Compact features for birdcall retrieval from environmental acoustic recordings

Xueyan Dong[1], Michael Towsey[2], Jinglan Zhang[2], Paul Roe[2]
School of Electrical Engineering and Computer Science
Queensland University of Technology (QUT)
Brisbane, Australia
[1]xueyan.dong@student.qut.edu.au
[2]m.towsey, jinglan.zhang, p.roe@qut.edu.au

***Abstract*—Bioacoustic data can be used for monitoring animal species diversity. The deployment of acoustic sensors enables acoustic monitoring at large temporal and spatial scales. We describe a content-based birdcall retrieval algorithm for the exploration of large data bases of acoustic recordings. In the algorithm, an event-based searching scheme and compact features are developed. In detail, ridge events are detected from audio files using event detection on spectral ridges. Then event alignment is used to search through audio files to locate candidate instances. A similarity measure is then applied to dimension-reduced spectral ridge feature vectors. The event-based searching method processes a smaller list of instances for faster retrieval. The experimental results demonstrate that our features achieve better success rate than existing methods and the feature dimension is greatly reduced.**

Keywords—birdcall retrieval; environmental acoustic recordings; spectral ridge; event detection; histrogram of oriented ridges

I. INTRODUCTION

Acoustic sensors provide a useful means to estimate bird species richness [1] by collecting avian sounds. The unattended acoustic sensors can operate continuously for several weeks in a wild field, such as an open forest. The collected bioacoustic data can be accumulated up to several terabytes of recordings which is equivalent to multiple years of audio data. It is a time-consuming task for ecologists to listen through all acoustic recordings or visually scan spectrograms (representation of audio signals). Automated analysis is required to assist the analysis of acoustic recordings to some extent [2].

Bioacoustic recordings can provide a persistent and verifiable record of the acoustic soundscape [3]. In terms of species level, many research efforts have been made on automated call detection and species recognition [4, 5]. There are two types of important tasks in birdsong recognition: call classification and call retrieval. In the former task, a classifier is trained to recognize a limited number of species or certain call types. The trained classifier will assign a label to a test instance. Such a classification system is constrained by the training species/call classes. In contrast, a retrieval system is able to response to an arbitrary query sound. Most importantly, the search database is allowed to contain multiple species rather than fixed classes.

To better explore the animal sounds, many acoustic libraries are established to store fauna sound archives and provide useful tools for processing and analyzing the bioacoustic data [6, 7]. In these platforms, navigating recordings in the library is an important function. One approach is to utilize the metadata describing the audio recordings, like species name, recording location and date [8]. Annotating a large volume of recordings is a laborious task which requires good knowledge of animal voices. There is an urgent need for automatically analyzing recordings by extracting higher level features [7]. An option is to conduct a content-based retrieval system by providing an audio clip as a query. The system will search through the audio files to find similar calls to the query in terms of particular features.

This paper aims to establish a birdcall retrieval system on environmental acoustic recordings where multiple species, geophony (wind and rain) and anthrophony (plane and motorbikes) are contained. In the implementation, two contributions are made: 1). an event-based search scheme is present to locate potential candidates from the searching recordings; 2). to characterize birdcalls using high level information, a low dimensional descriptor derived from spectral ridges is developed.

II. RELATED WORK

Generally, a content-based audio retrieval system contains three major procedures: segmentation, feature extraction and similarity measure. In this section, we discuss the methods on segmentation and features used in previous studies.

