IET Computer Vision

*Special Issue: Computer Vision for Animal Biometrics*

IET Journals
The Institution of Engineering and Technology

# Bird and whale species identification using sound images

ISSN 1751-9632
Received on 31st January 2017
Revised 9th November 2017
Accepted on 14th November 2017
E-First on 22nd December 2017
doi: 10.1049/iet-cvi.2017.0075
www.ietdl.org

*Loris Nanni[1], Rafael L. Aguiar[2,3], Yandre M.G. Costa[2], Sheryl Brahnam[4], Carlos N. Silla Jr.[3] ✉, Ricky L. Brattin[4], Zhao Zhao[5,6]*

[1]*DEI, University of Padua, Italy*
[2]*PCC/DIN, State University of Maringá, Maringá, Brazil*
[3]*PPGIA, Pontifical Catholic University of Paraná, Curitiba, Brazil*
[4]*CIS, Missouri State University, Springfield, USA*
[5]*School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, People's Republic of China*
[6]*Department of Forestry and Natural Resources, Purdue University, West Lafayette IN47907, USA*
✉ *E-mail: carlos.sillajr@gmail.com*

**Abstract:** Image identification of animals is mostly centred on identifying them based on their appearance, but there are other ways images can be used to identify animals, including by representing the sounds they make with images. In this study, the authors present a novel and effective approach for automated identification of birds and whales using some of the best texture descriptors in the computer vision literature. The visual features of sounds are built starting from the audio file and are taken from images constructed from different spectrograms and from harmonic and percussion images. These images are divided into sub-windows from which sets of texture descriptors are extracted. The experiments reported in this study using a dataset of Bird vocalisations targeted for species recognition and a dataset of right whale calls targeted for whale detection (as well as three well-known benchmarks for music genre classification) demonstrate that the fusion of different texture features enhances performance. The experiments also demonstrate that the fusion of different texture features with audio features is not only comparable with existing audio signal approaches but also statistically improves some of the stand-alone audio features. The code for the experiments will be publicly available at https://www.dropbox.com/s/bguw035yrqz0pwp/ElencoCode.docx?dl=0.

## 1 Introduction

Although automatic image identification of animals is mostly centred on identifying them based on their shapes, or appearance, there are other ways images can be used to identify animals, including by the tracks they make [1] and by images representing their sounds, which is the focus of this study.

Classifying sound using images is a rather recent development. In 2011, Costa *et al.* [2] started using spectrogram images to classify music genres using the grey level co-occurrence matrix (GLCM) [3]. This was quickly followed by a series of papers [4, 5] that explored using powerful texture descriptors, such as local binary patterns (LBP), local phase quantisation (LPQ) and Gabor filters (GFs), as well as ensembles of texture features extracted from spectrograms and their combination with features taken directly from the audio signal [6]. In [4] a novel way of splitting the spectrogram image into zones which correspond to frequency bands was also introduced. Recently, research into the visual-spectrogram techniques for audio classification has been extended to other domains, such as language identification [7], bird identification [8, 9], and whale detection and identification [10].

Research into the automatic classification of animals by the sounds they make began in the 1970s with the work of Deuser *et al.* [11] and Giryn *et al.* [12], both of whom explored machine recognition of the sounds fish and other sea creatures emit. A little over 15 years ago in 2001, research was extended to the sounds animals make. Chesmore [13], for instance, investigated the use of machine learning techniques for discriminating the sounds of different animal species within a group. For this purpose, the author trained an artificial neural network to classify features extracted from audio signal segmentation using an extension of time encoded speech, which he called time domain signal processing. Several experiments were performed that discriminated 13 different British species of Orthoptera under low noise conditions, 25 Orthoptera species under noisy conditions that included the sound of grasshoppers, and ten Japanese bird species. Results proved promising, which led Chesmore to conclude that it is possible for auditory systems to identify not only species, including individuals within a species but also the presence of other animals within various outdoors environments.

In [14] the original sound signals of 30 classes of frogs and 19 classes of crickets were segmented into a set of basic acoustic units called syllables. The audio content was then described using Mel-frequency cepstral coefficients (MFCCs), a well-known descriptor for audio content already in use in other application domains. The classification step was carried out using linear discriminant analysis. The average classification accuracy was up to 96.8% for the frog call database and 98.1% for the 19 cricket call dataset.

Molnar *et al.* [15] trained a Naïve Bayes classifier on a data pool containing 6000 dog barks recorded in six different communicative situations. The objective was not only to identify the individual dogs but also to distinguish between six different communicative contexts (fighting, walking, being alone, playing, catching a ball, and encountering a stranger). Individual identification involved classifying the barks recorded from 14 dogs. The descriptors used in these experiments were generated using an extractor discovery system, an audio signal processing system that produced descriptors adapted to a particular audio classification problem (for more details, see [16]). The Naïve Bayes classifier was able to categorise the barks according to their recorded situation with efficiency of 43% and according to identity with an efficiency of 52%.

In [17] classifiers were trained to identify bird species. Birds play important roles in the ecosystem, providing, for example, insect control and seed dispersion and pollination. As a result, it is vital to improving knowledge about different bird species, species evolution, and geographic distribution. Bardeli *et al.* [17] reported a study testing the feasibility of bird monitoring by automatically

