Ecological Informatics 52 (2019) 74–81

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Handcrafted features and late fusion with deep learning for bird sound classification

Jie Xie[a,c,*], Mingying Zhu[b]

[a] Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, PR China

[b] Department of Economics, University of Ottawa, Ontario K1N6N5, Canada

[c] Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment and Technology, Jiangnan University, PR China

ARTICLE INFO

Keywords:
Bird sound classification
Convolutional neural networks
Acoustic feature
Visual feature

ABSTRACT

Automated classification of calling bird species is useful for large-scale temporal and spatial environmental monitoring. In this paper, we investigate acoustic features, visual features, and deep learning for bird sound classification. For the deep learning approach, the Convolutional Neural Network layers are used for learning generalized features and dimension reduction, while a conventional fully connected layer is used for classification. Then, an unified end-to-end model is built by combing those three layers for classifying calling bird species. For visual and acoustic features, two traditional classifiers are compared to classify the bird sounds. Experimental results on 14 bird species indicate that our proposed deep learning method can achieve the best F1-score 94.36%, which is higher than using the acoustic features approach (88.97%) and using the visual features approach (88.87%). To further improve the classification performance, a class-based late fusion method is explored. Our final best classification F1-score is 95.95%, which is obtained by the late fusion of the acoustic features approach, the visual features approach, and deep learning.

1. Introduction

Birds are widely regarded as an excellent indicator of biodiversity for their important ecosystem services (Acevedo et al., 2009). However, the decrease in bird population has been noticed worldwide. To optimize the protection policy and increase the bird population, it is becoming ever more important to monitor birds (Bao and Cui, 2005). Traditional methods for surveying birds that require ecologists to conduct a bird census are both time-consuming and costly (Bardeli et al., 2010; Brandes, 2008).

Acoustics have been widely used to monitor animals' presence and activity, because most animals, such as insects, anurans, birds, and certain mammals are often heard rather than seen (Briggs et al., 2012; Brown, 1991). To fully utilize acoustic cues of animal calls, collecting animal sounds is in high demand. Recent advances in acoustic sensor techniques provide a novel way to record animal vocalizations over larger temporal and spatial scales (Costa et al., 2012; Farina and Gage, 2017). Since several gigabytes of compressed data can be generated by an acoustic sensor per day, enabling automating animal species identification in acoustic datasets has become increasingly important (Brandes, 2008).

Previous studies have proposed various methods for bird sound classification. (Gregory and van Strien, 2010) presented a novel method to recognize inharmonic and transient bird sounds efficiently, where wavelet decomposition was used for feature extraction. Two neural networks were then used for classifying eight bird species. The best classification result of the test sounds was 96%, which was obtained by the supervised multilayer perceptron. This result indicated that wavelet transform was an efficient method for bioacoustic signal feature extraction. (Han et al., 2011) compared three machine learning algorithms for the automated classification of nine frogs and three bird species. The best average true positive rate and false positive were 94.95% and 0.94% for Support Vector Machines (SVMs). The features used included minimum frequency, maximum frequency, maximum power and call duration. (Hinton et al., 2012) proposed a multi-instance multi-label (MIML) approach for classifying multiple simultaneous bird species. A bag of instance representation was used to characterize the audio signal, where a short audio recording was a bag, and the instances correspond to 2D segments in the time-frequency domain described by a vector of their acoustic properties. Then existing MIML

* Corresponding author at: Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, PR China; Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment and Technology, Jiangnan University, PR China.
E-mail addresses: xiej8734@gmail.com (J. Xie), mzhu089@uottawa.ca (M. Zhu).

https://doi.org/10.1016/j.ecoinf.2019.05.007
Received 21 November 2018; Received in revised form 24 April 2019; Accepted 8 May 2019
Available online 10 May 2019
1574-9541/

classifiers were applied for classifying simultaneously calling bird species. An accuracy of 96.1% (true positives/negatives) was achieved for 13 bird species. (Hirsch and Pearce, 2000) investigated unsupervised feature learning to improve the performance of automatic large-scale bird sound classification. Mel-Frequency Cepstral coefficients (MFCCs) were found to be worse than the raw Mel spectral data. In addition, unsupervised feature learning provided a substantial performance boost over MFCCs and Mel spectra procedures. Recently, (Huang et al., 2009) proposed two active learning methods for effectively reducing the need for expert human annotation. A kernel-based extreme learning machine was interated for bird sound classification, where an unweighted average recall of 80% was achieved. (Jiang et al., 2002) first proposes a Gaussian Mixture Model-based frame selection with an event-energy-based sifting procedure that selected representative acoustic events. Then, a Mel band-pass filter was employed on each event's spectrogram, where the output in each sub-band was parameterized by an auto-regressive model. Finally, a SVMs classifier was used for classification and achieved a F-score metric of 0.928 over other recent approaches (0.632).

Deep learning approaches have demonstrated great success on pattern recognition task (Kahl et al., 2017; Krizhevsky et al., 2012; Lasseck, 2018). (Lee et al., 2008) used a variety of CNN to generate features extracted from visual representations of field recordings. A mean average precision of 0.605 was obtained for the BirdCLEF2017,[1] and 0.786 was achieved when only foreground was considered. (Lee et al., 2009) proposed a fusion of shallow learning based on unsupervised dictionary learning with a deep CNN combined with data augmentation for bird species classification. The best classification accuracy was 0.96 for a dataset of 5428 flight calls spanning 43 species. (Ojala et al., 2002) presented deep learning techniques for audio-based bird identification at very large scale and Deep Convolutional Neural Networks (DCNNs) were fine-tuned to classify 1500 species. Various data augmentation techniques were applied to prevent over-fitting and to further improve model accuracy and generalization. The proposed method surpassed previous state-of-the-art by 15.8% identifying foreground species and 20.2% considering also background species.

