AUTOMATIC BIRD VOCALIZATION IDENTIFICATION BASED ON FUSION OF SPECTRAL PATTERN AND TEXTURE FEATURES

Sai-hua Zhang[a], *Zhao Zhao*[a,b,*], *Zhi-yong Xu*[a,*], *Kristen Bellisario*[b], *Bryan C. Pijanowski*[b]

[a] School of Electronic and Optical Engineering, Nanjing University Sci. & Technol., 210094, China
[b] Department of Forestry and Natural Resources, Purdue University, West Lafayette IN47907, USA
zhaozhao@njust.edu.cn, ezyxu@njust.edu.cn

ABSTRACT

Automatic bird species identification from audio field recordings is studied in this paper. We first used a Gaussian mixture model (GMM) based energy detector to select representative acoustic events. Two different feature sets consisting of spectral pattern and texture features were extracted for each event. Then, a ReliefF-based feature selection algorithm was employed to select distinguishing features. Finally, classification was performed using support vector machine (SVM). The main focus of the proposed method lies in the fusion of a spectral pattern feature with several texture descriptors, which extends our previous work. Experiments used an audio dataset comprised of field recordings of 11 bird species, containing 2762 bird acoustic events and 339 detected "unknown" events (corresponding to noise or unknown species vocalizations). Experimental results demonstrate superior classification performance compared with that of the state-of-the-art method, which renders the proposed method more suitable for real-field recording analysis.

Index Terms—bird species identification, spectral pattern feature, texture descriptor, feature selection, support vector machine

1. INTRODUCTION

Recent developments of technology have already provided considerable support to biodiversity monitoring. Birds have been used widely as indicators of biodiversity because they are distributed over a wide range of natural habitat and are relatively easy to detect. There is also a significant amount of existing knowledge on the biology of most of the species by expert ornithologists through field observations, providing enormous support to relevant research. As a complement to traditional human-observer-based survey methods, acoustic analysis of bird vocalizations can be used for automated species identification, leading to a promising non-intrusive method for bioacoustic monitoring [1, 2].

The first stage of acoustic bird species identification is usually to segment the continuous recordings into isolated acoustic events. Some studies involved manual segmentation [3, 4], which is extremely laborious and time-consuming. Recently, various methods have been employed for automated segmentation [5, 6]. In particular, modeling the distribution of short-term energy with a Gaussian mixture model (GMM) is a more sophisticated acoustic activity detection approach that has been widely used in noisy environments [7]. Note that bird vocalizations can be divided into two categories: calls and songs, where the former usually refer to isolated monosyllabic sounds and the latter contain a few syllables. In this paper, an acoustic event refers to either a call or a syllable.

After segmentation, each segment is represented by a set of features that can discriminate between different classes of bird sounds. Then, a recognition algorithm is employed to identify the bird species based on the extracted features. In general, features developed for acoustic bird species classification roughly include two categories: frame-level and event-level features [7]. Frame-level features are calculated in each frame, such as Mel-frequency cepstral coefficients (MFCCs), linear predictive coding coefficients (LPCCs), peak frequency, short-time bandwidth, as well as their changes between adjacent frames [8, 9, 10]. In contrast, event-level features focus on a whole acoustic event, rather than a single frame within it. A variety of event-level features have been investigated, including highest frequency, average or maximum bandwidth, duration, maximum power, and different combinations of these features [3, 11]. Besides, some more complex descriptors have also been proposed, such as harmonic structure, spectral peak tracks, spectrogram ridge, and MPEG angular radial transform (ART) descriptor [4, 12, 13, 14]. However, features of both categories were usually investigated using datasets that only involved the species of interest. Considering that the classifier will have to assign some acoustic events not well suited to any existing classes to an unknown class when working with real-world datasets, we proposed a novel spectral pattern feature in our most recent study [7]. This parameterized feature depicts the species-specific spectral pattern, contributing to superior robustness in real-world scenarios.

