Large-scale identification of birds in audio recordings

Notes on the winning solution of the LifeCLEF 2014 Bird Task

Mario Lasseck

Animal Sound Archive
Museum für Naturkunde Berlin
Mario.Lasseck@mfn-berlin.de

Abstract. The Bird Identification Task of the LifeCLEF 2014 lab is to automatically identify 501 different species in over 4000 audio recordings collected by amateur and expert ornithologists through a citizen sciences initiative. It is one of the biggest bird classification challenges so far considering the quality, quantity and variability of the recordings and the very large number of different species to be classified. The solution presented here achieves a Mean Average Precision of 51.1% on the test set and 53.9% on the training set with an Area Under the Curve of 91.5% during cross-validation.

Keywords: Bird Identification · Information Retrieval · Citizen Sciences · Image Segmentation · Median Clipping · Template Matching · Decision Trees

1 Introduction and Task Overview

The LifeCLEF 2014 Bird Identification challenge asks participants to automatically identify the vocalizing species in 4339 audio recordings with undetermined content. For training, 9688 audio recordings paired with metadata including dominant and background species are provided. A recording may contain only one or up to 11 simultaneously vocalizing birds. What makes this challenge unique but also quite difficult is the very large amount of data, the high variability of the recordings, both in quality and content and of course the large number of different species to be classified. The all in all 14,027 audio files, if added together 33.3 GB of data with over 4.5 days of acoustic material, are provided by Xeno-canto (http://www.xeno-canto.org/). The files were recorded between 1979 and 2013 in over 2000 different locations centered on Brazil by almost 250 amateur and expert ornithologists, using different combinations of microphones and portable recorders. The duration of the recordings varies from half a second to several minutes. Also the quality of the audio files is quite diverse and challenging. One has to deal with all kinds of background noise and in some cases artifacts due to lossy mp3 data compression.
An overview and further details about the LifeCLEF Bird Identification Task is given in [1]. The task is among others part of the CLEF 2014. A general overview of all tasks can be found in [2,3,4].

2 Feature Extraction

The features used for classification are taken from three different sources briefly described in the following sections.

2.1 Metadata

The first source for feature extraction is the provided metadata. Each audio file is paired with additional contextual information about the date, time, location and author of the recording. This information is used to extract 8 features per file:

- Year
- Month
- Time
- Latitude
- Longitude
- Elevation
- Locality Index
- Author Index

To use the provided metadata a few steps had to be taken for preparation. From the recording date only the year and month were extracted and considered as relevant features. The recording time was converted in minutes. Since only numeric values can be used as features, for locality and author a look up table was created and the corresponding index was used. All missing or none numeric values were replaced by the mean value of its category.

2.2 openSMILE

The openSMILE feature extraction tool [5] was used to extract a large number of features per audio recording. The framework was configured with the *emo_large.conf* configuration file written by Florian Eyben. It was originally designed for emotion detection in speech signals but was also recently applied in the field of audio scene analysis [6]. The here used configuration file first calculates 57 so called low-level descriptors (LLD) per frame, adds delta (velocity) and delta-delta (acceleration) coefficients to each LLD and finally applies 39 statistical functionals after moving average smoothing the feature trajectories.

The 57 LLDs consist of:

- 35 spectral features
  - Mel-Spectrum bins 0-25
  - zero crossing rate
  - 25%, 50%, 75% and 90% spectral roll-off points
  - spectral flux
  - spectral centroid
  - relative position of spectral minimum and maximum
- 13 ceptral features
  - MFCC 0-12
- 6 energy features
  - logarithmic energy
  - energy in frequency bands 0-250 Hz, 0-650 Hz, 250-650 Hz, 1000-4000 Hz and 3010-9123 Hz
- 3 voicing-related features
  - F0
  - F0 envelope
  - voicing probability

To describe an entire audio recording, statistics are calculated from all LLD, velocity and acceleration trajectories by 39 functionals including e.g. means, extremes, moments, percentiles and linear as well as quadratic regression. This sums up to 6669 (57 x 3 x 39) features per recording. Further details regarding openSMILE and the extracted features can be found in the openSMILE 1.0.1 manual and the *emo_large.conf* configuration file (http://opensmile.sourceforge.net/).

