Multiple Instance Learning-Based Birdsong Classification Using Unsupervised Recording Segmentation

J. F. Ruiz-Muñoz, Mauricio Orozco-Alzate, G. Castellanos-Dominguez
Universidad Nacional de Colombia - Sede Manizales
{jfruizmu, morozcoa, cgcastellanosd}@unal.edu.co

Abstract

Traditional techniques for monitoring wildlife populations are temporally and spatially limited. Alternatively, in order to quickly and accurately extract information about the current state of the environment, tools for processing and recognition of acoustic signals can be used. In the past, a number of research studies on automatic classification of species through their vocalizations have been undertaken. In many of them, however, the segmentation applied in the preprocessing stage either implies human effort or is insufficiently described to be reproduced. Therefore, it might be unfeasible in real conditions. Particularly, this paper is focused on the extraction of local information as units –called instances– from audio recordings. The methodology for instance extraction consists in the segmentation carried out using image processing techniques on spectrograms and the estimation of a needed threshold by the Otsu's method. The multiple instance classification (MIC) approach is used for the recognition of the sound units. A public data set was used for the experiments. The proposed unsupervised segmentation method has a practical advantage over the compared supervised method, which requires the training from manually segmented spectrograms. Results show that there is no significant difference between the proposed method and its baseline. Therefore, it is shown that the proposed approach is feasible to design an automatic recognition system of recordings which only requires, as training information, labeled examples of audio recordings.

1 Introduction

One of the biggest challenges in ecology and conservation biology is the assessment of biodiversity through effective monitoring techniques that allow covering large scales in both time and space [Depraetere *et al.*, 2012]. Nevertheless, traditional methods for assessing the presence and abundance of species are expensive as well as spatially and temporally limited since they typically consist in manual data collection through extensive transects [Aide *et al.*, 2013]. Such a task would be much easier by using technological tools, e.g., it is possible to implement a continuous monitoring of the sound-emitting wildlife —as birds, which have been traditionally used as a biodiversity indicator [Briggs *et al.*, 2012]— by installing microphones and recorders in the field. Moreover, in regions with high cloudiness, this technique may be even more effective than visual inspection.

Signal analyses and classification approaches used in bioacoustics range from trained humans listening to recordings or visually inspecting spectrograms, to autonomous classification systems based on digital signal processing and pattern recognition methods [Blumstein *et al.*, 2011]. However, considering the overwhelming amount of acoustic data that can be collected, relying on analyses made by human experts is limited and often unfeasible. Furthermore, according to [Hao *et al.*, 2012], the automation of analyses using signal processing techniques and pattern recognition algorithms is less expensive —in the long term— than the assessments made by experts and, potentially, even more accurate.

As indicated by Briggs *et al.* [2012], methods for acoustic classification of bird species can be categorized into two types: 1) those that classify individual syllables, and 2) those that classify recordings having sounds of multiple sources. The former ones require a detailed annotation of each segment while in the latter ones, for training, it is only required to label the presence of the species of interest in each recording. In either case, segmentation is a high-priority step [Neal *et al.*, 2011]. However, methods of the first type are more sensitive to the segmentation quality because omitted syllables become false negatives and those sounds incorrectly detected could become false positives. In contrast, for methods of the second type, it is possible to achieve correct classification even if the two above-mentioned miss-segmentation cases occur since every recording may contain multiple syllables.

In general, segmenting recordings into smaller recognition units is assumed as a part of the preprocessing stage and it is done either manually or automatically. Yet, automatic recognition should not require manual segmentation [Trifa *et al.*, 2008]. To this end, segmentation algorithms have been developed mostly using energy and entropy as criteria to identify onset and offset times of the regions of interest [Fagerlund, 2007]. Under ideal conditions, when the vocalization call is the only sound in the recording, an increase in energy clearly reveals a region of interest, making segmentation procedures simple enough [Neal *et al.*, 2011]. In real conditions, however, recorded signals are degraded due to the presence of many sound sources, e.g., wind streams, background noise from other animals and surrounding events. In spite of that, several research studies on automated species recognition clarify that their methods work well when the recognition units are correctly detected, often this issue is not discussed in depth (making only a brief description). Furthermore, as indicated in [Hao *et al.*, 2012], it should be taken into account that achieving perfectly segmented data is at least as difficult as the classification step.

