Ecological Informatics 48 (2018) 187–197

Contents lists available at ScienceDirect

Ecological Informatics

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

Bird species identification using spectrogram and dissimilarity approach

Rafael H.D. Zottesso[a,b], Yandre M.G. Costa[a,*], Diego Bertolini[a,c], Luiz E.S. Oliveira[d]

[a] PCC/DIN, State University of Maringa (UEM), Av. Colombo, 5790, Bloco C56, Jd. Universitario, Maringa, PR 87020-900, Brazil
[b] Federal Institute of Parana (IFPR), Rua Jose Felipe Tequinha, 1400 - Jd. das Nacoes, Paranavai, PR 87703-536, Brazil
[c] Federal Technological University of Parana (UTFPR), Via Rosalina Maria dos Santos, 1233, Campo Mourao, PR 87301-899, Brazil
[d] PPGInf, Federal University of Parana (UFPR), Av. Cel Francisco H. dos Santos, 100, Jd. das Americas, Curitiba, PR 80530-000, Brazil

ARTICLE INFO

Keywords:
Bird species identification
Dissimilarity
Spectrogram
Texture

ABSTRACT

In this work, we investigate bird species identification starting from audio record-ings on eight quite challenging subsets taken from the LifeClef 2015 bird task contest database, in which the number of classes ranges from 23 to 915. The classification was addressed using textural features taken from spectrogram im-ages and the dissimilarity framework. The rationale behind it is that by using dissimilarity the classification system is less sensitive to the increase in the number of classes. A comprehensive set of experiments confirms this hypothesis. Although they cannot be directly compared to other results already published because in this application domain the works, in general, are not developed ex-actly on the same dataset, they overcome the state-of-the-art when we consider the number of classes involved in similar works. In the hardest scenario, we obtained an identification rate of 71% considering 915 species. We hope the subsets proposed in this work will also make future benchmarking possible.

1. Introduction

The bird monitoring has an important role to play in the control of migratory flux of birds and in the bird species identification tasks. Regarding migratory flux, Negret, (1988) points out that each bird species has a particular migratory flux along the different seasons of the year, which makes its identification even more challenging. Faria et al., (2006) describe several monitoring methods aiming to provide identification of existing birds species: line transect direct observation, bird capture using mist nets, bird listening points, and identification based on bird vocalization.

The mist nets usage is among the most widely used strategies to perform bird species identification (Faria et al., 2006). Mist nets are commonly made of polyester or nylon mesh dangling between two poles, similar to a badminton net. If suitably installed, the net is supposed to be invisible, consisting of an important tool for several purposes, like monitoring species diversity, relative abundance, population size, and demography. Although the usage of mist nets is an efficient way to capture individuals in its own habitat, it can hurt the animals when they collide with the net. In extreme cases, fragile animals may even die. By this way, taking into account bird welfare concernments, experts suggest that non-invasive techniques must be used from the data collection up to the species identification. Besides that, it is very unlikely that all the aimed species would fly over the area where the mist nets are placed.

With the technological developments, several audio recording devices became accessible, smaller, and frequently used. By this way, the monitoring systems became able to capture bird calls and songs in their natural habitat, in a less invasive way, without the need for physical contact (Faria et al., 2006; Conway, 2011; Schuchmann et al., 2014).

The bird species identification starting from their vocalization is a time-consuming task, which can be divided into three main steps: equipment set up, sound recording, and data annotation (Conway, 2011). One example of how important is the use of technology in this kind of application can be found in the project called “Os Sons do Pantanal”[1] (from the Portuguese “the sounds of Pantanal”), developed in Brazil. In this project, researchers aim to perform bird acoustic monitoring in an important zone, which includes the Pantanal biome (Schuchmann et al., 2014). Taking into account that the project is developed on a quite huge area, the use of technological devices is crucial to make the project viable, especially regarding data collection.

Even with the difficulties and challenges to record the bird sounds, the bird sounds databases became more accessible to the research community, fostering the development of new investigations related to

* Corresponding author.
E-mail addresses: rafael.zottesso@ifpr.edu.br (R.H.D. Zottesso), yandre@din.uem.br (Y.M.G. Costa), diegobertolini@utfpr.edu.br (D. Bertolini), luiz.oliveira@ufpr.br (L.E.S. Oliveira).

[1] http://www.ufmt.br/ufmt/site/noticia/visualizar/16901/juliomuller.

https://doi.org/10.1016/j.ecoinf.2018.08.007
Received 2 April 2018; Received in revised form 5 July 2018; Accepted 29 August 2018
Available online 08 September 2018
1574-9541/

species identification using bird vocalizations. The Xeno-canto[2] project is one example of this. In that project, a large database of bird vocalization samples is shared such a way that researchers can make use of it for the development of their works. The database is fed by professional or amateurs ornithologists from all over the world. Recently, several works have been developed using datasets taken from Xeno-Canto (Lopes et al., 2011a; Lopes et al., 2011b; Marini et al., 2015; Lucio & Costa, 2015; Zottesso et al., 2016).

The bird species identification can be addressed as a typical pattern recognition problem, i.e., pre-processing, feature extraction (acoustic or visual) and classification (Fagerlund, 2007). To the best of our knowledge, the first works on bird species identification were reported in the 1990s (Anderson et al., 1996) (Kogan & Margoliash, 1998).

In this study, we extend some works previously developed aiming to perform bird species classification using spectrograms (Lucio & Costa, 2015; Zottesso et al., 2016). Spectrograms has already been successfully used on different audio classification tasks (Costa et al., 2011; Nanni et al., 2016; Costa et al., 2012a; Freitas et al., n.d.). This time, we address bird species classification taking into account a quite challenging scenario with a much larger number of classes. For this purpose, the dissimilarity framework will be used. One of the main advantages of using dissimilarity is that it is not necessary to retrain the classification model each time new classes (bird species) are considered in the classification problem.

The experiments were performed on the database used in the LifeCLEF bird identification task 2015.[3] Several subsets of the original database were used, and the number of classes in these subsets ranges from 23 to 915 species. The results obtained using the dissimilarity framework are comparable to the state-of-the-art, and they overcame the results without dissimilarity on all the datasets. Furthermore, the classifiers developed using dissimilarity demonstrated to be less sensitive to the increase in the number of classes. In the hardest scenario, we obtained an identification rate of 71% considering 915 species.

In order to encourage other authors to compare their approaches with this work, the list of audio recordings contained in each subset and the spectrogram images used here were made freely available.[4]

The reminding of this work is organized as follows: Section 2 presents some related work described in the literature, Section 4 describes the organization of the database used in the experiments and some details about preprocessing and feature extraction, Section 5 details the dissimilarity framework, Section 6 describes the experiments performed and some discussion about the obtained results. Finally, the main conclusions are drawn and some future works are pointed.

2. Related works

One of the first papers on bird species classification using sounds was pro-posed by Anderson et al. (Anderson et al., 1996). In that work, the authors employed Dynamic Time Warping (DTW) algorithm to perform classification using only 2 bird species. The identification rate reported was 98.1%. Kogan and Margoliash (Kogan & Margoliash, 1998) evaluated Hidden Markov Models (HMMs) and DTW on a database composed of samples belonging to two different species. The best accuracy was 92.5%. Cai et al., (2007) presented a work using Mel-Frequency Cepstral Coefficients (MFCC) extracted from bird song. In this case, the authors used only the corner known as a call. Neural networks were used in the classification step and the accuracies were 98.7% and 86.8% using 4 and 14 species, respectively.

Lopes et al., (2011b) performed several experiments varying features and classifiers. The used database consisted of sounds from three species that were divided into five folds to perform accuracy by cross-validation. The experiments were carried out on bird songs obtained from the Xeno-Canto website. The best identification rates were 79.2% using the full audio and 99.7% using pulses (i.e. small pieces of the sound where the amplitude is highlighted). The authors achieved this performance using Multilayer Perceptron (MLP) and features extracted with MARSYAS[5] framework based on timbre, including MFCC.

Marini et al., (2015) employed SVM classifier on a Xeno-Canto dataset with 422 audio samples labeled according to 50 species and divided into five folds. The identification rate was calculated according to a Top-N best hypothesis between 1 and 10, resulting in 45.9% on Top-1 and 86.97% on Top-10. The audio signals were preprocessed to remove the quiet spaces between songs.

