LifeCLEF Bird Identification Task 2017

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, Universit de Toulon, CNRS, ENSAM, LSIS UMR 7296, IUF, 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 challenge BirdCLEF offers a large-scale proving ground for system-oriented evaluation of bird species identification based on audio recordings of their sounds. One of its strengths is that it uses data collected through Xeno-canto, the worldwide community of bird sound recordists. This ensures that BirdCLEF is close to the conditions of real-world application, in particular with regard to the number of species in the training set (1500). The main novelty of the 2017 edition of BirdCLEF was the inclusion of *soundscape recordings* containing time-coded bird species annotations in addition to the usual Xeno-canto recordings that focus on a single foreground species. This paper reports an overview of the systems developed by the five participating research groups, the methodology of the evaluation of their performance, and an analysis and discussion of the results obtained.

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 BirdCLEF proposes to evaluate the state-of-the-art of audio-based bird identification systems at a very large scale. Before BirdCLEF started in 2014, three previous initiatives on the evaluation of acoustic bird species identification took place, including two from the SABIOD[6]

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

group [6,5,1]. In collaboration with the organizers of these previous challenges, the 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, these tasks were 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.). They therefore produced substantially lower scores and offered a better progression margin towards building real-world generalist identification tools.
The main novelty of the 2017 edition of the challenge with respect to the previous years was the inclusion of *soundscape recordings* containing time-coded bird species annotations. Usually xeno-canto recordings focus on a single foreground species and result from using mono-directional recording devices. Soundscapes, on the other hand, are generally based on omnidirectional recording devices that monitor a specific environment continuously over a long period. This new kind of recording reflects (possibly crowdsourced) passive acoustic monitoring scenarios that could soon augment the number of collected sound recordings by several orders of magnitude. In this paper, we report the methodology of the performance evaluation as well as an analysis and a discussion of the results achieved by the 5 participating groups.

2 Dataset

As the soundscapes appeared to be very challenging in 2015 and 2016 (with an accuracy below 15%), new soundscape recordings containing time-coded bird species annotations were integrated in the test set (so as to better understand what makes state-of-the-art methods fail on such contents). This new data was specifically created for BirdCLEF thanks to the work of three people: Paula Caycedo Rosales (ornithologist from the Biodiversa Foundation of Colombia and Instituto Alexander von Humboldt, Xeno-canto recordist), Herv Glotin (bioacoustician, co-author of this paper) and Lucio Pando (field guide and ornithologist). In total, about 6,5 hours of audio recordings were collected and annotated in the form of time-coded segments with associated species name. This dataset is composed of two main subsets:

Peru soundscapes, about 2 hours (1:57:08) 32 annotated segments: recorded in the summer of 2016 with the support of Amazon Explorama Lodges within the BRILAAM STIC-AmSud and SABIOD.org project. These recordings have been realized in the jungle canopy at 35 meters high (the highest point of the area), and at the level of the Amazon river, in the Peruvian basin. The recordings are sampled at 96 kHz, 24 bits PCM, stereo, dual -12 dB, using multiple systems: TASCAM DR, SONY PMC10, Zoom H1.

Colombia soundscapes, about 4,5 hours (4:25:55), 1990 annotated segments: These documents were annotated by Paula Caycedo Rosales, ornithologist from the Biodiversa Foundation of Colombia and an active Xeno-Canto member.

In addition to these newly introduced records, the test set still contained the 925 soundscapes and 8,596 single species recordings of BirdCLEF 2016 (collected by the members of Xeno-Canto[7] network, see [7] for more details).

As for the training data, we consistently enriched the training set of the 2016 edition of the task, in particular to cover the species of the newly introduced time-coded soundscapes. Therefore, we extended the covered geographical area to the union of Brazil, Colombia, Venezuela, Guyana, Suriname, French Guiana, Bolivia, Ecuador and Peru, and collected all Xeno-Canto records in these countries. We then kept only the 1500 species having the most recordings so as to get sufficient training samples per species (48,843 recordings in total). The training set has a massive class imbalance with a minimum of four recordings for *Laniocera rufescens* and a maximum of 160 recordings for *Henicorhina leucophrys*. Recordings are associated to various metadata such as the type of sound (call, song, alarm, flight, etc.), the date, the location, textual comments of the authors, multilingual common names and collaborative quality ratings. 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 run their system so as to identify all the actively vocalising birds species in each test recording (or in each test segment of 5 seconds for the soundscapes). Up to 4 *run files* per participant could be submitted to allow evaluating different systems or system configurations (a *run file* is a formatted text file containing the species predictions for all test items). Each species had to be associated with a normalized score in the range $[0, 1]$ reflecting the likelihood that this species is singing in the test sample. For each submitted run, participants had to signal if the run was performed fully automatically or with human assistance, and if they used a method based only on audio analysis only or with the use of the metadata.