We investigate acoustic features, visual features, and deep learning for bird species classification. We first classify each bird call using acoustic features and visual features with two traditional classifiers: a K-nearest neighbor (K-NN) classifier and a random forest (RF) classifier. Then, we investigate deep learning techniques to classify bird calls. To further improve the classification result, we finally use a class-based late fusion of selected classification frameworks.

This paper is organized as follows: In Section 2, we describe the proposed approach for bird sound recognition, which includes data description, feature extraction, and recognition. Section 3 gives the description of the late fusion method. Section 4 reports the experimental results. Section 6 presents conclusions and directions for future work.

2. Data and methods

In this study, we investigate acoustic features, visual features, and CNN for studying bird calls, and the conceptual framework of the investigated systems is shown in Fig. 1.

2.1. Dataset

We aim to classify 14 bird species, which are widespread in Queensland, Australia. Recordings are collected from Xeno-Canto website (https://www.xeno-canto.org/) and re-sampled at 11,025 Hz, because the spectrum of these sounds have shown that most frequencies of them are below 5 kHz. Then, all re-sampled recordings are mixed to mono and saved in WAV format (see Table 1).

2.2. Audio images

CNNs are often used for image classification, which is a task of classifying two-dimensional data. However, the sample audio signals are one-dimension, which needs to be transformed into two-dimensional data. We convert the audio signals to a log scale time-frequency representation using the Constant-Q Transform (CQT) (Pereira and Cooper, 2006). The CQT transform of a time-domain signal is defined as

$$X[k,n] = \sum_{q=n-\lfloor N_k/2 \rfloor}^{n+\lfloor N_k/2 \rfloor} x(q) a_k^*(q - N + N_k/2) \tag{1}$$

where $k = 1, 2, \cdots, K$ indexes the frequency bins of the CQT, $\lfloor . \rfloor$ denotes rounding towards negative infinity and $a_k^*(n)$ denotes the complex conjugate of $a_k(n)$. The basis functions $a_k(n)$ are complex-valued waveforms, here also called time-frequency atoms, and are defined by

$$a_k(n) = \frac{1}{N_k} \omega\left(\frac{n}{N_K}\right) exp\left(-j2\pi n \frac{f_k}{f_s}\right) \tag{2}$$

where f_k is the center frequency of bin k, f_s denotes the sampling rate, and $w(t)$, is a continuous window function sampled at points determined by t, N_k is the window length, which is inversely proportional to f_k in order to have the same Q-factor for all bins k. The center frequency f_k is defined as

$$f_k = f_1 2^{\frac{k-1}{B}} \tag{3}$$

where f_1 is the center frequency of the lowest-frequency bin, and B determines the number of bins per octave. In practice, B is the most important parameter of choice when using the CQT, because it determines the time-frequency resolution trade-off of the CQT.

2.3. Handcrafted features for bird species classification

2.3.1. Acoustic features

(1) Spectral centroid is the center point of spectrum distribution. With the magnitudes as the weight, it is calculated as the weighted mean of frequencies

$$Sc = \frac{\sum_{k=0}^{N-1} f_k X(k)}{\sum_{k=0}^{N-1} X(k)} \tag{4}$$

where $X(k)$ is the discrete Fourier transform (DFT) of the windowed signal of the k-th window, N is the half size of DFT.

(2) Spectral bandwidth can be used to represent the difference between the upper and lower cut-off frequencies.

$$Bw = \sqrt{\frac{\sum_{k=0}^{N-1} (k - Sc)^2 |x(n)|}{\sum_{k=0}^{N-1} X(k)}} \tag{5}$$

(3) Spectral contrast is defined as the decibel difference between peaks and valleys in the spectrum (Qian et al., 2017). First, the spectrum of each subband is calculated as $\{x_{k,1}, x_{k,2}, \ldots, x_{k,N}\}$. Then, all spectrum is sorted in a descending order, which can be represented as $\{x_{k,1}', x_{k,2}' \ldots, x_{k,N}'\}$, where $x_{k,1}' > x_{k,2}' > x_{k,N}'$. Next, the strength of spectral peaks and spectral valleys are estimated as:

$$Peak_k = log \frac{1}{\alpha N} \sum_{i=1}^{\alpha N} x_{k,i}' \tag{6}$$

[1] A bird identification task supported by the Xeno-Canto foundation for nature sounds as well as the French projects Floris'Tic(INRIA, CIRAD, Tela Botanica) and SABIOD and EADM MaDICS.

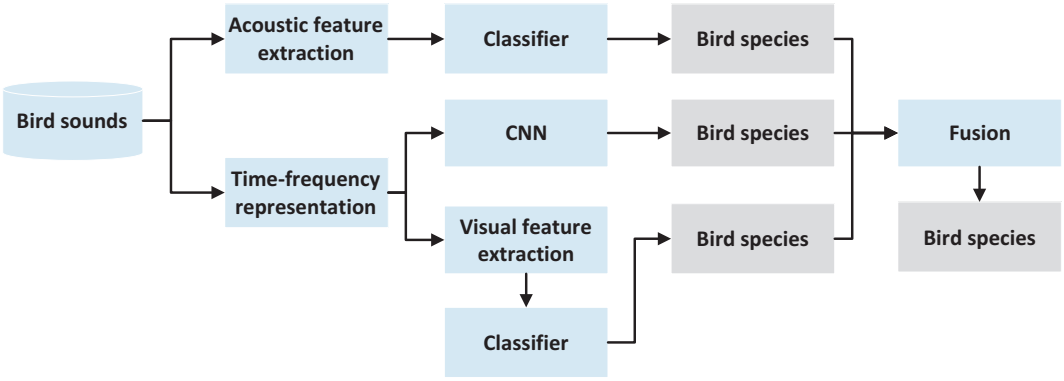


Fig. 1. The flow diagram of our proposed approach.

Table 1
The 14 bird species included in this study. Codes in the right column are abbreviations of common names used in this paper.

