BIRD SPECIES IDENTIFICATION USING SPECTROGRAMS AND CONVOLUTIONAL NEURAL NETWORKS

AYMEN SAAD

A project report submitted in partial fulfilment of the
Requirements for the award of the degree of
Master of Engineering (Computer and Microelectronic System)

School of Electrical Engineering
Faculty of Electrical Engineering
Universiti Teknologi Malaysia

JUNE 2020

DEDICATION

This thesis is dedicated to my father, who taught me that the best kind of knowledge to have is that which is learned for its own sake. It is also dedicated to my mother, who taught me that even the largest task could be accomplished if it is done one-step at a time.

ACKNOWLEDGEMENT

I am grateful to God for the health and wellbeing to complete this project.

I am deeply indebted to the support and affection of my parents and my wife who have made all this possible

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts. In particular, I wish to express my sincere appreciation to my main supervisor, Dr Shahidatul Sadiah, for encouragement, guidance, critics and friendship.

I am also very thankful to Professor Muhammad Mun'im bin Ahmad Zabidi and Dr Usman Ullah for their guidance, advice and motivation. Without their continued support and interest, this thesis would not have been the same as presented here.

I am also indebted to Universiti Teknologi Malaysia (UTM). Librarians at UTM for their assistance in supplying the relevant literature.

My sincere appreciation also extends to all my colleagues and others who have provided assistance at various occasions. Their views and tips are useful, indeed. Unfortunately, it is not possible to list all of them in this limited space.

ABSTRACT

Birds are particularly useful ecological indicators as they respond quickly to the changes in their environment. Thus, studies regarding the diversity of birds are indispensable. Domain experts classify birds manually to achieve accurate results, but the process is tedious with growing amounts of data. Meanwhile, bioacoustics monitoring employs automated recorders to collect large-scale audio data of fauna vocalization. Nevertheless, the analysis of large-scale audio is impossible to be done manually. Hence, machine learning is a more practical approach. Previously, Convolutional Neural Network (CNN) approach had achieved excellent results using the augmented spectrogram image of the audio. Varieties of CNN architectures such as Cube, Inception-v3, DenseNet, ResNet, and ConvNets are having advantage in high accuracy, but are disadvantage in high computational cost. These architectures are suitable for large-scale classification up to 1000 species due to the deep layer of neural network models to obtain high-level feature extraction from the spectrogram image. However, many devices intended for these models have limited computation resources and strict power consumption constraints. Therefore, the proposed study aims to optimize the CNN-based birdcall classifier model targeting embedded platforms. A low complexity CNN model, MobileNet-v2 was implied which is sufficient for a small-scale classification such as to identify ten bird species as inputs. The dataset used to train our model is from the Xeno-canto repository. Each audio data is amplified to 16 kHz and segmented into 1-second sample data. An algorithm to splice the audio according to the label is proposed. Then, each sample that contains the birdcall signal is augmented into three samples, and the noise only samples are removed from the dataset. The spectrogram image of the samples is obtained using STFT and MFCC conversion, and then all images are resized to 224×224×1, using Matlab 2019b. To verify our model, we compare it with the high complexity CNN model, ResNet-50. In the result, the MobileNet-v2 model has reduced the computational cost of ResNet-50 by 86% with a slight trade-off to the accuracy. Compared to ResNet-50, the accuracy of MobileNet-v2 dropped 12% if using STFT, but only dropped 2% if using MFCC, which made MobileNet-v2 model with MFCC conversion the best CNN model for device applications with small number of classifiers.