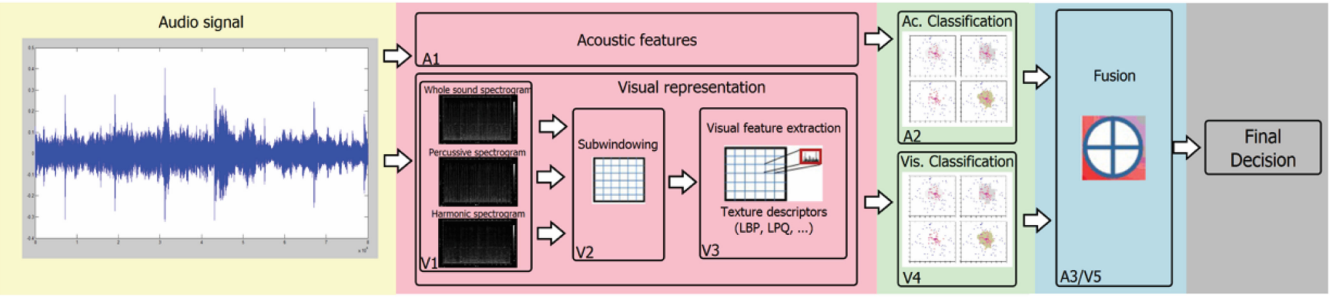


**Fig. 1** *Visual and acoustic features extraction and classification steps*

classifying bird songs. In that work, novel algorithms were developed for detecting the vocalisations of two endangered bird species; moreover, it was shown how this task could be used in automatic habitat mapping. The methods developed in [17] are based on detecting temporal patterns in a given frequency band typical for the species. The authors achieved high recognition rates even in real-world recording conditions, but they warned that special effort is needed in the suppression of the noise present in real-world audio scenes. Cheng *et al.* [18] performed individual recognition of birds using MFCCs and Gaussian mixture models across four passerine species. Their system obtained accuracies ranging from 89.1 to 92.5%, and the authors claim that the acoustic feature/classifier method they developed has excellent potential for individual animal recognition and can be easily applied to other species.

Relevant to the approach taken here, Lucio and Costa [8] started to investigate the use of time-frequency representations of the sound in bird species classification. The authors used powerful texture descriptors to describe the content of spectrogram images obtained from a bird species dataset composed of bird vocalisation recordings of 46 different species. The best accuracy rate was 77.65% obtained using GFs. Lucio and Costa [19] continued the investigations on the same dataset, this time including three different acoustic descriptors (statistical spectrum descriptors (SSD), rhythm patterns, and rhythm histogram (RH)) to extract features directly from the audio signal. The authors assessed the performance using only classifiers built with visual features (i.e. features taken from spectrograms), acoustic features, and features that combined both types. In the end, the best accuracy rate obtained was 91.08% using robust LBP (RLBP) and SSD (i.e. using both visual and acoustical descriptors). In Nanni *et al.* [9], the performance on the same bird song dataset was assessed using different ensembles of classifiers based on combinations of both visual features and acoustic features. In [9], local ternary phase quantisation, heterogeneous auto-similarities of characteristics, and an ensemble of variants of local binary pattern histogram Fourier were tested with spectrograms for the first time. The proposed set of descriptors outperformed the author's previous works on music classification based on visual features extracted from the spectrogram. The best accuracy rate on the Bird Songs 46 dataset was 94.5%.

Still, in the context of pattern recognition tasks applied to animal classification/identification, much research has been devoted in the last decade towards identifying whales, which is important for avoiding whale collisions with vessels. The goal is to reduce the whale mortality rate [20, 21]. One of these initiatives is 'The Marinexplore and Cornell University Whale Detection Challenge,' which provides a publicly available dataset [organised by The Marinexplore and Cornell University and available at https://www.kaggle.com/c/whale-detection-challenge/] containing 30,000 labelled audio clips that contain mixtures of right whale calls, non-biological noise, and other whale calls.

The goal of the study presented is to investigate additional texture descriptors for the purpose of identifying birds and right whales. The main contribution of this work is the design and evaluation of an ensemble of descriptors that includes the standard binarised statistical image features (BSIF) descriptor to capture the spectrogram content in audio. We also test the harmonic and percussion images before representing the audio files. We assess our approach to the whale dataset and two bird datasets described in Section 4.1. To demonstrate the strength of our new system compared with a larger number of other audio classification systems, we also evaluate our method on three well-known benchmark datasets for music genre classification (the Latin Music Database, the ISMIR 2004 database, and the GTZAN genre collection), where it is shown that the proposed image-based approach outperforms all other image-based approaches [9, 10]. Results demonstrate that the fusion between texture features and audio descriptors performs not only competitively against classification systems using audio descriptors but also statistically improves the performance of the stand-alone audio features.

## 2 Proposed approach

In Fig. 1, we present a scheme of our proposed approach. In the first step, an audio signal is represented using audio features (A1) and visual features (V1–3). In step 2 (A2 and V4) each of these features are classified. Finally, in steps A3 and V5, the results are combined for a final decision.

In steps V1–3, features are extracted from visual representations of an audio signal. Each of these visual features is classified in step V4 using a support vector machine (SVM). As shown in Fig. 1, visual feature extraction is a three step process:

Step V1: The audio signal is represented by three types of audio images:

1. Spectrogram images, which are created with the lower limits of the amplitudes set to −70, −90, and −120 dBFS, respectively, and both grey-scale and colour images are produced.
2. Percussion images.
3. Harmonic images; the latter two types of images are created using the median filtering technique proposed by FitzGerald [22].

Step V2: Each image is divided into a set of sub-windows.

Step V3: A set of local texture descriptors (described in Section 3.3) are extracted from the sub-windows and each descriptor is classified by a separate SVM.

Each of these three steps is described in more detail in Section 3. In step A1, acoustic features (described in Section 4) are extracted directly from the audio signal with each classified by an SVM.

The final decision (steps A3/V5) is obtained by combining all the above SVM scores for the acoustic features (A3) and the audio images (V5) using the weighed sum rule (described in Section 5).

## 3 Audio image representation

### *3.1 Step V1*

In Step V1, different strategies were used for each dataset in order to produce more promising samples from the audio signals. For the bird and whale vocalisation tasks, it is important to extract the most relevant parts of the signal, which in the literature are called *shots*. For the bird dataset, samples are built from shots that are extracted manually from the original bird samples, and each sample is concatenated with itself until the size is equal to 30 s. This procedure does not impact the results either positively or

negatively and was used only to standardise the size of the relevant content. Audio files in the whale dataset are divided into lengths of 2 s. Once the audio files are segmented, the spectrogram, harmonic, and percussion images in step V1 are produced.

*3.1.1 Spectrogram images:* The sample audio signals are converted into a spectrogram image showing the spectrum of frequencies along the vertical axis as they vary in time, which is represented on the horizontal axis. The intensity of each point in the image represents the signal's amplitude. Spectrograms are generated using the Hanning window function with the discrete Fourier transform (DFT) computed with a window size of 1024 samples. The audio sample rate is 22,050 Hz. In every case, even when the original audio signal was available in stereo mode, we used only the right channel since there is no important difference between the content of both channels.

