Instance-based bird species identification with undiscriminant features pruning - LifeCLEF 2014

Alexis Joly[1], Julien Champ[2], and Olivier Buisson[3]

[1] Inria, LIRMM, Montpellier, France
alexis.joly@inria.fr
[2] Inra, AMAP, LIRMM, Montpellier, France
julien.champ@inra.fr
[3] Institut National de l'Audiovisuel (INA), Bry-sur-Marne, France
olivier.buisson@inra.fr

Abstract. This paper reports the participation of Inria to the audio-based bird species identification challenge of LifeCLEF 2014 campaign. Inspired by recent works on fine-grained image classification, we introduce an instance-based classification scheme based on the dense indexing of MFCC features and the pruning of the non-discriminant ones. To make such strategy scalable to the 30M of MFCC features extracted from the tens of thousands audio recordings of the training set, we used high-dimensional hashing techniques coupled with an efficient approximate nearest neighbors search algorithm with controlled quality. Further improvements are obtained by (i) using a sliding classifier with max pooling (ii) weighting the query features according to their semantic coherence (iii) making use of the metadata to filter incoherent species. Results show the effectiveness of the proposed technique which ranked 3rd among the 10 participating groups.

1 Introduction

Building accurate knowledge of the identity, the geographic distribution and the evolution of living species is essential for a sustainable development of humanity as well as for biodiversity conservation. In this context, using multimedia identification tools is considered as one of the most promising solution to help bridging the taxonomic, i.e. the difficulty for common people to name observed living organisms and then produce or access to useful knowledge. The LifeCLEF [6] lab proposes to evaluate this challenge in the continuity of the image-based plant identification task was run within ImageCLEF the years before but with a broader scope (considering birds and fish in addition to plants and audio and video contents in addition to images). This paper particularly reports the participation of Inria ZENITH research group to the audio-based bird identification task. Inspired by some recent works on fine-grained image classification [8, 5], we introduce an instance-based classification scheme based on the dense indexing of MFCC features and the pruning of the non-discriminant ones. Section 2 describes the preliminary audio processings and features extraction steps. Section 3 then presents our raw instance-based classification scheme making use of high-dimensional hashing techniques coupled with an efficient approximate nearest neighbors search algorithm with controlled quality. Section 4 then introduces the additional semantic filtering and weighting rules that are used to improve the raw scheme. Section 6 finally reports our official results within LifeCLEF 2014 as well as few additional experiments conducted to evaluate the contribution of the different steps.

2 Pre-processing and features extraction

The dataset used for this challenge is composed of 14,027 audio recordings belonging to 501 bird species from Brazil area. As various recording devices are used, and because it is difficult to capture these sounds as birds are often far away from the recording devices, many recordings contains a lot of noise. To overcome this problem, we used SoX, the "Swiss Army knife of sound processing programs"[4]. As a first step, we used the *noisered specialised filter*, to filter out noise from the audio, and then we reduce the length of large (i.e. $> 0.1s$) silent passages from audio files to $0.1s$. In order to obtain audio files with ideally no more noise but still enough signal, we tried removing as much noise as possible (using the noisered amount parameter) while guaranteeing that the resulting audio file was at least 20% the size of the initial audio record. After this pre-processing step, we used an open source software framework, marsyas[5], to extract MFCC features with parameters based on the provided audio features in the Birdclef task : MFCC are computed on windows of 11.6 ms, each 3.9 ms, and we additionally derive their speed resulting in 26-dimensional feature vectors (13+13) for each frame.

3 Raw instance-based classification scheme

We consider a training set S of $|S|$ audio records weakly annotated with one or several labels among $|C|$ classes (corresponding to one or several bird species recognizable in the record). Each picture is described by a set of MFCC features forming a reference dataset of N features $\mathbf{x}_i$. Our instance-based classification scheme can be summarized as follows: MFCC features are extracted from the query record I_Q and searched independently in the reference set using an efficient k-NN search scheme. A sliding vote procedure followed by a max pooling step is then used to determine the best matching interval of each retrieved record and rank them accordingly. A strong classifier is finally derived from the ranked list of retrieved records via a simple top-K classifier (K=30 in all our experiments). The details of these different steps are provided below:

