Multimodal Identification of Birds from Visual and Acoustic Data

Arnav Bhavsar, Dileep A. D. , Padmanabhan Rajan*
Multimedia Analytics and Systems Group
School of Computing and Electrical Engineering
Indian Institute of Technology Mandi, India
{arnav, dileepad, padman}@iitmandi.ac.in

1 Introduction

The realization about the importance of biodiversity preservation has provided impeteus to the scientific study of ecological systems, including those for monitoring the diversity, populations, distributions etc. []. The development of such scientific studies occurred well before the common availability of advanced technological aids. Thus, traditionally, the observation and identification process had been manual requiring much time and effort [].

With the advent of the digital acquisition devices for capturing sights and sounds, the observation aspect has become easier, faster and more reliable. However, the identification component is still largely manual, with experts such as zoologists, ornithologists, botanists etc. manually labeling the acquired data.

On the other hand, automated identification, labeling, annotating etc. has become commonplace in various human technological endeavors such as biometrics, surveillance, entertainment, navigation, sports etc. Clearly, given the nature of technical progress in more human-centric domains, it is natural to question the relative lack of automated identification aids in study of ecological systems.

Perhaps, an important concern which makes such research problems technically challenging, is the very large amount of variations in the data. For instance, there are variations in different entities of the same species, and also spatial and temporal variations of the observed entities. This makes it difficult to model or represent each of the relevant classes in a specific problem, or to develop good classification approaches which can handle possible overlapping representations between classes.

Having described the challenges for pattern recognition research in the ecological domains, it must be acknowledged that the situation is not so grim. The evolution of machine learning, vision and speech-acoustics research has yielded some powerful mathematical tools have such as graphical models, sparse learning, and various advanced feature representation and classification frameworks. As hinted above, these have been applied successfully in various areas such as person identification, scene understanding, speech processing and recognition etc., which also involve some of the above discussed concerns about variability and resulting ambiguities.

Indeed, the applications and adaptations of such frameworks have recently been reported for automatic

*Author names appear in alphabetic order.

identification tasks in some domains of ecological interests. Some example domains that employ audio-visual data include identification and classification of birds, butterflies, plants etc., analysis of animal and insect behaviour [10, 11, 5, 1, 21].

However, in general, such research is still in its nascent stages, and has much scope for improvements in accuracy and variety of scenarios. At the same time it also provides much scope of intellectually challenging technical progress in terms of developing new tools and frameworks, or adapting existing ones, so as to address the new challenges. Another important perspective is that such research can lead to possibilities of some system development, which we will elaborate on subsequently. Some basic forms of such systems (e.g. website or applications on mobile devices for data collection and analysis) have been setup in recent years (e.g. [9]). However, from a regional point of view, such systems largely focus on the data from the western world, and there is enough scope to develop similar or better systems for the Indian scenario.

Having provided a general perspective, we acknowledge that each ecological domain and the problems therein requires quite different approaches. In this proposal, we focus on the problem of the detecting and classification of birds. Our interest in considering this problem stems from the following reasons:

- Our primary motivation to consider this problem is the challenge that it poses in terms of the variety and necessarily unstructured nature of the data. While, as indicated above, interest in this area has been growing, the performance on existing approaches is very modest due to such practical challenges.
- The varied nature of real data, in turn, brings about various technical challenges such as exploring ways to extract/learn the discriminative features, considering and modeling the local inter-relationships between visual bird-body parts, developing frameworks which yields high similarity between similar species, and making the failure cases to be graceful (i.e. more among similar species and less as the specific difference increases), xxxx, etc.
- We are also motivated to consider this problem from the perspective of using both visual and acoustic modalities, which to our knowledge, has not been reported. We believe that sight and sound, in the task of bird detection and identification will compliment each other well and can be an important towards improving the performance.
- Clearly, central to the problem is the availability of data in terms of images/videos and sounds. Some standard public datasets for this purpose are available for the purpose of benchmarking (e.g. [8]). However, these are rather limited and region-specific to the western world (e.g. North America). Also, there is further need for more realism in terms of scale, pose, illumination variations in terms of visual data and xxxx in terms of acoustic data. Thus, over the course of this this project, we also plan to collect data for Indian birds, both in terms of sight and sounds, and make it publicly available.