ABSTRAK

Burung adalah penunjuk ekologi penting yang cepat bertindak balas terhadap perubahan persekitaran. Kecekapan pakar domain mengelaskan burung secara manual memberi ketepatan pengelasan, tetapi prosesnya memakan masa. Pemantauan bioakustik menggunakan perakam automatik dapat mengumpul data audio berskala besar, tetapi analisisnya sukar dilakukan. Oleh itu, pembelajaran mesin adalah pendekatan yang lebih praktikal dimana kaedah terbaik adalah Convolutional Neural Network (CNN) yang menggunakan imej audio dalam bentuk spektogram. Pelbagai model CNN seperti Cube, Inception-v3, DensetNet, ResNet, dan ConvNets sesuai untuk pengelasan berskala besar sehingga 1000 spisis burung disebabkan oleh kedalaman lapisan model untuk mendapatkan pengekstrakan ciri berperingkat tinggi dari imej spektogram. Walau bagaimanapun, banyak kegunaan model ini memerlukan peranti yang mempunyai sumber pengiraan yang terhad dan kekangan penggunaan kuasa. Oleh itu, kajian ini mencadangkan untuk mengoptimumkan model pengelasan audio burung berasaskan CNN. Model MobileNet-v2 akan digunapakai untuk pengelasan berskala kecil sehingga sepuluh spisis burung. Set data yang digunakan untuk melatih model ini adalah dari repositori Xeno-canto. Setiap data audio dikuatkan kepada 16 kHz dan dibahagikan kepada sampel data berdurasi 1 saat. Satu algoritma untuk menyusun audio mengikut label dicadangkan. Setiap sampel yang mengandungi bunyi burung dijadikan kepada tiga sampel menggunakan imbuhan spektogram, dan sampel yang hanya mempunyai bunyi bising dikeluarkan dari dataset. Sampel imej spektogram diperoleh menggunakan penukaran STFT dan MFCC, dan kemudian semua saiz imej diubah menjadi 224 × 224 × 1, menggunakan Matlab 2019b. Untuk memastikan prestasi model ini, model ResNet-50 diambil sebagai perbandingan. Hasil mendapati model MobileNet-v2 telah menurunkan kos pengiraan ResNet-50 sebanyak 86% dan diseimbangkan dengan penurunan ketepatan pengelasan sebanyak 12% sekiranya menggunakan penukaran STFT, dan penurunan sebanyak 2% sekiranya menggunakan MFCC. Ini menunjukkan model MobileNet-v2 dengan penukaran MFCC merupakan model terbaik untuk peranti aplikasi pengelasan berskala kecil.

TABLE OF CONTENTS

TITLE		PAGE
DECLARATION		**3**
DEDICATION		**4**
ACKNOWLEDGEMENT		**5**
ABSTRACT		**6**
ABSTRAK		**7**
TABLE OF CONTENTS		**8**
LIST OF TABLES		**10**
LIST OF FIGURES		**11**
LIST OF ABBREVIATIONS		**12**
LIST OF SYMBOLS		**13**
CHAPTER 1	**INTRODUCTION**	**1**
1.1	Problem Background	3
1.2	Research Questions	3
1.3	Research Objectives	3
1.4	Project Scope	4
1.5	Project Outline	5
CHAPTER 2	**LITERATURE REVIEW**	**6**
2.1	Overview of Bird Classification	6
2.2	CNN Based Bird Classification	8
2.3	CNN Models	9
2.3.1	High Complexity Models	10
2.3.2	Low Complexity Models	13
2.4	Summary	15
CHAPTER 3	**RESEARCH METHODOLOGY**	**17**
3.1	Retrieve Dataset	18
3.2	Data Segmentation	18
3.3	Audio to Spectrograms Image Conversion	19

3.3.1 Short-Time Fourier Transform 20
3.3.2 Mel Frequency Cepstral Coefficients 20
3.4 Resize Image Spectrogram 21
3.5 Training and Testing 22
3.5.1 Calling the data 22
3.5.2 Checking the data 23
3.5.3 Load pre-trained Network 23
3.5.4 Prepare the data 23
3.6 Reduce output 24

CHAPTER 4 RESULT 25
4.1 Training Stage Outcome 25
4.1.1 High Complexity Model (ResNet-50) 25
4.1.2 Low complexity model (MobileNet-v2) 26
4.2 Testing Stage Outcome 28
4.3 Confusion Matrix. 28
4.4 The Accuracy and Time 32
4.5 Chapter Summary 32

CHAPTER 5 CONCLUSION AND RECOMMENDATIONS 33
5.1 Research Outcomes 33
5.2 Conclusion 33
5.3 Recommendations for Future Works 34

REFERENCES 35

LIST OF TABLES

TABLE NO.	TITLE	PAGE
Table 3.1	Computer Device Specification	17
Table 4.1a	ResNet-50 with STFT	26
Table 4.1b	ResNet-50 with MFCC	26
Table 4.2a	MobileNet-v2 with STFT	26
Table 4.2b	MobileNet-v2 with MFCC	27
Table 4.3	Summary of Results	32