Lucio & Costa, (2015) describe a bird species recognition approach using spectrograms generated by the corner of birds. In their work, 46 species taken from the Xeno-Canto database and three texture descriptors (LBP, LPQ and Gabor Filter) were considered. The best identification rate achieved was 77.65% using Gabor Filter and SVM classifier. However, it is important to note that all the audio signals used in this work were manually segmented in order to find the regions of interest with bird songs and without external noises. Zottesso et al. extended the work presented in (Lucio & Costa, 2015) by automatically segmenting the input signal. The same image texture descriptors were used. The authors reported an identification rate of 78.97% using SVM classifier and texture features extracted using Gabor Filters.

Albornoz et al., (2017) described experiments on a dataset composed of audio recordings from South America labeled into 25 different species. Part of these samples was taken from Xeno-canto database. The audio signal was preprocessed using Wiener Filter to obtain noise reduction. Moreover, the Rabiner and Schafer method was applied to detect acoustic activity in order to identify the sound of birds. OpenSMILE toolkit was used to extract features and different classifiers were evaluated. The best accuracy was 89.32%, achieved using MFCC features classified with Multilayer Perceptron.

Zhao et al., (2017) addressed bird species identification using samples of 11 bird species taken from Xeno-canto. The authors segmented the audio signals using a scheme based on Gaussian Mixture Model (GMM) to select the acoustic events more representative. The spectrograms of these events were submitted to a Mel band-pass filter bank. The output of each sub band was then parameterized by an Autoregressive (AR) model. Finally, it was used as features submitted to an SVM classifier. The performance achieved was 93.9% for Precision and 91.7% to Recall to the unknown classes.

Chou et al., (2007) addressed bird species classification on a dataset containing samples labeled on 420 different species taken from a commercial Compact Disk (CD). In this experiment, the corner was segmented into syllables and two-thirds of each vocalization were randomly selected to compose the training set and one-third to test set. Each set of syllables was modeled by an HMM to represent their features. The authors used the Viterbi algorithm to classify the test set. The best identification rate achieved was 78.3%.

Ntalampiras, (2018) proposed an approach using transfer learning. In this case, the author used music to build the probabilistic models. Ntalampiras, (2018) employed 10 bird species from Xeno-Canto database to evaluate the proposed approach, using only bird calls with a duration between 1249 and 1651 seconds and the identification rate achieved was 92.5%

In the LifeCLEF 2016[6] (Goëau et al., 2016) bird identification task, many teams of competitors employed concepts of deep learning in their proposals. Using the same database already used in BirdCLEF 2015 competition, the winner team Sprengel et al., (2016) achieved an improvement of 14.1 percentage points of MAP score when compared to

[2] http://www.xeno-canto.org/.

[3] http://www.imageclef.org/lifeclef/2015/bird.

[4] List of audio recordings and spectrogram images available at: https://sites.google.com/din.uem.br/lifeclef2015birdtasksubsets.

[5] Available at http://marsyas.info/.

[6] http://www.imageclef.org/lifeclef/2016/bird.

the winner of 2015 edition and this result was obtained by using deep learning in the proposed approach. The authors used a convolutional neural network with five convolutional layers and one dense layer for. In a preprocessing step, the audio was decomposed to separate regions of interest, which contain audible bird songs, and noisy regions, which are not supposed to have bird sounds. After that, spectrograms of signal from the regions of interest and from the noisy parts were computed. Later, the spectrograms are divided into 3-seconds pieces. In this way, these pieces of the signal part were used as samples in training/test set in the neural network. The proposed approach reached an average accuracy of 68.6% using the main species and 55.5% when all species are employed.

More recently, one can observe some impressive results obtained in LifeCLEF Bird Identification task 2017.[7] Particularly, one can remark some works that somehow used convolutional neural networks (CNN) on the proposed approach. Before briefly describing some details of those works, it is important to mention some details about the database used in that contest. The dataset provided for training consists of 36,496 audio recordings containing 1,500 different bird species.

In the aforementioned contest, Fritzler et al. (2018), Linda Cappellato and Ferro, 2017 used a pre-trained Inception-v3 convolutional neural network to identify bird species in BirdCLEF 2017 contest. For that purpose, the authors fine-tuned the network by using 36,492 audio recordings made available for the participants of the contest. After transforming the audio recordings into spectrograms, they applied bandpass filtering, noise filtering, and silent region removal. Data augmentation was also performed, and the authors claim that results obtained by fine-tuning a pre-trained CNN are better than those obtained by training a CNN from scratch. The obtained mean average precision (MAP) score was 56.7% for traditional records and the MAP score for records with background species on the test dataset was 49.6%.

Kahl et al. (2018) used a variety of CNNs to generate features extracted from spectrograms of field recordings. All the 36,496 audio recordings available on BirdCLEF 2017 training dataset were used. The authors also applied data augmentation by using vertical roll, Gaussian noise, noise samples addition, and batch augmentation. The best result was obtained by averaging the results of seven different CNN models created. They obtained a MAP score of 60.5% (official score) and 68.7% considering only foreground species.

Fazekas et al. (2018) use a multi-modal Deep Neural Network (DNN) starting from audio recordings and metadata as input. The audio is fed into a Convolutional Neural Network (CNN) using four convolutional layers. The additionally provided metadata is processed using fully connected layers. The authors also used data augmentation and, in the best case, they obtained a MAP score of 57.9% considering only the main species, and a MAP score of 51.1% on the traditional recordings considering also the background species.

Finally, we briefly describe the Soundception (Sevilla and Glotin, 2018), the classification scheme which scored highest on all tasks in the BirdClef2017 challenge. The creation of Soundception was based on the deep convolutional network Inception-v4 tailored aiming to boost its performance on bioacoustic classification problems. The authors also used a data augmentation strategy and two attention mechanisms: a temporal attention into the auxiliary branch, and a time-frequency attention in the main branch. Soundception obtained a MAP score of 71.4% on bird species task.

As we can observe, comparing these works is not straightforward, mainly because of the variation in the number of classes used. Some works employ 2 classes while another uses 1,500 classes. The number of samples used for training and testing also varies greatly. Hence, we may notice a huge variation on the identification rates, which range from 45.2% to 99.7%. Just in case, we describe in Table 1 some summarized information about related works described in this section in chronological order.

3. Proposed method

The general scheme of the proposed method is illustrated in Fig. 1. Through-out this section, we describe some details about the database used in this work and introduce details about the main steps contained in the approach proposed here.

Xeno-canto[8] is a website dedicated to sharing bird sounds from all around the world. It is also a collaborative project in which people can submit their recordings of bird vocalization and contribute in identifying species. In addition, it aims to popularize recordings with bird sounds, improve accessibility to corners and disclose information about bird songs.

Due to the great diversity of sounds made available by the Xeno-Canto project, the LifeClef 2015 Bird Task (competition in which the goal is to perform bird species identification based on bird vocalization) presented a database of bird sounds containing 999 species taken from Xeno-Canto repository, establishing some important requirements for the classification task aiming to make it as close as possible to real-world applications:

- The audio samples of the same species were obtained from different birds present in different regions of the planet;
- Sounds have been recorded by several users who may not have used the same combination of microphone and recording device;
- Audio signals were obtained from recordings made at various seasons of the year and at different times of the day, besides having a variety of noises in the environment (other birds, buzzing insects, etc.);
- Species with only one bird's song while others species have over 50 samples;
- In the same species, file sizes can range from 119 KB to 17.8 MB;
- There are many samples with duration of only one second;
- Occurrence of silent time interval where there is no bird song in the audio signal.

In addition to the audio signals, information about the samples was made available. Among them, we can highlight the species of bird, which will be used in the classification of samples, and the type of vocalization, which can be found as song or calls. Catchapole and Slater (Catchpole & Slater, 2003) explain the differences between song and calls. According to them, bird song tends to be longer, complex and usually produced by males. It also appears spontaneously and is often produced at long intervals during the day and more often at some times of the year. On the other hand, bird call tends to be shorter, simpler and produced by both genders throughout the year. The bird call is usually related to specific functions like fights, threats, alarms and other types of behavior. In this way, the samples of bird call were discarded because they are not so typical of a species as the bird songs are.