1.1 Objectives

Based on the above discussions, we formulate below, some specific objectives. We provide more details on these in section 3.

1. Research and algorithm development for detection and identification of birds from images and videos
2. Research and algorithm development for detection and identification of birds from acoustic data

3. Integration of 1. and 2. above, into a common approach which can process audio-visual data.

4. Constructing datasets of local bird species and making these available to the research community.

5. Eventually, establishing a web-based system which can use the automated approaches, with some manual intervention, for further data collection and management, resulting in continuous improvements.

2 Related work

As mentioned above, the last few years has seen some modest but conscious effort towards addressing the problem of automatic classification of birds. We briefly discuss below, some related work, in the visual and acoustic domains.

2.1 Fine-grained visual classification for bird identification

From a visual point of view, the automated identification of bird species is one of the important problems considered in an upcoming sub-area of fine-grained visual classification (FGVC) in computer vision. Here, the term 'fine-grained' signifies that such classification problems are different (and, arguably, harder) than the traditional visual classification problems. The latter consists of classification at a coarser level of different object entities (e.g. type of vehicle, objects, animals etc.), whereas FGVC is concerned about classification within different species of a particular entity (e.g. car models, bird species etc.).

One of the fundamental aspects in automated classification problems is the *representation* of the entities different classes, in some objective terms (or features) such as shape, texture etc. Clearly, in case of FGVC, global representations of an entity such as the overall shape, morphology, geometry, texture etc. will not be useful for the classification task, since entities belonging to different classes would yield similar global representations.

Hence, a more logical philosophy followed in exisiting approaches for FGVC is that of considering the representation at a local level. These are typically termed as part-based approaches [3, 5, 1, 7, 6], wherein representative features are extracted from local parts of birds, and these features are then used for further modelling or classification.

For instance, in the work of [7], a variety of rectangular patches of various widths and heights are sampled, and some features (e.g. SIFT, HOG [7, 1]) are extracted from these patches. Then, the approach involves capturing the most discriminative patches, or pairs of patches, using a random forest classifier. Another work [6] also extracts features from a variety of rectangular patches (templates), but then follows a different direction, of computing response maps by matching each training image with the templates from all the images. Some local regions from these response maps are then pooled to form histograms, which are then used for classification via an support vector machine (SVM). The rationale behind this is that templates from a bird species is more likely to match similar bird species than the other. In these methods, there is no consideration of the structure or the relationship of the parts of the object, and there is also a possibility of including some background regions in the rectangular patches, while computing the features.

In another work [5], a semi-automatic approach is followed to consider the structure of the object (e.g. bird). First, an method which uses a small manual interaction is used to segment the bird from the back-

ground. This foreground is then segmented into smaller regions, each one of which appximately covers semantic parts of the bird (e.g. beak, nape etc.), based on some offline landmark information provided for each part. Now, feature descriptors are computed from these segmented regions, and a codebook is learned by pooling the descriptors, which is then used for classification using an SVM.

Realizing the importance of learning the most discriminative local features (which to some extent was attempted in [5] on unstructured patches), the authors in [1], propose to do the same at the level of semantic features. Here, too the features (e.g. histogram of gradients etc.) are extracted from patches, but the patches are structured around the points denoted by landmarks. The discriminative regions are learned for every two classes, based on the features in those regions, using an SVM, where the SVM also assign weights to each of the region. Also, important in this work is the notion of alignment of the regions across images (by scaling and rotating), based on the landmark points. This is important so that the regions from which the features are extracted coincide approximately in each image.

The importance of image alignment for FGVC of birds is also stressed in [4]. In this work, the alignment is based on a global feature, viz. an ellipse fitting on the bird shape. The alignment is then used to transfer the local features and annotated landmarks of all images onto a common reference frame.

In another interesting work [2], the primary goal is not the classification, but to translate the objective and mathematically defined features used in FGVC to semantic discriminative body parts, so as to form a visual guidebook to know what discriminative parts to look for. The objective features used in this work are those proposed in [1].

2.2 Bio-acoustic signal analysis for bird identification

Birdsong is one of nature's most well known sounds. Acoustic communication in birds is one of the primary ways by which make their presence known to one another. Recently, automatic monitoring of bird populations by means of pattern classification algorithms have received much attention. Most birds have distinctive calls, which can be used to identify a particular species. Automatic birdsong analysis provides several problems for the development of novel machine learning and signal processing algorithms.