2.3 Segment-Probabilities

The idea of using the matching probabilities of segments as features or more precisely the maxima of the normalized cross-correlation [7] between segments, also referred to as region of interests (ROIs) or templates, and spectrogram images was previously used by Nick Kriedler in The Marinexplore and Cornell University Whale Detection Challenge, Fodor Gabor in the MLSP 2013 Bird Classification Challenge [8] and Ilyas Potamitis in the NIPS 2013 Bird Song Classification Challenge [9].
For the current competition an adaptation of this method was used which was already very successfully applied also in the NIPS 2013 Challenge [10]. It differs mainly in the way how segments are extracted and which subsets of segments and their probabilities are used during classification. It turned out that proper preprocessing and segmentation of the spectrogram images is a key element to improve classification performance. The number of segments should be rather small but still representative, capturing typical elements and combinations of sounds of the species to be identified.

The following sections give a brief overview of the feature extraction steps regarding Segment-Probabilities. Some additional details can be found in [10].

Preprocessing and Segmentation. As mentioned above the way of preprocessing and segmentation is crucial to gather a good repertoire of segments especially when dealing with unknown content and noisy recordings. The following steps were performed for each audio file in the training set:

- resample to 22050 Hz
- get spectrogram via STFT (512 samples, hanning window, 75% overlap)
- normalize spectrogram to 1.0
- remove 4 lowest and 24 highest spectrogram rows
- get binary image via *Median Clipping* per frequency band and time frame by setting each pixel to 1, if it is above 3 times the median of its corresponding row AND 3 times the median of its corresponding column, otherwise to 0
- apply closing, dilation and median filter for further noise reduction
- label all connected pixels exceeding a certain spatial extension as a segment
- define its size and position by a rectangle with a small area added to each direction

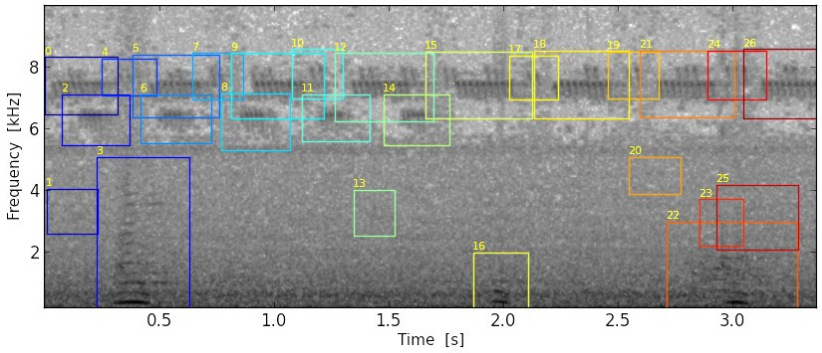

Fig. 1. Example of a spectrogram image (log) with marked segments (MediaId: 86)

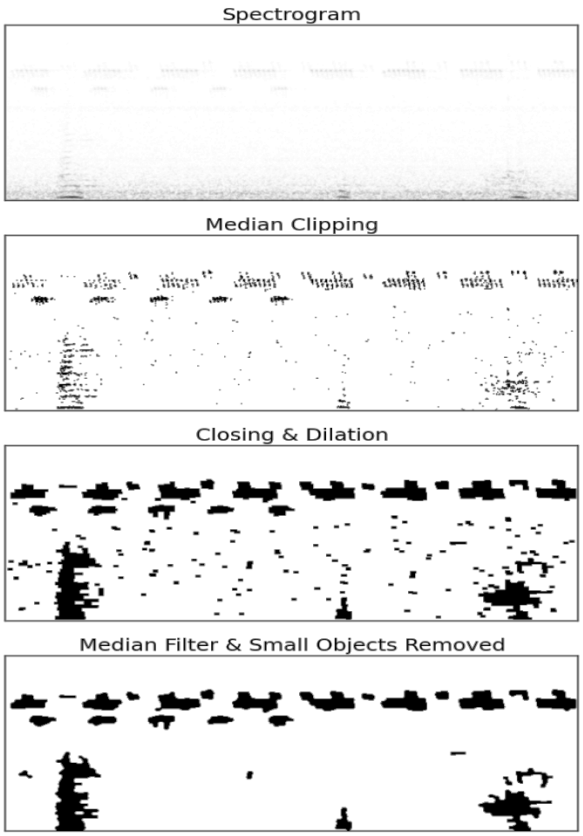

Fig. 2. Preprocessing steps for segmentation (MediaId: 86)

Selection of typical Segments per Species. In opposite to the Metadata and openSMILE feature sets that are species respectively class independent, Segment-Probabilities form individual feature sets for each species. In order to get a small but representative set of features per species, only segments from files without background species and very good quality (metadata: Quality = 1) were selected. For some species this condition was too strict, leading to none or too few segments. The following queries were applied successively for every target species until there was at least one file that met the conditions and the number of retrieved segments was greater than 40:

Select all segments of files WHERE: Species = target species AND:

1. BackgroundSpecies = {} AND Quality = 1
2. BackgroundSpecies = {} AND (Quality = 1 OR Quality = 2)
3. BackgroundSpecies = {} AND (Quality = 1 OR Quality = 2 OR Quality = 3)
4. Quality = 1 OR Quality = 2

The number of segments retrieved this way sums up to 492,753 for all training files with an average of approximately 984 segments per species.