Finally, these images are subjected to a battery of tests for finding some complementarity among the different representations. Results of these tests led us to select the following three values: −70, −90, and −120 dBFS.

*3.1.2 Harmonic and percussion images:* These images are generated using the harmonic percussion separation (HPSS) method proposed by Fitzgerald [22]. The method works by using a median filter across successive frames of the spectrogram of the audio signal. This procedure can generate two images. If median filtering is performed across the frequency bins, the percussive events are enhanced and the harmonic components are suppressed. On the other hand, by using the median filtering across the time axis, the percussive events are suppressed, and the harmonic components are enhanced. These median filtered spectrograms are applied to the original spectrogram as masks to separate the harmonic and percussive parts of the signal, thereby generating the harmonic and percussion Images. In this work, we have used the Librosa [23] implementation of the HPSS method.

### 3.2 Step V2: sub-windowing

As recommended in [2, 4, 24], it is best to employ a zoning mechanism to preserve local information about the extracted features. Zoning consists of simply subdividing the whole image into smaller zones (or windows) in such a way that one can obtain the descriptors and subsequently produce classifiers specialised for the different frequency bands. In [6] it was shown that a sub-windowing technique based on the Mel scale [25] outperformed a method where features were extracted from the entire image. In that work, 15 zones were selected for each spectrogram that varied in size as defined by the Mel scale. This produced a total of 45 zones since one spectrogram is created for each segment taken from the original signal. In [10], SVMs were trained on each of the zones, and all 45 results were combined by sum rule. In this work, we follow the method of sub-windowing proposed in [10]. However, to reduce the computation time of the expensive ensemble proposed in this study, we have simply divided the spectrogram into three zones.

### 3.3 Step V3: texture descriptors

In step V3, sets of texture descriptors are extracted from the sub-windows. The following texture descriptors are evaluated in this work:

- LBP [26], a multi-scale uniform LBP, where the final descriptor is obtained by concatenating the patterns at different radii $R$ and at different sampling points $P$: ($R = 1$, $P = 8$) and ($R = 2$, $P = 16$).
- LPQ [27], a multi-scale LPQ, an LBP variant that is based on quantising the Fourier transform phase in local neighbourhoods with radii 3 and 5.
- LBP-HF [28], a multi-scale LBP HF descriptor that is obtained from the concatenation of LBP-HF with values ($R = 1$, $P = 8$) and ($R = 2$, $P = 16$).
- RICLBP [29], a multi-scale rotation invariant co-occurrence of adjacent LBP with values ($R = 1$, $P = 8$), ($R = 2$, $P = 8$) and ($R = 4$, $P = 8$).
- MLPQ [30], an ensemble of LPQ that is based on a ternary encoding. Effective ensembles can be designed by combining sets of LPQ extracted by varying the parameters $r$ (the neighbourhood sizes, where $r \in [1, 3, 5]$), $a$ (the scalar frequency, where $a \in [0.8, 1, 1.2, 1.4, 1.6]$, and $\rho$ (the correlation coefficient between adjacent pixel values with $\rho \in [0.75, 0.95, 1.15, 1.35, 1.55, 1.75, 1.95]$). Each LPQ trains a different classifier whose results are combined by sum rule. This ensemble is built up with 105 descriptors, and the scores are summed and normalised by dividing the sum by 105.
- Heterogeneous auto-similarities of characteristics (HASC) [31] are applied to heterogeneous dense features maps.
- BSIF [23], where the standard BSIF descriptor is extracted by projecting sub-windows of the entire image onto sub-spaces. The images are binarised (using a threshold $th$) and a histogram is built (see [32] for details). The filters used in this approach are built by maximising the statistical independence of the filter responses on a set of sub-windows extracted from natural images by independent component analysis. To increase the performance of this descriptor, an ensemble, as in [33], is built by varying the parameters of the approach: the filter size ($size \in \{3, 5, 7, 9, 11\}$); and the threshold used for binarising ($th \in \{-9, -6, -3, 0, 3, 6, 9\}$). In total, the ensemble is built with 35 SVMs combined by sum rule, with each SVM trained using a different feature vector extracted with a possible ($size$, $th$) combination of the parameters of BSIF. We label this ensemble FullF.
- GABOR, where the GF features that are extracted using several different values for scale level and orientation that are experimentally evaluated; the best result obtained was with five different scale levels and 14 different orientations. The mean-squared energy and mean amplitude were calculated from each possible combination between scale and orientation. In this way, a feature vector of size $5 \times 14 \times 2$ is obtained.
- Adaptive hybrid pattern (AHP) [34] is an LBP variant that is noise robust because a quantisation algorithm is applied that uses an equal probability quantisation to maximise partition entropy. The vector quantisation thresholds are adaptive to the content of the local patch with little discriminant information loss. In our experiments, quantisation_level = 5; ($P = 8$, $R = 1$); ($P = 16$, $R = 2$).
- LEN is the LETRIST histogram in [35] that explicitly encodes the joint information within an image across feature and scale spaces. LEN is a two-step process: (i) a set of transform features is constructed by applying linear and non-linear operators on the extremum responses of directional Gaussian derivative filters in scale space and (ii) The scalar quantisation via binary or multi-level thresholding is adopted to quantise these transform features into texture codes. We use the default values in the available MATLAB toolbox.
- ELHF [36], an ensemble of variants of the local binary pattern histogram Fourier that is built from the following seven descriptors (each trained by an SVM and with SVM scores summed and normalised by dividing the sum by seven):

  - Fast Fourier (FF): the original method, where from each DFT, the first half of the coefficients are retained.
  - DC: an approach where the first half of the coefficients are retained from each discrete cosine transform (DCT).
  - An approach where the histogram is decomposed by Daubechies wavelet before DFT and then FF is performed.
  - An approach where the histogram is decomposed by Daubechies wavelet before DCT and then the method DC is performed;
  - An approach where the histogram is decomposed by Daubechies wavelet before DFT and then the method FF is performed, with all co-efficients, retained.
  - An approach where the histogram is decomposed by Daubechies wavelet before DCT and then the method DC is performed, with all coefficients, retained.