More recently, texture descriptors have been investigated and successfully applied to the task of face recognition [15] and image retrieval [16]. It is worth remarking that fusion of different feature sets has been used for automatic acoustic classification of anurans and proved efficient in performance improvement [9]. Based on these recent studies, we propose a new method by combining the spectral pattern feature with texture descriptors in this paper, extending our previous work. Note that considering the "curse of dimensionality" problem, it's not always appropriate to simply concatenate a series of features. Thus, the ReliefF algorithm [17] is employed for feature ranking and selection, helping to reduce the misclassification rate and computational demands. Experimental results demonstrate that a distinguishing subset of the features obtained by feature selection can improve performance for automatic bird species identification.

2. PROPOSED METHOD

In summary, for those numerous methods developed for automatic bird sound detection and species identification, a typical analysis

workflow could contain following three steps: automated segmentation, feature extraction and classification. In accordance with this workflow, the overall block diagram of our method is depicted in Fig. 1. The individual processing steps are described in the following subsections.

2.1. Automated segmentation

After each field recording is divided into frames, the distribution of log-energies of frames is modeled by a GMM of two mixtures. In this model, one mixture component is fitted to the distribution of the low-energy frames and the other works for the high-energy frames. Then, the crossing point of the two components is usually selected as the decision threshold [18].

Specifically, given the sampling frequency of 32kHz, a recording is divided into frames of 320 samples with an overlap of 160 samples between adjacent frames. Short-time Fourier transform (STFT) is then implemented to each frame using a Hamming window with length 512. Finally, the corresponding spectrogram $S(k,l)$ is fed into the subsequent step with k and l the indices of Fourier coefficient and frame number, respectively.

Considering a recording with spectrogram $S(k,l)$, the energy as well as the log-energy of the l-th frame is denoted as

$$e(l)=\sum_{k=N_L}^{N_H}\left|S(k,l)\right|^2 \qquad le(l)=\log_{10}\left(e(l)\right) \tag{1}$$

where $1\le l\le L$ and L is the total number of frames in the recording. The frequency range of interest is constructed by N_L and N_H corresponding to 1kHz and 16kHz, respectively. Assume that the log-energy le is generated by a two-component GMM and the corresponding probability density function can be written as

$$p(le;w,\mu,\sigma)=\sum_{m=1}^{2}\frac{w_m}{\sigma_m\sqrt{2\pi}}\exp\left(-\frac{(le-\mu_m)^2}{2\sigma_m^2}\right) \tag{2}$$

where $w_m, m=1,2$ are the mixing coefficients satisfying $0\le w_m\le 1$ and $\sum_{m=1}^{2}w_m=1$. μ_m and σ_m are the mean and standard deviation of the m-th Gaussian component, respectively. The maximum likelihood solution for the parameter set $\{w_m,\mu_m,\sigma_m\}, m=1,2$ can be obtained by the widely used expectation-maximization (EM) algorithm. With the threshold chosen as the crossing point of the two Gaussian components, most of the promising high-energy frames are selected. Then, every cluster of consecutive selected frames is grouped into a single event, i.e. a segment. Those events that are shorter than 20ms will be discarded.

Note that after the above GMM-based event detection step, there are still some events with faintish energy that cannot be identified with certainty even by ornithologists to be selected. Therefore, an event-energy-based post-processing procedure is needed to eliminate them. To be more specific, given an initial candidate events set $D=\{AE_1,AE_2,\cdots,AE_K\}$ with K events being detected, the k-th event-energy is calculated as

$$EAE_k=10\log_{10}\left(\sum_{l\in AE_k}e(l)\right), k=1,\cdots,K \tag{3}$$

The maximum is denoted as $ME=\max_k EAE_k$. As for the k-th event, if $ME-EAE_k\ge 20dB$, this event will be discarded. Finally, the remaining events constitute the set *RD*.

After the event-energy-based sifting and manual inspection on the output of automated segmentation, bird species events and "unknown" events (corresponding to noise or unknown species vocalizations) are obtained. Then, they are fed into the following feature extraction step.