Species name	Common name	Code
Macropygia phasianella	Brown Cuckoo-dove	BCD
Lichmera indistincta	Brown Honeyeater	BHE
Burhinus grallarius	Bush Stone-curlew	BSC
Psophodes olivaceus	Eastern Whipbird	EWB
Eopsaltria australis	Eastern Yellow Robin	EYR
Rhipidura albiscapa	Grey Fantail	GFT
Trichoglossus moluccanus	Rainbow Lorikeet	RLK
Pachycephala rufiventris	Rufous Whistler	RFW
Chrysococcyx lucidus	Shining Bronze Cuckoo	SBC
Zosterops lateralis	Silvereye	SVE
Pardalotus striatus	Striated Pardalote	SPD
Cacatua galerita	Sulphur-crested Cockatoo	SCC
Corvus orru	Torresian Crow	TRC
Melithreptus albogularis	White-throated Honeyeater	WTH

$$Valley_k = log\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} x'_{k,N-i+1} \tag{7}$$

Finally, spectral contrast is defined as

$$SC_k = Peak_k - Valley_k \tag{8}$$

where N is the total number in k-th sub-band.

(4) Spectral flatness provides a way to quantify the tonality of a sound. A higher spectral flatness indicates a similar amount of power of the spectrum in all spectral bands. Spectral flatness is measured by the ratio between the geometric mean and the arithmetic mean of the power spectrum and defined as

$$Sf = \frac{\sqrt{\frac{1}{N}\sum_{k=0}^{N-1} InX(k)}}{\frac{1}{N}\sum_{k=0}^{N-1} X(k)} \tag{9}$$

(5) Spectral roll-off is often used to measure the spectral shape, and defined as the frequency H. Here H is the value below which θ of the magnitude distribution is concentrated.

$$\sum_{k}^{H} X(k) = \theta \sum_{k=1}^{N-1} X(k) \tag{10}$$

where θ is set to 0.8.

(6) Zero-crossing rate denotes the rate of signal change along a signal. When adjacent signals have different signs, a zero-crossing occurs. The mathematical expression of zcr is shown as follows.

$$zcr = \frac{1}{2}\sum_{n=0}^{L-1} [sgn(x(n)) - sgn(x(n+1))] \tag{11}$$

where $x(n)$ is the framed signal, L is the length of the frame.

(7) The energy of a signal corresponds to the total magnitude of the signal, which roughly corresponds to how loud the signal is. The root-mean-square energy (RMSE) in a signal is defined as

$$RMSE = \sqrt{\frac{1}{N}\sum_{n} |x(n)|^2} \tag{12}$$

(8) Mel-frequency Cepstral coefficients (MFCCs), which are obtained by applying discrete cosine transform to a sub-band Mel-frequency spectrum within a short time, have been widely used in speech/speaker recognition (Rakotomamonjy and Gasso, 2015). Here, MFCCs are calculated based on the method of (Salamon et al., 2017).

Step 1: Band-pass filtering: The amplitude spectrum is filtered using a set of triangular band-pass filters.

$$E_j = \sum_{k=0}^{N/2-1} \phi_j(k)A_k,\ 0 \le j \le J-1 \tag{13}$$

where J is the number of filters, ϕ_j is the j^{th} filter, and A_k is the amplitude of $X(k)$.

$$A_k = |X(k)|^2,\ 0 \le k \le N/2 \tag{14}$$

Step 2: Discrete cosine transform: MFCCs for the i^{th} frame are computed by performing DCT on the logarithm of E_j.

$$C_m^j = \sum_{j=0}^{J-1} \cos\left(m\frac{\pi}{J}(j+0.5)\right) log_{10}(E_j),\ 0 \le m \le L-1 \tag{15}$$

where L is the number of MFCCs. The filter bank consists of 40 triangular filters, that is $J = 40$. The length of MFCCs of each frame is 12 (L = 12).

2.3.2. Visual features

(1) Local binary pattern.

LBP with circular neighborhoods allowing any radius and number of neighbors is introduced by (Selin et al., 2007) for dealing with the structure at various scales. For LBP, each pixel of an image is calculated by comparing each central pixel, g_c, with its neighboring pixels, g_p. Here the distance between central pixel and P neighboring pixel is denoted by R. The calculation of LBP is shown as

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \tag{16}$$

where

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{17}$$

Since an increased number of P will lead to a high cost of dimensionality, (Selin et al., 2007) also introduced uniform patterns, $LBP_{P,R}^{u2}$, rotation invariant, $LBP_{P,R}^{ri}$ and rotation invariant uniform patterns, $LBP_{P,R}^{riu2}$ for achieving effective and robust texture features. Rotation invariant, $LBP_{P,R}^{ri}$ and rotation invariant uniform patterns, $LBP_{P,R}^{riu2}$ are proposed to deal with a rotated textured input image, which can translate LBP patterns into a different location and to rotate them about their origin. This characteristic is suitable for texture classification, where the pattern value depends on the illumination, translation and rotational variance of the texture. However, it is unsuitable for acoustic scene classification without rotation variations or translations. Our spectrogram texture pattern is characterized by the intensity along frequency and time instances, which provides a uniform pattern representation. Therefore, uniform patterns, $LBP_{P,R}^{u2}$ are selected for our analysis. For $LBP_{P,R}^{u2}$, the value of P and R are chosen among [(8,1); (12,1); (12,2); (16,1)].

To capture more information, we divide the CQT representation into N linear zones of equal size, where LBP is calculated for each zone (Stowell and Plumbley, 2014). Then, an early fusion of those linear zones is used for combining LBP. Preliminary experiments indicate that the optimal value of N is 10.

(2) Histogram of oriented gradients.

HOG is used for emphasizing the fast spectral transitions between adjacent frames within an acoustic scene. Since HOG is calculated based on the histogram of gradient directions over the adjacent frequency bins in both temporal and frequency directions, the modulation of scene along the temporal axis can be captured.