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE
Figure 2.1	Bird species identification using a visual and audio classifier.	7
Figure 2.2	Inception-v4 model	8
Figure 2.3	Accuracy for Each Model Type	10
Figure 2.4	AlexNet Architecture	11
Figure 2.5	A building block of residual learning	12
Figure 2.6	InceptionV3	13
Figure 2.7	Basic Structure of MobileNetV2	14
Figure 2.7	SqueezeNet	14
Figure 3.1	Design Flow	17
Figure 3.2	Species of birds used in the data set.	18
Figure 3.3	VAD technology	19
Figure 3.4	Convert tone from time to spectrogram domain using STFT	20
Figure 3.5	Audio Conversion from Time to Spectrogram Domain using MFCC	21
Figure 3.6a	Resize STFT	22
Figure 3.6b	Resize MFCC	22
Figure 3.7	(Add-Ons) window	22
Figure 3.8	load images	22
Figure 3.9	Summary of the Number of Images Per Category	23
Figure 3.10	The splitting of training and testing datasets	24
Figure 3.11a	before reducing	24
Figure 3.11b	after reducing	24
Figure 4.1	Training Accuracy.	27
Figure 4.2	Confusion Matrix model (ResNet-50) with STFT.	29
Figure 4.3	Confusion Matrix model (ResNet-50) with MFCC	30
Figure 4.4	Confusion Matrix model (MobileNet-v2) with STFT	31
Figure 4.5	Confusion Matrix model (MobileNet-v2) with MFCC	31

LIST OF ABBREVIATIONS

ARU	-	Autonomous Recordings Units
C-MAP	-	Classification Mean Average Precision
CNN	-	Convolutional Neural Network
DL	-	Deep Learning
DNN	-	Deep Neural Network
GPU	-	Graphics Processing Unit
HOG	-	Histogram of Oriented Gradients
LBP	-	Local Binary Patterns
MAP	-	Mean Average Precision
ML	-	Machine Learning
MFCC	-	Mel Frequency Cepstral Coefficients
MLP	-	Multi-Layer Perceptron
RELU	-	Rectified Linear Unit
SGD	-	Stochastic Gradient Descent
SWIFT	-	Song Written and Recorded by American Singer-Songwriter Taylor
STFT	-	Short time Fourier Transform
SURF	-	Speeded-Up Robust Features
VAD	-	Voice Activity Detection

LIST OF SYMBOLS

δ	-	Minimal error
D, d	-	Diameter
F	-	Force
v	-	Velocity
p	-	Pressure
I	-	Moment of Inersia
r	-	Radius
Re	-	Reynold Number

CHAPTER 1

INTRODUCTION

Birds are an essential indicator of the biodiversity as the population and variation of their species in an ecosystem can directly reflect the ecosystem health and suitability of the habitat [1]. Meanwhile, audio-based bird identification system has proven to be particularly useful for biodiversity monitoring and education. This is because audio records are more accessible than bird photos since birds are in fact challenging to photograph, and birdcalls are easier to collect. Therefore, ecologist monitor birds through acoustic recordings of their sound.

The domain experts classify birds manually by staying at the field sites for weeks or months to record bird sounds. The process is not only time consuming, but also tedious with growing amounts of data. Thus, many automatic tools were initiated to help them in this process, such as autonomous recordings units (ARU) or SWIFT. This strategy helps professionals working with bioacoustics sounds by processing large audio collections in a fast and automated way, or help amateurs to identify birds in self-made recordings [2]–[7].

From 2013 onwards, the best method used to process audio data is Convolutional Neural Network (CNN), which can achieve excellent results using the augmented spectrogram image of the audio. The authors conclude in the 9th annual MLSP competition, that "convolutional neural networks have achieved excellent results without comprehensive function engineering. Therefore further research into these approaches is justified"[8]. Nonetheless, 2017, 2018 and 2019 Bird CLEF winners used a CNN, which was based on augmented spectrogram data derived from bird song audio files[9]–[12].

Given the considerable amount of research into automated bird species identification, there is still no adequate system for field recording and most of the works have limited their reach to inspect less noisy and carefully selected recordings due to birdcall identification challenges[13]. The main advantage in these architectures is their accuracy in solving image recognition problems. However, they have some disadvantages, such as high computational cost, and it is quite slow to train if without a functional General Processing Unit (GPU) [14].

