VISUAL AND ACOUSTIC IDENTIFICATION OF BIRD SPECIES

A. Marini, A. J. Turatti, A. S. Britto Jr., A. L. Koerich

Pontifical Catholic University of Paraná
Postgraduate Program in Informatics
Curitiba, PR, Brazil

ABSTRACT

This paper presents a novel approach for bird species identification that relies on both visual features extracted from unconstrained bird images and acoustic features extracted from bird vocalizations. The Scale Invariant Feature Transform (SIFT) detects local features in bird images, which are then used to train a support vector machine classifier. The instances that are not classified with a certain degree of certainty are then rejected and reclassified using Mel-frequency cepstral coefficients (MFCCs) extracted from the bird songs if available. Experiments conducted on a dataset of 50 bird species that comprise images from the CUB200-2011 and audio samples from Xeno-Canto have shown that improvements between 1.2 and 15.7 percentage points are achieved when using an acoustic classifier to re-process the instances rejected by the visual classifier, depending on the rejection level.

Index Terms— fusion of information, fine-grained classification, combination of classifiers, SIFT, MFCC

1. INTRODUCTION

Bird species identification arouses interest in different groups of admirers and experts whether by the beauty of birds and their song or by their ecological importance. Bird identification is a well-known problem to ornithologists and is considered as a scientific task since antiquity. Ornithologists study every aspect of birds life such as birds live in their environment, parts of birds, the songs that they produce, their distribution and ecological impact. There are some practical reasons to observe, study or monitor birds. Scientists often use birds to study and understand ecosystems due to many reasons: they are numerous, sensitive to environmental changes, easier to control than other species, they are everywhere and are relatively easy to be seen.

Several real-world applications can rely on birds such as monitoring of environmental pollution [1], assessing the quality of the environment [2] and estimating sustainability indicators. Therefore, the use of automated methods for bird identification is an effective way to assess the quantity and diversity of birds that appear in a region and may be useful in several practical applications. However, bird species identification is a challenging problem both for humans and for computational algorithms that aim to accomplish this task automatically.

Current approaches for bird species identification are either based on acoustic or visual information. Several approaches based on bioacoustics signals have been proposed [2, 3, 4, 5, 6, 7]. Such approaches have reached very interesting correct classification rates, between 78% and 95%, depending on the number of bird species taken into account. For instance, Lopes et al. [3] show that correct classification rate drops from 95.1% to 78.2% when the number of bird species increases from 3 to 20. Bird species identification based on their songs is challenging because there is also a high confusion between classes, background noise and overlapping between several bird songs and a high diversity in the acquisition conditions (devices, recordist uses, context diversity, etc.) [8]. On the other hand, the approaches based on image analysis [9, 10, 11, 12, 13, 14] have reached relatively low classification rates, between 2% and 30% in the Caltech-UCSD Birds 200 dataset (CUB-200) which contains over 6,000 images of 200 different birds species typically from the North America [9]. Bird species identification based on images is also challenging due to the variation of the background and illumination because most of the bird images are collected in their natural habitat. In these images one cannot control rotation, scale and viewing angle at the time of image acquisition. Using audio records rather than bird pictures is justified by current practices [8]. Birds are actually not easy to photograph; audio calls and songs have proven to be easier to collect and sufficiently species specific [8]. However, the visual properties such as color, shape, size, parts, among others, are important for the bird species recognition and can be very useful as the number of species taken into account increases to hundreds or thousands species.

This paper proposes a new approach for bird species identification that employs both visual and acoustic features. Given the complexity of the problem, a scenario which is visually and acoustically unconstrained, a high number of classes, a high visual and acoustic similarity between some bird species, background noise and a high diversity in the

Thanks to CNPq (grant 311.832/2013-0) and CAPES for funding.