The LifeClef 2015 Bird Task database is composed of bird songs taken from species of South America. The complete database contains 33,203 samples from 999 possible species. The audio signal was standardized at 44.1 kHz in 16-bit and it was made available in WAV format.

Due to the variation in the time and quantity of the samples available for each species in LifeClef 2015 database, in this work, we propose different subsets based on the duration time (in seconds) of the samples and in the number of samples per species. We believe these subsets can be used as benchmarks for further comparison.

To create the subset #1, we performed a search on the LifeClef 2015 Bird Task database by filtering only song vocalizations aiming to find

[7] http://www.imageclef.org/lifeclef/2017/bird.

[8] https://www.xeno-canto.org/.

Table 1
Summary of the works described in the state-of-the-art.

Reference	Year	Feature/input and classifier	Number of species	Database	Identification rate (%)
(Anderson et al., 1996)	1996	Spectrogram and DTW	2	Sounds from animals housed in wire cages	98.1[a]
(Kogan & Margoliash, 1998)	1998	DTW and HMM	2	Vocalizations recorded in laboratory	92.5[a]
(Cai et al., 2007)	2007	MFCC and MLP	14	Birds in Backyards, Australian Bird Calls, Voices of Subtropical Rainforests and Data collected from Samford sensors	86.8[a]
(Chou et al., 2007)	2007	HMM and Viterbi algorithm	420	Commercial Dataset	78.3[a]
(Lopes et al., 2011b)	2011	MFCC, KNN, SVM, MLP and j4.8	3	Xeno-Canto	99.7[b]
(Marini et al., 2015)	2015	MFCC and SVM	50	Xeno-Canto	45.9[a]
(Lucio & Costa, 2015)	2015	LBP, LPQ, Gabor filters and SVM	46	Xeno-Canto	77.6[a]
(Zottesso et al., 2016)	2016	LBP, SVM and Genetic algorithm	45	Xeno-Canto	78.9[a]
(Sprengel et al., 2016)	2016	Spectrogram and CNN	999	LifeCLEF 2016	68.8[*c] 55.5[**c]
(Albornoz et al., 2017)	2017	LLD, MFCC, SVM, MLP and random forest	25	Xeno-Canto and Birds of Argentina and Uruguay	89.3[a]
(Zhao et al., 2017)	2017	Gaussian Mixture Model and SVM	11	Xeno-Canto	93.9[***d]
(Fritzler et al., 2018)	2017	Spectrogram and Inception-v3 CNN	1500	LifeCLEF 2017	56.7[*c] 49.6[**c]
(Kahl et al., 2018)	2017	Spectrogram and CNN	1500	LifeCLEF 2017	68.7[*c] 60.5[**c]
(Fazekas et al., 2018)	2017	DNN from audio and metadata	1500	LifeCLEF 2017	57.9[*c] 51.1[**c]
(Sevilla and Glotin, 2018)	2017	Spectrogram and Soundseption	1500	LifeCLEF 2017	71.4[*c] 61.6[**c]
(Ntalampiras, 2018)	2018	HMM and Reservoir Network	10	Xeno-Canto	92.5[*a]

[*] Only foreground species.
[**] With background species.
[***] To the unknown classes.
[a] Accuracy
[b] F-measure.
[c] Mean Average Precision (MAP).
[d] Recall.

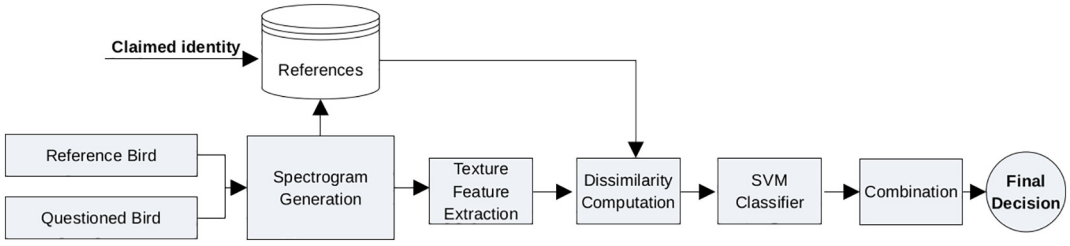

Fig. 1. Proposed method general scheme.

Table 2
Subsets of the LifeClef database proposed in this work.

Subset	Minimum time (s)	Samples by species[a]	Species found	Total samples
#1	30	10	23	230
#2	20	10	48	480
#3	15	10	88	880
#4	10	10	180	1800
#5	05	10	349	3490
#6	05	06	614	3684
#7	05	04	772	3088
#8	05	02	915	1830

[a] Randomly taken samples.

species which have at least 10 audio samples with at least 30-seconds, thus, 23 species were found. Then, we run a search for species with at least 10 audio samples that lasted for 20-seconds or more, creating the subset # 2, with 48 species. The same process was performed repeatedly with the minimum duration requirements of the audio samples and the quantity of them per species to create eight subsets. The number of species and samples found in each subset is presented in Table 2.

Using the criteria presented in Table 2, classes which had only one sample were discarded. Hence, it was not possible to keep all the 999 species found in the original database, but rather 915.

Using the same criteria to create other subsets, the species found in subset #1 also are part of subset #2, since the set of samples with 20 seconds or more also includes cases that are 30 seconds or longer. However, the samples are not necessarily the same, as we have used a random selection of samples. Thus, the species of subset #2 are part of subset #3, and so on. Table 3 describes the eight subsets proposed in this work.

These eight subsets were divided into folds, each one containing a single sample per species selected randomly. Thus, the training and test sets are balanced, avoiding that the trained model had more ability to classify some species than others.

3.1. Noise reduction

As aforementioned, the Life Clef 2015 Bird Task dataset was

Table 3
Description of the generated subsets.

Subset	Conjunction	Number of species
#1	23	23
#2	species in #1 + 25	48
#3	species in #2 + 40	88
#4	species in #3 + 92	180
#5	species in #4 + 169	349
#6	species in #5 + 265	614
#7	species in #6 + 158	772
#8	species in #7 + 143	915

composed of audio recordings taken from the Xeno-Canto repository. In Xeno-Canto database is common to note that the audio signals provided do not have a pat-tern regarding the environment in which they are recorded or recording devices used. There are audio recordings in both unpopulated and populated regions, in urban areas or close to civilization. In this way, other sources of sounds, such as winds, waterfalls, streams, overlapping of audio with other animals or insects, cars, people, and more can be noted alongside the birds singing. This becomes a big problem since most audio samples have a large amount of noise.

In order to minimize the occurrence of noise and highlight the sound of the birds, this work uses a strategy to reduce the prevalent noise in the audio signals used in the classification process similar to that used by Zottesso et al., (2016). First, a sample of the signal is collected in order to identify its noise profile. This sampling is done based on the first 400 milliseconds of the audio signal (size set empirically). Once the signal noise profile is defined, the signal noise reduction as a whole occurs. This reduction is based on the subtraction of the identified noisy profile of the original signal. To perform this step, we use the noise removal tool provided with the software Sound eXchange (SOX)[9] version 14.4.1. Fig. 2(a) and (b) illustrate respectively the spectrogram of the same audio signal before and after the noise reduction process used here.

In this work, all subsets defined in the previous subsection underwent a noise reduction step.

3.2. Automatic segmentation of audio signal

Besides the presence of noise, most of the audio signals available have some time stretches in which there is no sound of birds. This time interval elapses between the bird's song since it does not sing continuously throughout the entire audio signal. Thus, the application of a method to detect segments of interest becomes extremely important in order to obtain better results, because, according to Evangelista et al., (2014), it is necessary to use the most representative parts of the audio signal in order to obtain better results in the classification stage. Even though in some specific cases, this time interval could be used as a feature for bird species identification.