Template Matching. After selecting a set of segments for each species, template matching was performed to get an individual feature set per species. The highest matching probability was determined using normalized cross-correlation after applying a gaussian blur to segment and target image. Due to the large number of audio files and segments used as templates, the method described in [10] was way too time consuming and had to be modified. In order to speed up the process the following changes were applied:

- segments and target spectrogram images were calculated via STFT using only 50% overlap (instead of 75%)
- search range for segments within the target images along the frequency axes was set to ± 3 pixel (instead of 4 pixel)
- segments and target spectrogram images were converted to 8 bit unsigned integer before the template matching procedure (instead of 32 bit floating point)

Even with these modifications, the process of template matching (sliding almost half a million templates over 14,027 target images) took very long and kept four computers with regular hardware quite busy for several days.

3 Feature Selection

To cope with the large number of features and to improve and speed up the classification process a reduction of features was inevitable. It was performed in two phases, before and during classification.
The openSMILE features were reduced from 6669 to 1277 features per file before the actual classification step. This was done by recursive feature elimination with the scikit-learn [11] RFECV selector [12] and a support vector machine with linear kernel and 2-fold cross-validation. For this preselection only a small subset from the training data consisting of 50 species and good quality files was used.
During classification, furthermore the k highest scoring features were individually selected per species using univariate feature selection. This was done separately for each fold during classifier training with cross-validation. Different values for k were tested, ranging from 150 to 400 features per class.

4 Training and Classification

Since it was optional to use the information about background species, single- and multi-label approaches were tested. In both cases the classification problem was split up into 501 independent classification problems using one classifier for each species following the one-vs.-all resp. the binary relevance method. For the single-label approach only dominant species were considered as targets. In case of the multi-label approach background species (BS), if assigned, were also considered for each training file but were set to lower probabilities compared to the dominant species. The classification was done with the scikit-learn library (ExtraTreesRegressor) by training ensembles of randomized decision trees [13] with probabilistic outputs. Following variations were used for training:

- classification methods
  - single-label
  - multi-label with probabilities of dominant species set to 1.0
    - probabilities of BS set to 0.3
    - probabilities of BS set to 0.7
    - probabilities of BS set to 1.0 (equally weighted as dominant species)
- feature sets & feature set combinations
  - Meatadata Only
  - openSMILE Only
  - Segment-Probabilities[1] (Seg.Probs.)
  - Metadata + openSMILE
  - Metadata + openSMILE + Seg.Probs.
  - openSMILE + Seg.Probs. (Audio Only)
- number of features (univariate feature selection per species in each fold)
  - 150, 170, 180, 200, 250, 300, 400
- number of folds for cross-validation
  - 10, 12, 15

With following variations of tree parameters:

- number of estimators
  - 300, 400, 500
- max_features
  - 4, 5, 6, 7
- min_sample_split
  - 1, 2, 3, 4, 5

[1] By the time of the submission deadline Segment-Probabilities were extracted for 485 species. The remaining 16 species used Metadata + openSMILE features for classification.

During cross-validation using stratified folds the probability of each species in all test files was predicted and averaged. Additionally each species was predicted in the held out training files for validation. This way it was possible to choose a variation and/or parameter set separately per species and to increase the MAP score on the test files by optimizing the MAP score on the training files.

5 Results

In Table 1 the results of the four submitted runs are summarized using evaluation statistics based on the mean of the Area Under the Curve (AUC) calculated per species and the Mean Average Precision (MAP) on the public training and the private test set. All four runs outperformed the runs of the other participating teams.