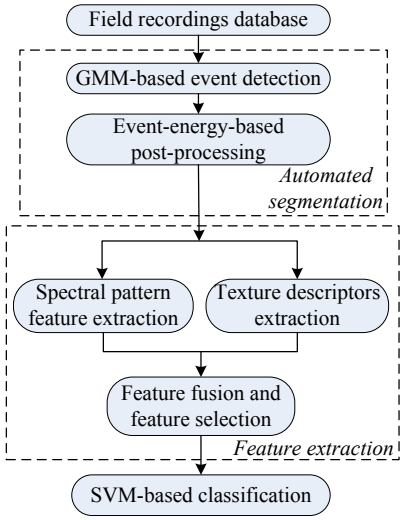

Fig. 1. Overall block diagram of the proposed method.

2.2. Feature extraction

In this step, a feature fusion using the spectral pattern feature and five texture descriptors is explored for bird species classification. Furthermore, a feature selection algorithm is incorporated to reduce feature dimension and increase accuracy.

2.2.1. Spectral pattern feature

The spectral pattern feature is first proposed in our previous work [7]. Given an event in *RD*, its spectrogram is first filtered by a Mel-scaled filter bank containing 32 equal-height triangular band-pass filters within the frequency range from 1kHz to 16kHz, leading to 32 subband time-series x_1, ..., x_{32}. Then, an AR model is used to characterize the spectral pattern property for each subband. As for the i-th subband, the AR model is

$$x_i(l)+\alpha_1^{(i)}x_i(l-1)+\cdots+\alpha_{M_i}^{(i)}x_i(l-M_i)=z_i(l) \tag{4}$$

where $z_i(l)$ is a zero-mean white noise excitation and the model order M_i is determined by the Akaike information criterion (AIC) with the maximum being experimentally set to 10 in this paper. The coefficients of the AR model constitute the subband feature

$$\mathbf{sp}_i=[\alpha_1^{(i)},\alpha_2^{(i)},\cdots,\alpha_{10}^{(i)}]^T, i=1,2,\ldots,32 \tag{5}$$

Afterwards, the corresponding spectral pattern feature vector is denoted as $\mathbf{sp}=[\mathbf{sp}_1^T,\mathbf{sp}_2^T,\cdots,\mathbf{sp}_{32}^T]^T$. It is worth noting that the species-specific feature with respect to certain acoustic event is conveyed by the model coefficients. This parameterization process can deal with a variety of bird acoustic events with either different durations or different sound unit shapes.

2.2.2. Texture descriptors

Since the spectrogram can be viewed as an image, it is feasible to use texture descriptors to depict bird sound events. The state-of-the-art texture descriptors [19] incorporated in this work are briefly described as follows:

(1) ULBP: a uniform local binary pattern (ULBP) descriptor denoted as **lbp**. The local binary pattern (LBP) is first computed at each pixel, considering the binary differences between the gray values of a central pixel and those of P pixels in a circular neighborhood with radius R pixels. In this work, R and P were set

to 1 and 8, respectively. Compared with LBP, ULBP is an improved descriptor in which the number of transitions in the binary code sequence between 0 to 1 or 1 to 0 is less than or equal to two, producing a feature vector with length 59.

(2) LBPHF [20]: a LBP histogram Fourier (LBPHF) descriptor denoted as **lbf**. LBPHF is a rotation-invariant image descriptor that is computed from discrete Fourier transforms (DFT) of LBP histograms. Given the histogram $h(U_P(n,r))$ where $U_P(n,r)$ represents a specific uniform LBP pattern with P neighboring sampling points, n denotes the row number and r defines the rotation of the pattern, we set $H=(n,\cdot)$ to be the DFT of the n-th row of histogram. The LBPHF feature with length 38 refers to corresponding magnitude spectrum.

(3) LPQ [21]: a local phase quantization (LPQ) descriptor denoted by **lpq**. Given a rectangular $M\times M$ (M was set to 3 in this work) neighborhood $N_{\mathbf{x}}$ at each pixel position $\mathbf{x}$ of the image $f(\mathbf{x})$, corresponding 2-D DFT is first computed. Then, the DFT coefficients are quantized into a two bit code: first bit for real part and second bit for imaginary part. This gives 8-bit code from four quantized coefficients. Finally, after decorrelation and quantization, a histogram of quantized coefficients from all image positions is composed and used as a 256-dimensional feature vector.

