Overview of BirdCLEF 2019: Large-Scale Bird Recognition in Soundscapes

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

[1] Chemnitz University of Technology, Germany, stefan.kahl@informatik.tu-chemnitz.de
[2] Inria/LIRMM ZENITH team, Montpellier, France, {fabian-robert.stoter, alexis.joly}@inria.fr
[3] CIRAD, UMR AMAP, Montpellier, France, herve.goeau@cirad.fr
[4] Université de Toulon, Aix Marseille Univ, CNRS, LIS, DYNI team, Marseille, France, herve.glotin@univ-tln.fr
[5] Xeno-canto Foundation, The Netherlands, {wp,bob}@xeno-canto.org

Abstract. The BirdCLEF challenge—as part of the 2019 LifeCLEF Lab [7]—offers a large-scale proving ground for system-oriented evaluation of bird species identification based on audio recordings. The challenge 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 (659). In 2019, the challenge was focused on the difficult task of recognizing all birds vocalizing in omni-directional soundscape recordings. Therefore, the dataset of the previous year was extended with more than 350 hours of manually annotated soundscapes that were recorded using 30 field recorders in Ithaca (NY, USA). This paper describes the methodology of the conducted evaluation as well as the synthesis of the main results and lessons learned.

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, especially so-called 'birders' as well as professionals such as park rangers, ecological consultants and of course ornithologists are potential users of an automated bird sound identification system

in the context of wider initiatives related to ecological surveillance or biodiversity conservation. The BirdCLEF challenge —as part of the 2019 LifeCLEF Lab [7]— evaluates 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] group [5,4,2]. In collaboration with the organizers of these previous challenges, the BirdCLEF 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 metadata 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.).

The main novelty of the 2017 and 2018 editions 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 passive acoustic monitoring scenarios that could soon augment the number of collected sound recordings by several orders of magnitude. Despite the technological progress in recent years, the results of the previous editions on this challenging soundscape task were quite low. We decided to shift the focus of the 2019 challenge to soundscape analysis only. We extend the previous dataset with North American bird species for which more annotated data was available. In particular, we built a dataset of 350 hours of soundscapes that were recorded and annotated by expert birders of the Cornell Lab of Ornithology in Ithaca, NY, USA (see Figure 1). This large volume of data allowed us to share a fully-annotated, three-day validation dataset to enable participants to thoroughly evaluate their systems.

2 Task description

The 2019 BirdCLEF challenge featured the largest, fully-annotated collection of soundscape recordings. With respect to real-world use cases, labels and metrics were chosen to reflect the vast diversity of bird vocalizations and high ambient noise levels in omnidirectional recordings.

2.1 Goal and evaluation protocol

The goal of the task is to localize and identify all audible birds within the provided soundscape test set. Each soundscape is divided into segments of 5

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

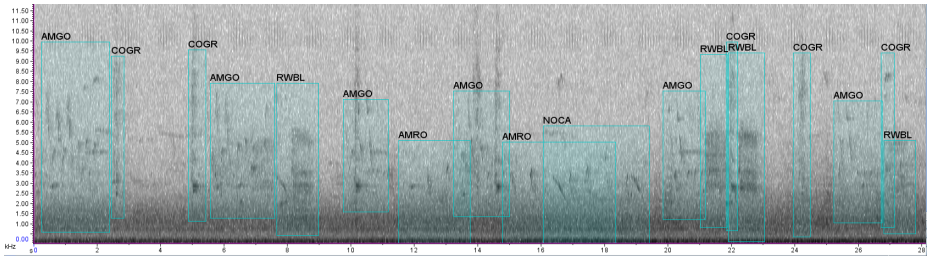


Fig. 1: Example of an annotated soundscape recording. Expert birders provided more than 80,000 bounding box annotations using the Raven Pro analysis software. For reasons of better comparability, those annotations were condensed into label lists for 5-second intervals.

seconds, and a list of species associated to probability scores had to be returned for each segment. The used evaluation metric is the classification mean Average Precision ($cmAP$), considering each class c of the ground truth as a query. This means that for each class c, all predictions with $ClassId = c$ are extracted from the run file and ranked by decreasing probability in order to compute the average precision for that class. The mean across all classes is computed as the main evaluation metric. More formally:

$$cmAP = \frac{\sum_{c=1}^{C} AveP(c)}{C}$$

where C is the number of classes (species) in the ground truth and $AveP(c)$ is the average precision for a given species c computed as:

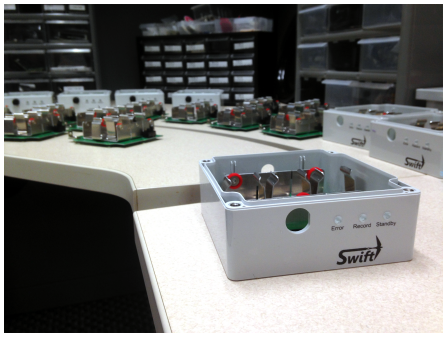
$$AveP(c) = \frac{\sum_{k=1}^{n_c} P(k) \times rel(k)}{n_{rel}(c)}.$$

where k is the rank of an item in the list of the predicted segments containing c, n_c is the total number of predicted segments containing c, $P(k)$ is the precision at cut-off k in the list, $rel(k)$ is an indicator function equaling 1 if the segment at rank k is a relevant one (*i.e.* is labeled as containing c in the ground truth) and $n_{rel}(c)$ is the total number of relevant segments for class c.

2.2 Dataset

The 2019 dataset contains about 350 hours of manually annotated soundscapes—most of which were recorded using field recorders between January and June of 2017 in Ithaca, NY, USA. We used SWIFT recording units provided by the Bioacoustics Research Program[7] of the Cornell Lab of Ornithology (Figure 2). These omnidirectional recorders capture audio over an array of 30 units spanning

[7] http://www.birds.cornell.edu/brp/



(a) SWIFT recorder assembly line

(b) SWIFT recorder in the field

Fig. 2: Autonomous recording units are a widely used sampling tool in ecological research. The SWIFT recorder provided by the Bioacoustics Research Program (BRP) of the Cornell Lab of Ornithology allows to record up to 30 consecutive days of audio. Optimizing the assembly of these weatherproof recorders reduces the costs per unit significantly. Images provided by the BRP.

one square mile of diverse vegetation and water bodies. We randomly selected one file for each hour of the day recorded with one of the 30 recorders to compile a data collection of 15 days. Each hour-long recording was then annotated by experts who provided more than 80,000 bounding boxes—one for each audible bird vocalization. For the sake of comparability with previous editions, these annotations were condensed into label lists for 5-second segments of audio.

In addition, we also re-used the soundscape data from the previous years of BirdCLEF. More specifically, this concerns about 4,5 hours of soundscapes recorded in Columbia by Paula Caycedo Rosales, ornithologist from the Biodiversa Foundation of Colombia and an active member of Xeno-canto. More details about this soundscape data (locations, authors, etc.) can be found in the overview working note of BirdCLEF 2018 [6].

As for training data, we provided an newly composed Xeno-Canto subset covering 659 species from South and North America (including all species annotated in the soundscapes). The vast collection of recordings provided by the Xeno-canto community often features multiple hundreds of recordings per (common) bird species. Especially North American species are well represented in the collection. Therefore, we limited the amount of audio files to 100 recordings per species. This way, we decreased data imbalance and provided a manageable amount of data. In total, the training data featured 50,153 files with a total run length of 608 hours. We selected recordings based on their community rating to preserve a high quality for most species. Each recording contained weak labels that state the presence of fore and background species.

Recordings are associated to various metadata such as the type of sound (call, song, alarm, flight, etc.), the date of recording, the location, textual comments

of the authors, multilingual common names and collaborative quality ratings. Additionally, we provided eBird.org frequency lists to enable participants to decide which species are plausible for a given time, date and location. Frequency estimations of bird species occurrences were compiled using eBird checklist data for the soundscape recording locations in the US and Colombia provided by the eBird API 1.1 (which was unfortunately discontinued in March 2019).

The shift in acoustic domains between mono-species, high quality recordings and omnidirectional soundscapes with high ambient noise levels is one of the major challenges in bird sound recognition for avian activity monitoring. Participants were required to submit at least one run that used training data only. Aside from that, participants were allowed to use validation data for training (despite the fact that this would require extensive annotation when switching recording locations in real-world applications).

3 Results

103 participants registered for the BirdCLEF 2019 challenge and downloaded the dataset. Five of them succeeded in submitting runs, but only four teams published their approaches. Details of the methods and systems used in the runs are synthesized in the individual working notes of the participants and are summarized in this section. In Figure 3 we report the performance achieved by the 25 collected runs, Table 1 provides more detailed insights and additional scores for each of the two soundscape recording locations.