The main steps for calculating HOG are listed as follows: (1) compute the gradient of the CQT representation; (2) compute angles of all pixel gradients; (3) split images into non-overlapping cells; (4) count the occurrence of gradient orientations in a given cell; (5) eventually normalize each cell histogram according to the histogram norm of neighboring cells. Following (Wimmer et al., 2013a), filtering and pooling are used for optimizing the HOG. Here the filter size is chosen among [1; 15], and the pooling size is chosen among [(1,64); (64,1)].

2.3.3. K-nearest neighbor

For the k-NN classifier, an object is classified to the majority class of its k nearest neighbors (Wimmer et al., 2013b). Specifically, in the training phase, bird feature vectors are stored with species labels. For the test phase, the simplest classification combination method is the voting method. The k closet vectors are selected for voting, then the classification for the input feature vector $f_{i,c}$ is assigned with the majority class.

The second classification combination method is to calculate the average distance between an input bird feature vector and k closest vectors. For example, the Euclidean distance between an input feature vector $f_{i,c}$ and one stored feature vector $f_{j,c}$ is calculated as

$$d(i,j) = \sqrt{\sum_{c=1}^{n} (f_{i,c} - f_{j,c})^2} \tag{18}$$

where i and j are indices of the feature vector, n means the dimension of the feature vector. Next, k nearest neighbors of feature vector i are selected based on the Euclidean distance for voting. If the following equation is satisfied

$$\frac{1}{k_1} \sum_{j \in s_1} d(i,j(s_1)) < \frac{1}{k_2} \sum_{j \in s_2} d(i,j(s_2)) \tag{19}$$

where $k = k_1 + k_2$, k_1 is the number of bird species s_1, k_2 is the number of bird species s_2. Here, the input feature vector i will be classified as bird species s_2.

The third classification combination method is to calculate the sum of similarity of k closest feature vectors. For a binary classification task with two classes: k_1 and k_2. If

$$\sum_{j \in s_1} d(i,j(s_1)) < \sum_{j \in s_2} d(i,j(s_2)) \tag{20}$$

Then the input feature vector i will be classified as belonging to class s_2. Following prior work (Xie et al., 2015; Xie et al., 2016), the distance function used for K-NN is the Euclidean function, and k is chosen among [1, 10, 20, 30, 40, 50].

2.3.4. Random forest

Random forest (RF) is a tree-based algorithm, which builds a specified number of classification trees without pruning. The nodes are split on a random drawing of m features from the entire feature set M. A bootstrapped random sample from the training set is used to build each tree. The advantage of RF is its ability to generate a metric to rank predictors based on their relative contribution to the model's predictive accuracy (Zhao et al., 2017). The prediction is defined as follows.

$$Pred = \frac{1}{K} \sum_{n=1}^{K} T_i \tag{21}$$

where T_i is the n-th tree response of the RF. In this work, the number of trees K is chosen among [100, 1000, 2000, 3000, 5000]. As for the predictor variables m, it is set at $\sqrt{N}$, where N is the feature dimension.

2.4. Deep learning for bird species classification

The CNN architecture consists of three convolutional layers. We use a receptive field of 5 × 5 followed by a max pooling operation for every convolutional layer. Rectified linear unit (ReLU) is used as an activation function. Dropout is employed in convolutional layers with rate 0.3 to address over-fitting. The CNN is optimized by employing back-propagation algorithm. Table 2 depicts CNN architecture used in this study. Since the duration of various bird species is different, the input size of CNN is resized to 500 × 128.

3. Late fusion

To further improve the classification performance, we apply a class-based late fusion method. Let's assume, the fusion of decisions from n models for a m-class problem. The sets of models and classes can be presented as $M = M_1, M_2, \ldots, M_n$ and $C = C_1, C_2, \ldots, C_m$. When classifying a test instance x, each model provides a predicted class label along with a posterior probability of the predicted label, which is a measure of the confidence of the decision from that model for that test

Table 2
CNN architecture. The data shape indicates time × frequency × number of filters.

Layer(type)	Filter/Stride	Output Shape	Param #
Conv1	5 × 5/2 × 2	64 × 250 × 128	3328
MaxPool1	5 × 5	32 × 125 × 128	
Conv2	5 × 5/2 × 2	16 × 63 × 128	400 K
MaxPool2	5 × 5	8 × 31 × 128	
Conv3	5 × 5/2 × 2	4 × 16 × 128	400 K
MaxPool3	5 × 5	2 × 8 × 128	
FC1		256	524 K
FC2		14	3598
Total			1351 K

instance. Let the predicted vector for that instance be $V(x) = V_1(x)$, $V_2(x), \ldots, Vn(x)$, where each $Vi(x) \in C$, and the posterior probabilities be $W_2(x) = W_{21}(x), W_{22}(x), \ldots, W_{2n}(x)$. A decision fusion technique provides a final prediction for x by combining individual predictions $V(x)$.

Our class-based fusion scheme considers the class-based weights $W_1(x)$ and current prediction vector $V(x)$ to make a final prediction for a test instance (x). This method calculates score for each class using following formula.

$$Score_k = \sum_{V_i(x)=C_k} (W_{1ik}),\ 1 \le k \le m,\ 1 \le i \le n \tag{22}$$

Finally, it selects the class label as final prediction, which has maximum score using Eq. (22).

$$Final_{label} = C_{argmax_{k=1}^{m}\, Score_k} \tag{23}$$

4. Evaluation rule

In this experiment, the dataset was divided into five folds. Four folds were used as the training data, and the rest was for testing. The performance of our proposed bird call classification system was evaluated using accuracy and a weighted F1-score which are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{24}$$

$$F1 - score = \sum_{i=1}^{n} 2 \cdot \frac{precision(i) \cdot recall(i)}{precision(i) + recall(i)} * r_i \tag{25}$$

where $F1 - score$ denotes the weighted F1-score, *precision* is defined as $\frac{TP}{TP+FP}$, and *recall* is defined as $\frac{TP}{TP+FN}$, *TP* is true positive, *TN* is true negative, *FP* is false positive, *FN* is false negative; *i* is the class index, r_i is the ratio between the number of samples of one class and total number of samples in all classes.