To extract these important segments, the segmentation technique used by Zottesso et al., (2016) was applied to all samples of the subsets used in this work. According to the authors, the process basically consists of:

- Extraction of two sequences with audio signal features, one based on Signal Energy and another in the Spectral Centroid;
- For each sequence, two thresholds are estimated dynamically using the histogram of sequence values and local maximum;
- A threshold criterion is applied to separate the segments that have meaningful sound content from the others with little or no sound content;
- Joining the segments identified in the previous step.

[9] http://sox.sourceforge.net/

Fig. 3 shows one example of the audio signal before and after the seg-mentation process.

In this work, when referring to the segmented database, it means that the audio samples that are part of it have passed through this segmentation approach.

3.3. Zoning approach

In the experiments carried out in this work, it was observed that the texture present in the Spectro of bird songs taken from different species does not present a uniform content along the time and frequency axes. Thus, a strategy was proposed to divide the spectrograms into zones so that it was possible to highlight information in specific regions of the spectrogram.

The idea of image zoning is to extract local information from each region and try to highlight the features of different frequency bands (Costa et al., 2011). One specific feature vector is taken from each region created by linear zoning and consequently, this vector will be used to train a classifier. One classifier is created individually for each zone, and their final scores can be combined based on some fusion rules proposed by (Kittler et al., 1998).

Two types of zoning are experimented in this work, vertical and horizontal. Vertical zoning aims to segment the spectrogram with respect to time. Horizontal zoning makes it possible to extract features in different frequency bands.

In the vertical zoning, zones of the same size are established in the image of the spectrogram, corresponding to periods of time with the same duration. The size of each zone depends on the length of the audio signal and the number of vertical zones established (3, 5 or 9). The Fig. 4(a) illustrates a division into three vertical zones.

The use of horizontal zoning allows describing contents of the signal which remains at specific wave frequencies. In other words, this strategy aims to capture local features.

The horizontal zoning can be performed in the linear way or taking into account the Mel scale of frequencies (non-linear way). Linear zones divide the image into regions of equal size. The limits depend on the number of zones that are created. Fig. 4(b) shows the zoning of the spectrogram into three linear zones. Some values of linear zones have been defined empirically in this work, they are 1, 3, 5, and 10.

In Mel-frequency, the divisions represent frequency bands that are directly related to the frequencies perceived by humans. There are 15 different frequency bands (regions) and each has its limits, which in Hertz (Hz) are: 40, 161, 200, 404, 693, 867, 1,000, 2,022, 3,000, 3,393, 4,109, 5,526, 6,500, 7,743 and 12,000 (Umesh et al., 1999). The higher limit in the zoning of the image depends on the frequency upper limit set in the generation of the spectrogram from the audio signal. Fig. 5 exemplifies a spectrogram with a frequency limit of 11,000 Hz and the creation of 15 regions according to the Mel-frequency division.

3.4. Feature extraction

Texture is notably the main visual content one can see in the spectrogram image. In light of this, we decided to use successful texture descriptors presented in the image processing literature. Taking into account the good performances obtained in previous works in which textural content of spectrograms have been used, in this work we decided to use Local Binary Pattern (LBP), Robust Lo-cal Binary Pattern (RLBP), and Local Phase Quantization (LPQ). Table 4 describes the dimensionality of the feature vectors produced by these texture descriptors.

In the following sub-sub-sections one can find a brief report about how these texture descriptors were used in this work.

3.4.1. Local binary pattern (LBP)

Local Binary Pattern is a well-known texture descriptor that have been successfully used in works developed on different application

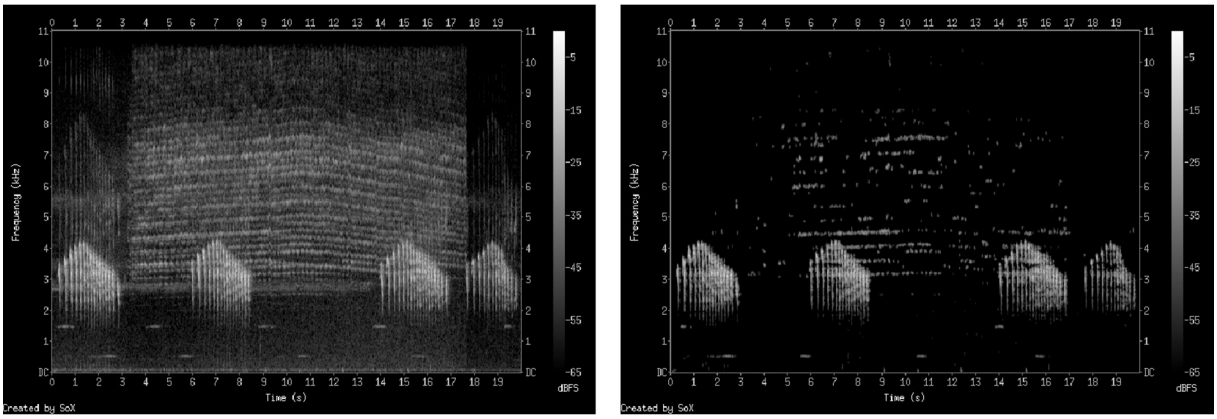

Fig. 2. Example of audio signal spectrogram before and after the noise reduction process.

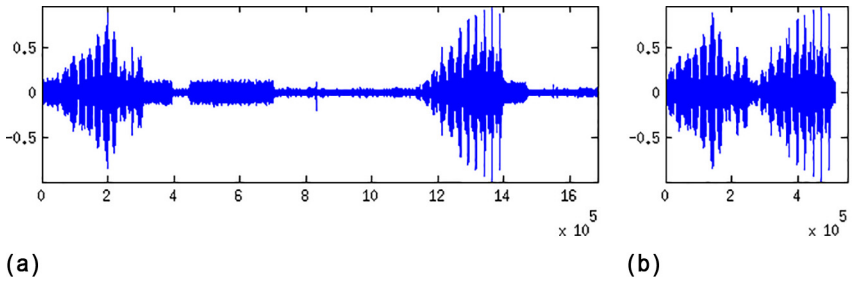

Fig. 3. Example of audio signal before and after automatic segmentation.

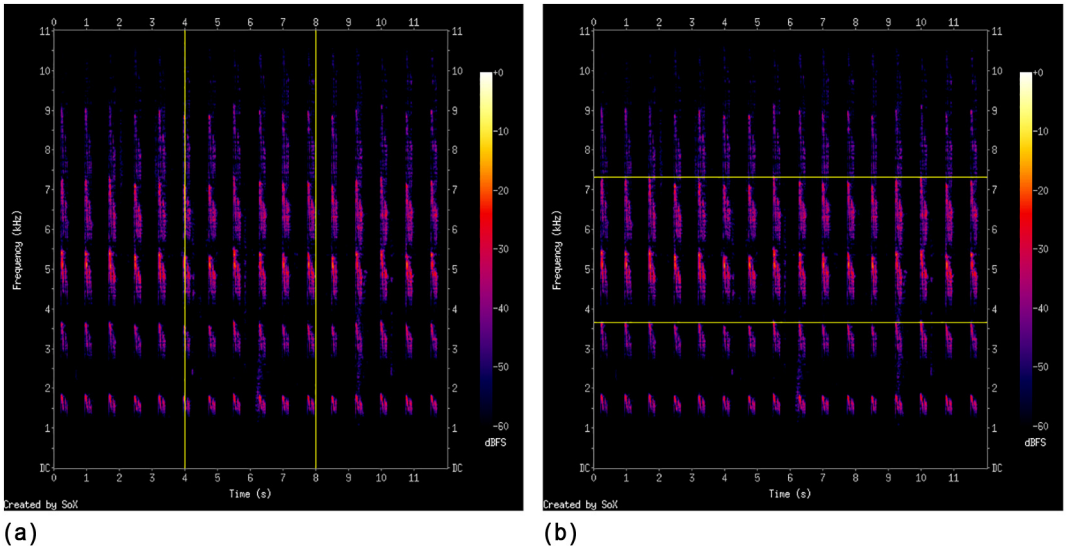

Fig. 4. Examples of vertical and horizontal zones.

domains, as: face recognition (Ahonen et al., 2006), music genre recognition (Costa et al., 2011), manuscript writer identification and bird species classification (Lucio & Costa, 2015; Zottesso et al., 2016). It is important to remark that LBP has obtained good performances in all these works.