Considering that the recognition of recordings has the advantage of not making the impractical assumption of requiring perfectly segmented data, in this study, it is proposed a classification methodology for audio recordings which only uses —in the training phase— labels from training recordings, that is, isolated and labeled vocalization segments are not required beforehand since the methodology includes a novel unsupervised segmentation method for birdsong recordings. Interest sounds are detected from the Short-time Fourier transform (STFT). In the segmentation method described in Sec. 2.1, the output is a matrix of the same size of the corresponding spectrogram, where interest elements (or pixels) are marked with “one” and non-interest elements with “zero”. Classification is carried out using the multiple instance classification (MIC) approach, as follows: 1) neighboring interest pixels are grouped into regions, 2) each region is described by a feature representation and 3) a classifier based on multiple instance learning (MIL) is trained considering each spectrogram as a bag of instances. Its classification performance is estimated when using the unsupervised segmentation method proposed in Sec. 2.1 and compared against the performance obtained when using the supervised segmentation method proposed by Briggs *et al.* [2012]. Both methods consist in the detection of regions in the spectrogram likely associated with vocalizations; however, in the latter, it is required to provide a set of manually annotated spectrograms where pixels have been labeled according to whether or not they correspond to bird sounds.

2 Material and methods

2.1 Segmentation of birdsong recordings

Time-frequency analysis of audio recordings is usually carried out through spectrograms representing power intensity at each time-frequency point. Particularly, spectrograms are considered as recognizable images to identify bird species [Dennis *et al.*, 2011], whose vocalizations are represented by intensity variations. Thus, under the assumption that segment vocalizations give a form of continuous regions holding the highest power values, our unsupervised segmentation methods consists in the following stages (see Fig. 1):

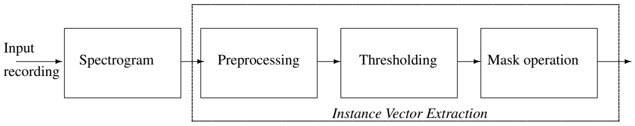


Figure 1: Flow diagram of the proposed segmentation method.

- **Spectrogram estimation:** Based on the STFT decomposition, we compute a spectrogram matrix $\boldsymbol{S}=\{s_{ij}: i=0,\dots,N-1; j=0,\dots,M-1\}$, with $\boldsymbol{S}\in\mathbb{R}^{N\times M}$, where indexes i and j stand for the frequency and time domains (i.e., N points in the frequency domain that are estimated in each one of M time frames). We use the Hann window lasting 512 samples and overlapping 256 samples as in [Briggs *et al.*, 2012].
- **Preprocessing:** After applying the two-dimensional Wiener filter, a denoised and smoothed spectrogram, $\widetilde{\boldsymbol{S}}=\{\tilde{s}_{ij}\}$ is estimated with elements:

$$\tilde{s}_{ij} = \mu + s_{ij}(\sigma^2 + \sigma_\eta^2)/\sigma^2$$

  where $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}$ are the local mean and variance, respectively, and $\sigma_\eta^2 \in \mathbb{R}$ is the noise variance. The first two values are estimated at a $k\times k$ neighborhood centered on each point (we set $k=5$ as in [Pourhomayoun *et al.*, 2013]), and σ_η^2 is the average over all local variances σ^2. Then, we suppress structures that are lighter than their surroundings and are connected to image borders by considering a value of 8 as connectivity parameter. An erosion [van den Boomgaard and van Balen, 1992] is performed before by estimating the lowest number of image elements that are not connected to the edges. The erosion procedure results in a smoothed image $\boldsymbol{I}\in\mathbb{R}^{N\times M}$, from which we perform morphological reconstruction of $\widetilde{\boldsymbol{S}}$ to remove all intensity fluctuations (except the intensity peak). As a result, we get the matrix $\boldsymbol{H}\in\mathbb{R}^{N\times M}$ that only holds objects with neighboring borders. At the same time, the difference matrix, $\widetilde{\boldsymbol{H}} = \widetilde{\boldsymbol{S}}-\boldsymbol{H}$ is also computed holding only those objects from the original image not having neighboring borders.
- **Thresholding:** We fix a threshold to binarize each image $\widetilde{\boldsymbol{H}}$ using the nonparametric and unsupervised Otsu's method of automatic threshold selection [Otsu, 1979]. Extracted only from a computed gray-level histogram, the optimal threshold, T_o, is selected by maximizing an introduced discriminant measure of separability among all resultant gray level classes, as follows:

$$\hat{k} = \arg\max_{k\in[1,L]}\{(\phi_T\omega(k)-\phi(k))^2/(\omega(k)(1-\omega(k)))\}$$

  where $\phi_T = \sum_{j=1}^{L} jp_j$, $\phi(k) = \sum_{j=1}^{k} jp_j$, $\omega(k) = \sum_{j=1}^{k} p_j$, $p_j = n_j/N$ are the values of the normalized gray level histogram, L is the number of gray levels, n_j is the number of pixels at level j, and N is the total number of pixels over the whole difference image $\widetilde{\boldsymbol{H}}$. Therefore, the optimal threshold is computed as $T_o = (\hat{k}-1)/(L-1)$.

