DEEP LEARNING FOR ENVIRONMENTAL SENSING TOWARD SOCIAL WILDLIFE DATABASE

Conference Paper · February 2020

CITATIONS
0

READS
188

7 authors, including:

Clément Duhart
Massachusetts Institute of Technology
17 PUBLICATIONS **62** CITATIONS
SEE PROFILE

Spencer Russell
Amazon
9 PUBLICATIONS **43** CITATIONS
SEE PROFILE

Felix Michaud
Université du Maine
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Gershon Dublon
Massachusetts Institute of Technology
30 PUBLICATIONS **520** CITATIONS
SEE PROFILE

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

DEEP LEARNING FOR ENVIRONMENTAL SENSING TOWARD SOCIAL WILDLIFE DATABASE

Clement Duhart[1], Spencer Russell[1], Felix Michaud[2], Gershon Dublon[1], Brian Mayton[1], Glorianna Davenport[1], Joseph Paradiso[1]

***Abstract*—Climate change and environmental degradation are causing species extinction worldwide. Automatic wildlife sensing is an urgent requirement to track biodiversity losses on Earth. Recent improvements in machine learning can accelerate the development of large-scale monitoring systems at high resolution that would help track conservation targets and outcomes. This would offer also unique opportunities for studying wildlife sociology at individual scale. In this paper, we present our efforts to develop suitable tools for building machine learning databases for wildlife detection, identification, acoustic source separation and geolocalization. These tools work on data collected at the Tidmarsh Wildlife Sanctuary, the site of the largest freshwater wetland restoration in Massachusetts.**

I. INTRODUCTION

Ubiquitous sensing technologies [1] can be used to capture aspects of ecosystem function and ecological transformation with minimal impact at high resolution over long periods of time. However, in part due to recognition challenges, automatic wildlife sensing remains mostly out of reach. In the ecological research community, wildlife surveys are still conducted by experts estimating a given species population at a specific time. Intensive manual effort is required, even with the help of recordings and modern signal processing tools. Field surveyors need to maintain perceptual awareness and attention to detail; they also need to conduct surveys at different times of day and on many days throughout the year as the animal populations migrate and breed.

Efforts to automate surveys are vital to gaining a real-time understanding of a massive wave of species extinction. This represents a significant opportunity for Artificial Intelligence (AI) systems, which thrive on big data, and might one day be able to analyze and characterize wildlife populations around the globe. Accurate and continuous wildlife detection, identification and geolocalization would transform wildlife surveys into high-resolution activity maps that can update in real time and at geographic scale. Recent advances in Deep Learning enable recognition of rare species that otherwise produce low-occurrence signals in evolving and noisy environments, and a distributed sensing approach to studying Wildlife might also uncover localized interactions and social behaviors that would be difficult to identify and track manually.

Optical and acoustic sensing provide complementary information. For example, the biophony is intrinsically complex in terms of vocalizations (e.g. birds) and diverse regarding species candidates. However, the biophony is mainly produced by creatures that are difficult to spot. A multi-modal sensing approach can help separate noisy geophony and anthrophony from the desired wildlife signal. One crucial requirement is a system's ability to detect new species in an area, especially in a dynamic restoration such as the one presented in Section II.

Some recent contributions have demonstrated the ability of Deep Learning to scale biologists' efforts to identify wildlife. For example, automatic animal identification from camera trap images using a VGG model trained on 1.4 million images over 48 classes was shown to have 96.8% accuracy [2]. AI can be leveraged to save time when used with human volunteers. For acoustic identification of wildlife, several new contributions focus on deep learning technology, including [3] for amphibians, [4] bats, [5] insects, and [6] bird vocalization segmentation. We expect interesting contributions in the future thanks to the Bird Audio Detection challenge [7], [8], [9]. Such lab-based results are essential to accelerating field deployments.

In this paper, we present our own efforts to monitor wildlife activity at the Tidmarsh Wildlife Sanctuary, the site of the one of the largest-ever freshwater wetland restorations in the northeastern United States. In this deployment, our Deep Learning models have been running 24/7 on data streaming from microphones and cameras in real-time over the last 3 years, and our sys-

Corresponding author: C Duhart, duhart@mit.edu [1]Responsive Environment, MediaLab, Massachusetts Institute of Technology, USA [2]Laboratoire d'Acoustique de l'Universite du Mans, FRANCE

tem has been used by biologists, restoration scientists, and other practitioners. These resources also provide an acoustic and visual database for machine learning researchers to build systems that are able to detect, identify and geo-localize wildlife interaction patterns at the individual level. As the capability improves, the system can provide unique scientific data for studying wildlife sociology in natural environments at this critical juncture for shrinking populations and ecosystems.