(4) HASC: the heterogeneous auto-similarities of characteristics (HASC) descriptor. HASC is applied to heterogeneous dense features maps that simultaneously encode linear relations by covariances (COV) and nonlinear associations through entropy and mutual information (EMI). For a given rectangular patch P in a d-dimensional feature image extracted from original image (d was set to 6 in this work), containing K pixels, the COV and EMI matrices are calculated. Then, we vectorize both matrices, obtaining **cov** and **emi**. Finally, HASC is defined by concatenation as $\mathbf{hac}=[\mathbf{cov}^T,\mathbf{emi}^T]^T$ with length 42.

(5) GF: the Gabor filter (GF) descriptor. We used 5 different scale levels and 8 different orientations. The mean-squared energy and the mean amplitude were calculated for each scale and orientation. In this way a feature vector **gf** of size 80 is obtained.

2.2.3. Feature fusion and feature selection

After all the features above are extracted, a fusion stage is further incorporated to obtain a compact feature representation for each acoustic event. Specifically, for the i-th event, the concatenated feature vector obtained is

$$\mathbf{cfv}_i=\left[\mathbf{sp}_i^T,\mathbf{lbp}_i^T,\mathbf{lbf}_i^T,\mathbf{lpq}_i^T,\mathbf{hac}_i^T,\mathbf{gf}_i^T\right]^T, i=1,2,\cdots,K \quad (6)$$

where K is the total number of events.

It is worth noting that this simple concatenation leads to a high-dimensional feature vector, which will increase the computational burden and potentially cast an adverse effect on subsequent classification performance. In this context, feature selection, aiming to choose a discernible subset of features to reduce the fusion feature length with the lowest information loss, can be employed to remove irrelevant and/or redundant features. As one of the filter based feature selection methods, the Relief algorithm [22] is an effective approach to feature weight estimation. ReliefF [17] extends two-class Relief algorithm to deal with multi-class problem. With the help of ReliefF algorithm, an attribute weight was calculated for each feature, ranging from -1 to 1 with a high positive weight assigned to an important attribute. Then we sorted out N most important features as the effective subset. According to our preliminary study, we selected N=400 according to the best performance for the following classification step.

2.3. SVM-based classification

SVM is a robust supervised learning method that has been extensively studied for classification and regression. SVM classifiers use a hyperplane and a kernel function for nonlinear classification of two class data. As for the multi-class classification, we employed the "one-versus-one" strategy [23]. Besides, we used the radial basis function (RBF) as the kernel function.

3. EXPERIMENTAL EVALUATIONS

3.1. Field recordings description

In order to evaluate our method, the field audio recordings used in this work were downloaded from the Xeno-canto Archive (http://www.xeno-canto.org/) and the details can be found in [7]. Note that these are all real-world recordings and each recording potentially contains vocalizations of several animal species and competing noise originating from wind, rain, or anthropogenic interference. The recordings were resampled to a uniform sampling frequency of 32 kHz. There are 11 bird species in the recordings that can be divided into five types based on sound unit shapes, including constant frequency (CF), frequency modulated whistles (FM), broadband pulses (BP), broadband with varying frequency components (BVF), and strong harmonics (SH) [1]. After manual inspection on the result of the automated segmentation described in subsection 2.1, a total of 2762 acoustic events for 11 bird species and simultaneously 339 "unknown" events were available. We provide the description of the dataset used in this study in Table 1.

3.2. Experimental setup

For each detected event, the spectral pattern feature and 5 texture descriptors were calculated and ReliefF algorithm was employed to select discriminative features. The two parameters of the RBF kernel, gamma and cost, were set to 0.0625 and 8. The baseline method only utilized the spectral pattern feature [7]. 10 trials were conducted to compare the performance of the proposed method and the baseline one. In each trial the dataset was split randomly into 60% training set and 40% testing set to acquire statistically relevant results. Meanwhile, the percentage of each class was kept as same as 60:40.