A typical birdsong consists of single or multiple syllables. Syllable segmentation is an important step in automated species classification. It can be achieved in time domain [9, 10] based on an assumption that syllables tend to have high energy values during a short interval (usually 20-30 ms). However, Briggs et al. [11] pointed out that energy-based algorithms are not suitable for recordings showing high noise and containing calls overlapping in time. To overcome the drawbacks, they employed a random forest classifier for syllable segmentation in time-frequency domain. This segmentation method can be effective for simultaneous birdcalls. However, as indicated in their conclusion, two difficulties are found when dealing with noisy recordings: (1). it is hard to obtain fully annotated segments for training the classifier. (2). high noise results in

inaccurate segmentation by detecting noise as syllables. Retrieval from large audio collections can utilize metadata (date, location and other annotations) attached to isolate recording segments [8] or it can perform a similarity-search using point matching on the acoustic content of the recordings. [7]. Most existing segmentation strategies were investigated for classifying birdsongs containing either individual syllable or single species, more work is required for real-world recordings consisting of multiple species and overlapping calls [12].

After segmentation, discriminative features are extracted to form a feature vector characterizing the segments so as to differentiate various syllables. Mel-frequency cepstral coefficients (MFCCs) are widely used features for describing birdsongs [13]. They are derived from individual frame of bird sound. To obtain a compact feature representation based on MFCCs, Lee et al. [14] averaged the frame features during a syllable of birdcall. However, MFCCs feature are not optimum for birdsong analysis as they may lose pitch information [10]. Since many birdcall exhibit modulated tonal structure, sinusoids models become a common approach in birdsong recognition [10, 15, 16]. Recently, many researchers began to use image processing techniques to animal sound analysis [7, 17]. Eight directional histogram of gradients (HOG) are calculated as partial features in bird species classification [11] but the contribution of HOG features to the final result is not investigated. Fourier transform is used to explore the shape property of birdcalls in spectrograms. The features are compressed into 12-bit for improving the retrieval speed in large audio collection [7].

This summary of approaches demonstrates that existing segmentation approaches may not be suitable for detecting overlapping calls that are commonly found in environmental acoustic recordings. Furthermore, a low dimensional feature representation is needed for large scale audio retrieval.

III. METHOD

The proposed retrieval system has four steps: 1. each one-minute segment of audio recording is converted to a noise-reduced spectrogram; 2. detection of acoustic ridges in four direction categories; 3. event-based search for candidates; 4. ranking of candidates according to similarity with the query.

A. Preprocessing

The preprocessing step includes spectrogram generation and noise reduction. Spectrograms are generated using the short-time Fourier transform (STFT) with a window of 512 samples and 50% overlap. We denote spectral values by $X_{(t, f)}$, where t indexes a window and f represents a discrete frequency bin. A spectrogram has 256 frequency bins, each spanning 43.07 Hz. Spectral amplitude values are converted to decibels (dB) using $dB = 20\log_{10}(X)$.

To reduce background noise and improve acoustic contrast, we applied the noise removal algorithm developed by Towsey et al. [2] which calculates a separate decibel threshold for each frequency bin assuming an additive noise model. This algorithm adapts to variable noise levels at one minute resolution.

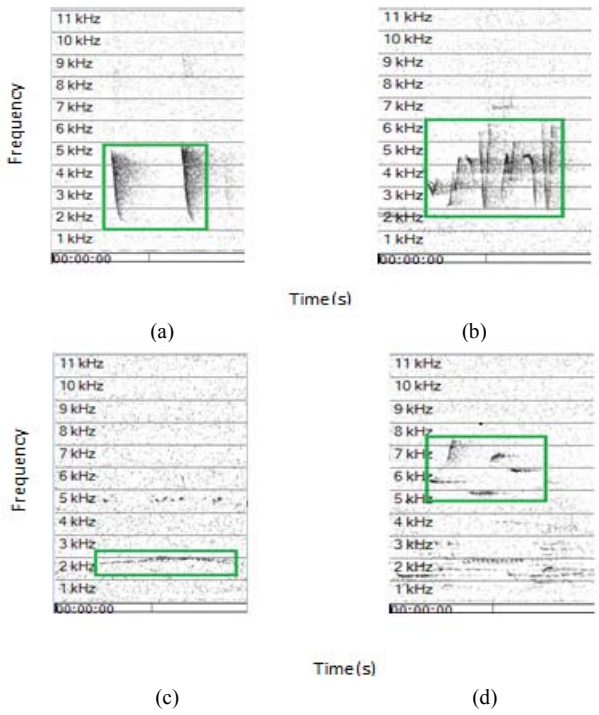


Fig.1. Example of birdcalls in the study

(a) Eastern yellow robin (two clicks), (b) Rufous whistler (chirps), (c) Bush Stone curlew (constant frequency), and (d) Grey Fantail (multiple acoustic elements). All spectrograms are 2 seconds duration.