II. TIDMARSH WILDLIFE SANCTUARY

The Tidmarsh Wildlife Sanctuary is a 485 acre former cranberry farm in south-eastern Massachusetts that was actively restored to a freshwater wetland (2010-2016) . Different types of sensors, illustrated in Figure 1, are permanently deployed on the site to monitor its evolution, including environmental changes to water quality and temperature, wetland surface, stream channels, soils, atmosphere, plants and animal life etc. Such a data collection provides a high resolution environmental map revealing multi-scale dynamic interactions over time. Relying solely on theoretical frameworks to model such a complex environment is challenging: how do we determine chains of cause and effect? For example, how might we link changes to animal behavior to new microbial populations in the soils through intermediate effects on the plant community? One opportunity may be to map an ecosystem across many variables over time, space and multi-scales.

Fig. 1: Sensors are fully autonomous nodes with wireless communication and solar energy harvesting for collecting data from ground and atmospheric probes.

A. Wildlife Sensing Framework

Our 'Tidzam' wildlife detection system monitors wildlife, leveraging 24 custom-designed microphones and 6 cameras deployed across four different areas at Tidmarsh, as illustrated in Figure 2.

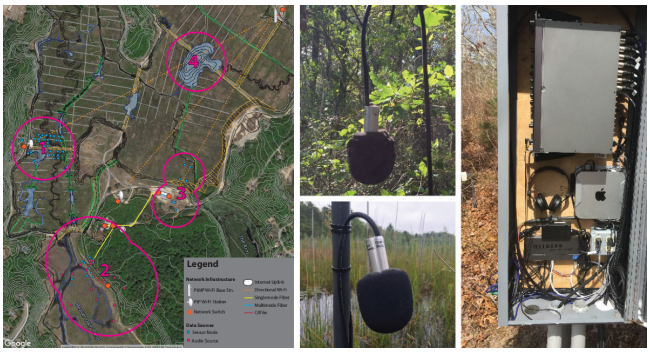

Fig. 2: Sensors, microphones, and cameras are deployed in four regions of interest. They have been specially designed to withstand wetland conditions year-round.

In the Tidzam framework, we implement and deploy Deep Learning techniques from the literature to detect, identify and geolocalize wildlife activities. Over the last four years, we have tested a number of different approaches leveraging bio-acoustics and computer vision.

1) Bio-Acoustic Classifiers: The Tidmarsh bio-acoustic ecosystem has evolved dramatically over years of restoration progress. Dynamic environments require continuous learning to make classifiers robust to both episodic and permanent acoustic changes – especially concerning the identification of as-yet unseen species. To that end, we developed a semi-automatic database augmentation mechanism using a confidence function detailed in [10]. A flow controller limits the recording volume and parameterizes the extraction balance between unidentified and uncertain predictions. Our 'Tidplay' platform, introduced in Section II-B, allows human experts to annotate and discuss these recordings while building a local acoustic database used to iteratively refine the classifiers. At the time of writing, the database is composed of 400,000 500 ms recordings distributed over 66 classes including system failure modes (e.g. microphone crackling due to water ingress), geophonic scenes (e.g. rain, wind, quiet), anthrophonic sounds (e.g. cars, airplanes, human voices), and finally, bio-acoustic events from insects (e.g. crickets, cicadas), amphibians (e.g. spring peepers, green frogs), and bird vocalizations across 42 species.

Several classifier models have been tested, presented in Table I. The classifier is retrained from scratch every 2 months, taking into consideration new recording

annotations. The average accuracy gain increases significantly at each training iteration, with the extent of the improvement depending on the number of new classes, diversity of vocalizations, and quality of the extracted recordings. Our current bio-acoustic classifier is based on a revisited expert architecture [11] running on one Titan X GPU. It continuously analyzes overlapped 500 ms Mel-Spectrogram windows from 24 discrete microphones and 3 on-camera microphones.