Sounds produced by birds can be complex and varied. A few broad categories of sounds have been identified, which are common to most bird sounds [21]. But there are complexities in the form of variations and combinations. There can be simple repetitions of the sounds, or random sequences with no repetition. Additionally, in field recordings, there are other sounds, including calls from other birds, as well as other natural and man-made sounds. All these factors make automatic birdsong classification a non-trivial task. In addition to identifing different species, automatated systems have shown some success in identifying individuals *within* species [22].

As in any pattern classification task, there are two stages in building an automated birdsong classifier: feature extraction, followed by classification. Features used in birdsong classification have ranged from simple features which directly measure properties of the audio signal, to features used in automatic human speech recognition and their variants [26, 27, 28, 29]. Similarly, the classification stage has used techniques like Eucledian distance, Bayesian classifiers [30], hidden Markov models [31, 32], information-theoretic measures [33], Gaussian mixture models [34], neural networks[35], decision trees [36], and support vector machines [29].

3 Proposed methodology and project plan

3.1 Visual classification

Based on the discussion of the related work, we notice some aspects which are clearly important with respect to identification of birds. From a visual identification point of view, these can be summarized as: a) Important of considering local parts, b) formulating some objective feature definitions on these parts, c) learning most discriminative local features, d) Considering the structure and the relationship between the local parts, e) Importance of alignment, f) considering the availability of landmarks or manual annotations.

Some of the these aspects (e.g. a, b) are considered in all of the exisiting works, as discussed above, while other aspects (e.g. c-f) are considered in some but not all. In any case, given the adolesence of the area, there is still much scope of further explorations on these aspects. Here, we propose and briefly discuss some directions which we would like to explore:

- Modelling relationships between all the parts via graphical models: While individual parts or pair-wise relationships between parts are considered in existing state-of-the-art, higher level relationships between parts are not. Thus, a possible direction for further exploration is modeling a higher level body structure using graphical models (Markov random fields or conditional random fields) [12, 14, 13, 15, 16]. Indeed, such higher order models between parts do exists in computer vision but in a different context of human pose/activity recognition [17]. We would like to explore this in our context of FGVC for birds.

- Considering cases with or without landmarks/annotations: In a real-world scenario, often we may not have annotations (or manual part marking). Thus, it would be useful to develop methods for such cases. Thus, the task of bird segmentation or automatic part identification are also involved in such a scenario. Interestingly, this issue has not been considered in any of the existing works. These may be seen as a unsupervised low-level feature detection or segmentation problem, or learning the segmentation labels or feature characteristics from a high confidence landmarked or annotated examples.

- With respect to human speech, automatic identification of speakers have demonstrated state-of-the-art performance using the so-called *subspace methods.* Here, the underlying idea is that most of the relevant information needed for classification lies in a relatively low-dimensional space. The basis vectors of this space can be learnt form training data [23]. Similar subspace techniques can be investigated for classification of bird sounds.

- Again with respect to human speech, an important auditory cue for the identification of speakers is the use of formant information. Formants are resonances of the vocal tract. Bird vocalizations also show characteristic formant information. Features that capture formant information (eg. group delay based representations [24, 25]) show great potential in automatic identification of bird calls.

- Feature selection and combination: Many kinds of feature descriptors can be used in both audio and visual data, and the effectiveness of feature descriptor varies with the problem domain. Thus, an interesting direction is that of selecting the best image features and sound features for the pool of features using the differnt techniques including multiple kernel learning and combine the selected

features from image domain and sound domain using informatin theoretic methods and multiple kernel leraning.

- Comparative analysis between low-level feature descriptor: There is no standardization or review yet for FGVC of birds. In would be interesting and informative, as a part of this project to work on such an exhaustive survey, especially also because the area is gaining popularity in the vision and speech community.

- Classifiers: There are various classifiers that can be explored for this task (e.g. [20]). Some well-known examples are Gaussian mixture models, neural networks, K-nearest neighbour classifieer, support vector machines using Gaussin kernel, kernels for images and kernels for sound/speech signals, sparse-representation based classifiers [18, 19]. Some of these classifiers can also be used to learn better discriminative features.