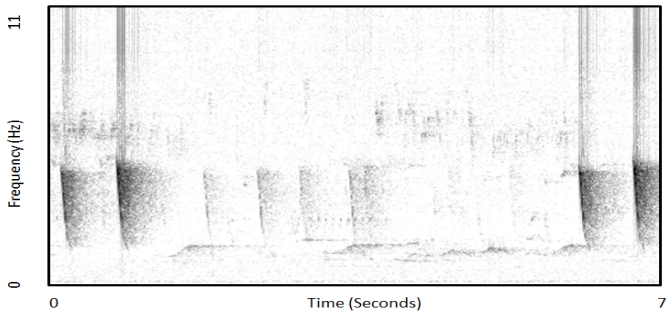
B. Ridge event detection

The syllables or basic components of many bird vocalisations appear in spectrogram images as ridges. *Ridge events* can therefore be used to identify bird calls. We detect spectral ridge events automatically in two steps: ridge detection and event detection.

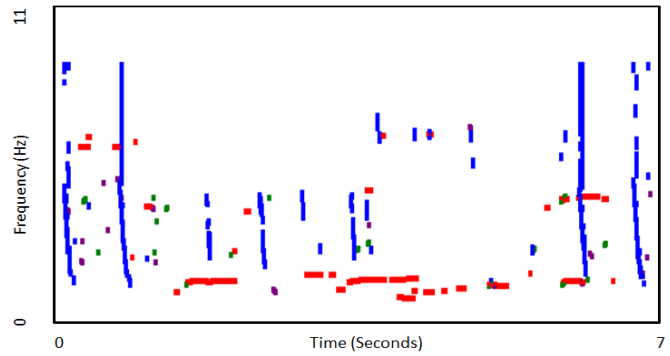
1) Spectral ridge detection: we implemented an existing ridge detection algorithm [18]. Specifically, spectral ridges are detected by convolving the spectrogram with 5×5 masks, one mask for each of four ridge orientations: 0, π/4, π/2, and 3π/4 radians. A spectrogram cell is assigned a ridge direction corresponding to the mask which yields maximum convolution score only if the score exceeds a threshold of 6 dB. Note that a constant frequency whistle (e.g. see Fig.1(c)) would show in the spectrogram as a ridge having direction 0 and a broadband click (Fig.1 (a)) would show as a ridge having direction π/2.

The ridge structure of a single syllable is often broken due to discontinuous intensity values in the spectrogram. Therefore, we smooth the ridge points by convolving with a 5×5 Gaussian kernel, σ = 1.4. Fig. 2.(b) shows the effect of smoothing – the ridges are marked with different colors to differentiate the four directions, blue- π/2, red-0, green- π/4, and purple-3π/4.

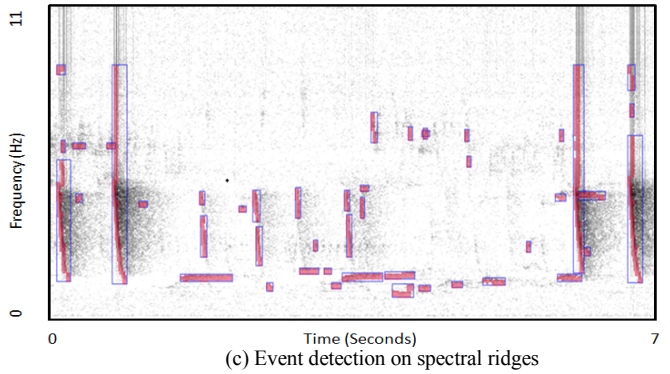
2) Event detection: after spectral ridge detection, an event detection algorithm is applied to automatically form ridge events. We apply event detection separately for each of the four ridge directions. The algorithm initiates an iterative spider