Participants were asked to run their system so as to identify all the actively vocalising birds species in each test recording (or in each test segment of 5

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

seconds for the soundscapes). The submission *run files* had to contain as many lines as the total number of identifications, with a maximum of 100 identifications per recording or per test segment). Each prediction had to be composed of a species name belonging to the training set and a normalized score in the range $[0, 1]$ reflecting the likelihood that this species is singing in the segment. The used evaluation metric used was the Mean Average Precision.

The evaluation metric used to compare the systems is the mean Average Precision (mAP) averaged across all queries, considering each audio file in 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

78 research groups registered for the BirdCLEF 2017 challenge. Five of them finally submitted run files and four of them submitted working notes describing their system. Details of the used methods and evaluated systems are synthesized below (by alphabetical order) and further developed in the working notes of the participants [9,4,10,3]:

Cynapse, Austria, 4 runs [3]: This system is based on a multi-modal deep neural network taking audio samples and metadata as input. The audio is fed into a convolutional neural network using four convolutional layers. The additionally provided metadata is processed using fully connected layers. The flattened convolutional layers and the fully connected layer of the metadata were joined and put into a large dense layer. For the sound pre-processing and data augmentation, they used a similar pipeline as the best system of BirdCLEF 2016 described in [2]. The two runs Cynapse Run 2 and 3 mainly differ in the FFT window size used for constructing the time-frequency representation passed as input to the CNN (respectively 512 and 256). Cynapse Run 4 is an average of Cynapse Run 2 and 3.

DYNI UTLN, France, 2 runs [10]: This system is based on an adaptation of the image classification model Inception V4 [11] extended with a time-frequency attention mechanism. The main steps of the processing pipeline are (i) the construction of a multi-scaled time-frequency representation passed as a RGB image

to the Inception model, (ii) data augmentation: random hue, contrast, brightness, saturation, random crop in time and frequency domain and (iii) the training phase relying on transfer learning from the initial weights of the Inception V4 model (learned in the visual domain using the ImageNet dataset).

FHDO BCSG, Germany, 4 runs [4]: Like the DYNI UTLN team, these participants based his system on an adaptation of an image classification model, *i.e* Inception V3 [12]. Audio records were encoded through spectrograms and further processed by applying bandpass filtering, noise filtering, and silent region removal. For data augmentation purposes, they intended to use time shifting, time stretching, pitch shifting, and pitch stretching. Unfortunately, the data augmentation was not properly executed and the learned models suffered from overfitting problems. The first three runs differ in term of preprocessing, while the Run 3 is an average of the runs: Run 2 manipulates binary pictures and Run 4 uses grayscale pictures. Run 1 exploited the 3 RGB channels: the original grayscale picture in one channel, its blurred and sharpened versions for the two other channels.

TUCMI, Germany, 4 runs [9]: This system is also based on convolutional neural networks (CNN) but using more classical architectures than the Inception model used by DYNI UTLN. The main steps of the processing pipeline are (i) the construction of magnitude spectrograms with a resolution of 512x256 pixels, which represent five-second chunks of audio signal, (ii) data augmentation (vertical roll, Gaussian noise, Batch Augmentation) and (iii) the training phase relying on either a classical categorical loss with a softmax activation (TUCMI Run 1), or on a set of binary cross entropy losses with sigmoid activations as an attempt to better handle the multi-labeling scenario of the soundscapes (TUCMI Run 2). TUCMI Run 3 is an ensemble of 7 CNN models including the ones of Run 1 and Run 2. TUCMI Run 4 was an attempt to use geo-coordinates and time as a way to reduce the list of species to be recognized in the soundscapes recordings. Therefore, the occurrences of the eBird initiative were used complementary to the data provided within BirdCLEF. More precisely, only the 100 species having the most occurrences in the Loreto/Peru area for the months of June, July and August were kept in the training set.

5 Results

Figure 1 reports the performance measured for the 18 submitted runs. For each run (*i.e.* each evaluated system), we report the Mean Average Precision for the three categories of queries: traditional mono-directional recordings (the same as the one used in 2016), non time-coded soundscape recordings (the same as the one used in 2016) and the newly introduced time-coded soundscape recordings. To measure the progress over last year, we also plot on the graph the perfor-

mance of last year's best system [2].