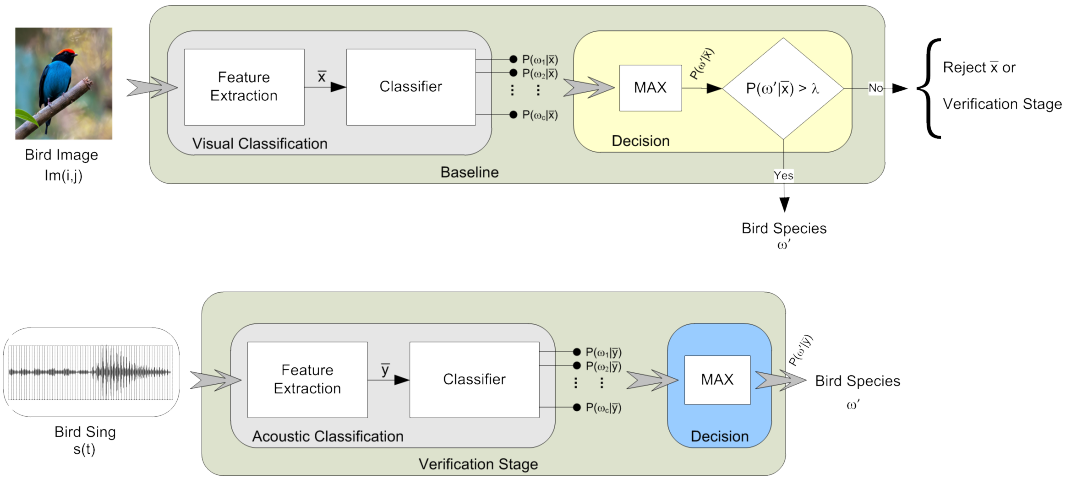

Fig. 1. Visual classification with rejection and acoustic verification.

acquisition conditions, there is a need of novel methods to deal with each step of the problem and to provide results that are more reliable than those achieved currently by both visual and acoustic approaches alone. Why is the fusion of visual and audio data important to deal with this problem? If we take into account a more realistic scenario where the number of bird species surpasses 9,000, the scalability of current approaches is limited [3, 14] and the fusion of information opens up an alternative for scalability.

This paper is organized as follows. Section 2 describes the proposed method for bird species identification. The experimental results on a dataset of 50 bird species that comprises images selected from the CUB200-2011 dataset and audio samples selected from Xeno-Canto[1] are presented in Section 3. Finally, conclusions and suggestions for future work are presented in the last section.

2. PROPOSED APPROACH

The problem of bird species identification can be defined as: given a bird image and/or bird song, assign a species among a fixed and large number of possibilities. However, the main question addressed in this paper is on how to combine both visual and acoustic information? Considering the situation where there is not enough audio data available, which means that for a given instance we may have only a visual representation (bird image) or both visual and acoustic representation (bird vocalization). In particular, only 1/4 of the instances have one-to-one correspondence between image and audio. Therefore both image and audio can be combined only at a post-processing level.

The strategy that we propose in this paper is to employ rejection at the output of the visual classifier. The concept of rejection admits the potential refusal of a bird species hypothesis if the classifier is not certain enough about the bird species hypothesis. In our case, the probabilities assigned to bird species should be used as a guide to establish a rejection criterion.

Fig. 1 shows an overview of the proposed approach for bird species identification. Given an instance, first, the visual features are extracted from the bird image and the resulting feature vector is classified by the multi-class SVM. The SVM assigns a probability of such a feature vector to belong to each one of the C classes. The MAX operator then chooses the class which provides the highest probability. However, depending on the value of such a probability as well as the availability of audio data, the instance can be either rejected or sent to a verification stage which employs acoustic features. At the verification stage, the acoustic feature vector is classified by the multi-class SVM and the MAX operator chooses the class which provides the highest probability.

The visual features are based on SIFT [15, 16] which detects and describes local features in bird images. SIFT transforms an image into a large collection of features which are invariant to image translation, scaling and rotation, robust to local geometric distortion and partially invariant to illumination changes. A Gaussian pyramid is constructed from the input image by repeated smoothing and subsampling, and a difference-of-Gaussians pyramid is computed from the differences between the adjacent levels in the Gaussian pyramid. Then, interest points are obtained from the points at which the difference-of-Gaussians values assume extrema with respect to both the spatial coordinates in the image domain and the scale level in the pyramid. Next, points with low contrast and points along edges are discarded and dominant orientations are assigned to the remaining points. The large amount of visual features are reduced to a small vocabulary of visual words using the Elkan algorithm [17].