Architecture	F_1
64RBM-16RBM + SAE + CE	73%
121C-2P-16C-2P-1024FC-1024FC + CE	85%
121C-2P-16C-2P-3EA(1024FC-1024FC) + CE	88%
121C-2P-16C-2P-1024FC-1024FC + T-Lost	87%

TABLE I: Testing F1 scores on the Tidmarsh dataset using a Restricted Boltzman Machine (RBM) with Stacked Auto-Encoder on Cross-Entropy (CE), Convolution (C) with Pooling (P) and Fully Connected (FC) layers, Expert Architecture (EA) and Triplet-Lost

2) Camera Trap Classifiers: Camera traps use movement detectors to trigger video recording. In an outdoor environment such as Tidmarsh, non-animal movements dominate the trigger. Common causes include rain, wind, and water flow, which together produce a large number of irrelevant video recordings.

Deep Learning can provide high level visual semantic descriptions, saving volunteer time. We have experimented with and deployed different types of computer vision models to pre-filter our motion video databases. These include CNN, Fast R-CNN [12], and Yolo v3 [13]. As illustrated in Figure 3, precise species identification is still a challenging task, given the number of possible classes and the lack of a sufficient training dataset for general-purpose wildlife recognition. It would be a massive challenge to build a database containing every species present on the planet, and it may even be impossible to build a corresponding classifier model. As a result, we use our Tidplay platform to build a locally-dependent visual database to refine the pre-trained classifier model. This platform allows volunteers to create new classes and add new bounding boxes to video frames automatically extracted by a confidence function similar to the one used in our bio-acoustic classifier. Our current system is based on the Yolo v3 model and analyzes video recordings coming from 6 network cameras at Tidmarsh.

B. Tidplay Annotation Platform

Tidplay is an open-source, crowd-sourcing annotation web platform that we have designed to build training databases from audio and video sources. Users can upload, download and share audio and video files, write down annotations and comments, and create their custom databases while learning about wildlife. Tidplay has two intended user bases. First, wildlife ecologists can use Tidplay to share data for collaborating on the construction of annotated databases. Second, a tutorial mode can be used for public engagement and student training. Users can learn how to distinguish different sounds coming from geophony, anthropophony and biophony, progressively developing their abilities to identify challenging bird calls, for example. The multiple training levels available allow users to extend their bio-acoustic skills by comparing their answers and discussing ambiguous recordings with other users ranging from novices to experts. Recordings extracted automatically by Tidzam classifiers are integrated into Tidplay for cross-validation by multiple wildlife experts before being integrated into training databases. The Tidplay platform can also be used for annotating new audio as shown in Figure 4, for drawing video bounding boxes around objects or wildlife of interest, and for sketching subjects' body pose from video frames, all for use by the machine learning system.

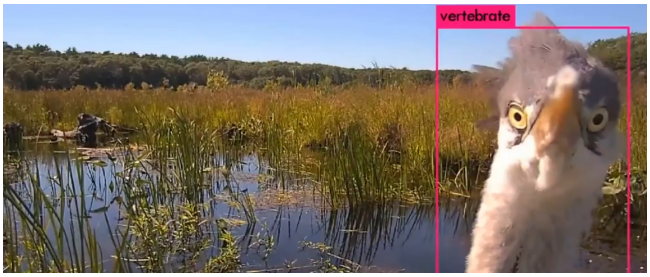

Fig. 3: Wildlife detection on the Herring site by a Yolo v3 model that has not been refined by a local database.

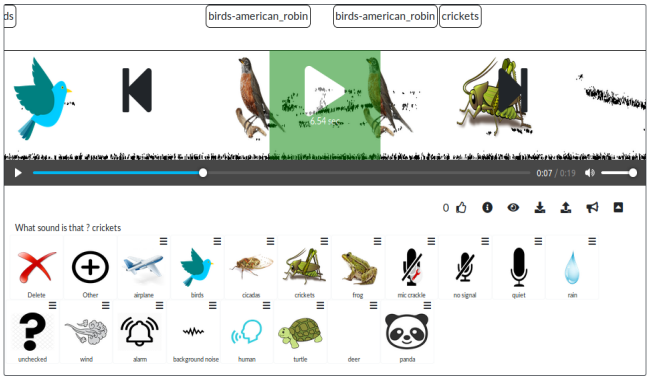

Fig. 4: The audio module of the Tidplay annotation platform shows the recording's spectrogram during listening to facilitate the annotation task.