○ An approach where all the bins of the histogram are retained.

- CLM: is the CodebookLess Model proposed in [37]. This method is a dense sampling approach similar to a bag of features, but an approach that is not based on a codebook since it represents the images with a single Gaussian model (see the original paper for details). In this work, we train three different CLM models as detailed in Table 1, which reports Raw_feature (the type of extracted feature), Reduction (the method for dimensionality reduction: linear SVM (LRSVM), principle component analysis, and LRSVM), and Redim (the size of the reduced feature vector).

The final score is given by the sum rule of the three CLM models in Table 1. We use two different approaches for classifying the features extracted by the above CLM method:

- CLMa, which uses the standard one-versus-all SVM.
- CLMb, where for both training and testing patterns we calculate the scores obtained by classifying them by one-versus-one SVM approach (the SVMs are trained using only the training data). Then the patterns described by the set of scores are classified by a one-versus-all SVM.

## 4 Acoustic features

For the acoustic feature sets, we selected the following:

- SSD [38], which is a set of statistical measures describing the audio content taken from the moments on the Sonogram (the Sone) of each of the 24 critical bands defined according to Bark scale.
- RH [38], where the magnitudes of each modulation frequency bin of the 24 critical bands defined according to the Bark scale are summed up to form a histogram of 'rhythmic energy' per modulation frequency. This histogram has 60 bins reflecting the modulation frequencies between 0 and 10 Hz, and the feature set is the median of the histograms of each of the 6 s extracted segments.
- Modulation frequency variance descriptor (MVD) [38], a descriptor that measures variations over the critical frequency bands for each modulation frequency. The MVD descriptor for the audio file is the mean of the MVDs taken from the 6 s segments and is a 420-dimensional vector.
- Temporal SSD [16, 39], a descriptor that incorporates temporal information from the SSD, such as timbre variations and changes in rhythm. Statistical measures are taken across the SSD measures extracted from segments at different time positions in an audio file.
- Temporal rhythm histograms ([38]), a descriptor that captures rhythmic changes in music over time.

We use the acoustic features of a commercial system for the music genre dataset. This system is based on a method proposed in [40], which was later improved in [6]. The latter version is used in our experiments. Each acoustic feature is used to train an SVM.

## 5 Classification and fusion

Unless otherwise mentioned, we use a one-versus-all SVM with a radial basis function kernel for classification. To avoid over fitting because of small training sets, we set $c = 1000$ and $\gamma = 0.1$ constant for all experiments (rather than perform parameter optimisation). Before the classification step, features are normalised to [0, 1].

**Table 1** Parameters of ensemble of CLM

| Raw_feature | Reduction | Redim |
|---|---|---|
| eL2EMG | LRSVM | 450 |
| eSIFT | principal component analysis | 64 |
| eSIFT | LRSVM | 64 |

Outputs of sets of SVMs are combined by the sum algebraic expression. The final ensemble decision is the class that receives the largest support, with the support of a given class $j$ of a pattern $x$ defined as

$$\sum_{cl=1}^{N} \text{score}(cl, j, x), \tag{1}$$

where $N$ is the number of classifiers that belong to the ensemble and score $(cl, j, x)$ is the output $x$ of the classifier $cl$ with respect to class $j$.

## 6 Experimental results

### 6.1 Datasets

As mentioned in the introduction, many researchers have recently started using bioacoustic techniques to identify birds starting from audio records collected from nature. It is very difficult, however, to make fair comparisons between the results of these works. This is because some studies are based on audio clips recorded using professional equipment resulting in little noise interference [39, 41, 42] and other recording collected in an amateurish fashion using devices such as smartphones [42]. In terms of numbers, an important milestone to community research was the creation of Xeno-canto [Available at www.xeno-canto.org.], a website for sharing recordings of sounds of birds in the wild from all across the world. Unfortunately, works using data taken from Xeno-Canto rarely use the same number of species or the same number of individual samples within each species, again making comparisons with methods using this dataset problematical.

For the problem of bird identification, we use the following two datasets:

BIRD: this is the Bird Songs 46 dataset in [9], which is freely available [Available at www.din.uem.br/yandre/birds/bird_songs_46.tar.gz] and developed as a subset of that used in [43]. All bird species with less than ten samples were removed. The Bird Songs 46 dataset has been used in [8–10] and is composed of 2814 audio samples of bird vocalisation taken from 46 different species found in the South of Brazil. The protocol used for this dataset is a stratified 10-fold cross validation strategy, which is the same protocol used in [8–10]. The Bird Songs 46 dataset is composed of bird songs only and, in some case, one can find noise related to other bird species in the background.

BIRDzhao: this is our label for the control and real-world audio dataset used in [44] that is comprised of field recordings of 11 bird species taken from the Xeno-canto Archive. The BIRDzhao dataset was selected because it lends itself to the comparison. This dataset contains 2762 bird acoustic events (11 classes) with 339 detected 'unknown' events corresponding to noise and unknown species vocalisations.

WHALE: this is 'The Marinexplore and Cornell University Whale Detection Challenge' [available at www.kaggle.com/c/whale-detection-challenge] dataset composed of 84,503 (2 s) audio clips that contain mixtures of right whale calls, non-biological noise, and other whale calls. Thirty thousand samples are available with class labels. In this work, we used the first 20,000 labelled samples for the training set and the last 10,000 samples for the testing set. The results on this dataset are described using the area under the receiver operating characteristic curve (AUC), the performance indicator that was used in the challenge.