- **Mask operation:** To select the most relevant pixels from the spectrogram at hand, a binary matrix $\boldsymbol{B} = \{b_{ij}\}$, with $\boldsymbol{B} \in \mathbb{R}^{N \times M}$, is obtained by thresholding as follows:
$$b_{ij} = \begin{cases} 1, & \text{if } \tilde{h}_{ij} > T_o; \\ 0, & \text{otherwise.} \end{cases}$$
- **Instance vector extraction:** From computed arrangements $\boldsymbol{B}$ and $\boldsymbol{S}$, we compute a spectrogram region set $\mathcal{R} = \{\boldsymbol{R}_i : i = 1, \dots, l\}$, holding the respective matrices of all-connecting points $\boldsymbol{R}_i \in \mathbb{R}^{N_i \times M_i}$, with $\boldsymbol{R}_i \subset \boldsymbol{S}$, being $N_i < N$ and $M_i < M$. The number of regions l is automatically fixed as indicated in [Zakaria *et al.*, 2012]. Lastly, each region $\boldsymbol{R}_i$ supplies a single feature vector (or instance), denoted as $\boldsymbol{x}_i \in \mathbb{R}^d$. The number of features d is fixed in accordance with the training scenarios explained in Sec. 3.1.

2.2 Multiple instance classification

Within the standard supervised classification framework, the training set consists of n feature vector examples or instances $\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{R}^d : i = 1, \dots, n\}$ and their labels, in a two-class problem, $\mathcal{Y} = \{y_i \in \{0, 1\} : i = 1, \dots, n\}$. Thus, any classifier function, $\mathcal{X} \to \mathcal{Y}$, is trained to predict labels for each novel instance.

In MIC, an object is represented by a set, or bag, $\boldsymbol{X}_i = \{\boldsymbol{x}_{ij} \in \mathbb{R}^d : j = 1, ..., m_i\}$ of m_i instances $\boldsymbol{x}_{ij}$ and a label $\tilde{y}_i$, i.e., each label is associated with the entire bag but labels of the individual instances are unknown. Therefore, the training set consists of n bags $\widetilde{\mathcal{X}} = \{\boldsymbol{X}_i \in \mathbb{R}^{m_i \times d} : i = 1, ..., n\}$ and their corresponding labels $\widetilde{\mathcal{Y}} = \{\tilde{y}_i \in \{0, 1\} : i = 1, \dots, n\}$. Then, the classifier function, $\widetilde{\mathcal{X}} \to \widetilde{\mathcal{Y}}$, is trained to predict labels for each novel bag of instances. We address the problem of classifying recordings using the MIC approach because, in our case, we have labels for recordings, i.e. bags, but they are represented for several instances (feature vectors extracted from each region detected after the segmentation stage).

Particularly, we use the MILES (MIL via Embedded Instance Selection) classification algorithm because it has been experimentally shown that it performs well with bioacustic signals [Cheplygina *et al.*, 2015]. MILES transforms the original MIC problem into a standard supervised learning framework injectively relating instances and labels [Chen *et al.*, 2006]. It maps each bag into a feature space defined by instances in the training bags using an introduced instance (dis)similarity measure. Thus, bags are represented by the maximum (dis)similarity to all other instances. On this (dis)similarity representation, a sparse linear classifier is trained [Tax, 2013].

2.3 Segmentation performance measures

Since the manual recording segmentation is a very fatiguing task, rather than directly comparing between automated and manual outputs we indirectly estimate the quality of the proposed segmentation method by the recording classification performance that must be strongly influenced by the used segmentation procedure, as discussed in [Briggs *et al.*, 2012]. The most common performance measures for a classifier are the following ones: accuracy $a = (T_P + T_N)/(P + N)$, specificity $s = T_N/N$, recall rate $r = T_P/P$ and precision rate $p = T_P/(T_P + F_P)$; where T_P is the number of recordings correctly classified as positives, T_N is the number of recordings correctly classified as negatives, F_P is the number of recordings incorrectly classified as positives, F_N is the number of recordings incorrectly classified as negatives, P and N are the total number of positives or negatives recordings, respectively.

However, these performance measures are affected by the relative size of the classes. Therefore, to overcome that drawback, the two-class F-score is used as the performance measure defined as follows:

$$F = 2T_P/(2T_P + F_P + F_N), \; F \in [0, 1]. \tag{1}$$

The F-score ranges from 0 to 1 where the higher its value, the better the classification performance. In this work, the one-against-all reduction from multi-class task to binary classification technique is used where each species is selected as objective (positive) class since the F-score is a two-class measure. As regards the classifier performance, we carry out validation using the Bootstrapping technique where input audio data are randomly split into two sets: one-half for training and one-half for testing. This procedure is carried out ten times.