5. Results

In this study, we evaluate various feature sets for bird sound classification. First, we compare the classification performance of acoustic features with various settings (Table 3). The feature dimension of acoustic features is 217. The *K* value for K-NN is 1 and the number of trees for RF is 5000. We can find that the best performance using acoustic feature is 88.97%, which is obtained with a window size of 256 samples and K-NN. For different window sizes, the classification performance is similar.

Secondly, we compare various visual feature sets of different settings. Table 4 shows the F1-score with different CQT settings for extracting LBP and HOG. Since K-NN achieves the better performance than RF using acoustic features, we further combine K-NN and visual feature sets for the classification. The feature dimension of visual features is 2126. The *K* value for K-NN is 1. Here, we compare the *B* value for both LBP and HPG. The best F1-score is 88.2%, when the value *B* is set to 96 and 32 for LBP and HOG, respectively. Here, *B* is used to determine the resolution of CQT representation. Therefore, a combination of LBP with a higher resolution CQT and HOG with a lower resolution CQT has the best feature discriminability. In addition, the filter size and pooling size for HOG is 15 and [64,1], respectively. The best averaged F1-score is 86.21%. Furthermore, we compare the setting of LBP as shown in Table 5. We find that the best F1-score is obtained when the LBP is set to (8,2) and the *B* value for HOG is 32.

Lastly, we report the classification results with deep learning. Since the duration of those recordings are different, all those representations are first resized to [500, 40]. Then, the resized spectrogram are used as an input to a designed CNN structure (see Table 2). Our best classification F1-score is 94.36%. Compared to acoustic and visual feature based classification, deep learning can achieve a better performance.

To fully understand the classification results, we plot the averaged confusion matrix of those three classification framework (Fig. 4). It is observed that seven *Eastern Yellow Robin* (EYR) samples are confused with *Striated Pardalote* (SPD) using the acoustic features approach. Fig. 2 shows the CQT of those two species. We can find that acoustic features do not capture the discriminant information. However, the visual representations of those two species are different, which explains the good performance using the visual features approach and deep learning. For the visual features approach, four *Shining Bronze Cuckoo* (SBC) samples are misclassified as *Grey Fantail* (GFT). Fig. 3 shows the time-frequency representations of *Shining Bronze Cuckoo* and *Grey Fantail*. We can find that the visual patterns of those two species are very similar.

Table 3
Comparison of acoustic features with different window sizes using K-NN and RF. Here, the window overlap is 50% for all sizes.

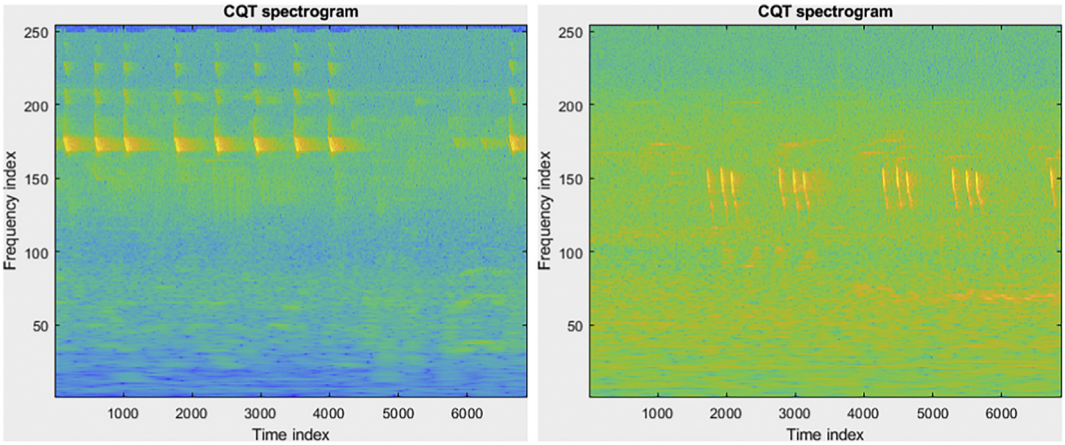
	Window size (sample)			
	256	512	1024	2048
K-NN	**88.97**	88.33	88.32	88.25
RF	88.93	88.74	88.90	88.76

Table 4
Comparison of visual features with different CQT settings: B value, filter size, and pooling size.

Filter size						Mean
		1	15	1	15	
B		Pooling size				
LBP	HOG	[64,1]	[64,1]	[64,1]	[64,1]	
32	32	84.51	88.00	87.04	84.85	86.10
32	96	84.19	86.59	85.11	81.78	84.42
32	128	83.71	85.86	84.35	80.85	83.69
96	32	84.85	**88.20**	87.13	84.66	**86.21**
96	96	84.33	86.60	85.16	81.84	84.48
96	128	84.23	86.48	84.54	81.93	84.30
128	32	84.83	88.23	87.41	84.64	86.28
128	96	83.78	86.51	84.98	81.92	84.30
128	128	83.64	86.04	84.29	81.22	83.80

Table 5
Comparison of visual features with different CQT settings. Here, the window overlap is 50% for all window sizes, the LBP setting is (12,2).