3.3. Performance metrics

Both methods were evaluated by means of three performance metrics for each class including precision (P), recall (R) and F-score which are denoted as

$$P=\frac{TP}{TP+FP} \qquad R=\frac{TP}{TP+FN} \qquad F\text{-}score=\frac{2PR}{P+R} \quad (7)$$

where TP is the number of detected true positive events for each class. FP and FN are the numbers of false positive and false negative events for each class, respectively. After 10 trials, averaged metrics for each method were calculated.

Meanwhile, the overall performance was measured in terms of the classification accuracy defined as $CA=\left(N_{CA}/N_{Te}\right)\times 100\%$ where N_{CA} is the number of correctly classified events and N_{Te} is the total number of testing events.

4. RESULTS AND DISCUSSION

In this section, we first present the comparative results among different features in Table 2 after 10 trials. In order to guarantee a fair comparison, all features were equipped with the same SVM classifier. It can be observed that the worst classification accuracy achieved is 86.6% when only those texture descriptors were

Table 1. Details of species and corresponding field recordings used in this work.

Bird species	Call/ Song	Sound unit shape	Number of events
Blue Jay (B-J)	Call	SH	251
Song Sparrow (S-S)	Call	SH	259
Marsh Wren (M-W)	Call	BP	249
Common Yellowthroat (C-YT)	Call	BP	256
Chipping Sparrow (C-S)	Call	FM	253
American Yellow Warbler (A-Y-W)	Call	FM	247
Great Blue Heron (G-B-H)	Call	BVF	247
American Crow (A-C)	Call	BVF	253
Cedar Waxwing (C-WW)	Call	CF	246
House Finch (H-F)	Song	SH	249
Indigo Bunting (I-BT)	Song	FM	252

Table 2. Comparison of various features in terms of classification accuracy (CA) .

Features	Dimension of feature vector	CA (%)
SP (baseline method)	320	93.7
HASC+LBP+LPQ+LBPHF+GF	475	86.6
SP+HASC+LBP+LPQ+LBPHF+GF without feature selection	795	94.0
SP+HASC+LBP+LPQ+LBPHF+GF with feature selection (this work)	400	96.7

combined. At the same time, fusion of spectral pattern feature with texture descriptors without feature selection achieved similar performance to that using only spectral pattern feature (baseline method) in terms of classification accuracy. However, the fusion of spectral pattern feature with texture descriptors plus ReliefF-based feature selection (this work) clearly outperformed the methods above, approximately 2.7% higher than the method without feature selection. This demonstrates that feature selection algorithm can sort out more robust and discriminating feature subset, which has a dimension of 400, only 50.3% of the original concatenated feature defined in Eq. (6).

In order to further investigate and compare the performance between the proposed method and baseline one, the averaged performance metrics of the two methods for each class are provided in Table 3. One should note that the performance metrics of the proposed method are almost all greater than those of the baseline method, except the precision of C-WW and Unknown class, which merely decreased by 0.3% and 0.4%. It is worth remarking that a simple comparison on these metrics, however, is not completely reliable since the winning algorithm may occasionally perform well due to the randomness in data split. Hypothesis test is usually employed for statistically reliable comparison between algorithms.

Here, we used Mann-Whitney test [24], which aims to assess the statistical significance of the differences between two groups. The Mann-Whitney test is the non-parametric equivalent of the independent samples *t* test. For both the proposed method and the baseline one, we calculated the classification accuracy values for 10 trials. Then, the Mann-Whitney test was employed to compare the statistical difference of the two methods. The significance level α was set to 0.05 and the corresponding estimate value

Table 3. Averaged precision and recall as well as the corresponding *F*-score for each species between the proposed method with the baseline method.

