SVM Candidates and Sparse Representation for Bird Identification

Rodrigo Martinez[1], Laura Silva [2], Esau Villarreal[2,3], Gibran Fuentes[3], and Ivan Meza[3]
[1]Facultad de Ciencias (FC)
http://ciencias.unam.mx
[2]Facultad de Estudios Superiores - Zaragoza (FES-Zaragoza)
http://www.zaragoza.unam.mx
[3]Instituto de Investigaciones en Matematicas Aplicadas y en Sistemas (IIMAS)
http://www.iimas.unam.mx
Universidad Nacional Autonoma de Mexico (UNAM)
http://www.unam.mx

No Institute Given

Abstract We present a description of our approach for the "Bird task Identification LifeCLEF 2014". Our approach consists of four stages: (1) a filtering stage for the filtering of audio bird recordings; (2) segmentation stage for the extraction of syllables; (3) a candidate generation based on HOG features from the syllables using SVM; and (4) a species identification using a Sparse Representation-based Classification of HOG and LBP features. Our approach ranked seventh team-wise in the challenge and showed a poor performance in the fourth stage.

1 Introduction

In this work we present the description of our system submitted to the *LifeCLEF 2014 Bird task* [2] part of the *LifeCLEF 2014 Laboratory* [3]. This task is concerned with the identification of bird species based on their signing. This setting has potential applications on ecological surveillance or biodiversity conservation. This year the task was formally defined as:

> The task will be focused on bird identification based on different types of audio records over 501 species from South America centered on Brazil. Additional information includes contextual meta-data (author, date, locality name, comment, quality rates). The main originality of this data is that it was specifically built through a citizen sciences initiative conducted by Xeno-canto, an international social network of amateur and expert ornithologists. This makes the task closer to the conditions of a real-world application: (i) audio records of the same species are coming from distinct birds living in distinct areas (ii) audio records by different users that might not used the same combination of microphones and portable recorders (iii) audio records are taken at different

periods in the year and different hours of a day involving different background noise (other bird species, insect chirping, etc).[1]

At the core of our approach is the Sparse Representation-based Classification (SRC) [5], a methodology that has been quite successful in face recognition. We adapted SRC to work at the syllable-level. In addition, our approach is composed of filtering, syllable extraction and candidate generation. In the filtering stage, the audio recordings are uniformly processed to be on the same bandwidth and to eliminate stationary noise. In the syllable extraction, the system identifies a set of syllables based on short time energy filter. Finally, we generate a set of candidate species based on the syllable information. For a recording, a set of candidates per syllable is ranked to generate a unique set.

The outline of this paper is as follows. Section 2 presents the architecture of our approach. Section 3 explains the preprocessing stage, section 4 the extraction of syllables stage, section 5 the candidates generation stage, section 5 the candidates generation stage, section 6 the identification stage. Section 7 presents our results. Finally, section 8 presents some conclusions and discusses about future work.

2 Architecture of the approach

Our approach is composed of four stages represented in Figure 1. The first stage filters the recording in the frequency domain. The second stage extracts bird syllables from the filtered signal. We have two settings for this, a coarse and fined grained setting. The third stage has the goal of creating a set of n candidates given a syllable. A final set of candidates for the recording is produced combining the candidate sets of each syllable. The model for the candidates is generated using the fine grained syllables. Finally, the four stage has the goal of doing the identification using the Sparse Representation-based Classification of the syllables. It tries to select from the candidate set the species with more resemblance to the examples from a dictionary based on hand picked syllables. The syllables are based on the coarse segmentation, and it relies on the representation of the syllable as a visual feature.

3 Filtering

There are two aspects we follow for the filtering of the signal. First we re-sample the original recordings from 44100hz to 16000hz. After this we apply a bandpass FIR filter between 500hz and 4500hz frequencies. Empirically we identified that most of the singing frequencies were located in this bandwidth. However, the low performance of our approach points to review this assumption. Figure 2 shows the effect of this filtering in one of the recordings.

4 Segmentation of syllables

The segmentation of syllables is done using a short time energy filter. A threshold defines what is considered activity. Each recording in the database is segmented after

[1] From http://www.imageclef.org/node/180 (May, 2014)

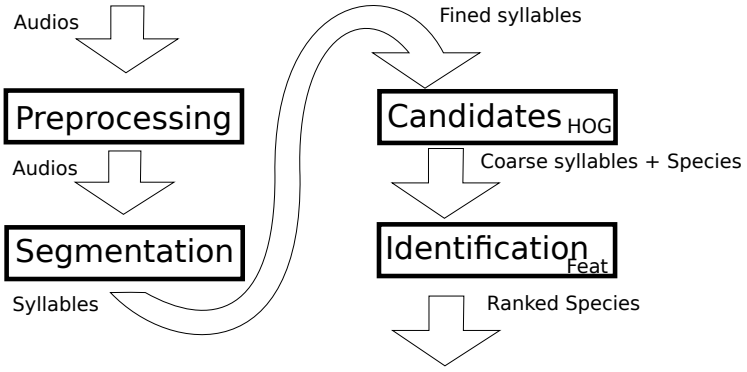

Figure 1. System architecture for the identification of species of bird signing.