Filter size						Mean
		1	15	1	15	
		Pooling size				
LBP	B(HOG)	[64,1]	[64,1]	[64,1]	[64,1]	
(8,1)	32	84.53	88.53	86.76	83.98	85.95
(8,1)	96	83.49	86.56	84.71	81.18	83.98
(8,1)	128	83.11	86.15	83.76	79.72	83.19
(8,2)	32	85.21	**88.87**	87.38	84.58	**86.51**
(8,2)	96	84.47	87.26	84.96	81.76	84.62
(8,2)	128	84.07	86.50	84.29	80.39	83.81
(12,2)	32	84.86	88.21	87.14	84.66	86.22
(12,2)	96	84.33	86.60	85.17	81.85	84.49
(12,2)	128	84.23	86.49	84.55	81.94	84.30
(16,1)	32	82.24	87.13	85.70	83.86	84.73
(16,1)	96	80.70	84.94	83.91	80.86	82.60
(16,1)	128	80.47	85.09	83.39	79.87	82.21
(16,2)	32	83.99	87.86	86.48	84.84	85.79
(16,2)	96	83.52	85.88	84.94	81.88	84.05
(16,2)	128	82.99	85.74	84.41	81.74	83.72



(a) Eastern Yellow Robin (b) Striated Pardalote

Fig. 2. Visual representation of *Eastern Yellow Robin* and *Striated Pardalote*. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

We further investigate a class-based late fusion method to improve the classification performance. The final classification result with late fusion is 95.95%, which is 1.59% higher than the deep learning method. The class-wise accuracy using four methods is shown in Fig. 5. We can find that the classification accuracy of all species using fusion is higher than 90% except for the *Striated Pardalote* (SPD). For *Striated Pardalote*, the classification accuracy is lower than 80%. Based on the late fusion of acoustic feature approach, visual feature approach, and deep learning, the fusion method achieves the highest classification accuracy for BHE, EWB, EYR, GFT, RLK, SVE, TRC.

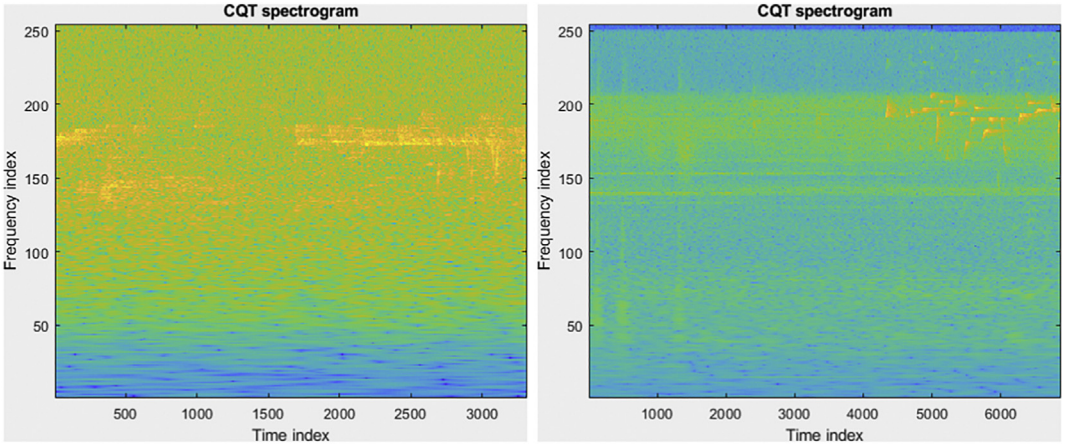
In previous studies, various feature sets have been investigated for bird sound classification, where MFCCs are widely used as the baseline for the comparison. Here, we use the statistical values of MFCCs (140) and K-NN as the baseline. The best classification F1-score is 73.8%, when the window size and overlap are 256 samples and 50%.

The statistical significance of the results is shown in Table 6. The classification F1-score of *fusion* is significantly higher than fusion2, CNN, acoustic, and visual approaches. Here, *fusion* denotes the integration of acoustic, visual feature approaches and deep learning; *fusion2* is fused based on the acoustic feature approach and deep learning.

6. Discussion and conclusion

In this study, three classification frameworks are presented to classify bird sounds: an acoustic features approach, a visual features approach, and a deep learning based method. Based on the experimental results, we find that a late fusion of the acoustic features approach, the visual features approach and deep learning achieves the best classification performance. We investigate the processes of extracting acoustic and visual features for bird sound classification with various settings. We also use deep learning for the classification of bird sounds. A late fusion process is further used to improve the classification performance. Since our used dataset is small, the performance with handcrafted features can be up to 88.97%. However, the deep learning method still achieves higher F1-score, which indicates the usefulness of deep learning for studying bird sounds.

For acoustic features, different window sizes achieve similar classification performance, and K-NN is more suitable for classifying bird sounds than RF using acoustic features. For visual features, both pooling and filter sizes are investigated for generating different audio images, and the best filter and pooling sizes are 15 and [64,1], respectively. The best classification F1-score with visual features is 88.87%. For LBP, the best performance is obtained when the value of P and R are 12 and 2, respectively. When the pooling and filter sizes are set to [64,1] and 15, HOG gives the best result. The best classification F1-score is 95.95% by fusing the acoustic features approach, the visual features approach, and deep learning.



(a) Shining Bronze Cuckoo (b) Grey Fantail

Fig. 3. Visual representation of *Shining Bronze Cuckoo* and *Grey Fantail*.