As suggested in [Bramer, 2013], each one of the eight considered feature sets over the 13 data sets (see Sec. 3.2), are compared in terms of the paired *t*-test, for which the null hypothesis states that the performance of two classification strategies can be statistically assumed as the same. The pseudo-code to determine whether to accept the null hypothesis is presented in Algorithm 1. Otherwise, when the null hypothesis is rejected, we select as the best strategy the one having the highest pair performance difference computed in average over all considered classification strategies.

3 Experimental Set-Up

3.1 Training strategy

Figure 2 shows the training scheme used through all experiments to test the proposed birdsong classification methodology based on unsupervised segmentation of audio recordings and multiple instance learning.

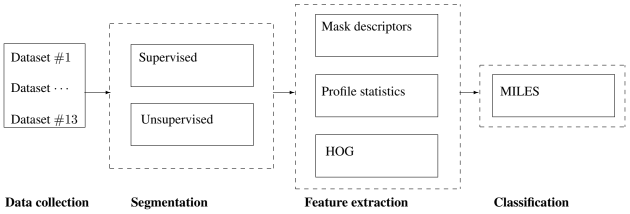


Figure 2: Training scheme used through all experiments to test the proposed birdsong classification methodology.

In the segmentation stage, only one approach, either the unsupervised (see Sec. 2.1) or the supervised (proposed in [Briggs *et al.*, 2012]), is used. Besides, the following four representation scenarios are separately considered: 1)

Algorithm 1 Process to determine whether or not to accept the null hypothesis of the paired *t*-test. Based on [Bramer, 2013].

```
Let $\boldsymbol{z} = [z_1,\ z_2,\ ...\ z_n]$ be the vector of the differences in classification performances between strategies 1 and 2 and let $n$ be the number of data sets.
Assign $t_{level}$ considering the following: for 13 degrees of freedom, $t \geq 1.771$ would only be expected to occur by chance with probability $0.10$ or less, and it is said that the hypothesis is rejected at $10\%$ level. Therefore, if $t_{level} = 1.771$, $t_{level} = 2.160$ or $t_{level} = 3.012$, and $|t| > t_{level}$, the null hypothesis is rejected at the $10\%$, $5\%$ or $1\%$ level, respectively.
procedure T-TEST($\boldsymbol{z}$, $n$, $t_{level}$)
    Compute $a = (\sum_{i=1}^{i=n} z_i^2) - (\sum_{i=1}^{i=n} z_i)^2/n$.
    The sample variance $s^2$ is equal to $a/(n-1)$.
    The sample standard deviation is the square root of $s^2$.
    Divide $s$ by the square root of $n$ to get the standard error $e_{std}$.
    The $t$ statistics is computed dividing the average value of $z$ by $e_{std}$.
    if $|t| \geq t_{level}$ then
        Return The hypothesis is rejected.
    else
        Return The hypothesis is assumed as true.
    end if
end procedure
```

Mask descriptors, denoted as "MD", that describe region shape; the following set of features are computed: minimum frequency, maximum frequency, bandwidth, duration, area, perimeter, non-compactness, and rectangularity. 2) **Profile statistics**, "PS", a set of fourteen features are computed, which are based on statistical segment properties in time and frequency: frequency-Gini, time-Gini, frequency-mean, frequency-variance, frequency-skewness, frequency-kurtosis, time-mean, time-variance, time-skewness, time-kurtosis, frequency-max, time-max, mask-mean, and mask-standard deviation. 3) **Histogram of Gradients**, "HOG", this set consists of 16 features characterizing shape and texture of each region where gradient directions over the pixels of the region are computed; each histogram holds 16 bins equally spaced over the angle range $[-\pi, \pi]$ and features are extracted from the normalized 1-D histograms [Dalal and Triggs, 2005]. 4) **All-features set**, "AF", that merges all above feature sets into a single one, as in [Briggs *et al.*, 2012].

Therefore, eight training strategies are tested for the feature extraction stage. In case of supervised segmentation training, the affix "-Br" (meaning Briggs) is added to the end of every feature notation, in accordance to the segmentation method proposed by Briggs *et al.* [2012]. Lastly, we use the MILES classification algorithm. We employ an exponential kernel $\exp\{-(||a-b||)/p\}$ as the instance (dis)similarity function, where parameter $p \in [2 \ldots 5]$ is heuristically fixed and notation $\|\cdot\|$ stands for Euclidean-norm.