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

[5] http://marsyas.info/

Hash-based k-NN search. The approximate k-Nearest Neighbors (k-NN) of each MFCC feature $\mathbf{x}_j^Q$ belonging to a query record I_Q are computed efficiently thanks to the hash-based multi-probe search method introduced in [3]. Its principle is to train an adaptive search model at indexing time through kernel density estimates computed on exact k-NN samples. This model is used at search time to select the most probable buckets to be visited so as to retrieve on average a fraction α of the real K-NN's (α was set to 0.80 in all our experiments). The advantage of this method compared to other state-of-the-art methods (such as PQ-code [7] or Soft-Assignment [9]) is that it allows controlling accurately the quality of the retrieved k-NN while being applicable to any quantization function and metrics.
In our case, the original scheme is transposed to the use of a more effective hash function and to the application of the Hamming Embedding principle [1]. More precisely, we used RMMH [4], a recent data-dependent hash function family, in order to embed the original feature vectors in compact binary hash codes of 128 bits (the parameter M of RMMH was fixed to $M = 32$). The $m = log_2(N)$ first bits of the hash codes are used as keys of a hash map containing the N 128-bits hash codes divided into the 2^m buckets. Still at indexing time, the search model is estimated based on the exact k-NN's in the 128-bits Hamming space rather than in the original space. It has actually be shown in [4] that RMMH somehow works as an unsupervised metric learning technique and can provide better matches in the embedded space than in the original one. At search time, the query feature $\mathbf{x}_j^Q$ is also embedded via RMMH and the resulting 128-bits hash code $\mathbf{h}_j^Q$ is passed to the probabilistic multi-probe algorithm (see [3] for more details). Once the most probable buckets have been selected, the k nearest neighbors are computed as the k nearest hash codes according to the Hamming distance between the query hash code $\mathbf{h}_j^Q$ and the hash codes belonging to the selected buckets. The value of k was trained by cross-validation and finally fixed to $k = 200$ in all our runs.

Sliding vote and max pooling. A voting scheme is used to merge the raw independent matches produced by the k-NN search of each MFCC feature and return a ranked list of records. To increase the robusteness of the matching and reduce the influence of outliers and background features, the score of each record is computed by first voting within a sliding window (centered around each frame) and then keeping the max score over the whole record. The size of the sliding window was trained by cross-validation and then fixed to $s = 1000$ frames in all our runs (resulting in a sliding window of 3.9 seconds).

Top-K classifier. The final instance-based classifier simply consists in summing the scores of the records of each class C_j within the top-K retrieved records. For a given test record I_Q, the classifier then returns a ranked list of species sorted by their score $S_Q(C_j)$.

4 Semantic pruning and weighting

As we are in the context of weakly annotated audio records with multiple classes (primary and secondary species) and highly cluttered contexts, we suggest improving the raw matching scheme by semantically pruning the non-discriminant training features (offline) and weighting the query features according to the semantic coherence of their k nearest neighbours (online).

Offline training features pruning. To select the most discriminant features of the training set, we first compute the full k-nearest neighbors graph of the training set by searching each individual MFCC feature $\mathbf{x}_i$ of each training audio record with the hash-based k-NN search technique described in section 3. We then affect a discrimination score $f(\mathbf{x}_i)$ to each feature computed as the percentage of k-nearest neighbors having the same weak label than the feature itself. We finally removed from the index all the features having a discrimination score equal to zero. Note that we did experiment several other strategies such as keeping all features in the index and using the discrimination scores $f(\mathbf{x}_i)$ as a weights within the vote of the search phase. This did not improve the results over the brute-force pruning of non-discriminant features we finally used (we experimented different regularization function and different values of k). This even degraded the performances over the raw instance-based classification scheme due to over-fitting. Too much weights are actually concentrated on very few features so that the contribution of the vast majority of useful features with moderate scores remains negligeable in the vote. Our pruning strategy allows keeping the contribution of these features by removing only the very bad ones with a null discrimination score. It is important to note that the fraction of such bad features is still large (about $XX\%$ of the $30M$ MFCC features). Removing them consistently reduce the noise of the dataset as well as the size of the index and the search complexity.

Besides, we also explored alternative discrimination scoring strategy considering the reverse k-nn's rather the direct k-nn's for the computation of $f(\mathbf{x}_i)$ (in a more Bayesian way). This lead to slightly worse results. Using the product of both the direct and the reverse discrimination scores lead to similar conclusions.

Online query features weighting. Similarly to the audio records of the training set, test records might also contain a lot of unuseful features corresponding to background noise or intervals of time where two much audible species overlap. We therefore also compute a discrimination score (referred as f_Q) for each MFCC feature $\mathbf{x}_j^Q$ belonging to a query record I_Q. A weak label l_j^Q is first estimated for each $\mathbf{x}_j^Q$ as the most represented label within the k-nearest neighbors of $\mathbf{x}_j^Q$ (still computed by the hash-based k-nn search method described in section 3). Similarly to $f()$, $f_Q()$ is then computed as the percentage of k-nearest neighbors having the same weak label than the feature itself (i.e. the percentage of k-nearest neighbors whose label is equal to l_j^Q).