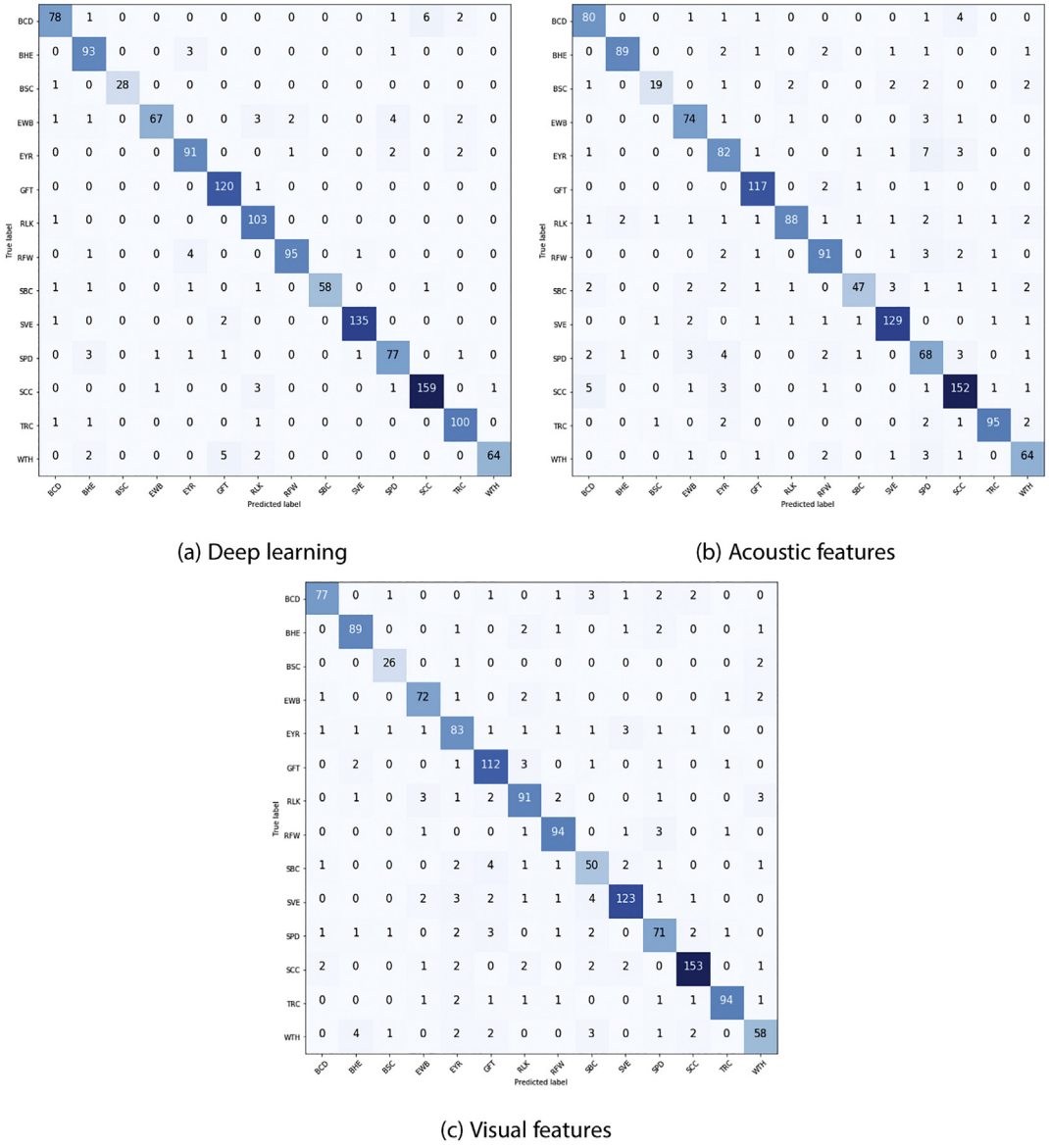


Fig. 4. Sum of confusion matrix of the best result for deep learning, acoustic features, and visual features. Here, the *x* and *y* axis denote the code of each bird species.

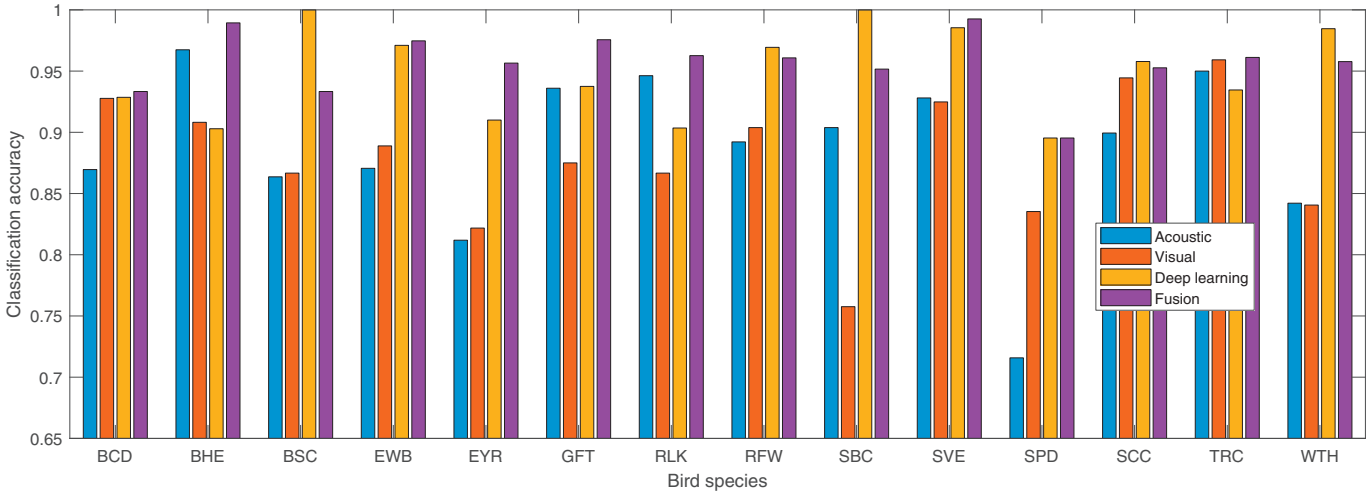


Fig. 5. Class-wise accuracy comparison of 14 bird species with four different classification methods. Here, the fusion is realized by acoustic feature approach, visual feature approach, and deep learning.

Table 6
Paired statistical analysis of the classification results. For the classification F1-score of each samples, the paired Student *t*-test was conducted.