3.2 Data set

For the sake of comparison, we perform experiments with the publicly available[1] data set used in [Briggs *et al.*, 2012]. This data collection holds 548 recordings sampled at 16 *kHz* that were manually labeled. The data set contains 13 bird species, often vocalizing simultaneously and perturbed with environmental noise, though each recording lasting ten-seconds holds between one and five species. Table 1 shows the amount of recordings holding each considered species.

Table 1: Data sets including the amount of recordings where each considered bird species (objective class) is labeled.

	Data set	*Labeled species name*	*Number of recordings*
1	BRCR	Brown Creeper	197
2	WIWR	Winter Wren	109
3	PSFL	Pacific-slope Flycatcher	165
4	RBNU	Red-breasted Nuthatch	82
5	DEJU	Dark-eyed Junco	20
6	OSFL	Olive-sided Flycatcher	90
7	HETH	Hermit Thrush	15
8	CBCH	Chestnut-backed Chickadee	117
9	VATH	Varied Thrush	89
10	HEWA	Hermit Warbler	63
11	SWTH	Swainson's Thrush	79
12	HAFL	Hammond's Flycatcher	103
13	WETA	Western Tanager	46

3.3 Results of compared classification strategies

Table 2 shows the estimated F-scores for the different considered feature sets; notice that the HAFL data set has the highest F-score (> 0.99). In contrast, both data sets DEJU and HETH achieve the lowest values (< 0.21) and get zero-value F-score for several features because those data sets hold very few recordings (see Table 1): 20 and 15, respectively.

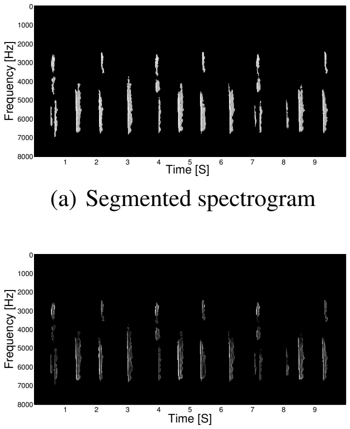
Table 2: Performed F-score values for all considered objective classes and each training scenario. The best reached F-score is marked in bold for each objective class. Besides, the notation "–" stands for null-value performance.

	Training scenario							
Class	*MD*	*MD-Br*	*PS*	*PS-Br*	*HOG*	*HOG-Br*	*AF*	*AF-Br*
BRCR	0.762	0.792	**0.848**	0.824	0.694	0.847	0.774	0.819
WIWR	**0.938**	0.870	0.917	0.870	0.896	0.900	0.933	0.905
PSFL	0.791	0.771	0.781	0.777	0.768	0.736	**0.817**	0.802
RBNU	0.853	0.742	0.793	0.725	0.759	0.784	**0.871**	0.831
DEJU	–	–	–	–	–	–	0.095	**0.214**
OSFL	0.701	0.704	0.718	0.681	0.667	0.694	**0.756**	0.738
HETH	–	–	–	–	–	–	–	–
CBCH	**0.742**	0.569	0.687	0.602	0.643	0.514	0.685	0.600
VATH	0.967	0.931	0.902	0.971	0.909	0.889	0.911	**0.983**
HEWA	0.729	0.718	0.643	**0.764**	0.641	0.652	0.715	0.741
SWTH	0.521	0.587	0.594	**0.813**	0.451	0.566	0.627	0.796
HAFL	**1**	0.996	0.999	0.996	**1**	0.994	**1**	0.996
WETA	**0.852**	0.762	0.795	0.757	0.340	0.372	0.840	0.823
Average	0.805	0.767	0.789	0.798	0.706	0.723	0.812	0.821

In terms of the considered feature sets, the use of AF-Br reaches the highest average performance value ($F = 0.821$) that is averaged over all considered objective classes, while

[1] Audio recordings of the data set are available at `http://www.miproblems.org/datasets/birds/`

HOG gets the lowest one, $F = 0.723$, for the baseline supervised approach. Once again, AF ($F = 0.812$) and HOG ($F = 0.706$) are the best and worst cases, respectively, for the unsupervised method. However, the MD feature set, in average, benefits the most from the use of the unsupervised segmentation, while the HOG set degrades the worst. It must be noted that the average performance is taken without DEJU and HETH data sets because of their accomplished anomaly values. This situation may be explained since the MD feature set encodes region shape attributes that are far from being easy to be manually computed, as seen in Fig. 3(a) showing a typical birdsong spectrogram. In contrast, the texture-based HOG set implies calculation of gradient directions over region pixels (see Fig. 3(b)); this procedure includes enhanced Gaussian filtering and histogram binning that are very sensitive to their parameter tuning.