In these previous studies, feature extraction using a spectrogram is essential. These architectures are suitable for large-scale classification up to 1000 species due to the deep layer of neural network models to obtain high-level feature extraction from the spectrogram image. However, many devices intended for these models have limited computation resources and strict power consumption constraints. Therefore, the proposed study aims to optimize the CNN-based birdcall classifier model targeting embedded platforms. We will use a MobileNetv2 CNN architecture [15] that is sufficient for a small-scale classification such as to identify ten bird species as inputs. The dataset used to train our model is from the Xeno-Canto repository. This website is exclusive for birdcall. Each audio files are resampled from 44,100 Hz to 16,000 Hz and segmented into 1-second sample data automatically by using Voice Activity Detection (VAD) technology [16]-[17]. Then, the spectrograms image of the sample data which is obtained using STFT and MFCC conversion, and then all images are resized to 224×224×1 to be used as input of the implemented CNN model. To verify our model, we compare it with the ResNet-50 model and expected to obtain between 80% to 85% accuracy.

1.1 Problem Background

Given the considerable amount of research into automated bird species recognition, there is still no adequate field recording tool and most studies have limited their reach to study less noisy and carefully chosen recordings due to birdcall recognition challenges[13]. One of the most critical challenges when using CNN architectures is the complicated and expensive algorithm, whereas many devices intended for these models have limited computation resources and strict power consumption constraints.

1.2 Research Questions

There are three research questions have been identified for this study.

Q1. What is the most cost-effective CNN model to classify birds?

Q2. What is the best method to implement the low-complexity CNN model?

Q3. What is the performance of low-complexity bird classifier compared to state-of-the-art approaches?

1.3 Research Objectives

The objectives of the research are:

1. To investigate the best architecture for the birdcall classifier model targeting embedded platforms.

2. To explore the low-complexity CNN model and implement the birdcall identification system.

3. To analyse the performance of bird classifier.

1.4 Project Scope

The project focuses on building a bird identification system of ten bird species using spectrograms and lightweight CNN architecture method and implementing this system on MATLAB 2019b.

The system can detect the sound of birds and identify to which class it belongs. The ten classes are;

1. *Mareca Penelope,*
2. *American Robin,*
3. *Calaudalla Cheleensis,*
4. *Cinnyris Asiaticus,*
5. *Zenaida Asiatica,*
6. *Lanius Collurio,*
7. *Vanellus Senegallus,*
8. *Phaethornis-longirostris,*
9. *Columba Oenas,*
10. *Terpsiphone Viridis*.

The dataset used to train our model is from the Xeno-canto repository (http://www.xeno-canto.org). This website is exclusive for birdcall. Each audio files are resampled from 44,100 Hz to 16,000 Hz and segmented into 1-second sample data automatically by using Voice Activity Detection (VAD) technology. Then, the spectrograms image of the sample data which is obtained using STFT and MFCC conversion, and then all images are resized to 224×224×1 to be used as input of the implemented CNN model.

We will use a MobileNetv2 CNN architecture that is sufficient for a small-scale classification such as to identify ten bird species as inputs. To verify our model, we compare it with the ResNet-50 model and expected to obtain between 80% to 85% accuracy.

Work on an embedded system is not in the scope even though the objective is to find a low-complexity solution. Therefore, all the work will be done as PC software.

1.5 Project Outline

We have clarified the context, problem statement, objectives and project scope in Chapter 1. In Chapter 2, we analyse the results provided during some of the most recent bird species identification challenges and present the theory of a state-of-the-art classifier of bird species, which will be used as the basis for this thesis. Chapter 3 consists of the entire process of the project and the methods for how the project is being carried out. We will present the theory for the processing of our data methodology, the deep neural network and the methods of assessment used in the thesis to generate the results. In Chapter 4, we will show the results of this project from an analysis of a new data set with ResNet-50 and MobileNet-v2. In Chapter 5, we summarized and discussed what we set out to do, what has been achieved, what problems arose, and propose possible routes for future work.

REFERENCES

[1] D. Konovalov and M. Sankupellay, “Bird Call Recognition using Deep Convolutional Neural Network , ResNet-50,” *Acoust. 2018*, no. November, 2018.

[2] S. Kahl and F. St, “Overview of BirdCLEF 2019 : Large-Scale Bird Recognition in Soundscapes,” pp. 9–12, 2019.

[3] M. Towsey, A. Nantes, and P. Roe, “A toolbox for animal call recognition,” no. June, 2012.

[4] J. Cai, D. Ee, B. Pham, P. Roe, and J. Zhang, “Sensor Network for the Monitoring of Ecosystem : Bird Species Recognition,” no. June 2014, 2008.

[5] V. M. Trifa, A. N. G. Kirschel, and C. E. Taylor, “Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models Automated species recognition of antbirds in a Mexican,” no. May 2014, 2008.