  Thus, based on the above discussion about various directions, we summarize the methodlogical aspects of our proposal:

  - Feature selection and feature detection/segmentation (the latter in case of non-annotated data) for image and sound data.
  - Classification of birds from images represented using the different features using different classifiers mentioned above.
  - Classify the birds from their sounds represented using the different features using different classifiers mentioned above.
  - Combine the classifier scores of the different classifiers built on both image and sound of birds.
  - Multimodal classification: Use the both image and sound data to build classifier.
  - Guaging how much role does features or classifiers play.
  - Review of features and classifiers.

4 Distribution of funds for this project

5 Discussion: Future work for a larger project grant

Further challenges:
Handling audio-visual clutter
Handling inter-class clutter for audio data
Handling finer specific variations

Research in improvements for algorithms:
Considering higher complexity by adding more classes

System development:

Implementing a real system for bird identification

A website where for maintaining a bird database and an online identification system

Considering for generalizing for other wildlife

References

[1] T. Berg, P. Belhumeur., *POOF: Part-based one-vs-one features for fine-grained categorization, face verification, and attribute estimation*, IEEE International Conference on Computer Vision and Pattern Recognition, (CVPR 2013), pp. 955–962, 2013.

[2] T. Berg, P. Belhumeur., *How do you tell a blackbird from a crow?*, IEEE International Conference on Computer Vision, (ICCV 2014), pp. 729–736, 2014.

[3] N. Zhang, R. Farrell, F. Iandola, T. Darrell., *Deformable part descriptors for fine-grained recognition and attribute prediction*, IEEE International Conference on Computer Vision, (ICCV 2013), 2014.

[4] E. Gavves, B. Fernando, C. Snoek, A. Smeulders, T. Tuytelaars., *Fine-grained categorization by alignments*, IEEE International Conference on Computer Vision, (ICCV 2013), 2013.

[5] L. Xie, Q. Tian, R. Hong, S. Yan, B. Zhang., *Hierarchical part matching for fine-grained visual categorization*, IEEE International Conference on Computer Vision, (ICCV 2013), 2013.

[6] B. Yao, G. Bradski, L. Fei-Fei., *A codebook-free and annotation-free approach for fine-grained image categorization*, IEEE International Conference on Computer Vision and Pattern Recognition, (CVPR 2012), 2012.

[7] B. Yao, A. Khosla, L. Fei-Fei., *Combining randomization and discrimination for fine-grained image categorization*, IEEE International Conference on Computer Vision and Pattern Recognition, (CVPR 2011), 2011.

[8] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, P. Perona. *Caltech-UCSD Birds 200*, Technical Report, California Institute of Technology, CNS-TR-2010-001, 2010.

[9] Columbia University and the University of Maryland. *BIRDSNAP: An Electronic Field Guide to Birds*, www.birdsnap.com, 2014.

[10] S.G. Wu, F.S. Bao, E.Y. Xu, Y. Wang, Y. Chang, Q. Xiang., *A leaf recognition algorithm for plant classification using probabilistic neural network*, International Symposium on Signal Processing and Information Technology, 2007.

[11] J. Wang, K. Markert, M. Everingham., *Learning models for object recognition from natural language descriptions*, British Machine Vision Conference, (BMVC 2009), 2009.

[12] S.Z. Li., *Markov random field modeling in computer vision*, Springer-Verlag, 1995.

[13] P. Kohli, C. Rother., *Higher-order models in Computer Vision*, In: O. Lezoray, L. Grady. (Eds.), Image Processing and Analysing Graphs: Theory and Practice, CRC Press, 2012.

[14] C. Wang., *Distributed and higher-order graphical models*, PhD Thesis, Ecole Centrale Paris, 2011.

[15] X. He, R. Zemel, M.Carreira-Perpinan., *Multiscale conditional random fields for image labeling*, IEEE International Conference on Computer Vision and Pattern Recognition, (CVPR 2004), pp. II-695–II-702, 2004.

[16] J. Verbeek and B. Triggs., *Scene segmentation with conditional random fields learned from partially labeled images*, Conference on Neural Information Processing Systems (NIPS 2007), 2007.