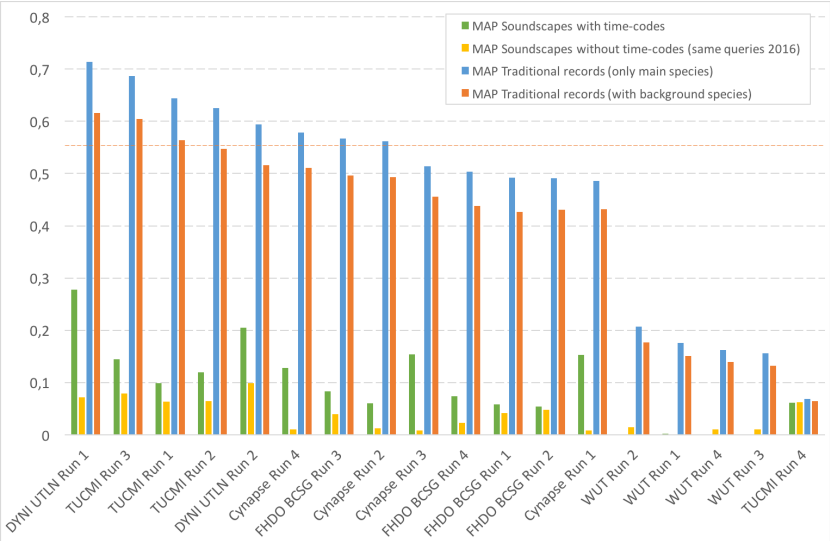

Fig. 1. BirdCLEF 2017 results overview - Mean Average Precision. The orange dot line represents the last year's best system obtained by the CUBE system [2].

It is remarked that all submitted runs were based on Convolutional Neural Networks (CNN) confirming the supremacy of this approach over previous methods (in particular the ones based on hand-crafted features which were performing the best until 2015). The best MAP of 0.71 (for the single species recordings) was achieved by the best system configuration of DYNI UTLN (Run 1). That rather similar to the MAP of 0.68 achieved last year by [2] but with 50% more species in the training set. Regarding the newly introduced time-coded soundscapes, the best system was also the one of DYNI UTLN (Run 1) whereas it did not introduce any specific features towards solving the multi-labeling issue. The main conclusions we can draw from the results are the following:

The network architecture plays a crucial role: Inception V4 that was known to be the state of the art in computer vision [11] also performed the best within the BirdCLEF 2017 challenge that is much different (time-frequency representations instead of images, a very imbalanced training set, mono- and multi-labeling scenarios, etc.). This shows that its architecture is intrinsically well-suited for a variety of machine-learning tasks across different domains. It also reveals a convergence of the methods to be used for machine learning tasks in the audio and the visual domain.

Ensembles of networks improve the performance consistently: This can be seen through Cynapse Run 4 and TUCMI Run 3 that outperform the other respective runs of these participants. The problem of such ensembles of networks is that their practical use in real-world applications is limited. They actually require a much higher GPU consumption so that their use in data intensive contexts is limited by cost issues. A promising solution towards this issue could be to rely on knowledge distilling [8]. Knowledge distilling consists in transferring the generalization ability of a cumbersome model to a small model by using the class probabilities produced by the cumbersome model as soft targets for training the small model. Alternatively, more efficient architectures and learning procedures should be devised.

The use of a multi-label loss function provides some improvements for soundscapes: The class-wise binary cross-entropy losses used in TUCMI Run 2 did allow a consistent performance gain on the time-coded soundscapes compared to the classical softmax loss in TUCMI Run 1. This was not enough to compensate the gain due to the Inception v4 architecture in the DYNI UTLN runs but we could expect a similar improvement with that architecture. Nevertheless, the multi-label loss function degrades the performance in the case of the traditional mono-directional recordings. Thus, it should be used only for records with many vocalizing birds such as the soundscapes.

The size of the FFT window used to construct the spectrograms plays an important role: This aspect was one of the main factor evaluated by the Cynapse team through their different runs. In particular, Cynapse Run 2 used a FFT window size of 512 whereas Cynapse Run 3 used a FFT window size of 256. As shown on Figure 1, the smaller FFT window size enables a considerable gain on the time-coded soundscapes probably because it reduces the overlap of different bird species within each chunk. On the other side, it degrades the performance on the traditional mono-directional recordings. On that content, a larger FFT window size helps recognizing the main foreground species.