Table 1. Performance of submitted runs (without / with background species)

	Public Training Set		Private Test Set
Run	**Mean AUC [%]**	**MAP [%]**	**MAP [%]**
1	91.4 / 85.0	53.7 / 48.6	50.9 / 45.1
2	91.1 / 84.9	49.4 / 44.6	49.2 / 43.7
3	91.5 / 85.1	53.9 / 48.7	51.1 / 45.3
4	91.4 / 85.3	50.1 / 45.3	50.4 / 44.9

For the first and the best performing third run a mix of parameter sets individually selected per species was used. As mentioned above the selection was based on how a particular set of trainings parameter was able to increase the overall MAP on the held out training files during cross-validation. A higher mean AUC score might be a hint of a generally good selection of training parameters but it is still possible that for some classes (species) a different selection works better. To give an example, in Fig. 3 AUC scores are visualized per species using one of the three different feature sets exclusively during training. On average the use of Segment-Probabilities outperforms the other feature sets but for some species the openSMILE and in rare cases even the Metadata feature set is a better choice.

For the winning third run a list of the parameters and feature sets used for each species, together with their individually achieved AUC scores can be downloaded from `http://www.animalsoundarchive.org/RefSys/LifeCLEF2014`. Here one can also find additional figures visualizing the preprocessing, segmentation and the most important segments used for classification.

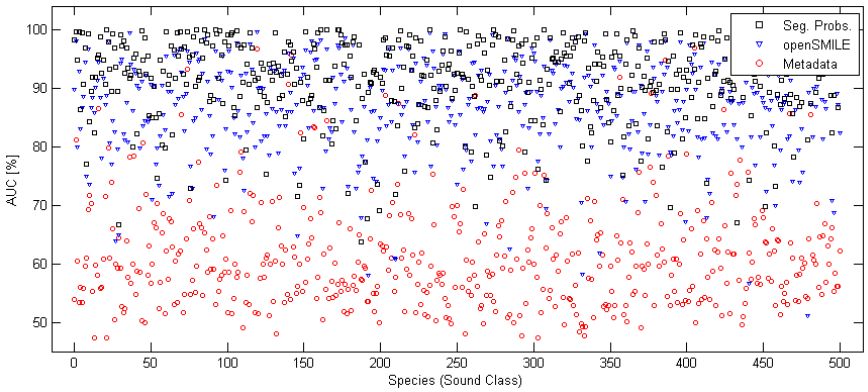

Fig. 3. AUC scores per species for individual feature sets calculated on left out training files during cross-validation (without background species)

To give an impression about the influence of different parameter settings on classification performance over the entire training set, run 4 was altered in several ways and corresponding evaluation statistics are visualized in Fig. 4 and 5.

Parameters used for Run 4:

- classification method: single-label
- feature set: Segment-Probabilities
- number of features: 150
- number of folds: 15
- number of estimators: 500
- max_features: 7
- min_sample_split: 4

Parameter variations of Run 4:

- Run 2 → max_features: 6 & min_sample_split: 3
- Run 5 → feature set: Metadata
- Run 6 → feature set: openSMILE
- Run 7 → feature set: openSMILE + Segment-Probabilities (Audio Only)
- Run 8 → method: multi-label and background species weighted with 0.3
- Run 9 → method: multi-label and background species weighted with 0.7
- Run 10 → method: multi-label and background species weighted equally