To demonstrate the power of the proposed approach, our method is also evaluated across the following three benchmark music datasets:

- LMD [45], also known as the Latin Music Database, is designed to evaluate music information retrieval systems and contains 3227 samples classified into ten musical genres: axe, bachata, bolero, forro, gaucha, merengue, pagode, salsa, sertaneja, and tango. The testing protocol for this dataset is the threefold cross-validation protocol, where an artist filter restriction is applied

**Table 2** Comparison of tested descriptors

| New descriptors | BIRD | WHALE | BIRDzhao | LMD | ISMIR | GTZAN |
|---|---|---|---|---|---|---|
| MLPQ | 87.5 | 92.1 | 88.8 | 85.1 | 84.4 | 85.4 |
| BSIF | 83.5 | 87.6 | 84.0 | 85.0 | 81.9 | 81.2 |
| FullF | 88.8 | 90.4 | 87.5 | 87.1 | 86.0 | 84.9 |
| CLMa | 89.9 | 87.2 | 60.4 | 86.7 | 84.9 | 86.3 |
| CLMb | 78.8 | 84.1 | 82.7 | 77.7 | 83.2 | 84.9 |
| AHP | 84.4 | 89.9 | 77.5 | 78.0 | 79.4 | 79.3 |
| LET | 67.7 | 90.3 | 75.6 | 83.1 | 80.9 | 82.9 |
| MLPQ + FullF | 89.2 | 91.7 | 88.7 | 86.8 | 85.4 | 86.7 |
| MLPQ + 2 × FullF | 89.3 | 91.4 | 88.4 | 87.2 | 86.0 | 86.8 |
| MLPQ + 2 × FullF + 0.5 × CLMa | 89.5 | 91.5 | 87.5 | 87.1 | 86.0 | 87.0 |
| 2 × FullF + CLMa | 89.8 | 90.8 | 81.8 | 87.8 | 86.7 | 86.8 |
| FusFULL | 85.0 | 92.8 | 88.2 | 89.0 | 85.5 | 86.3 |
| FusFULL_CLMa | 86.0 | 92.8 | 82.6 | 89.5 | 86.3 | 86.8 |

**Table 3** Comparison of ensemble of descriptors

| Descriptor | BIRD | WHALE | BIRDzhao | LMD | ISMIR | GTZAN |
|---|---|---|---|---|---|---|
| [6] | 85.9 | 87.1 | 83.1 | 86.1 | 81.6 | 83.8 |
| [10] | 89.2 | 92.2 | 85.1 | 86.2 | 82.2 | 86.1 |
| E_Full | **90.9** | 92.2 | 88.9 | 88.1 | **85.3** | 86.1 |
| E_Fullhp | 89.8 | 93.2 | 88.9 | 89.8 | **85.3** | 87.0 |
| New | 90.7 | 92.2 | 88.9 | 87.9 | 85.1 | 87.0 |
| NewFULL | 89.9 | **93.3** | **89.2** | **90.0** | **85.3** | **87.4** |

[46], i.e. where all the samples by a specific artist are included in only one fold. Since the distribution of samples per artist is far from uniform, only a subset of 900 samples is used for fold creation.

- ISMIR 2004 [47] is a dataset that contains 1458 samples (split evenly into testing and training sets) that are assigned to six different genres: classical, electronic, jazz/blues, metal/punk, rock/pop, and world.
- GTZAN [48] is a dataset that represents ten genre classes (Blues, Classical, Country, Disco, Hip Hop, Jazz, Metal, Popular, Reggae, and Rock). Each genre class contains 100 audio recordings. For fair comparisons with results reported in [40], we evaluate the performance using the same 10-fold split tested shared by [40].

For the music genre datasets, the original signal content is reduced to three 10 s segments that are taken from the beginning, middle, and end of the original audio signals to mitigate the possibility of unrepresentative samples. This strategy was inspired by [49].

The proposed approach is evaluated on the above datasets using the recognition rate as the performance indicator unless otherwise indicated in the above database descriptions.

### *6.2 Experiments*

As in [10], we built three spectrograms created with the lower limits set to the following three values: −70, −90, and −120 dBFS. SVMs are then trained with descriptors extracted from the three spectrograms, and the results are combined by sum rule. However, as noted in Section 3.2, although in [10] spectrograms are divided into 45 zones with 45 separate SVMs trained for each descriptor, in this work, spectrograms are divided into three zones to reduce computational time.

In Table 2, we report the performance obtained combining the nine descriptors extracted from the three zones of the three spectrograms. The row $A + K \times B$ indicates that we combine (by weighted sum rule) the SVM trained with $A$ (with weight = 1) with the SVM trained with $B$ (with the weight of $K$).

The row named FusFULL is the fusion of the SVMs trained with MLQP + 2 × FullF using the three spectrogram images, the harmonic image, and the percussion image (hence, five images for each of the three zones).

The row named FusFULL_CLM reports the fusion of the SVMs trained with 3 × FusFULL + 0.5 × CLMa.

Examining Table 3, it is clear that FullF outperforms BSIF, with the fusion of MLPQ + 2 × FullF outperforming both MLPQ and FullF. The method FusFULL works poorly only on BIRD (the Bird Song 46 dataset), due to the low performance obtained using harmonic and percussion images in that dataset. In BIRD, FullBSIF obtains a performance of 57.7%, which is much lower than FullBSIF based on the spectrogram images. In the other datasets, the performance of FullBSIF based on the harmonic and percussion images is BIRDzhao 83.7%, ISMIR 76.6%, GTZAN 86.8%, LMD 88.4%, and WHALE 92%.

To reduce the risk of over fitting, we used the same SVM parameters in all the datasets. It is probably the case that CLMa in BIRDzhao would improve more with more fine-tuning of SVM; another difference is that the audio patterns of BIRDzhao are very short. Since our aim is to propose an ensemble that works well in all the tested datasets, we do not use CLM in our set. It is clear that in some specific datasets, CLM works very well, and it can be used in those applications.

The best ensemble of the descriptors tested in [10] is the fusion by sum rule of SVMs trained using the following descriptors: LPQ, ELHF, LBP, RICLBP, HASC, and GF.

In Table 3 the following ensembles of texture descriptors are reported:

- New = $A + B$, where $A$ = MLPQ + 2 × FullF and $B$ is the best ensemble of the descriptors tested in [10] (named Old in this work); before fusion, the scores of $A$ and $B$ are normalised to mean 0 and standard deviation 1.
- NewFULL = $A$ + Old, where $A$ = FusFULL; before fusion, the scores of $A$ and Old are normalised to mean 0 and standard deviation 1.
- E_Full is the sum rule between Old and FullF (the computational expensive MLPQ is not used).
- E_Fullhp is the sum rule between Old and FullFhp.

