by blending (24 models

in Run 3). Run 4 also used blending to aggregate model predictions, but the predictions were included that after blending resulted in the highest possible mAP score calculated on the entire training set (13 models including the best model from 2015).

WUT, Poland, 4 runs [9]: as the Cube and the BME TMIT teams, they used a Convolutional Neural Network learning framework. Starting from denoised spectrograms, silent parts were removed with percentile thresholding, giving thus around 86,000 training segments varying in length and associated each with a single main species. As a data augmentation technique and for fitting the 5 seconds fixed input size of the CNN, segments were adjusted by either trimming or padding. The 3 first successive runs are produced by deeper and deeper, or/and, wider and wider filters. Run 4 is as an ensemble of neural networks averaging the predictions of the 3 first runs.

5 Results

Figure 1 and table 1 show the scores obtained by all the runs for the three distinct measured mean Average Precision (mAP) evaluation measures:

Table 1: Results of the LifeCLEF 2016 Bird Identification Task

RunNameShort	Official score: mAP (with background species)	mAP (only main species. same queries BirdCLEF2015)	mAP with background species (only queries 2016 "Soundscape")
Cube Run 4	0.555 (1)	0.686 (1)	0.072 (5)
Cube Run 3	0.536 (2)	0.660 (2)	0.078 (4)
Cube Run 2	0.522 (3)	0.649 (3)	0.066
MarioTsaBerlin Run 1	0.519 (4)	0.585 (4)	0.137 (1)
MarioTsaBerlin Run 4	0.472 (5)	0.551 (5)	0.129 (3)
WUT Run 4	0.412	0.529	0.036
MarioTsaBerlin Run 3	0.396	0.456	0.130 (3)
WUT Run 2	0.376	0.483	0.032
WUT Run 3	0.352	0.455	0.029
WUT Run 1	0.35	0.453	0.027
BME TMIT Run 2	0.338	0.426	0.053
BME TMIT Run 3	0.337	0.426	0.059
MarioTsaBerlin Run 2	0.336	0.399	0.000
BME TMIT Run 4	0.335	0.424	0.053
BME TMIT Run 1	0.323	0.407	0.054
Cube Run 1	0.284	0.364	0.020
LSIS naive MFCC Run 1	0.149	0.183	0.037
BIG Run 1	0.021	0.021	0.004
Best run BirdCLEF 2015	-	0.454	-

Figure 1 reports the performance measured for the 18 submitted runs. For each run (*i.e.* each evaluated system), we report the overall mean Average Precision (official metric) as well as the mAP for the two categories of queries: the soundscapes recordings (newly introduced) and the common observations (the same as the one used in 2015). To measure the progress over last year, we also plot on the graph the performance of the last year best system [7] (orange dotted line). The first noticeable conclusion is that, after two years of resistance of bird songs identification systems based on engineering features, convolutional neural networks finally managed to outperform them (as in many other domains). The best run based on CNN (Cube Run 4) actually reached an impressive mAP of 0.69 on the 2015 testbed to be compared to respectively 0.45 and 0.58 for the best systems based on hand-crafted features evaluated in 2015 and 2016. To our knowledge, BirdCLEF is the first comparative study reporting such an important performance gap in bioacoustic large-scale classification. A second important remark is that this performance of CNN's was achieved without any fine-tuning contrary to most computer vision challenges in which the CNN is generally pretrained on a large training data such as ImageNet. Thus, we could hope even better performance, e.g. by transferring knowledge from other bio-acoustic contexts or other domains. Now, it is important to notice that the other systems based on CNN (WUT and BME TMIT) did not perform as well as the Cube system and did not outperformed the system of TSA based on hand-crafted features. Looking at the detailed description of the three CNN architectures and their learning framework, it appears that the way in which audio segment extraction and data augmentation is performed does play a crucial role. Cube system does notably include a randomized background noise addition phase which makes it much more robust to the diversity of noise encountered in the test data. If we now look at the scores achieved by the evaluated systems on the soundscape recordings only (yellow plot), we can draw very different conclusions. First of all, we can observe that the performance on the soundscapes is much lower than on the classical queries, whatever the system. Although the classical recordings also include multiple species singing in the background, the soundscapes appear to be much more challenging. Several tens of species and even much more individual birds can actually be singing simultaneously. Separating all these sources seem to be beyond the scope of state-of-the-art audio representation learning methods. Interestingly, the best system on the soundscape queries was the one of TSA based on the extraction of very short species-specific spectrogram segments and a template matching approach. This very fine-grained approach allows the extracted audio patterns to be more robust to the species overlap problem. On the contrary, the CNN of Cube and WUT systems were optimized for the mono-species segment classification problem. The data augmentation method of the Cube system was in particular only designed for the single species case. It addressed the problem of several individual birds of the same species singing together (by mixing different segments of the same class) but it did not address the multi-label issue (i.e. several species singing simultaneously).