(a) Noise reduced spectrogram



(b) Smoothed spectral ridges



(c) Event detection on spectral ridges

Fig. 2. Spectral ridge event detection

search on the matrix of ridge points and generates a list of events. Each event, *e*, is parametrized as follows:

$$e = \{f_{max}, f_{min}, t_{s,} t_e, area, o\} \quad (1)$$

where f_{max} and f_{min} denote the maximum frequency (higher bound) and minimum frequency, t_s and t_e denote the time start and time end of the event, and *area* denotes the area of the event in spectrogram cells (or image pixels). Here *o* is the orientation of the ridges inside the event.

We set a small area threshold to remove small events. In our case, the threshold is set to 10% of the largest event in the one-minute spectrogram.

The final output of event detection is a list of events (Fig. 2. (c)). As can be seen in the figure, each event is bounded by a rectangle. Ridge events for 0 and π/2 directions are far more numerous than those for π/4 and 3π/4 directions.

C. *Searching for candidates*

Given a query birdcall, the procedure here searches through a one-minute spectrogram to locate matching candidates. For faster search, we proceed by event alignment rather than point matching as in Bardeli's work [7].

A *query* in this study is a *region* in the spectrogram defined by rectangle bounds that encloses all the syllables in a single call (see the examples in Fig 1). After event detection, the query is represented a set of events which fall within the query region. Among these events, we define the one with largest area as the *dominant* event. The dominant event is used as an anchor to align the query with potential candidates in the database. The query dominant event is aligned with each possible candidate event in a one minute recording. To limit the number of similarity calculations, we impose three conditions:

a) The candidate event must have same orientation as query anchor event, otherwise, move to the next candidate event.

b) If *fmin* of a candidate event is not within range *fmin*(query) ± 500Hz, then move to next candidate event.

c) The temporal center point of the anchor event is aligned with the temporal center of the candidate event. Calculate the fractional overlap ($area_{overlap}$) between the anchor and candidate events according to the equation (2), (3), and (4). If $area_{overlap}$ is < 0.5 move to next candidate event.

$$area_{overlap} = t_{overlap} \times f_{overlap} \quad (2)$$

where

$$t_{overlap} = Max(0, Min(c_{te}, d_{te}) - Max(c_{ts}, d_{ts}) \quad (3)$$

$$f_{overlap} = Max(0, Min(c_{fmax}, d_{fmax}) - d_{fmin} \quad (4)$$

If a candidate location survives these three tests, a candidate *region* is bounded corresponding to the bounds of the query.

D. *Feature extraction*

Since birdcalls have varied dimensions in frequency range and duration, finding a generalized feature set can be difficult. Therefore, we adapt the block descriptor technique described [19] for use in this study. The query and candidate call *regions* are divided into non-overlapping 11×11 blocks (Fig. 3). Block height is approximately 550 Hz and the width is 127 milliseconds. In a retrieval task, compact features are required to improve the search speed. To achieve this goal, we extract features that describe the distribution of ridge points in each column of blocks and each row of blocks.

First, to describe the distribution of ridge points in one row of blocks within a birdcall region, we use:

$$r_i = \{p_i, o_i\} \in Z^2 \quad (5)$$

where *i* is an index over rows in the birdcall region, p_i represents the fraction of row cells containing a ridge point and

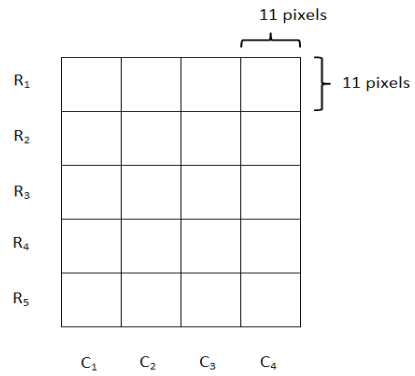


Fig. 3. A compact block descriptor for a simplified birdcall
A birdcall region (the outlier rectangle) is divided into non-overlapping 11×11 blocks, the features are calculated from each row (R) or col (C) rather than each block for compact representation.

o denotes the dominant orientation of the *i* th row. Likewise, column feature vectors are calculated the same way.