Clearly, the new set of descriptors improves the performance in [10] in all datasets.

In Table 4 we report the performance obtained by the ensembles combining acoustic and visual features, which are labelled $X$_Ac, where the fusion by sum rule is between $X$, the visual features, and

**Table 4** Comparison with the literature

| Work or method | ISMIR 2004 | GTZAN | LMD | BIRD | BIRDZhao |
|---|---|---|---|---|---|
| E_Full_Ac | **90.9** | **90.8** | 86.2 | 94.8 | 96.3 |
| E_Fullhp_Ac | 90.8 | **90.8** | 86.2 | 94.7 | 96.1 |
| New_Ac | 90.7 | **90.8** | 85.9 | **94.9** | **96.6** |
| NewFULL_Ac | **90.9** | 90.7 | 86.2 | 94.5 | 96.3 |
| [6] (fusion acoustic and visual features) | 90.2 | 89.8 | 85.1 | — | — |
| [10] (fusion acoustic and visual features) | **90.9** | 90.6 | 84.6 | 94.5 | — |
| [4] LBP (visual) | 82.1 | — | 82.3 | — | — |
| [44] | — | — | — | — | 93.6 |
| [24] GLCM (visual) | — | — | 70.7 | — | — |
| [50] GF (visual) | 82.2 | 82.1 | — | — | — |
| [50] Gaussian super vector (GSV) + GF | 86.1 | 86.1 | — | — | — |
| [5] LPQ (visual) | 80.8 | — | — | — | — |
| [5] GF (visual) | 74.7 | — | — | — | — |
| [48] MARSYAS features (acoustic) | — | 61.0 | — | — | — |
| [50] GSV-SVM + MFCC (acoustic) | 79.0 | 82.1 | 74.7 | — | — |
| [40] spectro-temporal features (acoustic) | 89.9 | 87.4 | — | — | — |
| [51] principal Mel-spectrum components (acoustic) | — | — | 82.3 | — | — |
| [52] time constrained sequential patterns (acoustic) | — | — | 77.0 | — | — |
| [53] rhythmic signatures + deep learning (acoustic) | — | — | 77.6 | — | — |
| [54] block-level (acoustic) | 88.3 | 79.9 | 79.9 | — | — |
| [55] joint sparse low rank representation (acoustic) | 85.5 | 89.4 | — | — | — |
| [56] CNN on harmonic, percussive and original spectrogram (visual) | — | 78.0 | — | — | — |
| [57] CNN + RLBP-SVM (visual) | 87.1 | — | **92.0** | — | — |

Ac, the acoustic features reviewed in Section 4 and used in [10] (note: before fusion, the scores of $X$ and Ac are normalised to mean 0 and standard deviation 1). In Table 4 we also compare our best ensemble approaches proposed here with the following:

- [10] (visual features) outperforms [6] (visual features) with a $p$-value of 0.05.
- NewFULL outperforms [10] (visual features) with a $p$-value of 0.05.

For the Bird dataset, we report the performance obtained by the suggested system in [10], i.e. the average best system among all the datasets. An *ad hoc* ensemble for that dataset obtained a better performance.

Unfortunately, even though the new set of texture descriptors proposed here outperforms the set proposed in [10], when these are coupled with acoustic features, the performance improvement, with respect to the full system proposed in [10], is less impressive. Nonetheless, our system performs well, and the method E_Full_Ac is on average the best system. To further validate our ensemble, we compare the different sets of descriptors using the Wilcoxon signed rank test [58]

- [10] (visual features) outperforms [43] (visual features) with a $p$-value of 0.05;
- NewFULL outperforms [10] (visual features) with a $p$-value of 0.05.

To check the error independence of the different SVMs, trained with different descriptors, we have used the Yule's Q-statistic [59]. The values of $Q$ are $[-1, 1]$, where classifiers that erroneously classify the same patterns have $Q < 0$ and those which correctly classify the same pattern have $Q > 0$. The average $Q$-statistic among the proposed ensemble of visual descriptors is 0.853; while the average $Q$-statistic between visual and acoustic features is 0.812. These values show that the SVMs trained with different descriptors provide different information; for this reason, their fusion improves the performance of the stand-alone approaches.

It is important to note that the results of the WHALE dataset are not directly comparable with those reported in the original competition since the only data available to the public is the training data. In the challenge, there was additional data available for the testing set. Furthermore, in the original whale challenge, the winner of the competition achieved an AUC of 0.938 based on an approach that was custom built and focused on feature development via trial and error, iterating many times with the competition dataset. Although the original challenge results and the results from our experiments are not directly comparable, it is interesting to note that our approach, which can be applied to different audio classification tasks, also achieves a good performance.

## 7 Conclusion

In this work, we present a novel system for audio classification tasks that combines powerful texture descriptors that were extracted from different images derived from the audio file and then compared. Some acoustic feature vectors were also evaluated and compared. The experiments reported in this study demonstrate that the fusion of different texture features results in improved performance; however, not all fusions of texture features combine equally well with audio features to improve performance. Nonetheless, our proposed ensemble that combines texture features with audio features managed to obtain results comparable with existing audio signal approaches.

In the future, we plan on adding other datasets to those used in the experiments reported here for a more extensive validation of the proposed ensemble. We also plan on testing this system with more animal and creature sounds. To improve performance further, our next step will be to test deep learning systems and heterogeneous ensembles of classifiers in place of the stand-alone SVM.

## 8 Acknowledgments

We thank the Brazilian Research-support agencies CAPES, CNPq and Fundação Araucária. We also thank the anonymous reviewers for their valuable feedback on the earlier versions of this manuscript.