III. Discussions and Future Work

Our Tidzam framework shows how Deep Learning technology can be used to detect and identify wildlife activities. However, its effectiveness for identifying or tracking individual animals and achieving accurate density estimation is an open question requiring additional data collection and validation. Currently, ecologists at Tidmarsh use correlations between Tidzam's detection density maps and periodic field surveys to estimate the wildlife population dynamics over years of restoration. Our current development effort is focused on how deep learning frameworks can help in tracking individuals through acoustic source separation, and how subjects can be localized more precisely. This work is summarized in the next subsection. Our main goal is to publish a complete database for ecological scientists and machine learning practitioners to study as is and apply in their own settings. The database will contain all the audio recordings from our microphone array with corresponding acoustic event annotations, their source separation masks, and estimated location coordinates.

A. Species Acoustic Source Separation

Recent enhancements have been made in acoustic source separation thanks to Deep Learning technology especially with the U-NET architecture [14]. Our early works based on such approach applied on our recordings is promising as illustrated in Figure 5. We are still investigating model limits regarding acoustic outdoor environment and a procedure for the database constitution regarding of detected species from Tidzam. Live deployment is also a concern regarding computation costs and the various acoustic streams collected from Tidmarsh. On demand analyze should be suitable based on Tidzam species identification.

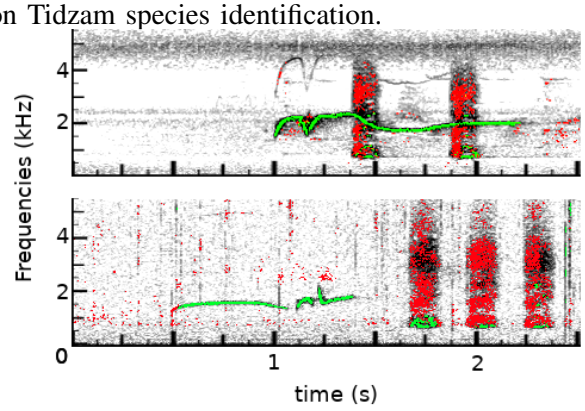

Fig. 5: Two examples of source separation masks when Tidzam has detected two birds : one american crow (red) and one eastern wood pewee (green) with cicadas (top) and rain (bottom) acoustic backgrounds.

B. Individual Acoustic Source Localization

While Tidzam is focused on classifying the the sounds recorded at our microphones, we are also working on methods to use all microphones jointly to localize wildlife within the environment. In addition to giving valuable ecological information, estimating the location of the sound can also inform the source separation process. The vast majority of existing source-separation research operates either in the single-channel context or with a small array with sources in the far-field. For instance, recent work has extended single-channel Deep Clustering using the phase information from an array [15]. However, spaced microphones provide a number of additional challenges because of the longer inter-microphone time delays and varying source-to-microphone propagation paths. Our preliminary work applying the framework of the Spatial Likelihood Function (SLF) [16] has been promising, as you can see in the map in Figure 6. This figure shows a heat map of where a given source (in this case a crow) is likely to be. While so far this work has been focused on classical digital signal processing (DSP) techniques, we are working on several approaches to integrating the DSP and AI frameworks to improve our localization estimates, as well as designing experiments to more rigorously evaluate the performance of our algorithms.

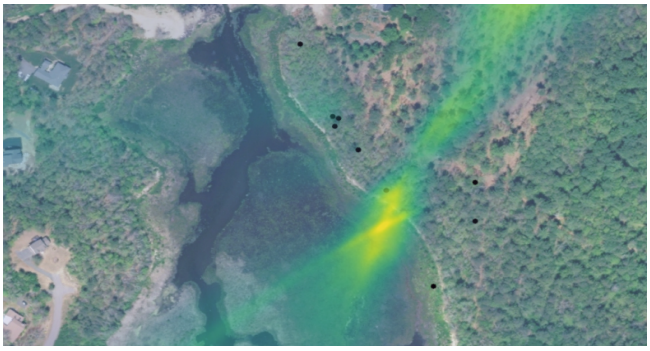

Fig. 6: Spatial Likelihood Function based on audio at a subset of the deployed microphones (black dots). The displayed area is roughly 500m x 350m

IV. Conclusion

We have presented an ongoing effort to deploy Deep Learning tools for automatic wildlife surveying Our work shows how Deep Learning can advance significant opportunities for ecological research efforts, restoration science, and public engagement. Our long term goal is to distribute an annotated database for machine learning practitioners, providing an unprecedented view towards analyzing wildlife in a restoration setting at the individual and species interaction level.

