# 3D CONVOLUTIONAL RECURRENT NEURAL NETWORKS FOR BIRD SOUND DETECTION

Technical Report

*Ivan Himawan, Michael Towsey, Paul Roe*

Science and Engineering Faculty, Queensland University of Technology
Brisbane, Australia
{i.himawan, m.towsey, p.roe}@qut.edu.au

## ABSTRACT

With the increasing use of a high quality acoustic devices to monitor wildlife population, it has become imperative to develop techniques for analyzing animals' calls automatically. Bird sound detection is one example of a long-term monitoring project where data are collected in continuous periods, often cover multiple sites at the same time. Inspired by the success of deep learning approaches in various audio classification tasks, this paper first review previous works exploiting deep learning for bird audio detection, and then proposes a novel 3-dimensional (3D) convolutional and recurrent neural networks. We employed 3D convolutions for extracting spatial and temporal information simultaneously. In order to leverage powerful and compact features of 3D convolution, we employ separate RNNs, acting on each filter of the last convolutional layers rather than stacking the feature maps in the typical combined CNN and RNN architectures.

***Index Terms—*** bird sound detection, deep learning, 3D CNN, GRU, biodiversity

## 1. INTRODUCTION

There has been growing interest to assess the wide-ranging impacts on biodiversity currently occurring around the globe. With the rapid decline in global wildlife populations due to environmental pollution, there has been a progressive effort over the years for monitoring vocalizing species as valid indicators of biodiversity. Monitoring avian population in their habitats is one of such efforts since birds are good ecological indicators of environmental changes [1]. For example, it enables researchers to obtain valuable information such as habitat change, migration pattern, pollution, and disease outbreaks in the environments. Because birds play a crucial role to the environment, there is a considerable effort has been devoted to focus for the conservation of birds.

In order to collect data at large spatio-temporal scales, ecologists often deploy acoustic monitoring devices to cover a large area of the land. As a result, a large quantity of recordings is being generated. These recordings, constitutes more than years of environmental monitoring, can not be analyzed manually. In this regard, ecoacoustics research [2, 3] has become one of the "big data" research area and may benefit substantially from "big data" analysis. Detecting bird sounds in audio recordings is one research problem example where data are continuously collected from various sources in a wide range of locations and environments, including mobile phones [4, 5].

In recent years, deep learning techniques have revolutionized the applicability of machine learning in speech, vision, and text processing. Significant improvements in many classification tasks are reported using deep architectures, where deep convolutional neural networks (CNN) have been used extensively in computer vision tasks. Since CNN learn filters that are shifted in both frequency and time, it addresses the limitation of deep neural networks (DNN), which lacks both time and frequency invariance. The use of deeper and more efficient CNN (e.g., GoogLeNet, ResNet, DenseNet) is also becoming popular and has shown state-of-the-art performance in object detection and image classification challenges [6, 7, 8]. The use of CNN is also popular in audio classification and speech recognition applications where audio signal is often converted into a spectrogram and treated as an input image to CNN.

Our novel contribution in this paper is the extension of conventional convolutional recurrent neural networks using 3-dimensional (3D) convolutional architecture for bird sound detection. The 3D CNN architecture has been employed in video processing application such as human action classification [9], audio-visual matching [10], and recently text-independent speaker verification [11]. In this work, we use 3D CNN to capture spatial information in time from audio data stream. Also, 3D CNN is assumed to produce powerful and compact features compared to 2D CNN [12]. In order to receive the most benefits from these features, we employ separate RNNs, acting on each filter of the last convolutional layers rather than stacking the feature maps in the typical combined CNN and RNN architectures.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 describes data and methods for bird sound detection. Experimental results are presented and discussed in Section 4. Finally, Section 5 concludes the paper.