3.1 MfN [11], Best overall

Lasseck managed to achieve top scores in most of the past editions in BirdCLEF. Most notably, his 2018 performance topped all previous results in the mono-species recording domain [10] and led to the observation that this task can be considered solved. This year, MfN build upon the results of past editions and managed to outperform all other participating teams with his very deep Inception and ResNet architectures that were pre-trained on ImageNet, as a continuation of [12]. Lasseck used 5-second spectrograms with mel-compressed frequency and dB amplitude scale. Sophisticated data augmentation methods lead to consistent improvements and can be considered a major contribution to the field of bioacoustics. Additionally, the use of validation data to fine-tune the pre-trained networks has a significant effect on the overall scores. Considering this, annotating soundscapes to adapt neural networks to specific recording conditions appears to be well worth the costs.

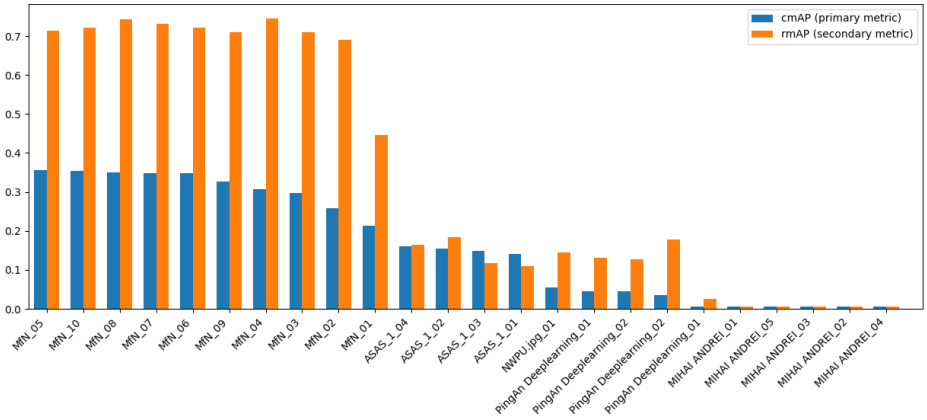


Fig. 3: Scores achieved by all systems evaluated within the bird identification task of LifeCLEF 2019. MfN scored best overall with outstanding results in both soundscape domains. The difference between runs that only used training data and those which also used validation samples for training is significant and raises the question whether local adaption to recording conditions is worth the manual annotation effort.

3.2 ASAS [9]

This team also used Inception and ResNet architectures to conduct their experiments. Again, task-specific data augmentation was key to achieve higher scores. ASAS followed the spectrogram extraction approach of the 2018 Baseline Repository [8]. The results however–although very competitive—do not outperform the approach of Lasseck despite the similarities in deep neural network design. This leads to the assumption that sophisticated augmentation strategies are of particular importance since they provide the needed variance to the input data distribution which prevents overfitting. The authors state that pre-processing of training data could have a significant impact on the overall performance due to the difficulties of weakly labeled data.

3.3 NWPU [1]

Consistent with all other approaches, these participants used mel-scale spectrogram extracted from the provided training data to build a Inception-v3 feature extractor and classifier. The model was pre-trained on ImageNet and a number data augmentation methods were applied. However, training deep neural networks is costly and subsequently, the participants were not able to conduct the amount of experiments needed to achieve higher scores. The results state once again the most notable observation across all submission: An elaborate training regime is key to good overall performance. It appears that this applies independent of the underlying network architecture.

Table 1: Detailed results of runs submitted by the participants. Lasseck submitted the best performing run based on our primary metric (MfN run 5). In both domains, Colombia and North America, results show string improvements compared to previous years. Team names shortened for brevity.

		Overall		North America		Colombia	
TEAM	RUN	cmAP	mAP	cmAP	mAP	cmAP	mAP
MfN	1	0.213	0.446	0.231	0.446	0.252	0.451
MfN	2*	0.258	0.690	0.320	0.697	0.238	0.463
MfN	3*	0.297	0.710	0.343	0.718	0.239	0.455
MfN	4*	0.308	**0.745**	0.364	**0.755**	0.239	0.443
MfN	5*	**0.356**	0.714	**0.407**	0.723	0.292	0.451
MfN	6*	0.348	0.721	0.398	0.728	0.287	0.493
MfN	7*	0.348	0.732	0.401	0.740	0.283	0.483
MfN	8*	0.350	0.743	0.404	0.751	0.283	0.483
MfN	9*	0.327	0.710	0.404	0.718	0.282	0.476
MfN	10*	0.354	0.722	0.403	0.729	**0.293**	**0.509**
MIHAI	1	0.005	0.006	0.000	0.006	0.011	0.003
MIHAI	2	0.000	0.000	0.000	0.000	0.000	0.001
MIHAI	3	0.000	0.001	0.000	0.001	0.001	0.000
MIHAI	4	0.000	0.000	0.000	0.000	0.000	0.000
MIHAI	5	0.000	0.001	0.000	0.001	0.000	0.000
PingAn	1	0.046	0.132	0.039	0.130	0.079	0.174
PingAn	2	0.046	0.128	0.034	0.127	0.075	0.167
PingAn	3	0.005	0.026	0.004	0.027	0.007	0.017
PingAn	4*	0.035	0.179	0.024	0.181	0.055	0.125
NWPU	1	0.054	0.145	0.067	0.145	0.059	0.140
ASAS	1	0.140	0.110	0.161	0.108	0.113	0.192
ASAS	2	0.154	0.184	0.183	0.183	0.116	0.206
ASAS	3	0.149	0.117	0.171	0.114	0.120	0.204
ASAS	4	0.160	0.164	0.178	0.163	0.137	0.204

