Deep Learning for Wildlife Conservation and Restoration Efforts

Clement Duhart [1] **Gershon Dublon** [1] **Brian Mayton** [1] **Glorianna Davenport** [1] **Joseph A. 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 that would help track conservation outcomes and target efforts. In this paper, we present one such system we developed. 'Tidzam' is a Deep Learning framework for wildlife detection, identification, and geolocalization, designed for the Tidmarsh Wildlife Sanctuary, the site of the largest freshwater wetland restoration in Massachusetts.

1. Introduction

Ubiquitous sensing technologies (Paradiso, 2016) 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 efforts are required, even with the help of recordings and modern signal processing tools. Field surveyors need to maintain perceptual awareness and attention to detail.

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 would update in real time and at large scale. Recent advances in Deep Learning enable recognition of rare species that otherwise produce low-occurrence signals in evolving and noisy environments.

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 non-visible creatures. 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 program such as the one presented in Section 2.

Regions of interest for wildlife sensing are often difficult to access and lack physical infrastructure. Such environments are generally hostile to technology. For example, there may be large gradients of temperature and humidity. Also, outdoor deployments are subject to unpredictable events like lighting strikes, ant colony intrusion, and rodent mastication of system wiring. In a well-designed real-world deployment, the AI infrastructure can be turned inwards to flag system failures, e.g. water intrusion, blurry images, lost signal, etc.

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 (Norouzzadeh et al., 2018). 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, such as (Strout et al., 2017) for amphibians, (Mac Aodha et al., 2018) for bats, (Kiskin et al., 2017) for insects, and (Potamitis, 2016) for bird vocalization segmentation. We expect interesting contributions in the future thanks to the Bird Audio Detection challenge (Stowell et al., 2019; Cakir et al., 2017; Adavanne et al., 2017). 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

[1]Responsive Environments Group, MIT Media Lab, Cambridge, USA. Correspondence to: Clement Duhart <duhart@mit.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

streaming from microphones and cameras in real-time over the last 3 years, and our system has been used by biologists, restoration scientists, and other practitioners.

2. Tidmarsh Wildlife Sanctuary

The Tidmarsh Wildlife Sanctuary is a 600-acre former cranberry farm in southern Massachusetts in the midst of a freshwater wetland restoration. Different types of sensors are permanently deployed on the site to monitor its evolution, including ground/topography/hydrology restoration and wildlife activity. Our 'Tidzam' wildlife detection system monitors wildlife, leveraging 24 custom microphones and 6 cameras deployed across four different areas at Tidmarsh, as illustrated in Figure 1.

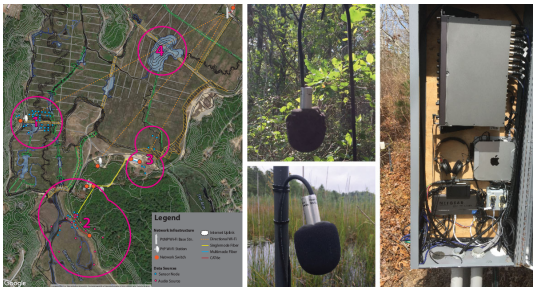

Figure 1. Sensors, microphones, and cameras are deployed in four regions of interest at Tidmarsh. Custom sensors have been specially designed to withstand wetland conditions year-round.

2.1. Tidzam Wildlife Sensing Framework

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.

2.1.1. 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 (Duhart et al., 2019). A flow controller limits the recording volume and parameterizes the extraction balance between unidentified and uncertain predictions. Our 'Tidplay' platform, introduced in Section 2.2, allows bio-acoustic human experts to annotate and discuss these recordings while building a local acoustic database used to iteratively refine the classifiers. The database is – at the time of writing – composed of 400,000 500 ms recordings distributed over 66 classes including system failure modes (e.g. microphone crackling or offline), 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 1. 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 (Jordan & Jacobs, 1992) running on one Titan X GPU. It continuously analyzes overlapped 500 ms Mel-Spectrogram windows from 24 discrete microphones and 3 on-camera microphones deployed on site.

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 1. 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

2.1.2. 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 a high level of visual semantic description, 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 (Girshick, 2015), and Yolo v3 (Redmon & Farhadi, 2018). As illustrated in Figure 2, precise species identification is still a challenging task in regards to 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.

Based on this assumption, 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.

Figure 2. Wildlife detection on the Herring site by a Yolo v3 model that has not been refined by a local training database.

2.2. 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 be used for timestamped annotations of audio as shown in Figure 3, for drawing video bounding boxes, and for pose estimation.