According to Ojala et al., (2002), LBP operates on the local neighborhood of a central pixel to find a local binary pattern. The feature vector which describes the textural content of the image corresponds to the histogram of local binary patterns found in all pixels of the image. There are two main parameters that can be changed to capture the LBP from an image. The first one is the number of neighboring pixels that will be taken into account for the central pixel, the second one is related to the distance between the central pixel and its neighbors. These values are respectively known as P and R.

In this work, we decided to use 8 neighbors at a distance equal to 2, since with this setup good results have been obtained by several researchers in dif-ferent application domains, including on works were audio classification tasks using spectrograms were assessed (Lucio & Costa, 2015; Bertolini et al., 2013; Costa et al., 2012b). This particular setup is commonly described as $LBP_{8;2}$, and in its most well successful form, in which only uniform patterns are discerned in the histogram, it is composed of 59 features.

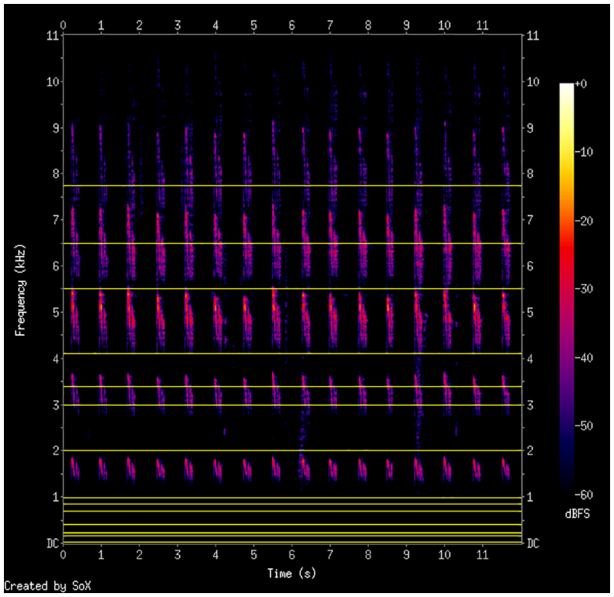

Fig. 5. Spectrogram zoning according to Mel-frequency scale.

Table 4
Dimensionality of texture descriptors vectors.

Texture descriptor	Feature vector length
LBP	59
RLBP	59
LPQ	256

3.4.2. Robust local binary pattern (RLBP)

Aiming to make the LBP texture descriptor even more efficient, Chen et al., (2013) proposed a slight change in the way on how the uniform patterns are considered to make the LBP histogram. The rationale behind it is that if one, and just one, bit in the binary pattern taken from a central pixel makes the pattern nonuniform according to LBP definition, this binary pattern should also be considered as a uniform pattern on RLBP. It makes the binary pattern occurrence a bit more flexible. According to the authors, it is typically related to the occurrence of some noise in the image.

Considering that the database used in this work is deeply affected by the occurrence of noise, we decided to assess the performance of the RLBP texture descriptor in the classification task investigated here.

Similarly to LBP, we have used 8 neighbors at a distance 2 of the central pixel. Therefore, it is referred as $RLBP_{8;2}$ and the feature vector generated is 59-dimensional.

3.4.3. Local phase quantization (LPQ)

LPQ was originally proposed intending to be a texture descriptor robust to the blurring occurrence. However, surprisingly it has achieved a good performance even in situations where the images are not blurred. In addition, several works already published has demonstrated the good performance of this descriptor in texture classification tasks (Bertolini et al., 2013; Costa et al., 2013).

In this work, the features were extracted by using a 3 3 sized window, the correlation coefficient was set to 0:90 and the Short-Term Fourier Transform (STFT) was used with a uniform window. By this way, the obtained features vector corresponds to a histogram composed of 256 values (features).

4. The dissimilarity approach

In this work we have used the dissimilarity framework, presented Cha & Srihari, (2002), Pavelec et al., (2008) and Hanusiak et al., (2011).This approach has been successfully used in the solution of problems related to identification and verification tasks, especially when many classes are involved in the problem.

The dissimilarity is a dichotomy model in which an n-class problem is reduced to a binary problem. As far as we know, the bird species classification problem (a typical multi-class problem) is being addressed using dissimilarity for the first time in this work, and this is one of the main contributions of this work.

This dichotomic transformation is illustrated in Fig. 6(a) and (b). The former one shows several samples labeled on five different classes distributed in a two-dimensional space, in which each sample is represented by a feature vector (f1; f2). The latter one shows the distribution of dissimilarity vectors, obtained by calculating the difference between the feature vectors of two samples.

As one can see in Fig. 6, the dissimilarity vectors are labeled according to two different classes: positive (+) or negative (*). The positive label is assigned to dissimilarity vectors obtained from two feature vectors of samples belonging to the same class. On the opposite way, the negative label is associated to dissimilarity vectors obtained from feature vectors belonging to different classes (Bertolini et al., 2013). Suppose there are two vectors V_i and Q_i in the feature space, labeled l_V and l_Q respectively. Assume further that Z_i is the dissimilarity feature vector obtained from $Z_i = |V_i\ Q_i|$ where $|.|$ refers to the absolute value. Observe that the dimensionality of Z_i is the same of V_i and Q_i.

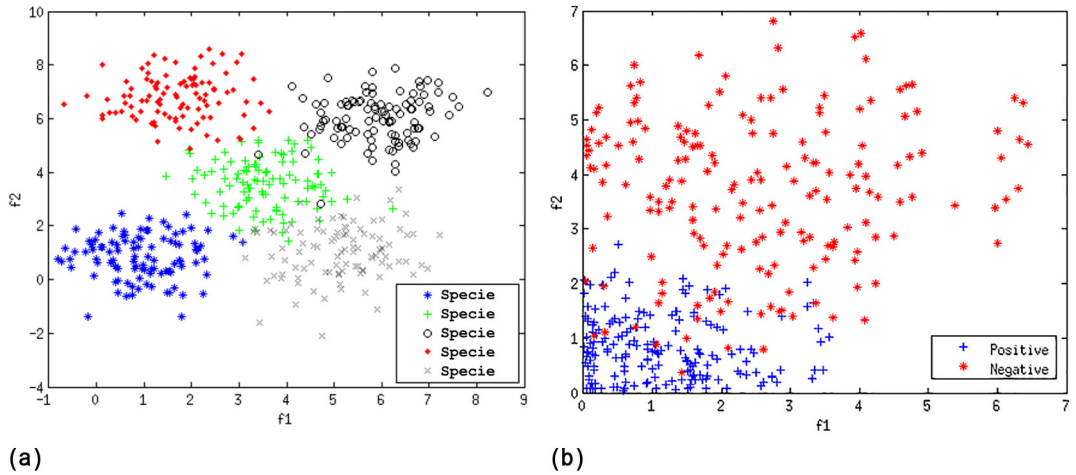

Fig. 6. Fig. (a): Samples in the feature space in a problem with five classes. Fig. (b): Samples in the dissimilarity space where (+) stands for the vectors associated to the within class and (*) stands for the vectors associated to the between class.

In this work, we propose to extract texture feature vectors from spectrograms obtained starting from bird call recordings. Therefore, we aim to use these feature vectors to create positive and negative dissimilarity vectors combining both feature vectors associated to the within class and to the between class, respectively. In this way, we assume that dissimilarity vectors obtained using feature vectors from the same class should have their values close to zero because those vectors are supposed to be similar. In the opposite way, dissimilarity vectors obtained using feature vectors from different classes should have their values far from zero (Bertolini et al., 2013).

4.1. Generation of dissimilarity feature vectors

The dissimilarity approach falls on binary classifiers to discriminate between positive and negative classes. It is worth reminding that positive samples are obtained between feature vectors belonging to the same class whereas negative examples are obtained from samples of different classes.

Aiming to generate the positive samples, we computed the dissimilarity vectors among R positive samples (references) of each species. In this case, the value of R may vary considering the number of segments of texture extracted from each spectrogram. This number varies according to the number of horizontal zones (Z_h) and vertical zones (Z_v), which resulted in ($R = Z_h \times Z_v \times$ Class) different combinations. The same number of negative samples can be generated by computing the dissimilarity between references of one species against references from others species.