The audio samples were manually segmented (hand-trimmed bird vocalizations) to retain only the excerpts where

[1] www.xeno-canto.org

bird sing was present. The segmentation process consists in eliminating the silence intervals and concatenating the sing intervals as shown in Fig. 2. Hence audio data is converted to a spectrogram-like representation, i.e. the magnitudes of short-time Fourier transformed (STFT) frames of audio, around 10 ms duration per frame. The STFT spectrum has the frequency axis transformed to the Mel scale. A convention, originating from speech processing, is to transform the Mel spectrum using a cepstral analysis and then to keep the lower coefficients which typically contain most of the energy. These coefficients became widespread in applications of machine learning to audio, including bird vocalizations. MFCCs have some advantages, including that the feature values are approximately decorrelated from each other, and they give a substantially dimension-reduced summary of spectral data. We treat the full-length audio as a single unit for training/testing purposes. Therefore the MFCCs are summarized over time using the mean and standard deviation which generates a 52-dimensional feature vector.

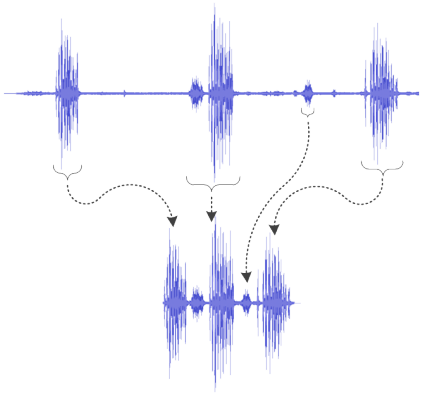

Fig. 2. Sample of an acoustic signal from a bird sing: the original sample and pre-processed sample where the "silence" intervals were removed.

The visual and acoustic features are used to train a multi-class SVM with linear kernel and a multi-class SVM with a Gaussian kernel respectively. The cost and gamma parameters were found by a 5-fold cross validation on the training dataset. Pairwise coupling is used to handle multi-class classification.

The fusion of the visual and acoustic information is carried out at classification level and it depends on whether an instance is accepted or rejected. The task of the rejection mechanism is to, based on the output vector $[P(\omega_1|\hat{x}), \ldots, P(\omega_c|\hat{x})]$ provided at the output of the visual classifier, which is ordered in a decreasing order according to the probability, decide whether the best bird species hypothesis, which is so far called the TOP 1, can be accepted or not. Now, the highest *a posteriori* probability provided by the MAX operator is not simply accepted, but it is compared with a rejection threshold (λ). If such a probability is greater than λ then the bird species is assigned to the instance, otherwise, no label is assigned to the instance $\hat{x}$ and it is rejected. In summary, the rejection rule is given by: (i) the TOP 1 bird species hypothesis is accepted whenever $P(\omega'|\hat{x}) \geq \lambda$; (ii) the TOP 1 bird species hypothesis is rejected whenever $P(\omega'|\hat{x}) < \lambda$.

Two fusion schemes are proposed for those rejected instances: (i) samples rejected at the visual classification are re-classified using the acoustic data if it is available. The decision is taken based solely on the verification stage; (ii) samples rejected at the visual classification are re-classified using the acoustic data if it is available. Further the output of both the visual and acoustic classifiers are combined through conventional rules such as SUM, $PROD$ and MAX.

3. EXPERIMENTAL RESULTS

The performance of the proposed approach was evaluated on a subset of 50 bird species out of 200 available in CUB200-2011 dataset [18]. The choice of particular bird species to make up this subset was driven by the availability of bird sing audio. However the complexity of fine-grained classification problem remains in this subset. The training set is made up of 1,499 images and 448 audio samples with a homogenous distribution among the 50 classes with an average of 30 images and 9 audio samples per class respectively. We have followed the same proportions in the CUB200-2011 dataset, therefore the testing set has 1,480 images and 422 audio samples with an average of 30 images and 8 audio samples per class respectively. Therefore, only 28.5% of the instances have both visual and acoustic information.

The setup for visual feature extraction was 1,200 visual words and two zoning schemes (2×2 and 4×4). From each zone is computed a histogram which is quantized in 1,200 visual words. Therefore, we have a 24,000-dimensional feature vector to represent each bird image. The setup for acoustic feature extraction is for every 512 samples, a window of 512 samples is produced, the window is multiplied with the hamming function, power spectrum is taken, MFCC is computed and the MFCC coefficient is obtained. Afterwards, the mean of 20 past MFCC with overlap of 19 is computed, then mean of 5 of the previous means with 0 overlap is computed. The same procedure is used to compute all the 13 MFCCs.