[6] F. Briggs *et al.*, “Acoustic classification of multiple simultaneous bird species : A multi-instance multi-label approach,” vol. 131, no. 6, 2012.

[7] P. Fallgren, Z. Malisz, and J. Edlund, “Towards fast browsing of found audio data : 11 presidents.”

[8] Y. Huang, F. Briggs, R. Raich, K. Eftaxias, and Z. Lei, “THE NINTH ANNUAL MLSP DATA COMPETITION Yonghong Huang Intel Corporation Portland , OR USA Konstantinos Eftaxias University of Surrey Surrey , UK Forrest Briggs , Raviv Raich Oregon State Unversity Corvallis , OR USA Vesta Corporation Portland , OR USA,” *2013 IEEE Int. Work. Mach. Learn. Signal Process.*, pp. 1–4, 2013.

[9] S. Kahl, T. Wilhelm-Stein, H. Klinck, D. Kowerko, and M. Eibl, “Recognizing Birds from Sound - The 2018 BirdCLEF Baseline System,” 2018.

[10] B. Fazekas, A. Schindler, T. Lidy, and A. Rauber, “A Multi-modal Deep Neural Network approach to Bird-song identification.”

[11] M. Lasseck, “Bird Species Identification in Soundscapes *,” no. September, pp. 9–12, 2019.

[12] J. Schlüter, “Bird Identification from Timestamped , Geotagged Audio Recordings,” no. 1, 2018.

[13] A. Technology, P. North, E. Group, and P. North, “Avian biology,” pp. 1–27, 2018.

[14] Z. Alom *et al.*, “A State-of-the-Art Survey on Deep Learning Theory and Architectures,” pp. 1–67, 2019.

[15] M. Sandler, M. Zhu, A. Zhmoginov, and C. V Mar, “MobileNetV2: Inverted Residuals and Linear Bottlenecks.”

[16] M. H. Moattar and M. M. Homayounpour, “A SIMPLE BUT EFFICIENT REAL-TIME VOICE ACTIVITY DETECTION ALGORITHM,” no. Eusipco, pp. 2549–2553, 2009.

[17] S. Ding, Q. Wang, S. Chang, and L. Wan, “Personal VAD: Speaker-Conditioned Voice Activity Detection.”

[18] A. Marini, A. J. Turatti, A. S. Britto, and A. L. Koerich, “VISUAL AND ACOUSTIC IDENTIFICATION OF BIRD SPECIES A . Marini , A . J . Turatti , A . S . Britto Jr ., A . L . Koerich Pontifical Catholic University of Paran’ a Postgraduate Program in Informatics Curitiba , PR , Brazil,” *2015 IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 2309–2313, 2015.

[19] T. Baba, “Time-Frequency Analysis Using Short Time Fourier Transform,” pp. 32–38, 2012.

[20] S. Time, F. Transform, and D. F. Transform, “B3 . Short Time Fourier Transform (STFT).”

[21] A. Sevilla and H. Glotin, “Audio bird classification with inception-v4 extended with time and time-frequency attention mechanisms,” *CEUR Workshop Proc.*, vol. 1866, 2017.

[22] A. Fritzler, S. Koitka, and C. M. Friedrich, “Recognizing Bird Species in Audio Files Using Transfer Learning.”

[23] S. Bianco, R. Cadene, L. Celona, and P. Napoletano, “Benchmark analysis of representative deep neural network architectures,” IEEE Access, vol. 6, no. October, pp. 64270–64277, 2018.

[24] Y. Zhong, C. Liqin, and L. Zhang, “Pre-Trained AlexNet Architecture with Pyramid Pooling and Supervision for High Spatial Resolution Remote Sensing Image Scene Classification,” no. August, 2017.

[25] Z. Alom *et al.*, “The History Began from AlexNet : A Comprehensive Survey on Deep Learning Approaches.”

[26] C. Szegedy and S. G. Com, “Batch Normalization : Accelerating Deep Network Training by Reducing Internal Covariate Shift,” vol. 37, 2015.

[27] A. G. Howard and W. Wang, “Applications,” 2012.

[28] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “50 X FEWER PARAMETERS AND < 0 . 5MB MODEL SIZE,” pp. 1–13, 2017.

[29] S. Time, F. Transform, and D. F. Transform, “B3. Short Time Fourier Transform (STFT).”

[30] M. Xu, L. Duan, J. Cai, and L. Chia, “HMM-Based Audio Keyword Generation,” pp. 566–574, 2004.