5 Metadata

To further improve the identification performances, we explored the use of metadata as a way to re-rank the returned species. The basic principle is to reduce the score $S_Q(C_j)$ of the species whose metadata values distribution in the training set does not fit the value of the metadata of the query record. More particularly, we focused on three types of metadata that did bring some improvements in our cross-validation tests: geo-location, altitude and time-of-day. Revised scores $S'_Q(C_j)$ are computed from the original ones according to:

$$S'_Q(C_j) = S_Q(C_j) \times e^{-\frac{\delta^2_{alt}}{2\sigma^2_{alt}}} \times e^{-\frac{\delta^2_{geo}}{2\sigma^2_{geo}}} \times e^{-\frac{\delta^2_{hour}}{2\sigma^2_{hour}}} \quad (1)$$

where the δ_y's are equal to the difference between the metadata value y_q of the query (e.g. the altitude of the query observation) and the closest metadata value of all training records belonging to class C_j (e.g. the closest observation in terms of altitude). The radial basis function allows preserving the score of the species having at least one observation close to the query. On the other side, the score of the species that do not have any hit close to the query tends to zero when the distance to the closest observation grows. Note that we could have sum the RBF weights on all training records rather than considering only the closest one. This would have lead to a classical kernel density estimation and the weight affected to each species whould have been equivalent to its likelihood according to the considered metadata. But it is important to note that the distribution of the observations in the training set is biased in many ways. The density of the observations is typically more correlated to the observer's displacements than to the density of plants. Our model rather rely on the proved presence of a species at a given place (or time of day) independently of the density of the observations of that species at that place (or time of day). The values of the $sigma_y$'s normalization factors were set proportionally to the standard deviation of the distance to the nearest neighbor of any observation in the dataset.

6 Experiments and results

6.1 Dataset and task

The LifeCLEF 2014 bird dataset [2] is built from the Xeno-canto collaborative database [6] involving at the time of writing more than 140k audio records covering 8700 bird species observed all around the world thanks to the active work of more than 1400 contributors. The subset of Xeno-canto data used for the 2014 edition of the task contains 14,027 audio recordings belonging to 501 bird species in Brazil area, i.e. the ones having the more recordings in Xeno-canto database. The dataset contains minimally 15 recordings per species (maximum 91) and minimally 10 different recordists, maximally 42, per species.Audio records are

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

associated to various metadata such as the type of sound (call, song, alarm, flight, etc.), the date and localization of the observations (from which rich statistics on species distribution can be derived), some textual comments of the authors, multilingual common names and collaborative quality ratings (more details can be found in [2]). The task will be evaluated as a bird species retrieval task. A part of the collection will be delivered as a training set available a couple of months before the remaining data is delivered. The goal will be to retrieve the singing species among the top-k returned for each of the undetermined observation of the test set. Participants will be allowed to use any of the provided metadata complementary to the audio content.

Table 1. Cross-validation retrieval tests

Method	**mAP (records)**
Raw scheme	0.003
Raw scheme + audio pre-proc.	0.0414
Raw scheme + audio pre-proc. + sliding vote	0.051
Raw scheme + audio pre-proc. + sliding vote + semantic p. & w.	0.0760

6.2 Preliminary cross-validation experiments

To tune the parameters of our instance-based retrieval scheme, we randomly split the training set in 3/4 for training and 1/4 for testing. We then measured the Mean Average Precision of our method at the audio records level, i.e. before applying the final top-K classifier at the species level. Some of the intermediate results are displayed in Table 1 to illustrate the contribution of the different steps of our method.

6.3 Submitted runs and official results

We submitted two runs to the official competition:

INRIA Zenith Run 1: Our instance-based classification scheme with semantic pruning & weighting but WITHOUT the metadata oriented reranking

INRIA Zenith Run 2: Our full instance-based classification scheme with both the semantic pruning & weighting AND the metadata oriented reranking

The official results are reported in Figure 1 and Table 2. We globally ranked as the third team over the ten participants and as the second one if we consider only the purely audio-based runs. Our best run achieved a mAP of $0,365$ when considering only the primary species of each test recording (vs. 0.317 when considering the background species as well). Compared to the mAP of our second