** used validation data for training*

3.4 MIHAI [3]

The submission results of this participant confirm this thought. The author states that he was able to confirm that deeper architectures do not necessarily lead to better performance, especially when computational constraints limit the choice of hyperparameters. Despite very low scores across all runs, the observation that task-specific training and model layouts matter was consistent with the submissions of other teams.

4 Conclusion

In this edition of the BirdCLEF challenge, participants built on established systems from previous years, all submitted runs featured a CNN classifier trained on spectrograms—very deep networks once again performed best. Participants were able to significantly improve the detection performance. In fact, we saw an increase of more than 180% for the best performing runs (2018: 0.193 - 2019: 0.356). This result is probably largely due to the high number of North American soundscapes that are less complex than their South American counterparts. However, the recognition performance for South American soundscapes also increased significantly compared to 2018 with a cmAP of 0.293 in 2019 over 0.222 from last year. Participants were allowed to use any publicly available metadata and even the provided validation data to improve the performance of their systems. Although expert annotations are not an adequate (or even easy-to-acquire) addition for the training of a recognition system for unseen habitats, the increase in overall performance is considerable. The highest scoring run submitted by MfN achieved a sample-wise mean average precision (our secondary metric) of 0.446 without the use of validation samples and 0.745 when validation data was used for training. These scores imply that domain adaption to new acoustic environments (and recorder characteristics) plays a crucial role and should be subject of investigation in future editions.

Acknowledgements The organization of the BirdCLEF task is supported by the Xeno-canto Foundation, the European Union and the European Social Fund (ESF) for Germany, as well as by the French CNRS project SABIOD.ORG and EADM GDR CNRS MADICS, BRILAAM STIC-AmSud, and Floris'Tic. The annotations of some soundscapes were prepared by the wonderful Lucio Pando of Explorama Lodges, with the support of Pam Bucur, H. Glotin and Marie Trone. We want to thank all expert birders who annotated North American soundscapes with incredible effort: Cullen Hanks, Jay McGowan, Matt Young, Randy Little, and Sarah Dzielski.

References

1. Bai, J., Wang, B., Chen, C., Fu, Z., Chen, J.: Inception-v3 based method of lifeclef 2019 bird recognition. In: CLEF working notes 2019 (2019)
2. 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)
3. Costandache, M.C.: Bird species identification using neural networks. In: CLEF working notes 2019 (2019)
4. 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`

5. 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
6. Goëau, H., Glotin, H., Planqué, R., Vellinga, W.P., Kahl, S., Joly, A.: Overview of birdclef 2018: monophone vs. soundscape bird identification. In: CLEF working notes 2018 (2018)
7. Joly, A., Goëau, H., Botella, C., Kahl, S., Servajean, M., Glotin, H., Bonnet, P., Vellinga, W.P., Planqué, R., Stöter, F.R., Müller, H.: Overview of lifeclef 2019: Identification of amazonian plants, south & north american birds, and niche prediction. In: Proceedings of CLEF 2019 (2019)
8. Kahl, S., Wilhelm-Stein, T., Klinck, H., Kowerko, D., Eibl, M.: Recognizing birds from sound - the 2018 birdclef baseline system. arXiv preprint arXiv:1804.07177 (2018)
9. Koh, C.Y., Chang, J.Y., Tai, C.L., Huang, D.Y., Hsieh, H.H.: Bird sound classification using convolutional neural networks. In: CLEF working notes 2019 (2019)
10. Lasseck, M.: Audio-based bird species identification with deep convolutional neural networks. In: Working Notes of CLEF 2018 (Cross Language Evaluation Forum) (2018)
11. Lasseck, M.: Bird species identification in soundscapes. In: CLEF working notes 2019 (2019)
12. Sevilla, A., Glotin, H.: Audio bird classification with inception v4 joint to an attention mechanism. In: Working Notes of CLEF 2017 (Cross Language Evaluation Forum) (2017)