## 9 References

[1] Russell, J.C., Hasler, N., Klette, R., *et al.*: 'Automatic track recognition of footprints for identifying cryptic species', *Ecology*, 2009, **90**, (7), pp. 2007–2013

[2] Costa, Y.M.G., Oliveira, L.E.S., Koerich, A.L., *et al.*: 'Music genre recognition using spectrograms'. 18th Int. Conf. on Systems, Signals and Image Processing, 2011, pp. 151–154

[3] Haralick, R.M., Shanmugam, K., Dinstein, I.: 'Textural features for image classification', *IEEE Trans. Syst. Man Cybern.*, 1973, **3**, (6), pp. 610–621

[4] Costa, Y.M.G., Oliveira, L.E.S., Koerich, A.L., *et al.*: 'Music genre classification using LBP textural features', *Signal Process.*, 2012, **92**, pp. 2723–2737

[5] Costa, Y.M.G., Oliveira, L.E.S., Koerich, A.L., *et al.*: 'Music genre recognition using Gabor filters and LPQ texture descriptors'. 18th Iberoamerican Congress on Pattern Recognition, 2013, pp. 67–74

[6] Nanni, L., Costa, Y.M.G., Lumini, A., *et al.*: 'Combining visual and acoustic features for music genre classification', *Expert Syst. Appl.*, 2016, **45**, pp. 108–117

[7] Montalvo, A., Costa, Y.M.G., Calvo, J.R.: 'Language identification using spectrogram texture', in Cancela, H., Cuadros-Vargas, A., Cuadros-Vargas, E. (Eds.): '*Progress in pattern recognition, image analysis, computer vision, and applications*' (Springer, Berlin, 2015), pp. 543–550

[8] Lucio, D.R., Costa, Y.M.G.: 'Bird species classification using spectrograms'. The XLI Latin American Computing Conf. (CLEI), Arequipa, Peru, 2015

[9] Nanni, L., Costa, Y.M.G., Lucio, D.R., *et al.*: 'Combining visual and acoustic features for bird species classification'. 28th IEEE Int. Conf. on Tools with Artificial Intelligence, 2016

[10] Nanni, L., Costa, Y.M.G., Lucio, D.R., *et al.*: 'Combining visual and acoustic features for audio classification tasks', *Pattern Recognit. Lett.*, 2017, **88**, (March), pp. 49–56

[11] Deuser, L.M., Middleton, D., Plemonset, T.D., *et al.*: 'On the classification of underwater acoustic signals. II. Experimental applications involving fish', *J. Acoust. Soc. Am.*, 1979, **65**, (2), pp. 444–455

[12] Giryn, A., Rojewski, M., Somla, K.: 'About the possibility of sea creature species identification on the basis of applying pattern recognition to echo-sounder signals'. Meeting on Hydroacoustical Methods for the Estimation of Marine Fish Population, 1979, pp. 455–466

[13] Chesmore, E.D.: 'Application of time domain signal coding and artificial neural networks to passive acoustical identification of animals', *Appl. Acoust.*, 2001, **62**, pp. 1359–1374

[14] Lee, C., Chou, C., Han, C., *et al.*: 'Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis', *Pattern Recognit. Lett.*, 2006, **27**, pp. 93–101

[15] Molnár, C., Kaplan, F., Roy, P., *et al.*: 'Classification of dog barks: a machine learning approach', *Animal Cogn.*, 2008, **11**, pp. 389–400

[16] Pachet, F., Zils, A.: 'Automatic extraction of music descriptors from acoustic signals'. 5th Int. Conf. on Music Information Retrieval (ISMIR), 2004

[17] Bardeli, R., Wolff, D., Kurth, F., *et al.*: 'Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring', *Pattern Recognit. Lett.*, 2010, **31**, pp. 1524–1534

[18] Cheng, J., Sun, Y., Ji, L.: 'A call-independent and automatic acoustic system for the individual recognition of animals: a novel model using four passerines', *Pattern Recognit.*, 2010, **43**, pp. 3846–3852

[19] Lucio, D.R., Costa, Y.M.G.: 'Bird species classification using visual and acoustic features extracted from audio signal'. Int. Conf. of the Chilean Computer Science Society, Valparaiso, Chile, 2016

[20] Urazghildiiev, I.R., Clark, C.W., Krein, T.P., *et al.*: 'Detection and recognition of north atlantic right whale contact calls in the presence of ambient noise', *IEEE J. Ocean. Eng.*, 2009, **34**, (3), pp. 358–368

[21] Spaulding, E., Robbins, M., Calupca, T., *et al.*: 'An autonomous, near-real-time buoy system for automatic detection of North Atlantic right whale calls'. 157th Meeting of the Acoustical Society of America, 2009

[22] Fitzgerald, D.: 'Harmonic/Percussive separation using median filtering'. 13th Int. Conf. on Digital Audio Effects (DAFx-10), Graz, Austria, 2010

[23] McAfee, B., Raffel, C., Liang, D.: 'Librosa: audio and music signal analysis in python'. Proc. 14th Python in Science Conf. (SCIPY), Austin, Texas, 2015

[24] Costa, Y.M.G., Oliveira, L.E.S., Koerich, A.L., *et al.*: 'Comparing textural features for music genre classification'. IEEE World Congress on Computational Intelligence, 2012, pp. 1867–1872

[25] Umesh, S., Cohen, L., Nelson, D.: 'Fitting the mel scale'. Int. Conf. on Acoustics, Speech, and Signal Processing, 1999, pp. 217–220

[26] Ojala, T., Pietikainen, M., Maeenpaa, T.: 'Multiresolution gray-scale and rotation invariant texture classification with local binary patterns', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2002, **24**, (7), pp. 971–987

[27] Ojansivu, V., Heikkila, J.: 'Blur insensitive texture classification using local phase quantization'. Int. Conf. on Image and Signal Processing, 2008, pp. 236–243

[28] Zhao, G., Ahonen, T., Matas, J., *et al.*: 'Rotation-invariant image and video description with local binary pattern features', *IEEE Trans. Image Process.*, 2012, **21**, (4), pp. 1465–1467

[29] Nosaka, R., Suryanto, C.H., Fukui, K.: 'Rotation invariant co-occurrence among adjacent LBPs'. ACCV Workshops, 2012, pp. 15–25