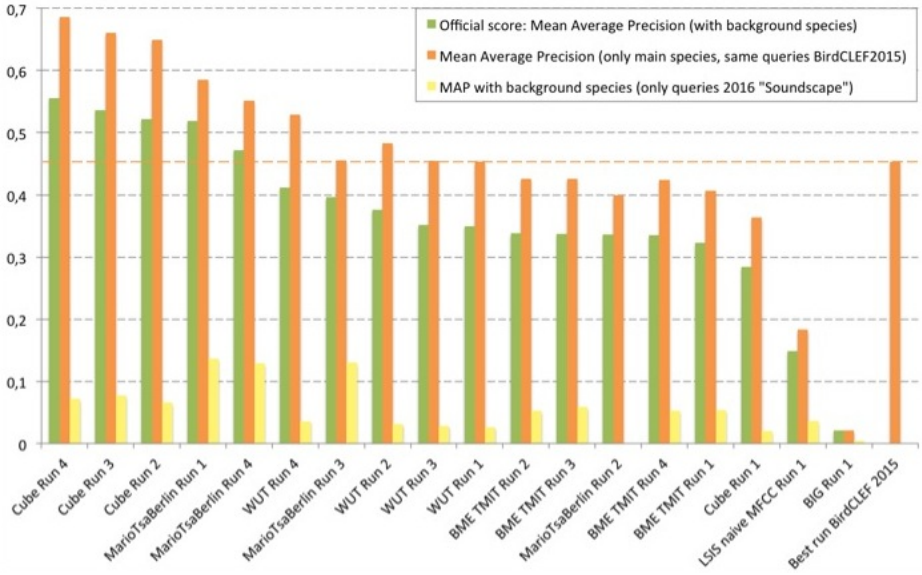

Fig. 1. Official scores of the LifeCLEF Bird Identification challenge 2016.

6 Complementary results

To study in more details the dynamic of the identification performance across the diversity of species, Figure 2 presents the scores achieved by the best system of each team on a selection of 3x10 species: (i) the top-10 best recognized ones (according to the performance of the best system *Cube Run 4*), (ii) 10 species of intermediate difficulties and (iii) the worst-10 recognized ones (still based on the performance of *Cube Run 4*). For a better interpretation of the chart, we also included for each of the 30 selected species, the number of audio recordings in the training set (ranging from 10 to 37 recordings). The graph first shows that there is a huge performance gap between the best recognized species and the worst cases. Some species are actually perfectly classified by 4 of the 6 systems whereas some others are never recognized by none of the systems. Interestingly, one can see that the performance does not seem to be correlated to the number of training samples. In the same way, we did observed that it is not correlated to the average length of the recordings in the class. This means that the high variability in performance is more related to other factors such as (i) the bird sounds variability (some birds are more audible than others), (ii) the acquisition difficulty (some birds are easier to record than others), (iii) the degree of confusion across close species. Another interesting remark is that two of the species that are not recognized at all by the CNN are comparatively pretty well recognized by the template matching kernel approach of MNB TSA. Thus, it would be interesting to study in more details the kind of audio patterns that have been

matched by their method so as to understand what the CNN missed and how such patterns could be automatically learned as well.