## 2. RELATED WORKS

Currently, the state-of-the-art results for bird sound detection, and also recognition are obtained with the use of CNN. Specifically, CNN can act as a feature extractor which is shown to be superior over hand-crafted features in many classification tasks [13]. Thus, a mid-level representation of audio (i.e., a spectrogram) is popular as an input feature since it contains high-dimensional information (e.g., channel, environment). Despite promising detection results when using sophisticated classifiers such as CNN, state-of-the-art results can only be obtained if CNN are tuned carefully. This often requires domain knowledge and the interpretation of model that are well suited for bird data. The typical workflow for large scale bird

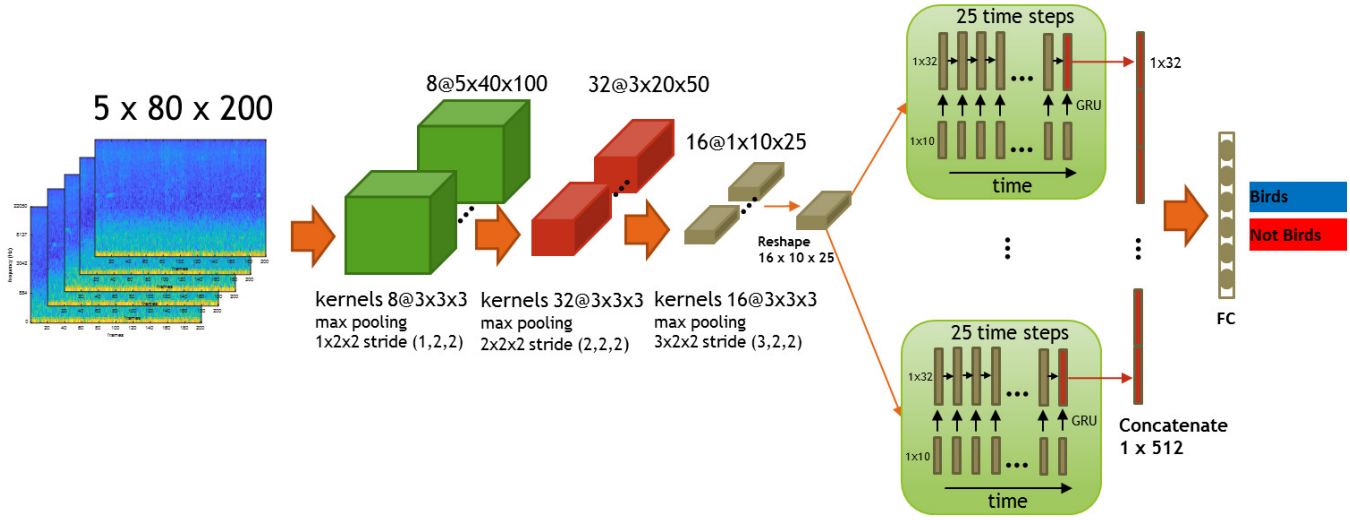


Figure 1: 3D-CNN architecture for bird sound detection. A 3D convolutional neural networks with three convolutional layers followed six teen recurrent layers and at the end one fully connected (FC) layer followed by softmax output layer. Input is a stack of 2-second audio clip.

sound detection and recognition using CNN consists of spectrogram feature extraction from audio recordings, and model training and evaluation. There is a considerable amount of works for predicting the location of bird sound within the spectrogram. The aim is to remove background noise and extract only parts contain bird singing/calling [14]. This includes spectral enhancement stage and image processing heuristics to discard non-bird sounds [15]. Even though noise reduction techniques may work well for certain dataset, bird sound localization is still a challenging task when there are dominant man-made noises (e.g., traffic, human singing, vehicles) in the audio clip.

A variety of CNN architectures have been explored for bird audio detection and recognition tasks. In general, very deep CNNs such as ResNet [7] and DenseNet [8] architectures achieve better performance compared to the standard CNN model [16]. However, as shown in the previous BAD challenge, using a wide receptive field in a conventional CNN configuration can also achieved state-of-the-art results (*bulbul* submission). Other notable deep learning architecture employed in BAD challenge is the combination of CNN and RNN architectures (CNN+RNN) [17, 18]. In this case, the CNN is used for local feature extraction and the recurrent layers to model the long-term dependencies. For example, [18] used bidirectional RNN (BRNN) to process feature maps of the last CNN layer and achieved 88.41% AUC measure on the evaluation data. Data augmentation strategy (i.e., frequency and time shift) to improve the generalization of the network is also employed by many teams, albeit with marginal improvement [17]. We also tested our proposed 3D-CNN+RNN in the previous BAD evaluation set (post-challenge submission) and achieved 88.95% AUC score, comparable to the official state-of-the-art results published in the first challenge.