Considering, for example, 10 species in the training step, with three horizontal zones ($Z_h = 3$) and three vertical zones ($Z_v = 3$), we would have 90 ($10 \times 3 \times 3$) positive samples and 90 negative samples ($10 \times 3 \times 3$). Fig. 7 illustrates this process. In the top of the Fig. 7, positive samples were created using three samples (bird corners) from the same species, thus, feature vectors are extracted from the reference images, in this example one per image ($Z_v = Z_h = 1$). Based on these three vectors, three dissimilarity vectors are computed (positive samples). These are positive dissimilarity vectors, which are expected to have components close to 0. A similar process is depicted in the bottom of the Fig. 7, in the case feature vectors taken from different classes, are used to create the negative dissimilarity vectors. In such case, it is expected that their components will be far from 0.

5. Experiments and discussion

In this section we describe the experiments and results obtained using the proposed approach on the eight subsets described in Section 4. Table 5 describes some details about the eight subsets used in this work.

Table 5
Number of samples in the training and test sets used in different subsets.

Subset	Species	Samples in train set	Samples in test set	Number of samples
		Validation		
#1	23	05	05	10
#2	48	05	05	10
#3	88	05	05	10
		Test		
#4	180	05	05	10
#5	349	05	05	10
#6	614	03	03	06
#7	772	02	02	04
#8	915	01	01	02

In all experiments, regardless of the subset used, the data was split into 50-50 for training and testing. In the testing set, the samples were divided into folds, each one containing at least one sample per species. The identification rates presented following were obtained by calculating the average between these folds.

In order to reduce the time taken to train the SVM models, since the amount of dissimilarity vectors is quite huge, some experiments were performed to find suitable values for C and Gamma (γ).

In order to find favorable values for these parameters, we have performed training using the subsets #1, #2 and #3.

In this work, various kernels functions were evaluated, and the best results were achieved using the Gaussian kernel. Thus, we have used in all experiments the Gaussian kernel with $C = 8$ and $\gamma = 2$.

In order to compute the Top-N identification rates, was have performed the fusion between the predictions scores of the classifiers obtained from different zones by using the Sum Rule. Since by comparing the Sum, Max, Product, Average and Median rules, the Sum Rule showed the best results in most cases. Fig. 8 depicts the combination strategy, proposed by Kittler et al. in (Kittler et al., 1998).

This section is divided into two Subsections. Subsection 6.1 describes results obtained with all the texture descriptors and zoning approaches assessed in this work. In the Subsection 6.2 we present the results obtained in different subsets employing the texture descriptor and the zoning approach with the best performances.

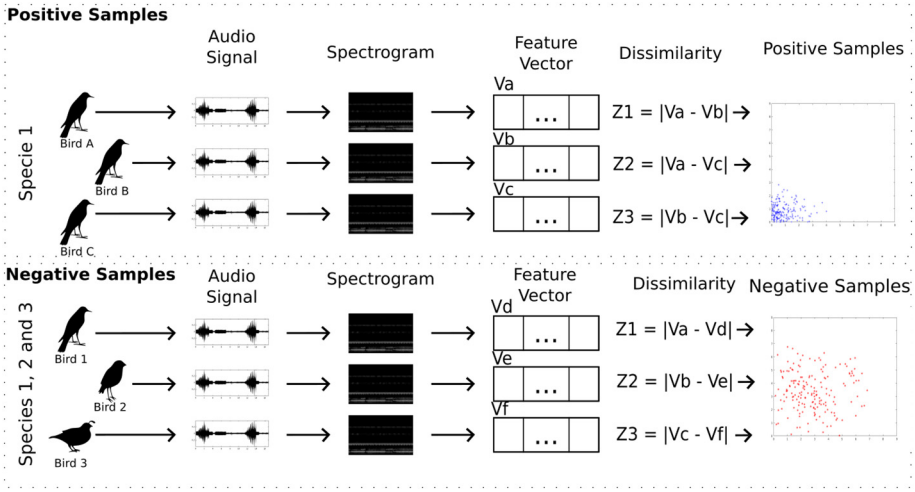

Fig. 7. Dissimilarity Framework. Positive Samples: dissimilarity among samples of the same specie to generate the positive samples. Negative Samples: dissimilarity among samples from different species to generate the negative samples.

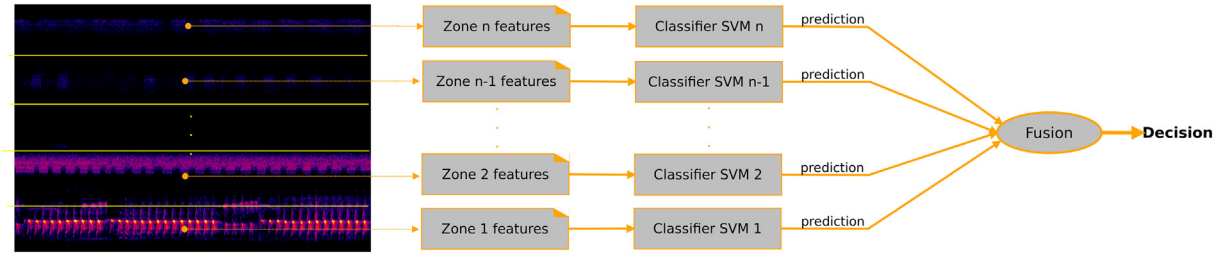

Fig. 8. Methodology used for combining classifiers.

Table 6
Evaluation of the texture descriptors LBP, RLBP and LPQ in the subset #3.

Descritor	Top 01	Top 05	Top 10
LBP	0.914	0.982	0.986
RLBP	0.905	0.980	0.986
LPQ	0.911	0.977	0.982

5.1. Evaluation of texture descriptors and zoning schemes

The texture descriptor used in the feature extraction step can strongly influence system performance. In this work, we firstly performed some experiments in order to compare the results obtained with LBP, RLBP, and LPQ. Table 6 describes the results obtained for each descriptor. In these experiments we used the subset # 3, with 88 species and spectrograms split into 3 vertical zones and 15 horizontal zones created according to the Mel scale frequency bands. The subset #3 was chosen in this experiment because it contains the subsets #1 and #2 and contains a representative number of species. In addition, the SVM classifier was performed using the RBF kernel, the parameters C and γ were set to 8 and 2, respectively. By analyzing the obtained results, we can observe that LBP performed better than RLBP and LPQ. In addition, the number of features extracted using LBP is equal to the number of features of RLBP and smaller than the number of features of the LPQ, implying in a shorter processing time. Thus, the LBP was chosen as the texture descriptor to be used in the following experiments.

The number of vertical and horizontal zones influences the amount of pos-itive and negative dissimilarity vectors that can be created. Table 7 presents the results obtained varying the number of vertical zones in 3, 5 and 9, and the horizontal zones were assessed without division (none zoning) and with 15 non-linear zones defined according to the Mel scale. For this, the features of the subset #3 (88 species) were extracted using the LBP texture descriptor. The SVM classifier was used with the same parameters used in the previous experiment. The best results were obtained when three vertical zones were used. The experiments with five and nine zones did not present satisfactory results.

One can suppose that the low performance obtained using 5 and 9 vertical zones may have occurred because of the lack of content in many of the created zones.

Thus, the texture descriptors generated from these slices do not describe any content of interest, leading to dissimilarity and classification fail.

Table 8 summarizes the results obtained by varying the number of horizontal zones, setting three vertical zones. The subset #3, the LBP and SVM classifier configured with C = 8, γ = 2 and RBF kernel were used. The best results were obtained with the Mel-scale, using frequency bands of different sizes related to those perceived by humans.

Table 7
Identification rates obtained from the variation of vertical and horizontal zones.

Vertical	Horizontal	Top-1	Top-5	Top-10
03	None	0.570	0.857	0.914
03	15 (Mel)	0.914	0.982	0.986
05	None	0.041	0.077	0.116
05	15 (Mel)	0.018	0.061	0.148
09	None	0.032	0.068	0.127
09	15 (Mel)	0.025	0.059	0.157

Table 8
Identification rates varying the number of horizontal zones.