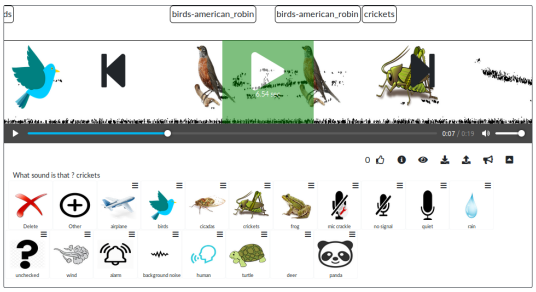

Figure 3. The audio module of the Tidplay annotation platform shows the recording's spectrogram during listening to facilitate the annotation task.

3. 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. More work is needed to develop an accurate and automatic wildlife survey system, but Tidzam already provides a powerful tool for ecological research.

3.1. Managed Care at San Diego Zoo

Wildlife behavior comparison between managed care and natural environments is another important aspect of measuring human impacts on animal quality of life. In collaboration with the San Diego Zoo, we are exploring how 2D/3D pose estimation could provide benefits in comparison studies, with the unique benefit of collecting data around-the-clock without intrusive deployment. Figure 4 shows an example frame of a panda from one of our early efforts.

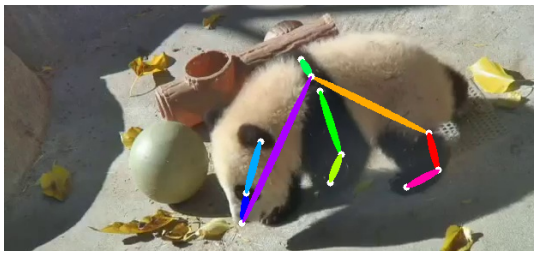

Figure 4. Tidzam 2D pose estimation in San Diego Zoo.

3.2. Amazon Conservation Area

Deployment in a rainforest, at high altitude, or in other remote environments requires the design of fully autonomous systems that are self-sufficient in terms of energy and computation. In collaboration with Instituto Nacional de Pesquisas da Amazonia (INPA), we are beginning a project aimed at the design of embedded Deep Learning nodes for long-term deployment in remote environments.

4. Conclusion

We have presented our Tidzam framework, an ongoing effort to deploy Deep Learning tools for automatic wildlife surveying in restoration, conservation and managed care environments. Our work shows how Deep Learning can advance significant opportunities for ecological research efforts, restoration science, and public engagement.

References

Adavanne, S., Drossos, K., akir, E., and Virtanen, T. Stacked convolutional and recurrent neural networks for bird audio detection. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 1729–1733, Aug 2017. doi: 10.23919/EUSIPCO.2017.8081505.

Cakir, E., Adavanne, S., Parascandolo, G., Drossos, K., and Virtanen, T. Convolutional recurrent neural networks for bird audio detection. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 1744–1748, Aug 2017. doi: 10.23919/EUSIPCO.2017.8081508.

Duhart, C., Dublon, G., Mayton, B., and Paradiso, J. Deep learning locally trained wildlife sensing in real acoustic wetland environment. In Thampi, S. M., Marques, O., Krishnan, S., Li, K.-C., Ciuonzo, D., and Kolekar, M. H. (eds.), *Advances in Signal Processing and Intelligent Recognition Systems*, pp. 3–14, Singapore, 2019. Springer Singapore. ISBN 978-981-13-5758-9.

Girshick, R. Fast r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

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

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

Mac Aodha, O., Gibb, R., Barlow, K. E., Browning, E., Firman, M., Freeman, R., Harder, B., Kinsey, L., Mead, G. R., Newson, S. E., Pandourski, I., Parsons, S., Russ, J., Szodoray-Paradi, A., Szodoray-Paradi, F., Tilova, E., Girolami, M., Brostow, G., and Jones, K. E. Bat detectivedeep learning tools for bat acoustic signal detection. *PLOS Computational Biology*, 14(3):1–19, 03 2018. doi: 10.1371/journal.pcbi.1005995.

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., and Clune, J. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25):E5716–E5725, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1719367115.

Paradiso, J. Our extended sensoria - how humans will connect with the internet of things. *The Next Step: Exponential Life, Open Mind Collection*, 1(1):47–75, 2016.

Potamitis, I. Deep learning for detection of bird vocalisations. *CoRR*, abs/1609.08408, 2016.

Redmon, J. and Farhadi, A. Yolov3: An incremental improvement, 2018.

Stowell, D., Wood, M. D., Pamua, H., Stylianou, Y., and Glotin, H. Automatic acoustic detection of birds through deep learning: The first bird audio detection challenge. *Methods in Ecology and Evolution*, 10(3):368–380, 2019. doi: 10.1111/2041-210X.13103.

Strout, J., Rogan, B., Seyednezhad, S. M. M., Smart, K., Bush, M., and Ribeiro, E. Anuran call classification with deep learning. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2662–2665, March 2017. doi: 10.1109/ICASSP.2017.7952639.

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.