Table 1: Bird audio detection challenge 2 statistics in the development set.

| Dataset | present | absent | total |
|---|---|---|---|
| freefield1010 | 1,935 | 5,755 | 7,690 |
| warblrb10k | 6,045 | 1,955 | 8,000 |
| BirdVox | 10,017 | 9,983 | 20,000 |
| Total | 17,997 | 17,693 | 35,690 |

## 3. DATA AND METHODS

### 3.1. Datasets

The bird audio detection challenge 2 used datasets released in previous BAD challenge with the addition of new datasets: (a) BirdVox (BirdVox-DCASE-20k), and (b) Poland (PolandNFC). Each audio clip is 10-second long and sampled at 44.1 kHz. The total number of audio recordings for development and evaluation set are 35690 and 12620, respectively. The label for development set is 1 if any bird sound is present and 0 if none. The statistics of the development and evaluation sets are presented in Table 2.

### 3.2. Feature Extraction

We split 10-second audio clip into $5 \times$ 2-second clips. A spectrogram (from 2-second clip) computed from sequences of Short-Time Fourier Transform (STFT) of overlapping windowed signals is used as the sound representation. A signal is framed using a window of 20 ms (882 samples). The STFT analysis is carried out using a Hamming window, 50% overlap, 1024 FFT bins by zero padding. Given the audio signal $s(t)$, the complex spectrum can be expressed as, $S(n, f) = |S(n, f)|e^{j\theta(n,f)}$, where $|S(n, f)|$ and $\theta(n, f)$ are the magnitude and phase spectrum at frame $n$ and frequency $f$, respectively. We constructed triangular-shape filters linearly spaced in mel scale to convert a spectrogram to a Mel-spectrogram with the

number of filters set to 80. The magnitude values are then converted into log magnitude, $S(n, f) = log(|S(n, f)|$. The input feature shape for spectrogram is $5 \times 80 \times 200$. The features were normalized as input to 3D-CNN.

### 3.3. 3D convolutional recurrent neural networks

In essence, the 3D convolution is the extension of 2d convolution. The 3D-CNN+RNN architecture proposed in this work consists of 3 convolutional layers. We use a receptive field of $3 \times 3 \times 3$ followed by a max pooling operation for every convolutional layer. The activation function is Rectified linear unit (ReLU). A batch normalization layer [19] was employed for all the convolutional layers. Dropout with rate of 0.5 was employed in convolutional layers. The weights are initialized with Xavier initialization [20]. We employed multiple GRUs where each feature map of the last convolutional layer is fed to the GRU. Hence, we had a total of 16 separate GRUs for 16 filters used at the last convolutional layer. We constructed 25 recurrent layers for each feature map, where 25 is the number of time steps mapped from the 200 time steps in the original spectrogram. We used recurrent networks with 32 GRU cells. The output for each RNN (many-to-one configuration) is concatenated and then fed into a fully connected layer. The combined 3D-CNN and RNN are optimized jointly by employing backpropagation algorithm. A softmax layer with two nodes is used (bird vs non-bird). The network is trained using RMSProp optimizer [21] with momentum of 0.9 and initial learning rate of $10^{-3}$. We used batches of 8 training example to train our models. The binary cross-entropy is used as a loss function. Tensorflow [22] is used to implement the models. The system codes with the proposed methods is made available in `https://github.com/himaivan/BAD2`.

## 4. EXPERIMENTS

### 4.1. Evaluation metric

The perfomance evaluation metric for bird sound classification is reported in terms of Area Under the ROC curve (AUC) as suggested in the evaluation plan.