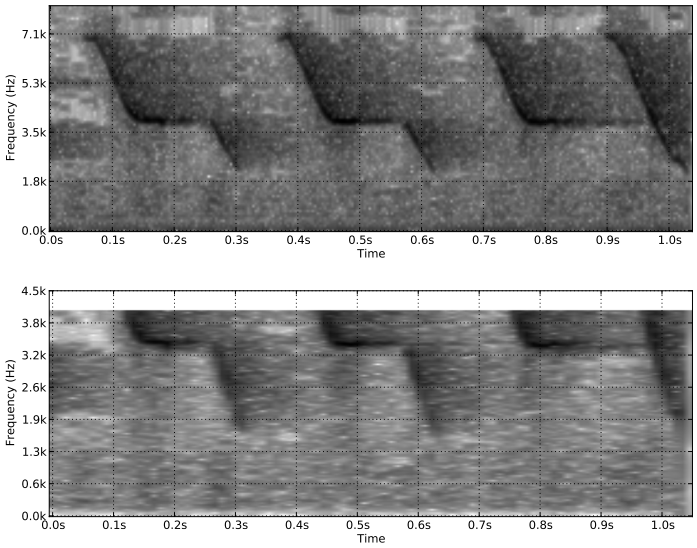

Figure 2. Example of the filtering of a recording. The spectrogram above represents the original recording. The spectrogram above represents the recording after filtering.

being filtered from the previous stage. Figure 3 shows the syllable identified by this stage.

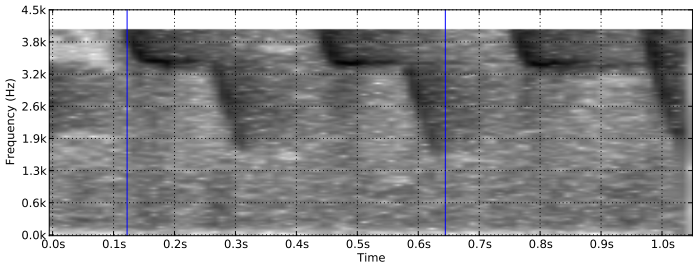

Figure 3. Segmented syllables for a recording using a short time energy filter.

Syllables are normalized by resizing them into a specific size in the frequency domain (100x100 pixels). Figure 4 shows a syllable after the normalization process. We have defined two thresholds for fined and coarse segmentation. Table 3 summarizes the number of syllables extracted for each type of segmentation in both databases.

Table 1. Number of syllables extracted from the recordings.

	Train	Test
Fined grained	$52,666$	$23,953$
Coarse grained	$47,899$	

5 Candidates generation

A set of candidates of the possible species is generated using a Support Vector Machine. For this we extract visual features from the segmented syllables. In particular we experimented with Histograms of Oriented Gradients (HOG) [1] and Local Binary Patterns (LBP) [4]. For each syllable in a recording we extract the candidates. These candidate sets are agglomerated into a final set. Figure 5 shows the effect on the size of the set containing the target species. For our experiments we define the size of the set to be 200 candidate species.

6 Sparse Representation-based Classification

For the identification stage we follow the method of *Wright et. al.* [5], originally proposed for the face recognition problem in which has been successful. We adapted the methodology to work at the syllable level on this bird task.

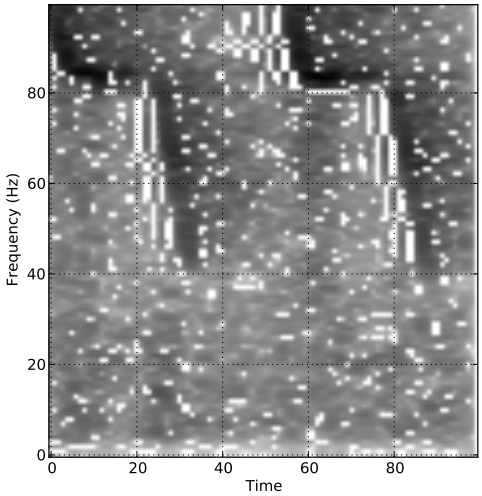

Figure 4. Normalized syllable.

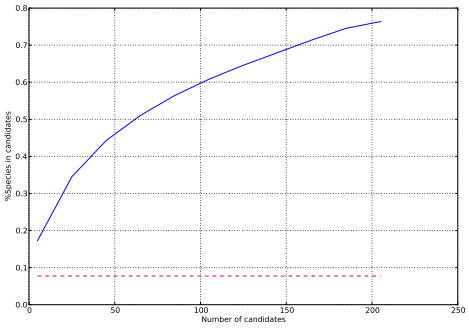

Figure 5. Percentage of times the target species belongs to the candidates (blue curve). Percentage of times the target species is the first candidate (red line)

The method relies on a dictionary representation of the syllables of the i species. Each specie is represented by j instances of species' syllables, which is given by a vector of arity m. Together candidates and instances define the dictionary matrix A of MXN dimension (where $M = ixj$). Given an unknown instance of a syllable y the goal of the method is to identify the vector x which represents the contributions of elements of the dictionary A to generate the syllable y. In other words, the contribution of each syllable to generate the unknown syllable. Once the contribution of each element of the dictionary is identified, it is a matter of quantifying the contribution by each candidate and decide if the contribution is enough to conclude that they represent the same person.