(a) Segmented spectrogram

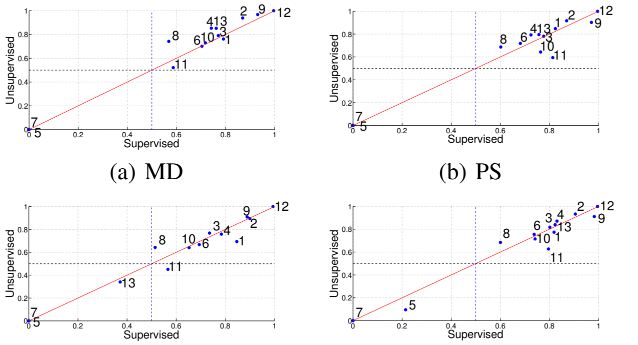
(b) Gradient directions

Figure 3: Example of a segmented spectrogram from a given audio birdsong recording.

In order to provide an intuitive illustration, all F-score values estimated in Table 2 are graphically presented in Fig. 4 where the vertical axis represents the performance obtained by the proposed unsupervised-based segmentation method, while the horizontal axis corresponds to the one obtained by the supervised reference method. Since the larger the F-score – the better the performance, any point above the diagonal line indicates that the proposed method outperforms the reference. In most of the cases, no remarkable difference in terms of performance is observed between both segmentation methods tested for the same representation scenario.

In case of the worst performance when the classifier guesses at random with equal frequency (i.e., $T_p = P/2$, $F_n = P/2$, $F_N = N/2$, and $T_n = N/2$), then one may calculate a threshold $F_t = P/(1.5P + 0.5N)$ value, under which the classifier is just guessing. In our case, $F_t = 0.42$ that is achieved for the BRCR data set. As shown in Fig. 4, most of the computed F-scores overcome by far the F_t threshold.

Table 3 shows several performance measures (namely, recall, r, specificity, s, precision, p, accuracy, a, and the F-score) obtained for the whole feature sets, i.e., AF and AF-Br data sets. Typically, both training strategies may be differently influenced by each considered class. Particularly, the former training strategy gets the better specificity for DEJU, HETH, VATH, HAFL, and WETA classes, while the latter strategy instead of WETA the HETH class also gets the highest value. At the same time, only the HALF class remains the best in terms of accuracy.



(a) MD

(b) PS

(c) HOG

(d) AF

Figure 4: Relationship of performed F scores for each considered feature set between both methods: the proposed unsupervised-based segmentation (vertical axis) and supervised reference (horizontal axis).

Table 3: Performance classification measures for the feature sets: AF (above) and AF-Br (below). The best column-wise measures achieved overall data-sets are highlighted in bold.

	Class	r	s	p	a	F
AF	BRCR	0.91	0.76	0.68	0.81	0.77
	WIWR	0.92	0.99	0.94	0.97	0.93
	PSFL	0.87	0.89	0.77	0.88	0.82
	RBNU	0.86	0.98	0.88	0.96	0.87
	DEJU	0.05	**1**	**1**	0.96	0.10
	OSFL	0.86	0.92	0.67	0.91	0.76
	HETH	0	**1**	0/0	0.97	0
	CBCH	0.74	0.89	0.64	0.86	0.68
	VATH	0.84	**1**	**1**	0.97	0.91
	HEWA	0.57	**1**	0.95	0.95	0.72
	SWTH	0.47	0.99	0.93	0.92	0.63
	HAFL	**1**	**1**	**1**	**1**	**1**
	WETA	0.77	0.99	0.92	0.98	0.84
AF-Br	BRCR	0.87	0.85	0.77	0.86	0.82
	WIWR	0.88	0.98	0.94	0.96	0.90
	PSFL	0.78	0.93	0.83	0.88	0.80
	RBNU	0.82	0.97	0.84	0.95	0.83
	DEJU	0.12	**1**	**1**	0.97	0.21
	OSFL	0.85	0.91	0.65	0.90	0.74
	HETH	0	**1**	0/0	0.97	0
	CBCH	0.46	0.98	0.85	0.87	0.60
	VATH	0.97	**1**	**1**	**1**	0.98
	HEWA	0.80	0.95	0.69	0.94	0.74
	SWTH	0.70	0.99	0.91	0.95	0.80
	HAFL	0.99	**1**	**1**	**1**	**1**
	WETA	0.72	**1**	0.97	0.97	0.82

To provide a better illustration, Table 4 shows the obtained results of the paired F-score difference, ΔF, computed between the AF and AF-Br features. Since the latter set is the reference, the estimated ΔF value gets negative sign when the reference feature set is better, otherwise ΔF becomes a positive value. The best and worst achieved ΔF values are marked in bold. Particularly, the SWTH gets the lowest difference (-0.169), that is, that class is the most negatively influenced by the proposed strategy while the CBCH class achieves the best influence (0.084). However, though the average value is $\Delta F = -0.017$ in support of AF-Br set, the corresponding estimated t value gets as high as 0.905, meaning that neither of the considered feature sets are statistically different at levels of 10%, 5%, and 1%.