### 4.2. Training

We tested different parameter combinations to decide the final architecture to be used in the evaluation which include the number of CNN layers $\{3, 4\}$ and drop-out rates $\{0.5, 0.7\}$. For the first training strategy, we trained our baseline model using 97% of the total data. The 3% validation split is used to monitor the training process and for selecting final models. We stopped the training after 150 epochs to avoid overfitting. Since the training data is large, we did not perform with data augmentation strategy. We then selected 5 models from different epoch with the highest accuracy on the validation split and averaged the predictions. We also trained our model using 3-way cross-validation strategy where in each fold two sets were used for training and the other one for testing, and averaged the predictions (hence, 15 networks were selected, five models for each cross-validation fold).

## 5. RESULTS

Our proposed 3D-CNN+RNN obtained a preview score of 87.13% when model is trained using the combined data. The 3-way cross-validation results where in each fold two sets were used for training and the other one for testing obtained 88.70% AUC score on the unseen evaluation data (via model ensemble method). Building a robust deep learning model typically requires a large amount of labeled training data. However, obtaining many labeled data is an expensive task and not always feasible. In future work, we will investigate the method to generate labeled data via pseudo-labeling method where approximate labels are produced from unlabeled data using trained models.

Table 2: Stratified 3-way cross-validation results.

| Train Configuration | Test | AUC |
|---|---|---|
| freefield1010 + warblrb10k | BirdVox | 63.1% |
| freefield1010 + BirdVox | warblrb10k | 85.9% |
| warblrb10k + BirdVox | freefield1010 | 79.4% |
| model ensemble | Evaluation data | 88.7% |

## 6. REFERENCES


[1] G. R. Walther et al., "Ecological responses to recent climate change," *Nature*, vol. 416, no. 6879, pp. 389–395, 2002.

[2] M. Towsey et al., "The use of acoustic indices to determine avian species richness in audio-recordings of the environments," *Ecological Informatics*, vol. 21, pp. 110–119, 2014.

[3] J. Sueur and A. Farina, "Ecoacoustics: the ecological investigation and interpretation of environmental sound," *Biosemiotics*, vol. 8, no. 3, pp. 493–502, 2015.

[4] D. Stowell et al., "Bird detection in audio: a survey and a challenge," in *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing*, 2016, pp. 1–6.

[5] ——, "Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge," *Methods in Ecology and Evolution*, 2018.

[6] C. Szegedy et al., "Going deeper with convolutions," in *Proceedings of Computer Vision and Pattern Recognition*, 2014, pp. 1–9.

[7] K. He et al., "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, no. 770-778, 2016.

[8] G. Huang et al., "Densely connected convolutional networks," in *Proceedings of Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[9] S. Ji et al., "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[10] A. Torfi et al., "3D convolutional neural networks for cross audio-visual matching recognition," *IEEE Access*, vol. 5, pp. 22 081–22 091, 2017.

[11] A. Torfi, J. Dawson, and N. M. Nasrabadi, "Text-independent speaker verification using 3D convolutional neural networks," in *Proceedings of IEEE International Conference on Multimedia and Expo*, 2018.

[12] I. Teivas, "Video event classification using 3D convolutional neural networks," Master's thesis, Tampere University of Technology, 2016.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[14] M. Lasseck, "Towards automatic large-scale identification of birds in audio recordings," in *Proceedings of Experimental IR Meets Multilinguality, Multimodality, and Interaction - 6th International Conference of the CLEF Association, CLEF*, 2015, pp. 364–375.

[15] I. Potamitis, "Unsupervised dictionary extraction of bird vocalisations and new tools on assessing and visualising bird activity," *Ecological Informatics*, vol. 26, pp. 6–17, 2015.

[16] T. Pellegrini, "Densely connected cnns for bird audio detection," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2017, pp. 1784–1788.

[17] E. Cakir et al., "Convolutional recurrent neural networks for bird audio detection," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2107, pp. 1794–1798.

[18] S. Adavanne et al., "Stacked convolutional and recurrent neural networks for bird audio detection," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2017, pp. 1779–1783.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[20] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256, 2010.

[21] T. Tieleman and G. Hinton, "Lecture 6.5—RMSProp: Divide the gradient by a running average of its recent magnitude," COURSERA: Neural Networks for Machine Learning, 2012.

[22] M. Abadi et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv:1603.04467*, 2016.