Pairs	Test results
fusion-fusion2	$t = 1.60$ ($p < .05$, df = 1342)
fusion2-CNN	$t = 1.71$ ($p < .05$, df = 1342)
CNN-acoustic	$t = 5.61$ ($p < .01$, df = 1342)
acoustic-visual	$t = 2.67$ ($p < .01$, df = 1342)

Acknowledgement

This work is supported by the 111 Project. This work is also supported by "the Fundamental Research Funds for the Central Universities JUSRP11924" and Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment & Technology FM-2019-06.

References

Acevedo, M.A., Corrada-Bravo, C.J., Corrada-Bravo, H., Villanueva-Rivera, L.J., Aide, T.M., 2009. Automated classification of bird and amphibian calls using machine learning: a comparison of methods. Ecol. Informa. 4 (4), 206–214.
Bao, L., Cui, Y., 2005. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. Bioinformatics 21 (10), 2185–2190.
Bardeli, R., Wolff, D., Kurth, F., Koch, M., Tauchert, K.-H., Frommolt, K.-H., 2010. Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. Pattern Recogn. Lett. 31 (12), 1524–1534.
Brandes, T.S., 2008. Automated sound recording and analysis techniques for bird surveys and conservation. Bird Conserv. Int. 18 (S1), S163–S173.
Briggs, F., Lakshminarayanan, B., Neal, L., Fern, X.Z., Raich, R., Hadley, S.J., Hadley, A.S., Betts, M.G., 2012. Acoustic classification of multiple simultaneous bird species: a multi-instance multi-label approach. J. Acoust. Soc. Am. 131 (6), 4640–4650.
Brown, J.C., 1991. Calculation of a constant q spectral transform. J. Acoust. Soc. Am. 89 (1), 425–434.
Costa, Y.M., Oliveira, L., Koerich, A.L., Gouyon, F., Martins, J., 2012. Music genre classification using lbp textural features. Signal Process. 92 (11), 2723–2737.
Farina, A., Gage, S.H., 2017. Ecoacoustics: The Ecological Role of Sounds. Wiley, Hoboken, NJ, USA.
Gregory, R.D., van Strien, A., 2010. Wild bird indicators: using composite population trends of birds as measures of environmental health. Ornithol. Sci. 9 (1), 3–22.
Han, N.C., Muniandy, S.V., Dayou, J., 2011. Acoustic classification of australian anurans based on hybrid spectral-entropy approach. Appl. Acoust. 72 (9), 639–645.
Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al., 2012. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Mag. 29 (6), 82–97.
Hirsch, H.-G., Pearce, D., 2000. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. ASR2000-Automatic Speech Recognition: Challenges for the New Millenium ISCA Tutorial and Research Workshop (ITRW).
Huang, C.-J., Yang, Y.-J., Yang, D.-X., Chen, Y.-J., 2009. Frog classification using machine learning techniques. Expert Syst. Appl. 36 (2), 3737–3743.
Jiang, D.-N., Lu, L., Zhang, H.-J., Tao, J.-H., Cai, L.-H., 2002. Music type classification by spectral contrast feature. In: Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on. vol. 1. pp. 113–116 IEEE.
Kahl, S., Wilhelm-Stein, T., Hussein, H., Klinck, H., Kowerko, D., Ritter, M., Eibl, M., 2017. Large-Scale Bird Sound Classification Using Convolutional Neural Networks. (Working notes of CLEF).
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105.
Lasseck, M., 2018. Audio-Based Bird Species Identification with Deep Convolutional Neural Networks, Working Notes of CLEF.
Lee, C.-H., Han, C.-C., Chuang, C.-C., 2008. Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients. IEEE Trans. Audio Speech Lang. Process. 16 (8), 1541–1550.
Lee, H., Pham, P., Largman, Y., Ng, A.Y., 2009. Unsupervised feature learning for audio classification using convolutional deep belief networks. In: Advances in Neural Information Processing Systems, pp. 1096–1104.
Ojala, T., Pietikainen, M., Maenpaa, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. 24 (7), 971–987.
Pereira, H.M., Cooper, H.D., 2006. Towards the global monitoring of biodiversity change. Trends Ecol. Evol. 21 (3), 123–129.
Qian, K., Zhang, Z., Baird, A., Schuller, B., 2017. Active learning for bird sound classification via a kernel-based extreme learning machine. J. Acoust. Soc. Am. 142 (4), 1796–1804.
Rakotomamonjy, A., Gasso, G., 2015. Histogram of gradients of time-frequency representations for audio scene classification. IEEE/ACM Trans. Audio Speech Language Process. 23 (1), 142–153 TASLP).
Salamon, J., Bello, J.P., Farnsworth, A., Kelling, S., 2017. Fusing shallow and deep learning for bioacoustic bird species classification. In: Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on, IEEE, pp. 141–145.
Selin, A., Turunen, J., Tanttu, J.T., 2007. Wavelets in recognition of bird sounds. EURASIP J. Appl. Signal Process. 2007 (1), 051806.
Stowell, D., Plumbley, M.D., 2014. Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. PeerJ 2, e488.
Wimmer, J., Towsey, M., Planitz, B., Williamson, I., Roe, P., 2013a. Analysing environmental acoustic data through collaboration and automation. Futur. Gener. Comput. Syst. 29 (2), 560–568.
Wimmer, J., Towsey, M., Roe, P., Williamson, I., 2013b. Sampling environmental acoustic recordings to determine bird species richness. Ecol. Appl. 23 (6), 1419–1428.
Xie, J., Towsey, M., Truskinger, A., Eichinski, P., Zhang, J., Roe, P., 2015. Acoustic classification of australian anurans using syllable features. 2015 IEEE Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing, Singapore, Singapore.
Xie, J., Towsey, M., Zhang, J., Roe, P., 2016. Adaptive frequency scaled wavelet packet decomposition for frog call classification. Ecol. Informa. 32, 134–144.
Zhao, Z., Zhang, S.-h., Xu, Z.-y., Bellisario, K., Dai, N.-h., Omrani, H., Pijanowski, B.C., 2017. Automated bird acoustic event detection and robust species classification. Ecol. Informa. 39, 99–108.