Table 4: Computation example of the paired F-score difference, ΔF, that gets negative sign when the reference feature set is better, otherwise ΔF becomes positive.

Dataset	*AF*	*AF-Br*	ΔF
BRCR	0.774	0.819	-0.045
WIWR	0.932	0.905	0.027
PSFL	0.817	0.801	0.016
RBNU	0.871	0.831	0.040
DEJU	0.095	0.214	-0.119
OSFL	0.756	0.738	0.018
HETH	0	0	0
CBCH	0.685	0.600	**0.084**
VATH	0.911	0.983	-0.072
HEWA	0.715	0.741	-0.026
SWTH	0.627	0.796	-**0.169**
HAFL	1	0.996	0.004
WETA	0.840	0.823	0.017
Average	0.694	0.711	-0.017

Lastly, the null-hypothesis values are shown in Table 5 in order to make clear the influence of each considered segmentation strategy, in terms of performed F-score classification measure. Values are computed at 10% level (a strong assumption) to either admit (value 0) or deny (value ± 1) the null hypothesis about their statistical similarity between each pair of contrasted feature sets (columns stand for the reference supervised feature sets and rows for the proposed unsupervised sets). In case the statistical similarity is rejected, the null hypothesis gets the value 1 if the column-wise feature set reaches a better performance than the the row-wise set. Otherwise, the null hypothesis gets the value -1.

As seen from the obtained main diagonal matrix values, one can infer that each compared feature set, except MD, has no statistical difference regardless of the used segmentation approach. In case of the MD set, its unsupervised version turns out to be better. This situation should be expected due to the above-given explanation about the advantage of automated MD feature extraction. As a result, the proposed birdsong classification methodology based on unsupervised segmentation of audio recordings and multiple instance learning has no statistical difference with the baseline supervised version, at least, when using the three extracted feature sets: PS, HOG, and AF.

Table 5: Values of null-hypothesis test computed at 10% level for both considered training segmentation strategies: supervised and unsupervised. Main diagonal elements marked in bold.

	MD	*PS*	*HOG*	*AF*
MD-Br	**1**	0	0	1
PS-Br	0	**0**	-1	0
HOG-Br	0	0	**0**	1
AF-Br	0	-1	-1	**0**

4 Discussion and Concluding Remarks

In this paper, the use of unsupervised segmentation of audio birdsong recordings is investigated along with multiple instance learning to classify among a given number of bird species. The proposed unsupervised segmentation of audio birdsong recordings is contrasted against its baseline reference supervised version requiring manual annotation of properly computed spectrograms, as described in [Briggs *et al.*, 2012]. Yet, since this manual recording segmentation poses as a very fatiguing task, we indirectly estimate the quality of the proposed segmentation method by the introduced two-class F-score as classification performance measure that is not affected by the relative class size. Afterwards, each one of the considered feature sets are compared in terms of the paired *t*-test, for which the null hypothesis states that the achieved F-score performance of two given classification sets can be statistically assumed as the same. The univariate paired *t*-test is preferred due to its simple interpretation though other multivariate tests may be used, for example, the multivariate paired Hotelling's T^2 test that provides similar results in our work. Both segmentation approaches are validated for the four feature sets: MD, PS, HOG, as well as the all-features set. In average, the MD feature set benefits the most from the use of the unsupervised segmentation, while the HOG set degrades the worst, as seen in Table 2. The main reason for the latter results is the fact that parameter tuning of the automated HOG feature extraction should be improved. However, this procedure is out of the scope of the present work. Nonetheless, according to the accomplished values of the null-hypothesis test shown in Table 5, the introduction of the unsupervised segmentation of audio recordings has no statistical difference with the baseline supervised version, at least, when using the following three extracted feature sets: PS, HOG, and AF.

This work provides a birdsong recognition framework using the MILES classifier, which in turn uses an exponential kernel as instance (dis)similarity measure. Even though performed F-score values are high, this classifier is sensitive to a low number of training recordings. As a conclusion, the proposed unsupervised recording segmentation of audio birdsong recordings improves species classification with the benefit of easier implementation since no manual handling of recordings is required, making feasible the design of fully automatic birdsong recognition systems.

As future work, the authors consider that more studies must be undertaken on improving the feature extraction stage. Besides, the use of classifiers demanding less input training sets remains an important issue since, in practice, collecting labeled audio birdsong recordings is very costly.