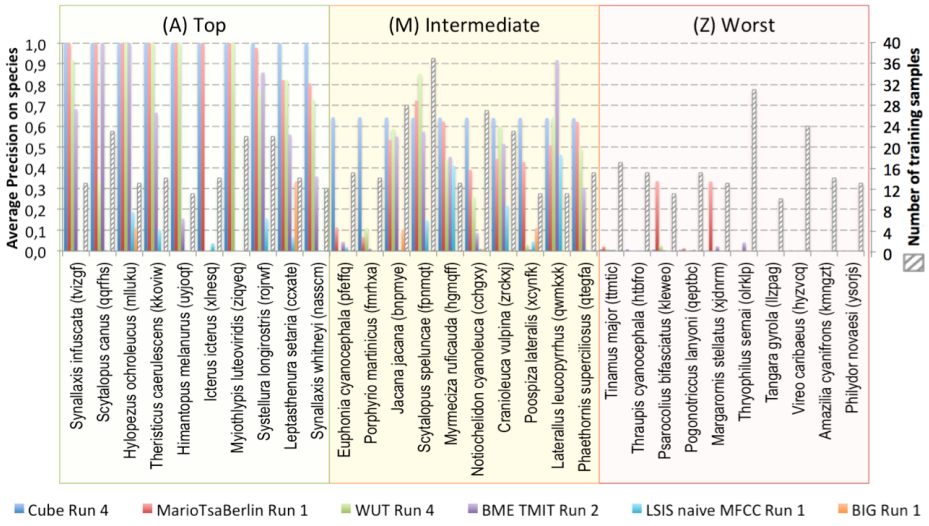

Fig. 2. Average Precision detailed on a selection of 3x10 species for the best run of each team, following the ranking given by the best overall system *Cube Run 4*: (A) the top-10 species, (M) intermediate species (species ranked from 495 to 505), and (Z) species with the lowest APs.

7 Conclusion

This paper presented the overview and the results of the LifeCLEF bird identification challenge 2016. The main outcome was that after two years of resistance of bird song identification systems based on engineering features, convolutional neural networks finally managed to outperform them with a significant margin. It is noticeable that the best performing CNN did not used any fine-tuning so that it did not benefit from the transfer learning capacities of that techniques. We could thus expect even better performances. Also, the used CNN architecture was mostly inspired by the ones which perform the best on computer vision tasks. Our detailed analysis of the results tend to show that some audio patterns might not be learned accurately through such network whereas they are detected through template matching techniques. Anyway, it is obvious that, as in many domains beforehand, deep learning is redefining the boundaries of the state-of-the-art and opens the door to further progress in the next years.

References

1. Briggs, F., Huang, Y., Raich, R., Eftaxias, K., et al., Z.L.: The 9th mlsp competition: New methods for acoustic classification of multiple simultaneous bird species in noisy environment. In: IEEE Workshop on Machine Learning for Signal Processing (MLSP). pp. 1–8 (2013)
2. 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), http://www.intechopen.com/books/soundscape-semiotics-localisation-and-categorisation
3. Glotin, H., Clark, C., LeCun, Y., Dugan, P., Halkias, X., Sueur, J.: Bioacoustic challenges in icml4b. In: in Proc. of 1st workshop on Machine Learning for Bioacoustics. No. USA, ISSN 979-10-90821-02-6 (2013), http://sabiod.org/ICML4B2013_proceedings.pdf
4. Glotin, H., Dufour, O., Bas, Y.: Overview of the 2nd challenge on acoustic bird classification. In: Proc. Neural Information Processing Scaled for Bioacoustics. NIPS Int. Conf., Ed. Glotin H., LeCun Y., Artières T., Mallat S., Tchernichovski O., Halkias X., USA (2013), http://sabiod.univ-tln.fr/nips4b
5. Goëau, H., Glotin, H., Vellinga, W.P., Planque, R., Rauber, A., Joly, A.: Lifeclef bird identification task 2015. In: CLEF working notes 2015 (2015)
6. 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)
7. Lasseck, M.: Improved automatic bird identification through decision tree based feature selection and bagging. In: Working notes of CLEF 2015 conference (2015)
8. Lasseck, M.: Improving bird identification using multiresolution template matching and feature selection during training. In: Working notes of CLEF conference (2016)
9. Piczak, K.: Recognizing bird species in audio recordings using deep convolutional neural networks. In: Working notes of CLEF 2016 conference (2016)
10. Ricard, J., Glotin, H.: Bag of mfcc-based words for bird identification. In: Working notes of CLEF 2016 conference (2016)
11. Tóth, B.P., Czeba, B.: Convolutional neural networks for large-scale bird song classification in noisy environment. In: Working notes of CLEF conference (2016)