Tab. 1 shows the performance of the individual visual and acoustic classifiers on the test subsets taken into account if the correct bird species is among the TOP N best hypotheses. The correct classification rate which is defined as the ratio between number of samples correctly classified and the number of samples tested. For instance, the correct bird species is among the TOP 6 best hypotheses for more than 57% of the cases. Tab. 1 also shows that the acoustic features are more discriminate than the visual ones since the correct classification rate is about 20 percentage points higher than that achieved with the visual features. However a direct comparison is not fair due to the different number of samples used to

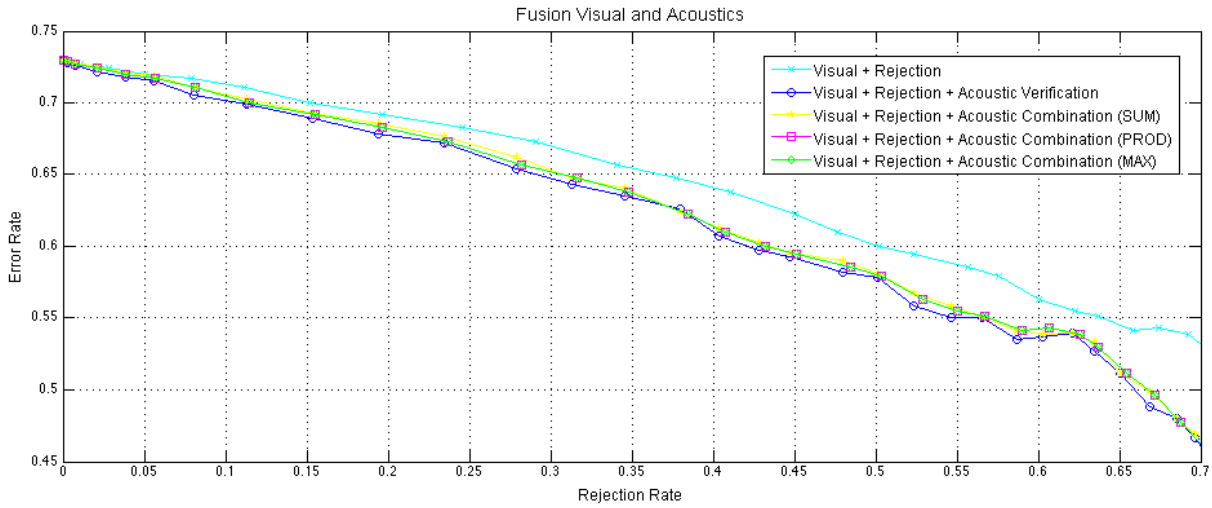

Fig. 3. Error rate versus rejection rate.

N best hypotheses	Correct Classification Rate (%) Visual	Acoustic
TOP 1	27.03	45.97
TOP 2	36.76	57.58
TOP 3	43.78	64.69
TOP 4	48.92	72.04
TOP 5	54.26	75.83
TOP 6	57.77	79.62
TOP 7	60.88	81.75
TOP 8	64.05	84.36
TOP 9	66.76	86.49
TOP 10	68.72	86.97

Table 1. Correct classification rates for the visual and acoustic classifiers on 1,480 images and 422 audio samples of the testing set at 0% rejection level.

train and test the visual and acoustic classifiers.

Tab. 2 shows the performance of the different combination strategies proposed to fuse the outputs provided by both the visual and acoustic classifiers under different rejection rates. Different rejection levels are achieved by varying the rejection threshold between $[0, 1]$ and comparing it with the highest probability assigned by the SVM classifier to the instance. Recalling that the aim of the acoustic features is to aid the visual classification of bird species, the best results were achieved by simply re-classifying the samples rejected by the visual classifier using the acoustic features. The improvements are between 1.2 and 2.9 percentage points relative to the visual classifier alone depending on the rejection level and between 3.07 and 15.17 percentage points relative to the visual classifier without rejection. Fig. 3 compares different variations of the proposed approach with the baseline that employs only visual information to classify bird species. In general, the improvement brought about the acoustic features tends to increase with the rejection rate.

Strategy	Rejection Rate 10%	30%	50%
Visual	28.89	32.70	40.02
Visual and Acous.	**30.10**	**35.65**	**42.20**
Visual and Acous. (SUM)	29.71	35.22	41.90
Visual and Acous. ($PROD$)	29.96	35.25	42.04
Visual and Acous. (MAX)	29.96	35.25	42.04

Table 2. Correct classification rates for the fusion between visual and acoustic classifiers on the testing set at 10%, 30% and 50% rejection level.