Table 2. Official results of LifeCLEF 2014 Bird Task

Run name	Type	mAP 1	mAP 2
		with bckg species	without bkg species
MNB TSA Run 3	AUDIO & META	0,453	0,511
MNB TSA Run 1	AUDIO & META	0,451	0,509
MNB TSA Run 4	AUDIO&META	0,449	0,504
MNB TSA Run 2	AUDIO & META	0,437	0,492
QMUL Run 3	AUDIO	0,355	0,429
QMUL Run 4	AUDIO	0,345	0,414
QMUL Run 2	AUDIO	0,325	0,389
QMUL Run 1	AUDIO	0,08	0,369
INRIA Zenith Run 2	**AUDIO & META**	**0,317**	**0,365**
INRIA Zenith Run 1	**AUDIO**	**0,281**	**0,328**
HLT Run 3	AUDIO & META	0,289	0,272
HLT Run 2	AUDIO & META	0,284	0,267
HLT Run 1	AUDIO & META	0,166	0,159
BirdSPec Run 2	AUDIO	0,119	0,144
Utrecht Univ. Run 1	AUDIO	0,123	0,14
Golem Run 1	AUDIO	0,105	0,129
Golem Run 2	AUDIO	0,104	0,128
BirdSPec Run 1	AUDIO	0,08	0,092
BirdSPec Run 4	AUDIO	0,074	0,089
Golem Run 3	AUDIO	0,074	0,089
BirdSPec Run 3	AUDIO	0,062	0,075
Yellow Jackets Run 1	AUDIO	0,003	0,003
Randall Run 1	AUDIO	0,002	0,002
SCS Run 1	AUDIO	0	0
SCS Run 2	AUDIO	0	0
SCS Run 3	AUDIO	0	0

run equals to 0.328 (0.281 with the background species), it shows that the contribution of our reranking scheme making use of the metatata is about 3.6 points of mAP.

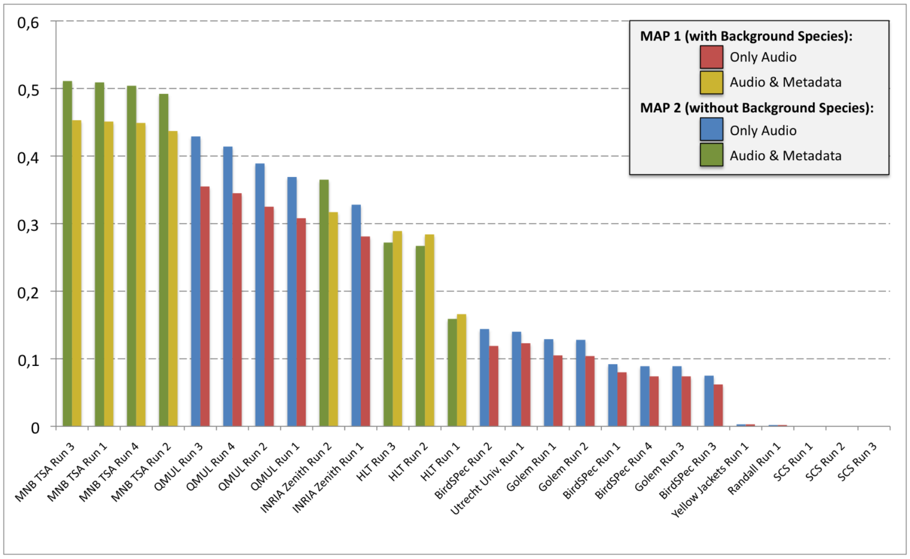

Fig. 1. Official results of LifeCLEF 2014 Bird Task - Our runs are referred as **INRIA Zenith Run 1** and **INRIA Zenith Run 2**

References

1. Andoni, A., Indyk, P.: Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In: Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on. pp. 459–468. Ieee (2006)
2. Goëau, H., Glotin, H., Vellinga, W.P., Rauber, A.: Lifeclef bird identification task 2014
3. Joly, A., Buisson, O.: A posteriori multi-probe locality sensitive hashing. In: El-Saddik, A., Vuong, S., Griwodz, C., Bimbo, A.D., Candan, K.S., Jaimes, A. (eds.) ACM International Conference on Multimedia (MM'08). pp. 209–218. ACM (10 2008)
4. Joly, A., Buisson, O.: Random maximum margin hashing. In: CVPR. IEEE, Colorado springs, United States (Jun 2011)
5. Joly, A., Goëau, H., Bonnet, P., Bakić, V., Barbe, J., Selmi, S., Yahiaoui, I., Carré, J., Mouysset, E., Molino, J.F., et al.: Interactive plant identification based on social image data. Ecological Informatics (2013)
6. 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

7. Jégou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. Pattern Analysis and Machine Intelligence, IEEE Transactions on 33(1), 117–128 (2011)
8. Krapac, J., Perronnin, F., Furon, T., Jégou, H.: Instance classification with prototype selection. In: ICMR - ACM International Conference on Multimedia Retrieval. Glasgow, Royaume-Uni (Feb 2014), http://hal.inria.fr/hal-00942275
9. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2008)