In the end, a birdcall region *R* is represented using (6):

$$R[f_{max}, t_s] = \{r_1 \cdots r_{Nr}, c_1, \cdots c_{Nc}\} \quad (6)$$

where the r_i are the features for the N_r rows and the c_i are the features for the N_c columns) in the region. N_r and N_c will of course vary with the size of the birdcall region but will be identical for query and candidate. The region is represented by $R[f_{max}, t_s]$ where f_{max} is the maximum frequency and t_s is the time start index.

Feature normalization is required prior to calculation of region similarity. The discrete ridge orientations (0, π/4, π/2, and 3π/4 radians) were 'normalized' as 0, +0.5, 1, and -0.5, respectively. The other features were normalized to the range of [0, 1] by (7).

$$p_i = \frac{p_i - \min(p_i)}{\max(p_i) - \min(p_i)} \quad (7)$$

E. Similarity score

Euclidean distance is applied to the feature vectors and yields a distance which is converted to a similarity score. The candidates are ordered by the similarity score and the top N matches displayed as the retrieval results.

IV. Results

Our algorithm was tested on 100 one-minute recordings. The recordings were manually selected from 20 days of recordings collected in an open forest from 13-17 th of October, 2010. Most of the test audio files were collected at dawn and contain target species as well as other bird species and other sound sources. For query selection, 20 species are chosen, each of which has five representative calls. The queries in the experiment were not taken from any of the test audio files.

To evaluate the retrieval performance, the birdcalls of interest in the test recordings were annotated with enclosing marquee (denoting the temporal and spectral bounds of the call) and a label made by experienced birders as well as the audio file name.

Given 100 queries (20 birdcall classes×5 individual calls), our system will find similar birdcalls to the query. The returned calls are ranked in a descending order according to the similarity score. An item in the returned list includes the boundary information of the call and audio file name. To check the retrieval results, the annotation data is utilized to check the match by comparing the boundaries and document name with ground truth data.

A. Experimental results

Rank of first correct match is a useful measure for evaluating a retrieval system. The best case is that the queries obtain correct retrieval at top one. Table 1 shows the results of average of first rank per class from 20 query call classes.

For baseline methods, we investigated two histogram of oriented ridges (HOR) based approaches. One approach is a global descriptor, which utilizes the overall distribution of four directional ridges to describe a birdcall region on the spectrogram. The descriptor counts the occurrences of oriented ridges in a birdcall region, and its feature dimension is fixed to four no matter what size of a birdcall region is.

TABLE 1. AVERAGE RANK FOR 20 QUERY CALL CLASSES

Query class	Query Species scientific name	Species common name	Av. rank of first rank
1	*Macropygia amboinensis*	Brown Cuckoo-dove	1.6
2	*Lichmera indistincta*	Brown Honeyeater	2.2
3	*Burhinus grallarius*	Bush Stone-curlew	2.4
4	*Psophodes olivaceus*	Eastern Whipbird	1.0
5	*Eopsaltria australis*	Eastern Yellow Robin	1.8
6	*Rhipidura albiscapa*	Grey Fantail	1.4
7	*Colluricincla harmonica*	Grey Shrike-thrush	1.2
8	*Pachycephala pectorails*	Golden Whistler	3.2
9	*Myiagra rubecula*	Leaden Flycatcher	4.4
10	*Oriolus sagittatus*	Olive-backed Oriole	2.4
11	*Pachycephala rufiventris*	Rufous Whistler	2.4
12	*Trichoglossus haematodus*	Rainbow Lorikeet	6.2
13	*Chrysococcyx lucidus*	Shining Bronze-cuckoo	1.2
14	*Cacatua galerita*	Sulphur-crested Cockatoo	1.8
15	*Zosterops laterails*	Silvereye	4.8
16	*Myzonmela sanguinolenta*	Scarlet Honeyeater (call)	1.6
17	*Myzonmela sanguinolenta*	Scarlet Honeyeater (song)	2.4
18	*Pardalotus striatus*	Striated Pardalote	1.2
19	*Corvus orru*	Torresian Crow	1.2
20	*Melithreptus albogularis*	White-throated Honeyeater	2.2
Average	-	-	**2.28±1.40**