[17] M. Bray, P. Kohli, P. Torr., *PoseCut: Simultaneous segmentation and 3D pose estimation of humans using dynamic graph-cuts*, European Conference on Computer Vision, (ECCV 2006), 2006.

[18] J. Wright, A.Y. Yang, A. Ganesh, S. Sastry, Y. Ma., *Robust face recognition via sparse representation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, (PAMI 2009), 2009.

[19] R. Rigamonti, M.A. Brown, V. Lepetit., *Are sparse representations really relevant for image classification ?*, IEEE International Conference on Computer Vision and Pattern Recognition, (CVPR 2011), 2011.

[20] R. Duda, P. Hart, D. Stark., *Pattern Classification*, John Wiley and Sons Inc., 2000.

[21] Scott Brandes, T, Automated sound recording and analysis techniques for bird surveys and conservation, *Bird Conservation International*, S1:18, 2008

[22] Kirschel, Alexander NG and Cody, Martin L and Harlow, Zachary T and Promponas, Vasilis J and Vallejo, Edgar E and Taylor, Charles E, Territorial dynamics of Mexican Ant-thrushes Formicarius moniliger revealed by individual recognition of their songs, *Ibis*, 2:153, 2010

[23] Dehak, N. and Kenny, P. and Dehak, R. and Dumouchel, P. and Ouellet, P., *IEEE Trans. Audio, Speech Lang. Process.*, 19:4, 2011

[24] Rajan, Padmanabhan and Kinnunen, Tomi and Hanilci, Cemal and Pohjalainen, J and Alku, P Using group delay functions from all-pole models for speaker recognition, *Proc. Interspeech*, 2013

[25] , Significance of the Modified Group Delay Feature in Speech Recognition, Hegde, R. M. and Murthy, H. A. and Gadde, V. R., *IEEE Trans. Audio, Speech, Lang. Process.*, 15:1, 2007

[26] Somervuo, Panu and Harma, Aki and Fagerlund, Seppo, Parametric representations of bird sounds for automatic species recognition, *Audio, Speech, and Language Processing, IEEE Transactions on*, 14:6, 2006

[27] Tyagi, Hemant and Hegde, Rajesh M and Murthy, Hema A and Prabhakar, Anil Automatic identification of bird calls using spectral ensemble average voice prints *Proceedings of the Thirteenth European Signal Processing Conference 2006*

[28] Graciarena, Martin and Delplanche, Michelle and Shriberg, Elizabeth and Stolcke, Andreas and Ferrer, Luciana Acoustic front-end optimization for bird species recognition *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on 2010*

[29] Tan, Lee Ngee and Kaewtip, Kantapon and Cody, Martin L and Taylor, Charles E and Alwan, Abeer Evaluation of a Sparse Representation-Based Classifier For Bird Phrase Classification Under Limited Data Conditions. *Proc. Interspeech 2012*

[30] Lopes, Marcelo Teider and Gioppo, Lucas L and Higushi, Thiago T and Kaestner, Celso AA and Silla, CN and Koerich, Alessandro L Automatic bird species identification for large number of species *Multimedia (ISM), 2011 IEEE International Symposium on 2011*

[31] Chu, Wei and Blumstein, Daniel T Noise robust bird song detection using syllable pattern-based hidden Markov models *Proc. ICASSP 2011*

[32] Graciarena, Martin and Delplanche, Michelle and Shriberg, Elizabeth and Stolcke, Andreas Bird species recognition combining acoustic and sequence modeling *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*

[33] Wang, Ni-Chun and Hudson, Ralph E and Tan, Lee Ngee and Taylor, Charles E and Alwan, Abeer and Yao, Kung Bird phrase segmentation by entropy-driven change point detection *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on 2013*

[34] Lee, Chang-Hsing and Hsu, Sheng-Bin and Shih, Jau-Ling and Chou, Chih-Hsun Continuous birdsong recognition using Gaussian mixture modeling of image shape features *Multimedia, IEEE Transactions on 2013*

[35] Mporas, Iosif and Ganchev, Todor and Kocsis, Otilia and Fakotakis, Nikos and Jahn, Olaf and Riede, Klaus and Schuchmann, Karl-L Automated Acoustic Classification of Bird Species from Real-Field Recordings *Tools with Artificial Intelligence (ICTAI), 2012 IEEE 24th International Conference on*

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