Classes	Recall (%)		Precision (%)		*F*-score	
	This work	Baseline method	This work	Baseline method	This work	Baseline method
B-J	**98.7**	97.9	**99.0**	97.9	**0.989**	0.979
S-S	**97.0**	92.6	**95.4**	88.4	**0.962**	0.904
M-W	**96.8**	90.1	**97.5**	94.5	**0.971**	0.922
C-YT	**95.6**	91.0	**95.1**	89.6	**0.953**	0.903
C-S	**97.7**	93.8	**97.0**	94.2	**0.973**	0.940
A-Y-W	**97.7**	93.8	**99.0**	90.7	**0.983**	0.922
G-B-H	**96.3**	91.4	**98.1**	94.8	**0.972**	0.930
A-C	**99.2**	97.9	**99.5**	96.7	**0.994**	0.973
C-WW	**96.8**	95.6	98.7	**99.0**	**0.977**	0.973
H-F	**95.1**	94.7	**95.3**	92.4	**0.951**	0.935
I-BT	**95.7**	94.0	**94.5**	93.3	**0.950**	0.936
Unknown	**95.2**	91.7	94.0	**94.4**	**0.945**	0.930

$p = 1.55 \times 10^{-4} < 0.05$, suggesting that there is significant difference between the two methods. Most recent study in cognitive science confirmed that spectral shape is the primary cue to bird sounds recognition [25, 26]. Therefore, in our previous study [7], the spectral pattern feature was proposed to describe the spectral shape information of bird vocalizations, achieving comparable identification performance with respect to the species of interest and superior robustness in real-world scenarios when compared with other recent approaches. In this work, considering the results of classification accuracy (Table 2), performance metrics for each class (Table 3) and the hypothesis test, we can conclude that the proposed method provides comparable robustness in real-field environments as well as superior identification performance regarding the species of interest—that is, the proposed approach outperforms the baseline method. This improvement can be attributed to the fact that the feature subset obtained by feature fusion and feature selection can depict the species-specific spectral shape information of the time-frequency distribution in each sub-band, as well as the spatial variation and arrangement of orientations information in the full-band.

5. CONCLUSIONS

Aiming to improve the audio parameterization process in bird species identification tasks, we proposed an automatic acoustic classification method based on feature fusion in this work. After GMM-based acoustic event detection and event-energy-based post-processing procedure, representative acoustic events were selected. For each event, two different feature sets, the spectral pattern feature and texture descriptors, were extracted. As for the combination of the two sets, ReliefF-based feature selection algorithm was employed to select a distinguishing feature subset. Experimental results using real-world recordings showed that the proposed method outperformed the state-of-the-art robust approach. The improved performance makes the proposed method more effective for the application of acoustic monitoring in terrestrial environments.

6. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundations of P.R. China under Grants No. 61401203 and No. 61171167; and the State Scholarship Fund of China [grant number 201606840023].

7. REFERENCES

[1] T.S. Brandes, "Automated sound recording and analysis techniques for bird surveys and conservation," *Bird Conservation International*, vol. 18, no. S1, pp. S163-S173, 2008.

[2] K. Kaewtip, L.N. Tan, A. Alwan and C.E. Taylor, "A robust automatic bird phrase classifier using dynamic time-warping with prominent region identification," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, pp. 768-772, 2013.

[3] M.A. Acevedo, C.J. Corrada-Bravo, H. Corrada-Bravo, L.J. Villanueva-Rivera and T.M. Aide, "Automated classification of bird and amphibian calls using machine learning: a comparison of methods," *Ecological Informatics*, vol. 4, no. 4, pp. 206-214, 2009.

[4] C.H. Lee, S.B. Hsu, J.L. Shi and C.H. Chou, "Continuous birdsong recognition using Gaussian mixture modeling of image shape features," *IEEE Transactions on Multimedia*, vol. 15, no. 2, pp. 454-464, 2012.

[5] L. Neal, F. Briggs, R. Raich, and X.Z. Fern, "Time-frequency segmentation of bird song in noisy acoustic environments," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Prague, Czech Republic,* pp. 2012-2015, 2011.

[6] A.G. Oliveira, T.M. Ventura, T.D. Ganchev, J.M. Figueiredo, O. Jahn, M.I. Marques and K.L. Schuchmann. "Bird acoustic activity detection based on morphological filtering of the spectrogram," *Applied Acoustics*, vol. 98, pp. 34-42, 2015.