Zones	Top-1	Top-5	Top-10
None	0.570	0.857	0.914
03	0.755	0.911	0.950
05	0.852	0.941	0.968
10	0.902	0.966	0.977
15 (Mel)	0.914	0.982	0.986

Corroborating results already obtained in music genre classification (Costa et al., 2012a) and in bird species identification tasks, the use of a suitable division in the creation of horizontal zones seems to be decisive to achieve the best possible identification rates.

The experiments carried out so far were developed aiming to find the ideal parameters to evaluate the process of bird species identification in different subsets. After performing several tests and analyzing the results described in the previous tables of this subsection, the following definitions were considered: 3 vertical zones, 15 horizontal zones (Mel scale), LBP texture descriptor, C = 8, γ = 2 and RBF kernel. Thus, the Subsection 6.2 shows the performance achieved on 8 different subsets using these configuration settings.

5.2. Subsets evaluation

Once we have defined the optimal parameters, now we describe experiments conducted aiming to evaluate the impact of the audio sample duration and number of classes on system performance.

Table 9 shows the results obtained using the eight subsets proposed in this work. Note that when the dissimilarity approach is used, it is possible to achieve good identification rates even with a significant increase in the number of classes. Moreover, it is possible to observe that even decreasing the duration of the audio samples to five seconds (on subsets # 5 to # 8) and increasing significantly the number of classes (from 23 to 915), the proposed approach keeps identification rates above 70%.

Hyperparameters C and γ found using samples randomly taken from

Table 9
Identification rates using the proposed approach in the eight different subsets.

Train and test	Classes	Top-1 (σ)	Top-5	Top-10
#1	23	0.895 ± 0.059	0.991	1.000
#2	48	0.875 ± 0.029	0.975	0.991
#3	88	0.920 ± 0.036	0.981	0.990
#4	180	0.848 ± 0.011	0.935	0.954
#5	349	0.793 ± 0.027	0.900	0.928
#6	614	0.749 ± 0.012	0.872	0.902
#7	772	0.722 ± 0.005	0.858	0.896
#8	915	0.701 ± 0.000	0.824	0.865

Table 10
Identification rates using different models to classify different subsets.

TrainnTest	#1	#2	#3	#4	#5	#6	#7	#8
#1 (23)	–	0.86	0.85	0.72	0.52	0.39	0.38	0.32
#2 (48)	0.93	–	0.91	0.75	0.59	0.45	0.49	0.44
#3 (88)	0.91	0.88	–	0.82	0.71	0.53	0.60	0.48
#4 (180)	0.90	0.89	0.92	–	0.78	0.67	0.68	0.63
#5 (349)	0.90	0.88	0.90	0.85	–	0.69	0.71	0.66
#6 (614)	0.86	0.88	0.90	0.85	0.80	–	0.73	0.71
#7 (772)	0.90	0.89	0.91	0.85	0.80	0.74	–	0.71
#8 (915)	0.87	0.86	0.90	0.85	0.79	0.72	0.72	–

subsets #1, #2 and #3.

From Table 9, we also can note that on subsets from #1 to #5 all the results are above 79% for Top-1. In these cases, there are five samples for each species both in training and testing sets. Moreover, taking into account subsets from #1 to #4, the approach presented even better results. In this second case, all samples have at least 10 samples and a minimum time of 10 seconds.

Taking into account other experiments previously reported in other works, e.g., on music genre classification or even on bird species classification, we have empirically defined the number of columns per second on the spectrogram image. In this sense, we have used 27 columns on the image for each second of the audio signal, hence, a sample with 10 seconds generates an image with 270 pixels wide (time). Using three vertical zones and minimum time of 10 seconds, each slice has at least 90 pixels wide. Initially, we believed that the factors that contribute to the robustness of the system are the minimum audio duration (10 seconds), and the minimum number of training samples (5 samples). However, since the number of samples from one subset to another increases considerably, it is difficult to conclude that the time and number of samples are the only factors that influence system performance.

We can observe in Table 9 that the performance on the subset # 3 is 0.92 for Top-1 and on the subset # 4 the identification rate is 0.848. In this case, the number of species practically doubled, and the performance dropped only 7.2 percentage points. From the subset #5 to #7, we also doubled the number of classes, and the drop was 7.1 percentage points. Thus, we can conclude that in addition to the impact of time and number of samples, the number of species used can have a high impact on system performance.

The main contribution of the dissimilarity approach is that the model does not need to be retrained whenever new classes are added to the classification system. Table 10 describes the results achieved using the eight different models to classify the eight subsets proposed in this work.

We can note that not always a model trained using more classes will reach the best results. Another interesting point is that when using the subset # 6 for training, we got the highest accuracy in databases with the highest number of classes. Another interesting point is that the best identification rates have always been obtained by using one subset for training and another subset for testing. Besides, if we have a model with a greater diversity of classes we probably have more impact than when we have audio signals with a longer duration.

It is worth of noticing that these results were obtained using a larger number of species from Xeno-Canto database, while the best results in the state-of-the-art use a reduced number of classes. As aforementioned, it is difficult to compare the results reported in this study with other works, because they do not necessarily use the same datasets. However, we can point out that the winner of the LifeClef 2015 bird task obtained a mean average precision close to 45% on the whole database.

6. Conclusion

In this work, we have addressed bird species identification starting from audio recordings using spectrograms and the dissimilarity framework. The experiments were carried out on eight different subsets of the LifeClef bird task 2015 contest, in which the number of classes ranges from 23 to 915.

Spectrogram was chosen as the source to extract the feature because it has been successfully used in many other audio classification tasks. Regarding the dissimilarity framework, it was used because the potential benefits using this strategy are twofold: this framework removes the need for retraining the model each time a new class is introduced in the classification system; dissimilarity has shown to keep good performance rates even when the classification problem involves a large number of classes.

Although the obtained results cannot be directly compared to other results, because the subsets used here were assessed for the first time in this work, the results lead us to believe that the proposed method is among the best ever presented. In the most challenging scenario evaluated, with 915 classes, we have obtained an identification rate of 71%. In order to encourage other researchers to compare those approaches with the method the list of the audio clips used in each dataset used here and also the spectrogram images extracted from the audio were made available.

As a future work, we aim to develop experiments using features obtained with deep learning. We also intend to investigate the complementarity between those features with handcrafted (i.e. LBP) features under the dissimilarity based method proposed here. In addition, we aim to evaluate our approach on the more recent version of LifeCLEF database, composed of 1,500 bird species.

Acknowledgment

We thank the Brazilian Research-support agencies Coordination for the Improvment of Higher Education Personnel (CAPES) and the Brazilian National Council for Scientific and Technological Development (CNPq).

References

T. Ahonen, A. Hadid, M. Pietikainen, Face description with local binary patterns: Application to face recognition, Pattern Analysis and Machine Intelligence, IEEE Transactions on 28 (12) (2006) 2037–2041.

Albornoz, E.M., Vignolo, L.D., Sarquis, J.A., Leon, E., 2017. Automatic clas-sification of furnariidae species from the paranaense littoral region using speech-related features and machine learning. Ecol. Informatics 38, 39–49.

Anderson, S.E., Dave, A.S., Margoliash, D., 1996. Template-based automatic recognition of birdsong syllables from continuous recordings. J. Acoustic. Soc. Am. 100 (2), 1209–1219.

Bertolini, D., Oliveira, L.S., Justino, E., Sabourin, R., 2013. Texture-based de-scriptors for writer identification and verification. Expert Syst. Appl. 40 (6), 2069–2080.

Cai, J., Ee, D., Pham, B., Roe, P., Zhang, J., 2007. Sensor network for the monitor-ing of ecosystem: Bird species recognition. In: Intelligent Sensors, Sensor Networks and Information, 2007. ISSNIP 2007. 3rd International Confer-ence on, IEEE, pp. 293–298.

Catchpole, C.K., Slater, P.J., 2003. Bird Song: Biological Themes and Variations. Cambridge university press.

Cha, S.-H., Srihari, S.N., 2002. On measuring the distance between histograms. Pattern Recogn. 35 (6), 1355–1370.

Chen, J., Kellokumpu, V., Zhao, G., Pietikäinen, M., 2013. RLBP: Robust local binary pattern. In: Proceedings of the British Machine Vision Conference.