[30] Nanni, L., Brahnam, S., Lumini, A., *et al.*: 'Ensemble of local phase quantization variants with ternary encoding', in '*Local binary patterns: new variants and applications*' (Springer, Berlin, 2014)

[31] San Biagio, M., Crocco, M., Cristani, M., *et al.*: 'Heterogeneous auto-similarities of characteristics (HASC): exploiting relational information for classification'. IEEE Computer Vision (ICCV'13), 2013, pp. 809–816

[32] Kannala, J., Rahtu, E.: 'Bsif: binarized statistical image features'. 21st Int. Conf. on Pattern Recognition (ICPR 2012), Tsukuba, Japan, 2012, pp. 1363–1366

[33] Nanni, L., Paci, M., Santos, F.L.C., *et al.*: 'Texture descriptors ensembles enable image-based classification of maturation of human stem cell-derived retinal pigmented epithelium', *PLoS One*, 2016, **11**, (2) p. e0149399

[34] Zhu, Z., You, X., Chen, C.L.P., *et al.*: 'An adaptive hybrid pattern for noise-robust texture analysis', *Pattern Recognit.*, 2015, **48**, pp. 2592–2608

[35] Song, T., Meng, F.: 'Letrist: locally encoded transform feature histogram for rotation-invariant texture classification', *IEEE Trans. Circuits Syst. Video Technol.*, 2017, **PP**, (99)

[36] Nanni, L., Brahnam, S., Lumini, A.: 'Combining different local binary pattern variants to boost performance', *Expert Syst. Appl.*, 2011, **38**, (5), pp. 6209–6216

[37] Wang, Q., Li, P., Zhang, L., *et al.*: 'Towards effective codebookless model for image classification', *Pattern Recognit.*, 2016, **59**, pp. 63–71

[38] Schroeder, M.R., Atal, B.S., Hall, J.L.: 'Optimizing digital speech coders by exploiting masking properties of the human ear', *J. Acoust. Soc. Am.*, 1979, **66**, (6), pp. 1647–1652

[39] Fagerlund, S.: 'Bird species recognition using support vector machines', *EURASIP J. Appl. Signal Process.*, 2007, **2007**, pp. 1–8

[40] Lim, S.-C., Lee, J.-S., Jang, S.-J., *et al.*: 'Music-genre classification system based on spectro-temporal features and feature selection', *IEEE Trans. Consum. Electron.*, 2012, **58**, (4), pp. 1262–1268

[41] Vilches, E., Escobar, I.A., Vallejo, E.E., *et al.*: 'Data mining applied to acoustic bird species recognition'. Int. Conf. on Pattern Recognition, Hong Kong, 2006, pp. 400–403

[42] Chou, C.-H., Liu, P.-H.: 'Bird species recognition by wavelet transformation of a section of birdsong'. Symp. and Workshops on Ubiquitous, Autonomic and Trusted Computing, 2009, pp. 189–193

[43] Lopes, M.T., Gioppo, L.L., Higushi, T.T., *et al.*: 'Automatic bird species identification for large number of species'. IEEE Int. Symp. On Multimedia (ISM), 2011

[44] Zhao, Z., Zhang, S.-H., Xu, Z.-Y., *et al.*: 'Automated bird acoustic event detection and robust species classification', *Ecological Inf.*, 2017, **39**, pp. 99–108

[45] Silla, C.N.Jr., Koerich, A.L., Kaestner, C.A.A.: 'The latin music database'. 9th Int. Conf. on Music Information Retrieval, Philadelphia, USA, 2008, pp. 451–456

[46] Flexer, A.: 'A closer look on artist filters for musical genre classification', *World*, 2007, **19**, (122), pp. 16–17

[47] Ong, B., Serra, X., Streich, S., *et al.*: '*ISMIR 2004 audio description contest*' (Music Technology Group-Universitat Pompeu Fabra, Barcelona, Spain, 2006)

[48] Tzanetakis, G., Cook, P.: 'Musical genre classification of audio signals', *IEEE Trans. Speech Audio Process.*, 2002, **10**, (5), pp. 293–302

[49] Costa, C.H.L., Valle, J.D.Jr., Koerich, A.L.: 'Automatic classification of audio data'. Int. Conf. on Systems, Man, and Cybernetics, 2004, pp. 562–567

[50] Wu, M.-J., Chen, Z.-S., Jang, J.-S.R., *et al.*: 'Combining visual and acoustic features for music genre classification'. Int. Conf. on Machine Learning and Applications, 2011

[51] Hamel, P.: 'Pooled features classification'. Submission to Audio Train/Test Task of MIREX, 2011

[52] Ren, J.-M., Jang, J.-S.R.: 'Discovering time-constrained sequential patterns for music genre classification', *IEEE Trans. Audio Speech Lang. Process.*, 2012, **20**, (4), pp. 1134–1144

[53] Pikrakis, A.: 'Audio latin music genre classification: a MIREX submission based on a deep learning approach to rhythm modelling', 2013

[54] Seyerlehner, K., Schedl, M., Pohle, T., *et al.*: 'Using block-level features for genre classification, tag classification and music similarity estimation'. 6th Annual Music Information Retrieval Evaluation eXchange (MIREX-2010), Utrecht, The Netherlands, 2010

[55] Panagakis, Y., Kotropoulos, C., Arce, G.R.: 'Music genre classification using locality preserving non-negative tensor factorization and sparse representations'. 10th Int. Conf. on Music Information Retrieval, 2009, pp. 249–254

[56] Gwardys, G., Grzywczak, D.: 'Deep image features in music information retrieval', *Int. J. Electron. Telecommun.*, 2014, **60**, (4), pp. 321–326

[57] Costa, Y.M.G., Oliveira, L.E.S., Silla, C.N.Jr.: 'An evaluation of convolutional neural networks for music classification using spectrograms', *Appl. Soft Comput.*, 2017, **52**, pp. 28–38

[58] Demšar, J.: 'Statistical comparisons of classifiers over multiple data sets', *J. Mach. Learn. Res.*, 2006, **7**, pp. 1–30

[59] Kuncheva, L.I., Whitaker, C.J.: 'Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy', *Mach. Learn.*, 2003, **51**, (2), pp. 181–207