TABLE 2. RETRIEVAL PERFORMANCE ON DIFFERENT FEATURE SETS

Method	Top-5 SR	Top-10 SR	Average number of candidates from 1-min file	Average number of blocks per region	Feature dimension per block
Global HOR	45%	62%	300	1	4
Local HOR [20]	81%	94%	100	48	6
Compact Local HOR	87%	100%	30	16	2

The other approach is a block-based local descriptor designed by [20]. In the method, a birdcall region is divided into non-overlapping blocks each of which is represented by a 6 dimensional (6-d) feature vector. In [20] a block is referred to as a neighborhood. A neighborhood feature vector consists of the four bins of HOR (4-d) and the frequency entropy (1-d) and frame entropy (1-d). In the end, a list of 6-d feature vectors is concatenated to describe the birdcall. This approach aims to capture the local property rather than the global one in the birdcall. Notice that the feature length is varied depending on the birdcall.

We compare our features with these two methods. The results in terms of top-5 and top-10 retrieval success rate and feature dimension are shown in Table 2. The top-N success rate (SR) [21] refers to the percentage of all queries (100 in this case) in which a correct match was returned within the first N returned instances. It is a single number which can evaluate the overall retrieval performance for multiple queries.

B. Discussion

Table 1 illustrates that our method achieves an average first rank for 20 call classes of less than 4 except for three call classes, *Leaden Flycatcher*, *Rainbow Lorikeet* and *Silvereye*. We investigated the reason for poor performance in retrieving silvereye is that the call varied a lot, and one component in silvereye sometimes is lost. And they are often overlapping with other calls. The poor performance for *Leaden Flycatcher* is that it has many variations in the syllables. The duration of *Rainbow lorikeet* can be varied. So the ridge detection tend to yield vertical lines or broken.

Table 2 compares the retrieval results of three methods. In particular, our method obtains the best result, 0.87 top-5 success rate. In contrast with global descriptor, local descriptor performs better. In terms of feature dimension, while the global histogram method yields a 4-d feature, it is not sufficient to differentiate the birdcall classes, which is shown as the low success rate. Compared to the local descriptor with a 48×6 feature, our method (Compact local HOR) is more compact with a 32-d feature set and the top-5 and top-10 success rate are also improved. Remember that the local descriptor and our method yields variable feature vector length and the values shown in the table is calculated based on the average value over all birdcalls (100 queries) in the study.

V. CONCLUSION

This paper presents a compact feature representation for birdcall retrieval from environmental recordings. The features are derived from spectral ridges and perform well in characterizing birdcalls. Event detection is before feature extraction so as to improve the search speed. The established retrieval system has a potential ability to assist the species presence or absence study. Due to limited annotation data, the retrieval algorithm is tested on a small dataset. In the future, we will investigate its application to large audio recordings.

REFERENCES

[1] J. Wimmer, M. Towsey, P. Roe, and I. Williamson, "Sampling environmental acoustic recordings to determine bird species richness," Ecological Applications, 2013.

[2] M. Towsey, J. Wimmer, I. Williamson, and P. Roe, "The use of acoustic indices to determine avian species richness in audio-recordings of the environment," Ecological Informatics, vol. 21, pp. 110-119, 2014.