Location-based and time-based species filtering is promising: TUCMI Run 4, that restricted the training set to the 100 most likely species according to the probability of eBird's occurrences did perform consistently worse than the other runs of TUCMI (N.B: only the time-coded soundscapes have to be considered here, *i.e.* the green bar in Figure 1). This reveals an unfitting selection of bird species. The period June-August in the Loreto/Peru area they used for selecting the most likely species is actually fitting only the Peruvian subset of the soundscapes not the Colombian one that is much larger. To better evaluate the benefit of the filtering strategy of TUCMI team, Table 1 provides the results of their submitted runs detailed by country. It shows that on the Peru's subset, the location-based and time-based filtering (Run 4) is very effective. On the other side, it degraded the performance on the Colombia subset because of the unfitting selection. Overall, we believe such location-based and time-based filtering is very promising for improving the performance. However, a finer and

more accurate species distribution model should probably be used. Using occurrence data solely for learning species distribution model is actually often not possible because of strong sampling bias. Thus, in ecology, the prediction of the presence or absence of a given species at a given location, is usually based on environmental variables that characterize the environment encountered at that location (e.g. climatic data, topological data, occupancy data, etc.).

Table 1: Results of TUCMI runs detailed by country (time-coded soundscapes only)

	All	Colombia	Peru
TUCMI Run 3	0,144	0,146	0,026
TUCMI Run 1	0,099	0,101	0,003
TUCMI Run 2	0,119	0,121	0,007
TUCMI Run 4	0,061	0,059	**0,158**

Learning from metadata was not really conclusive: The attempt of Cynapse to use metadata as a context information passed to the neural network did not allow to outperform the purely audio-based runs of DYNI and TUCMI systems. However, as they used a less advanced network architecture it is difficult to conclude on the real benefit of metadata learning. A run without the use of metadata would have been required.

6 Conclusion

This paper presented the overview and the results of the LifeCLEF bird identification challenge 2017. The main outcome was that the best performing system was based on a purely image-based convolutional neural network architecture (Inception V4) applied to a standard time-frequency representation. This shows the convergence of the best performing methods whatever the targeted domain. As in many challenges, ensembles of networks also improved the performance consistently even if their practical use in real-world applications is still limited. Concerning the soundscapes-based passive monitoring scenario that was evaluated this year, few additional conclusions came out: (i) the performance improved over last year but remains globally low, (ii) few design considerations specific to that contents allow consistent improvements such as a lower FFT window size to construct the spectrograms or the use of a multi-label loss function instead of a softmax. Finally, it was shown by one of the participant that the use of location-based and time-based species filtering could be beneficial for a real-world monitoring device that would be fixed at a given place. Such approach is now facilitated by the huge volume of occurrences collected and shared by the eBird citizen science project. Even the raw spatial frequency of the occurrences gives a rather good estimate of the observable species at a given place. However,

this might not help identifying the less abundant species that are often the ones that need a further follow-up.

Acknowledgements The organization of the BirdCLEF task is supported by the Xeno-Canto foundation for nature sounds as well as the French CNRS project SABIOD.ORG and EADM GDR CNRS MADICS, BRILAAM STIC-AmSud, and Floris'Tic. The annotations of some soundscape were prepared with regreted wonderful Lucio Pando at Explorama Lodges, with the support of Pam Bucur, H. Glotin and Marie Trone.

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. Elias Sprengel, Martin Jaggi, Y.K., Hofmann, T.: Audio based bird species identification using deep learning techniques. In: Working notes of CLEF (2016)
3. Fazekas, B., Schindler, A., Lidy, T.: A multi-modal deep neural network approach to bird-song identication. In: Working Notes of CLEF 2017 (Cross Language Evaluation Forum) (2017)
4. Fritzler, A., Koitka, S., Friedrich, C.M.: Recognizing bird species in audio files using transfer learning. In: Working Notes of CLEF 2017 (Cross Language Evaluation Forum) (2017)
5. 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`
6. 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`
7. Goëau, H., Glotin, H., Planqué, R., Vellinga, W.P., Joly, A.: Lifeclef bird identification task 2016. In: CLEF 2016 (2016)
8. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
9. Kahl, S., Wilhelm-Stein, T., Hussein, H., Klinck, H., Kowerko, D., Ritter, M., Eibl, M.: Large-scale bird sound classification using convolutional neural networks. In: CLEF 2017 (2017)
10. Sevilla, A., Bessonne, L., Glotin, H.: Audio bird classification with inception-v4 extended with time and time-frequency attention mechanisms. In: Working Notes of CLEF 2017 (Cross Language Evaluation Forum) (2017)
11. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning. arXiv preprint arXiv:1602.07261 (2016)
12. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. CoRR abs/1512.00567 (2015), `http://arxiv.org/abs/1512.00567`