Acknowledgment

This work is supported by *"Programa de formación doctoral en Colombia"* COLCIENCIAS and "Programa de Internacionalización del Conocimiento - Modalidad 3" from Universidad Nacional de Colombia.

References

[Aide *et al.*, 2013] T. Mitchell Aide, Carlos Corrada-Bravo, Marconi Campos-Cerqueira, Carlos Milan, Giovany Vega, and Rafael Alvarez. Real-time bioacoustics monitoring and automated species identification. *PeerJ*, 1:e103, 2013.

[Blumstein *et al.*, 2011] Daniel T. Blumstein, Daniel J. Mennill, Patrick Clemins, Lewis Girod, Kung Yao, Gail Patricelli, Jill L. Deppe, Alan H. Krakauer, Christopher Clark, Kathryn A. Cortopassi, Sean F. Hanser, Brenda McCowan, Andreas M. Ali, and Alexander N. G. Kirschel. Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus. *Journal of Applied Ecology*, 48(3):758–767, 2011.

[Bramer, 2013] Max Bramer. *Principles of Data Mining*. Springer, second edition, 2013.

[Briggs *et al.*, 2012] Forrest Briggs, Balaji Lakshminarayanan, Lawrence Neal, Xiaoli Z Fern, Raviv Raich, Sarah J K Hadley, Adam S Hadley, and Matthew G Betts. Acoustic classification of multiple simultaneous bird species: a multi-instance multi-label approach. *The Journal of the Acoustical Society of America*, 131(6):4640–4650, June 2012.

[Chen *et al.*, 2006] Yixin Chen, Jinbo Bi, and James Z Wang. MILES: Multiple-Instance Learning via Embedded Instance Selection. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 28(12):1931–1947, 2006.

[Cheplygina *et al.*, 2015] Veronika Cheplygina, David M.J. Tax, and Marco Loog. Multiple instance learning with bag dissimilarities. *Pattern Recognition*, 48(1):264 – 275, 2015.

[Dalal and Triggs, 2005] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.

[Dennis *et al.*, 2011] Jonathan Dennis, Huy Dat Tran, and Haizhou Li. Spectrogram Image Feature for Sound Event Classification in Mismatched Conditions. *Signal Processing Letters, IEEE*, 18(2):130–133, 2011.

[Depraetere *et al.*, 2012] Marion Depraetere, Sandrine Pavoine, Fréderic Jiguet, Amandine Gasc, Stéphanie Duvail, and Jerôme Sueur. Monitoring animal diversity using acoustic indices: Implementation in a temperate woodland. *Ecological Indicators*, 13(1):46 – 54, 2012.

[Fagerlund, 2007] Seppo Fagerlund. Bird species recognition using support vector machines. *EURASIP Journal on Advances in Signal Processing*, 2007(1):64–64, 2007.

[Hao *et al.*, 2012] Yuan Hao, Bilson Campana, and Eamonn Keogh. Monitoring and mining animal sounds in visual space. *Journal of Insect Behavior*, 26(4):466–493, November 2012.

[Neal *et al.*, 2011] Lawrence Neal, Forrest Briggs, Raviv Raich, and Xiaoli Z. Fern. Time-frequency segmentation of bird song in noisy acoustic environments. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2012–2015, 2011.

[Otsu, 1979] Nobuyuki Otsu. A Threshold Selection Method from Gray-level Histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66, 1979.

[Pourhomayoun *et al.*, 2013] Mohammad Pourhomayoun, Peter J. Dugan, Marian Popescu, and Christopher W. Clark. Bioacoustic signal classification based on continuous region processing, grid masking and artificial neural network. *CoRR*, abs/1305.3635, 2013.

[Tax, 2013] David M.J. Tax. MIL, a Matlab toolbox for multiple instance learning, Mar 2013. version 0.8.1.

[Trifa *et al.*, 2008] Vlad M Trifa, Alexander N G Kirschel, Charles E Taylor, and Edgar E Vallejo. Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models. *The Journal of the Acoustical Society of America*, 123(4):2424–2431, 2008.

[van den Boomgaard and van Balen, 1992] Rein van den Boomgaard and Richard van Balen. Methods for fast morphological image transforms using bitmapped binary images. *Graphical Models and Image Processing*, 54(3):252–258, 1992.

[Zakaria *et al.*, 2012] Jesin Zakaria, Sarah Rotschafer, Abdullah Mueen, Khaleel Razak, and Eamonn Keogh. Mining Massive Archives of Mice Sounds with Symbolized Representations. In *SIAM International Conference on Data Mining (SDM)*, pages 588–599, 2012.