[7] Z. Zhao, S.H. Zhang, Z.Y. Xu and K. Bellisario, N.H. Dai, H. Omrani and B.C. Pijanowski, "Automated bird acoustic event detection and robust species classification," *Ecological Informatics*, vol. 39, pp. 99-108, 2017.

[8] T.S. Brandes, "Feature vector selection and use with hidden Markov models to identify frequency-modulated bioacoustic signals amidst noise," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1173-1180, 2008.

[9] J.J. Noda, C.M. Travieso, and D. Sánchez-Rodríguez, "Methodology for automatic bioacoustic classification of anurans based on feature fusion," *Expert Systems with Applications*, vol. 50, pp. 100-106, 2016.

[10] I. Potamitis, S. Ntalampiras, O. Jahn and K. Riede, "Automatic bird sound detection in long real-field recordings: Applications and tools," *Applied Acoustics*, vol. 80, pp. 1-9, 2014.

[11] T. Schrama, M. Poot, M. Robb and H. Slabbekoorn, "Automated monitoring of avian flight calls during nocturnal migration," *International Expert Meeting on IT-based Detection of Bioacoustical Patterns*, pp. 131-134, 2007.

[12] P. Jančovič and M. Köküer, "Acoustic Recognition of Multiple Bird Species Based on Penalized Maximum Likelihood," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1585-1589, 2015.

[13] A. Harma and P. Somervuo, "Classification of the harmonic structure in bird vocalization," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, Canada,* pp. 701-704, 2004.

[14] X.Y Dong, M. Towsey, J.L. Zhang, J. Banks and P. Roe, "A novel representation of bioacoustic events for content-based search in field audio data," *IEEE International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1-6, 2013.

[15] A. Lumini, L. Nanni and S. Brahnam, "Ensemble of texture descriptors and classifiers for face recognition," *Applied Computing and Informatics*, vol. 13, no. 1, pp. 79-91, 2017.

[16] B. Jyothi, Y. MadhaveeLatha and P.G.K. Mohan, "Region based texture descriptor for content based medical image retrieval using second order moments," *IEEE 2nd International Conference on Innovations in Information Embedded and Communication Systems (ICIIECS)*, pp. 1-4, 2015.

[17] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine Learning*, vol. 53, no. 1-2, pp. 23-69, 2003.

[18] M. Sahidullah and G. Saha, "Comparison of speech activity detection techniques for speaker recognition," *Journal of Immunotherapy*, vol. 33, no. 33, pp. 609-617, 2012.

[19] L. Nanni, Y.M.G. Costa, D.R. Lucio, C.N. Silla Jr, and S. Brahnam, "Combining visual and acoustic features for audio classification tasks," *Pattern Recognition Letters*, vol. 88, pp. 49-56, 2017.

[20] G.Y. Zhao, T. Ahonen, J. Matas and M. Pietikäinen, "Rotation-invariant image and video description with local binary pattern features," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1465-1477, 2012.

[21] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," *International Conference on Image and Signal Processing (ICISP)*, pp. 236-243, 2008.

[22] I. Kononenko, "Estimating attributes: analysis and extensions of RELIEF," *European Conference on Machine Learning, Springer, Berlin*, pp. 171-182, 1994.

[23] C.W. Hsu and C.J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415-425, 2002.

[24] A. Hart, "Mann-Whitney test is not just a test of medians: differences in spread can be important," *British Medical Journal*, vol. 323, no. 7309, pp. 391-393, 2001.

[25] M.R. Bregman, A.D. Patel and T.Q. Gentner, "Songbirds use spectral shape, not pitch, for sound pattern recognition," *Proceedings of the National Academy of Sciences*, vol. 113, no. 6, pp. 1666-1671, 2016.

[26] R.V. Shannon, "Is birdsong more like speech or music?," *Trends in Cognitive Sciences*, vol. 20, no. 4, pp. 245-247, 2016.