In order to identify the contribution of each candidate, SRC uses the ℓ_1 minimization:

$$\begin{aligned} &\text{minimize} \quad x = argmin||x||_1 \\ &\text{subject to} \quad\quad\quad Ax = y \end{aligned} \tag{1}$$

This minimizes the sum of the individual contributions of a candidate such that the multiplication of the metric dictionary A and the contributions x generate the unknown instance. To perform this minimization we use the homotopy method since it is fast to create a good approximation of the vector x [6].

Once the vector x is identified the method proposes to calculate the square residuals per instance in the dictionary:

$$r_i = |y - Ax_i| \tag{2}$$

In which x_i is the contribution vector with the values for the candidates different to i zeroed. In this way, the r_i represents a score of the contribution of the instances of the candidate i. After calculating the residual per candidate, we look to identify the candidate which has the lower residual, this represent the one which is less different from the candidate and it is our identified person.

We adapted this setting for the identification of bird species. First, the candidate species were extracted from the third stage of our system. Given the list of candidates from this stage we generated the matrix A. For our experiments i was variable but j was set to 5. In particular the instances of A were hand picked for experts in the field as good syllables examples for a species. The size m for the vector depended on the representation HOG or LBP features. At this stage we used the coarse syllables extracted in the second stage.

We performed the methodology explained above for each syllable. We collected the identified species and ranked them by the probability obtained in the stage of candidate generation. This sorted list was used to produce the output required by the challenge.

7 Experimental Results

We submitted three configurations of our system:

100 HOG + HOG The sparse system used the 100 top candidates generated with HOG features, and the HOG features to identify the species.

50 HOG + HOG The sparse system used the 50 top candidates with HOG features, and the HOG features to identify the species.

50 HOG + LBP The sparse system used the 50 top candidates generated with HOG features, and the LBP features to identify the species.

Table 2. Mean average precision of identification of bird species in testing.

	Background species	without backbround species
100 HOG + HOG	10.5%	12.9%
50 HOG + HOG	10.4%	12.8%
50 HOG + LBP	7.4%	8.9%

As you can notice the performance of our system was poor. To reduce the amount of candidates did not have a significant improvement. On the other hand, the use of LBP affected the performance.

In order to analyse the labellings produced by our system, we analyse the results over a subset of the training corpus, those marked with more than one bird singing. We found that only 209 bird species were correctly recover. Table **??** shows the 10 most successful species. However, our performance is so poor that it is hard to account for the errors on the rest of the species at the moment.

Table 3. List of best identified species in recording with more than one species registered.

Laterallus viridis	Emberizoides herbicola
Cercomacra melanaria	Cyanocorax cristatellus
Nyctibius griseus	Setopagis parvulus
Hypocnemis hypoxantha	Procnias nudicollis
Synallaxis cinerascens	Crypturellus tataupa

8 Conclusions and Future work

These working notes present our system proposal for the identification of bird species through singing. This proposal was built in the context of the *LifeCLEF 2014 Bird task* [2], a part of the *LifeCLEF 2014 Laboratory*[3]. Our approach first generates candidates using SVM and the identification of the species at the syllable level using a sparse representation. As result of the challenge we identify several problems with our setting which had a poor performance in the challenge 25% of the best. We have several hypothesis of what could had been wrong. First, our filtering of the recording was to aggressive. Second, the segmentation of the syllables was not at the level and in many cases segmented more than syllable. Third, the identification stage fail to identify the species at a good rate 17% with 100 candidates. Fourth, we did not use information of the metadata or at the song level of the species.

In the future we aim to generate a better setting for the filtering and syllable segmentation, maybe by incorporating elements of other approaches. We also would like to continue experimenting with the setting of identification through SRC to fully discarded or to find the correct way to set it up in the task of bird identification.

References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. Conference on Computer Vision and Pattern Recognition, San Diego, USA (Junio 2005)
2. Goëau, H., Glotin, H., Vellinga, W.P., Rauber, A.: Lifeclef bird identification task 2014. In: CLEF working notes 2014 (2014)
3. Joly, A., Müller, H., Goëau, H., Glotin, H., Spampinato, C., Rauber, A., Bonnet, P., Vellinga, W.P., Fisher, B.: Lifeclef 2014: multimedia life species identification challenges. In: Proceedings of CLEF 2014 (2014)
4. Wang, L., He, D.: Texture classification using texture spectrum. Pattern Recognition (8), 905 – 910 (1990)
5. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. Pattern Analysis and Machine Intelligence, IEEE Transactions on 31(2), 210–227 (2009)
6. Yang, A., Zhou, Z., Balasubramanian, A., Sastry, S., Ma, Y.: Fast ℓ_1 -minimization algorithms for robust face recognition. Image Processing, IEEE Transactions on 22(8), 3234–3246 (Aug 2013)