[3] S. H. Gage and A. C. Axel, "Visualization of temporal change in soundscape power of a Michigan lake habitat over a 4-year period," Ecological Informatics, vol. 21, pp. 100-109, 2014.

[4] I. Agranat, "Automatically identifying animal species from their vocalizations," in Fifth International Conference on Bio-Acoustics, Holywell Park, 2009.

[5] N. Giret, P. Roy, and A. Albert, "Finding good acoustic features for parrot vocalizations: The feature generation approach," Journal of the Acoustical Society of America, vol. 129, pp. 1089-1099, 2011.

[6] J. Wimmer, M. Towsey, B. Planitz, I. Williamson, and P. Roe, "Analysing environmental acoustic data through collaboration and automation," Future Generation Computer Systems, vol. 29, pp. 560-568, 2013.

[7] R. Bardeli, "Similarity search in animal sound databases," IEEE Transactions on Multimedia vol. 11, pp. 68-76, 2009.

[8] E. P. Kasten, S. H. Gage, J. Fox, and W. Joo, "The remote environmental assessment laboratory's acoustic library: An archive for studying soundscape ecology," Ecological Informatics, vol. 12, pp. 50-67, 2012.

[9] S. Fagerlund, "Bird Species Recognition Using Support Vector Machines," EURASIP Journal on Advances in Signal Processing, vol. 2007, 2007.

[10] P. Somervuo, A. Harma, and S. Fagerlund, "Parametric representations of bird sounds for automatic species recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, pp. 2252-2263, 2006.

[11] F. Briggs, B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. Hadley, A. S. Hadley, and M. G. Betts, "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach," The Journal of the Acoustical Society of America, vol. 131, pp. 4640-4650, 2012.

[12] D. Stowell and M. D. Plumbley, "Birdsong and C4DM: A survey of UK birdsong and machine recognition for music researchers," Tech. Rep. C4DM-TR-09-12, Queen Mary University of London2010.

[13] J. Cai, D. Ee, B. Pham, P. Roe, and J. Zhang, "Sensor network for the monitoring of ecosystem: Bird species recognition," in 2007 Intelligent Sensors, Sensor Networks and Information. ISSNIP 2007. 3rd International Conference on, 2007, pp. 293-298.

[14] C.-H. Lee, Y.-K. Lee, and R.-Z. Huang, "Automatic recognition of bird songs using cepstral coefficients," Journal of Information Technology and Applications, vol. 1, pp. 17-23, 2006.

[15] A. Harma and P. Somervuo, "Classification of the harmonic structure in bird vocalization," in 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings.(ICASSP'04), 2004, pp. V-701-4 vol. 5.

[16] P. Jančovič, M. Köküer, and M. Russell, "Bird species recognition from field recordings using HMM-based modelling of frequency

tracks," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 8252-8256.

[17] Brandes, "Techniques for bioacoustic signal detection using image processing," Computational bioacoustics for assessing biodiversity. Bundesamt fur Naturschutz, Bonn, Germany, pp. 103-110, 2008.

[18] X. Dong, M. Towsey, J. Zhang, J. Banks, and P. Roe, "A Novel Representation of Bioacoustic Events for Content-Based Search in Field Audio Data," in 2013 International Conference on Digital Image Computing: Techniques and Applications (DICTA), 2013, pp. 1-6.

[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005, 2005, pp. 886-893.

[20] X. Dong, M. Towsey, A. Truskinger, M. Cottman-Fields, J. Zhang, and P. Roe, "Similarity-based birdcall retrieval from environmental audio," Ecological Informatics, vol. 29, pp. 66-76, 2015.

[21] Y.-D. Wu, Y. Li, and B.-L. Liu, "A new method for approximate melody matching," in 2003 International Conference on Machine Learning and Cybernetics, 2003, pp. 2687-2691.

[22]