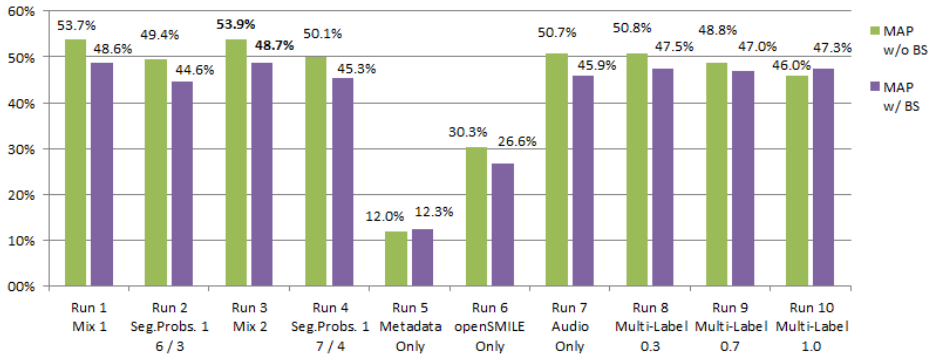

Fig. 4. Mean Average Precision (MAP) of Runs

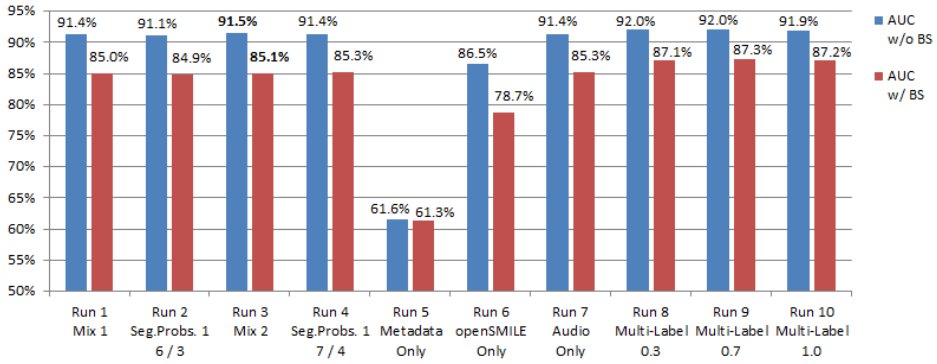

Fig. 5. Area Under the Curve (AUC) of Runs

6 Discussion and Future Work

To use matching probabilities of segments as features was once again a good choice. The drawback of this method is that the template matching procedure to calculate the feature sets takes quite a long time especially if the number of species and audio files is as large as in the current challenge. To use image-pyramids could help to speed up the process and would be worth to investigate in the near future. The features derived from metadata and the ones calculated with openSMILE did not perform as well but they could increase the overall classification performance by improving the results for individual species. Considering the use of the openSMILE tool there is still a lot of room for improvement. The configuration file could be altered to better capture the characteristics of bird sounds.

Furthermore a preselection of features on a per species basis to get individually designed feature sets for each species class, like done for Segment-Probabilities, could be advantageous. Also worth considering is windowing the audio files, classifying the fixed length sections and averaging the results via majority voting.

Acknowledgments. I would like to thank Hervé Glotin, Hervé Goëau, Andreas Raube and Willem-Pier Vellinga for organizing this competition, the Xeno-Canto foundation for nature sounds for providing the audio recordings and the French projects Pl@ntNet (INRIA, CIRAD, Tela Botanica) and SABIOD Mastodons for supporting this task. I also want to thank Dr. Karl-Heinz Frommolt for supporting my work and providing me with the access to the resources of the Animal Sound Archive [14]. The research was supported by the DFG (grant no. FR 1028/4-1).

References

1. Glotin H., Goëau H., Vellinga W.-P., Rauber A. (2014) LifeCLEF Bird Identification Task 2014, In: CLEF working notes 2014
2. Joly A., Müller H., Goëau H. et al. (2014) LifeCLEF 2014: multimedia life species identification challenges, In: Proceedings of CLEF 2014
3. Caputo B., Müller H., Martinez-Gomez J. et al. (2014) ImageCLEF 2014: Overview and analysis of the results, In: CLEF proceedings, 2014, Springer Berlin Heidelberg, Lecture Notes in Computer Science
4. Cappellato L., Ferro N., Halvey M., and Kraaij W., editors (2014). CLEF 2014 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-1180/
5. Eyben F., Wöllmer M., Schuller B. (2010) openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor, In: Proc. ACM Multimedia (MM), ACM, Florence, Italy, ACM, ISBN 978-1-60558-933-6, pp. 1459-1462, doi:10.1145/1873951.1874246
6. Geiger J.T., Schuller B., Rigoll G. (2013) Large-Scale Audio Feature Extraction and SVM for Acoustic Scenes Classification, In: Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on. IEEE
7. Lewis J.P. (1995) Fast Normalized Cross-Correlation, Industrial Light and Magic
8. Fodor G. (2013) The Ninth Annual MLSP Competition: First place. Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on, Digital Object Identifier: 10.1109/MLSP.2013.6661932 Publication Year: 2013, Page(s): 1- 2
9. Potamitis I. (2014) Automatic Classification of Taxon-Rich Community Recorded in the Wild. PLoS ONE 9(5): e96936. doi: 10.1371/journal.pone.0096936
10. Lasseck M. (2013) Bird Song Classification in Field Recordings: Winning Solution for NIPS4B 2013 Competition, In: Glotin H. et al. (eds.). Proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013: 176-181
11. Pedregosa F. et al. (2011) Scikit-learn: Machine learning in Python. JMLR 12, pp. 2825-2830
12. Guyon I. et al. (2002) Gene selection for cancer classification using support vector machines, Machine Learning, 46(1-3), 389–422
13. Geurts P. et al. (2006) Extremely randomized trees, Machine Learning, 63(1), 3-42
14. Animal Sound Archive, `http://www.animalsoundarchive.org`