4. CONCLUSIONS

This paper shows that the acoustic features are relevant to improve the identification of bird species based on bird image. The proposed approach has show to be useful in situations where partial acoustic information is available. The visual classifier alone achieved 27.03% correct classification rate. Under the condition of a perfect rejection rule, that rejects only the wrongly classified images, which are further classified through the audio, 37.33% of correct classification rate is achieved. Considering a realistic condition where a rejection rate should be established, rejecting 30% of the samples, the combination of visual and acoustic features achieve 35% of correct classification rate. Such a result proves the relevancy of the acoustic information in the image classification task.

A comparison of the proposed approach with other related works is difficult because this is the first approach that proposes the combination of visual and acoustic information for bird species classification. In spite of the good results achieved, the proposed approach could be improved in several ways such optimizing both the visual and acoustic feature extraction, or even by evaluating other strategies to combine these complementary information. This will be the subject of our future work.

5. REFERENCES

[1] R. A. Lovet, "How birds are used to monitor pollution," *Nature News*, Nov. 2012.

[2] R. Bardeli, D. Wolff, F. Kurth, M. Koch, K-H. Tauchert, and K-H. Frommolt, "Detecting bird songs in a complex acoustic environment and application to bioacoustic monitoring," *Patt Recog Letters*, vol. 31, pp. pp.1524–1534, 2010.

[3] M.T. Lopes, L.L. Gioppo, T.T. Higushi, C.A.A. Kaestner, C.N. Silla, and A.L. Koerich, "Automatic bird species identification for large number of species," in *IEEE Int'l Symp Multimedia*, 2011, pp. 117–122.

[4] M.T. Lopes, C.N. Silla Jr., A.L. Koerich, and C.A.A. Kaestner, "Feature set comparison for automatic bird species identification," in *IEEE Int'l Conf Sys Man Cybernetics*, 2011, pp. 965–970.

[5] F. Briggs, B. Lakshminarayanan, L. Neil, X. Z. Fern, and R. Raich, "Acoustic classification of multiple simultaneous bird species: a multi-instance multi-label approach," *J. Acoust. Soc. Am.*, vol. 131, no. 6, pp. 4640–4650, 2012.

[6] E. Stattner, W. Segretier, M. Collard, P. Hunel, and N. Vidot, "Song-based classification techniques for endangered bird conservation," in *Workshop Machine Learning for Bioacoustics*, Jun. 2013, pp. 67 –72.

[7] D. Stowell and M. D. Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning," Tech. Rep., PeerJ, 2014.

[8] A. Joly, H. Muller, H. Goeau, H. Glotin, C. Spampinato, A. Rauber, P. Bonnet, W.-P. Vellinga, B. Fisher, and R. Planque, "Lifeclef: Multimedia life species identification," in *Environm. Multim. Retr. Workshop*, Apr. 2014, pp. 7 –13.

[9] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, "Caltech-UCSD Birds 200," Tech. Rep. CNS-TR-2010-001, California Institute of Technology, 2010.

[10] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie, "Visual recognition with humans in the loop," Tech. Rep., Univ. California, San Diego - California Inst. Techn., 2010.

[11] C. Rother, V. Kolmogorov, and A. Blake, ""grabcut": interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, Aug. 2004.

[12] Y. Chai, V. Lempitsky, and A. Zisserman, "Bicos: A bi-level co-segmentation method for image classification," in *IEEE Int'l Conf. Comp. Vision*, nov. 2011, pp. 2579 –2586.

[13] Y. Chai, E. Rahtu, V. Lempitsky, L. Gool, and A. Zisserman, "Tricos: A tri-level class-discriminative co-segmentation method for image classification," in *IEEE Int'l Conf Comp Vision*, vol. 7572 of *LNCS*, pp. 794–807. Springer, 2012.

[14] A. Marini, J. Facon, and A. L. Koerich, "Bird species classification based on color features," in *IEEE Int'l Conf Sys Man Cybernetics*, Oct 2013, pp. 4336–4341.

[15] D. G. Lowe, "Object recognition from local scale-invariant features," in *IEEE Int'l Conf Comp Vision*, 1999, vol. 2, pp. 1150 –1157.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[17] C. Elkan, "Using the triangle inequlity to accelerate k-means," in *Int'l Conf on Man Learning*, 2003, pp. 147–153.

[18] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," Tech. Rep. CNS-TR-2011-001, California Inst Tech, 2011.