REFERENCES

[1] J. Paradiso, "Our extended sensoria - how humans will connect with the internet of things," *The Next Step: Exponential Life, Open Mind Collection*, vol. 1, no. 1, p. 47–75, 2016.

[2] M. S. Norouzzadeh, A. Nguyen, M. Kosmala, A. Swanson, M. S. Palmer, C. Packer, and J. Clune, "Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning," *Proceedings of the National Academy of Sciences*, vol. 115, no. 25, pp. E5716–E5725, 2018.

[3] J. Strout, B. Rogan, S. M. M. Seyednezhad, K. Smart, M. Bush, and E. Ribeiro, "Anuran call classification with deep learning," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2662–2665, March 2017.

[4] O. Mac Aodha, R. Gibb, K. E. Barlow, E. Browning, M. Firman, R. Freeman, B. Harder, L. Kinsey, G. R. Mead, S. E. Newson, I. Pandourski, S. Parsons, J. Russ, A. Szodoray-Paradi, F. Szodoray-Paradi, E. Tilova, M. Girolami, G. Brostow, and K. E. Jones, "Bat detectivedeep learning tools for bat acoustic signal detection," *PLOS Computational Biology*, vol. 14, pp. 1–19, 03 2018.

[5] I. Kiskin, B. P. Orozco, T. Windebank, D. Zilli, M. Sinka, K. J. Willis, and S. J. Roberts, "Mosquito detection with neural networks: The buzz of deep learning," *CoRR*, vol. abs/1705.05180, 2017.

[6] I. Potamitis, "Deep learning for detection of bird vocalisations," *CoRR*, vol. abs/1609.08408, 2016.

[7] D. Stowell, M. D. Wood, H. Pamua, Y. Stylianou, and H. Glotin, "Automatic acoustic detection of birds through deep learning: The first bird audio detection challenge," *Methods in Ecology and Evolution*, vol. 10, no. 3, pp. 368–380, 2019.

[8] E. Cakir, S. Adavanne, G. Parascandolo, K. Drossos, and T. Virtanen, "Convolutional recurrent neural networks for bird audio detection," in *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 1744–1748, Aug 2017.

[9] S. Adavanne, K. Drossos, E. akir, and T. Virtanen, "Stacked convolutional and recurrent neural networks for bird audio detection," in *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 1729–1733, Aug 2017.

[10] C. Duhart, G. Dublon, B. Mayton, and J. Paradiso, "Deep learning locally trained wildlife sensing in real acoustic wetland environment," in *Advances in Signal Processing and Intelligent Recognition Systems* (S. M. Thampi, O. Marques, S. Krishnan, K.-C. Li, D. Ciuonzo, and M. H. Kolekar, eds.), (Singapore), pp. 3–14, Springer Singapore, 2019.

[11] M. I. Jordan and R. A. Jacobs, "Hierarchies of adaptive experts," in *Advances in Neural Information Processing Systems*, pp. 985–992, Morgan-Kaufmann, 1992.

[12] R. Girshick, "Fast r-cnn," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[13] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018.

[14] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," *In proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[15] Z.-Q. Wang, X. Zhang, and D. Wang, "Robust TDOA estimation based on time-frequency masking and deep neural networks," in *Interspeech 2018*, p. 322326, ISCA, Sep 2018.

[16] P. Aarabi, "The fusion of distributed microphone arrays for sound localization," *EURASIP Journal on Applied Signal Processing*, vol. 2003, p. 338347, 2003.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Living Observatory and the Mass Audubon Tidmarsh Wildlife Sanctuary for the opportunity to realize the audio deployment at this location. The NVIDIA GPU Grant Program has provided the two TITAN X which are used by Tidzam. Clement DUHART has been supported by the PRESTIGE Fellowship of Campus France and the Pôle Léonard de Vinci. We also thank the Elements Collaborative and the sponsors of the MIT Media Lab for their support of this work.