Chou, C.-H., Lee, C.-H., Ni, H.-W., 2007. Bird species recognition by comparing the HMMs of the syllables. In: Innovative Computing, Information and Control, 2007. ICICIC'07. Second International Conference on, IEEE, pp. 143. https://doi.org/10.1109/ICICIC.2007.199.

Conway, C.J., 2011. Standardized north American marsh bird monitoring proto-col. Waterbirds 34 (3), 319–346.

Costa, Y.M.G., Oliveira, L.E.S., Koerich, A.L., Gouyon, F., 2011. Music genre recognition using spectrograms. In: Systems, Signals and Image Processing (IWSSIP), 2011 18th International Conference on, IEEE, pp. 1–4.

Costa, Y.M.G., Oliveira, L.E.S., Koerich, A.L., Gouyon, F., Martins, J., 2012a. Music genre classification using LBP textural features. Signal Process. 92 (11), 2723–2737.

Costa, Y.M.G., Oliveira, L.E.S., Koerich, A.L., Gouyon, F., 2012b. Comparing textural features for music genre classification. In: Neural Networks (IJCNN), The 2012 International Joint Conference on, IEEE, pp. 1–6.

Costa, Y.M.G., Oliveira, L.E.S., Koerich, A. L., Gouyon, F., Music genre recognition based on visual features with dynamic ensemble of classifiers selection, in: Systems, Signals

and image Processing (IWSSIP), 2013 20th International Conference on, IEEE, 2013, pp. 55–58.
Evangelista, T.L., Priolli, T.M., Silla, C.N., Angelico, B.A., Kaestner, C.A., 2014. Automatic segmentation of audio signals for bird species identification. In: Multimedia (ISM), 2014 IEEE International Symposium on, IEEE, pp. 223–228.
Fagerlund, S., 2007. Bird species recognition using support vector machines. EURASIP J. Appl. Signal Process. 2007 (1), 64. https://doi.org/10.1155/2007/38637.
Faria, C.M., Rodrigues, M., do Amaral, F.Q., Módena, É., Fernandes, A.M., 2006. Aves de um fragmento de mata atlântica no alto rio doce, minas ger minas gerais: colonização e extinção. Revista Brasileira de Zoologia 23 (4), 1217–1230.
Fazekas, B., Schindler, A., Lidy, T., Rauber, A., 2017. A Multi-Modal Deep Neural Network Approach To Bird-Song Identification. In: Linda Cappellato. 41. pp. 1–6 (URL http://ceur-ws.org/Vol-1866/paper_179.pdf).
G. K. Freitas, Y. M. G. Costa, R. L. Aguiar, Using spectrogram to detect North Atlantic right whale calls from audio recordings, in: Computer Science Society (SCCC), 2016 35th International Conference of the Chilean, IEEE, 2016, pp. 1–6.
Fritzler, A., Koitka, S., Friedrich, C.M., 2018. Recognizing Bird Species In Audio Files using Transfer Learning, in: Linda Cappellato. 41. pp. 1–14 (URL http://ceur-ws.org/Vol-1866/paper_169.pdf).
Goëau, H., Glotin, H., Vellinga, W.-P., Planqué, R., Joly, A., 2016. LifeCLEF Bird Identification Task 2016: The arrival of Deep learning. In: Working Notes of CLEF 2016 - Conference and Labs of the Evaluation Forum, Evora, Portugal, pp. 440–449 (URL https://hal.archives-ouvertes.fr/hal-01373779).
Hanusiak, R., Oliveira, L., Justino, E., Sabourin, R., 2011. Writer verification using texture-based features. Int. J. Doc. Anal. Recog. 1–1410 (1007/s10032-011-0166-4 URL http://dx.doi.org/10.1007/s10032-011-0166-4).
Kahl, S., Wilhelm-Stein, T., Hussein, H., Klinck, H., Kowerko, D., Ritter, M., Eibl, M., 2017. Large-Scale Bird Sound Classification Using Convolutional Neural Networks, in: Linda Cappellato. 41. pp. 1–14 (URL http://ceur-ws.org/Vol-1866/paper_143.pdf).
J. Kittler, M. Hatef, R. P. Duin, J. Matas, On combining classifiers, Pattern Analysis and Machine Intelligence, IEEE Transactions on 20 (3) (1998) 226–239.
Kogan, J.A., Margoliash, D., 1998. Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden markov models: a comparative study. J. Acoustic. Soc. Am. 103 (4), 2185–2196.
Linda Cappellato, L.G.T.M., Ferro, Nicola (Eds.), 2017. CEUR Workshop Proceedings 1866. http://ceur-ws.org/Vol-1866/.
Lopes, M.T., Gioppo, L.L., Higushi, T.T., Kaestner, C.A., Silla Jr., C.N., Koerich, A.L., 2011a. Automatic bird species identification for large number of species. In: Multimedia (ISM), 2011 IEEE International Symposium on, IEEE0, pp. 117–122.
Lopes, M.T., Koerich, A.L., Nascimento Silla, C., Kaestner, C.A.A., 2011b. Feature set comparison for automatic bird species identification. In: Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on, IEEE, pp. 965–970.
Lucio, D.R., Costa, Y.M.G., 2015. Bird species classification using spectrograms. In: Computing Conference (CLEI), 2015 Latin American, IEEE, pp. 1–11.
Marini, A., Turatti, A., Britto, A., Koerich, A., 2015. Visual and acoustic identifica-tion of bird species. In: Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, IEEE, pp. 2309–2313.
Nanni, L., Costa, Y.M.G., Lumini, A., Kim, M.Y., Baek, S.R., 2016. Combining visual and acoustic features for music genre classification. Expert Syst. Appl. 45, 108–117.
Negret, Á., 1988. Fluxos migratórios na avifauna da reserva ecológica do IBGE, brasília, DF, brasil. Revista Brasileira de Zoologia 5 (2), 209–214.
Ntalampiras, S., 2018. Bird species identification via transfer learning from music genres. Ecol. Informatics 44, 76–81. https://doi.org/10.1016/j.ecoinf.2018.01.006. (URL http://www.sciencedirect.com/science/article/pii/S1574954117302467).
T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and ro-tation invariant texture classification with local binary patterns, Pattern Analysis and Machine Intelligence, IEEE Transactions on 24 (7) (2002) 971–987.
Pavelec, D., Justino, E., Batista, L.V., Oliveira, L.S., 2008. Author identification using writer-dependent and writer-independent strategies. In: Proceedings of the 2008 ACM Symposium on Applied Computing, SAC '08. ACM, New York, NY, USA, pp. 414–418. https://doi.org/10.1145/1363686.1363788.
Schuchmann, K.-L., Marques, M.I., Jahn, O., Ganchev, T., Figueiredo, J., 2014. Os sons do pantanal: Um projeto de monitoramento acústico automatizado da biodiversidade. Boletim Informativo Sociedade Brasileira de Zoologia 108, 11–12.
Sevilla, A., Glotin, H., 2017. Audio Bird Classification With Inception-V4 Extended with Time and Time-Frequency Attention mechanisms, in: Linda Cappellato. 41. pp. 1–8 (URL http://ceur-ws.org/Vol-1866/paper_177.pdf).
Sprengel, E., Jaggi, M., Kilcher, Y., Hofmann, T., 2016. Audio based bird species identification using deep learning techniques. LifeCLEF 2016, 547–559.
Umesh, S., Cohen, L., Nelson, D., 1999. Fitting the mel scale. In: Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on, Vol. 1, IEEE, pp. 217–220.
Zhao, Z., Zhang, S.-h., Xu, Z.-y., Bellisario, K., Dai, N.-h., Omrani, H., Pijanowski, B.C., 2017. Automated bird acoustic event detection and robust species classification. Ecol. Informatics 39, 99–108.
Zottesso, R.H.D., Matsushita, G.H.G., Lucio, D.R., Costa, Y.M.G., 2016. Automatic segmentation of audio signal in bird species identification. In: Computer Science Society (SCCC), 2016 35th International Conference of the Chilean, IEEE, pp. 1–11. https://doi.org/10.1109/SCCC.2016.7836062.