Collaborative Learning of Fine-grained Visual Data

Tapabrata Chakraborty

a thesis submitted for the degree of

Doctor of Philosophy

at the University of Otago, Dunedin,
New Zealand.

12 December, 2018

Abstract

Problem: Deep learning based vision systems have achieved near human accuracy in recognizing coarse object categories from visual data. But recognizing fine-grained sub-categories remains an open problem. Tasks like fine-grained species recognition poses further challenges: significant background variation compared to subtle difference between objects, high class imbalance due to scarcity of samples for endangered species, cost of domain expert annotations and labeling, etc.

Methodology: The existing approaches, like transfer learning, to solve the problem of learning small specialized datasets are still inadequate in case of fine-grained sub-categories. The hypothesis of this work is that collaborative filters should be incorporated into the present learning frameworks to better address these challenges. The intuition comes from the fact that collaborative representation based classifiers have been earlier used for face recognition problems which present similar challenges.

Outcomes: Keeping the above hypothesis in mind, the thesis achieves the following objectives: 1) It demonstrates the suitability of collaborative classifiers for fine-grained recognition 2) It expands the state-of-the-art by incorporating automated background suppression into collaborative classification formulation 3) It incorporates the collaborative cost function into supervised learning (deep convolutional network) and unsupervised learning (clustering algorithms) 4) Lastly, during the work several benchmark fine-grained image datasets have been introduced on NZ and Indian butterflies and bird species recognition.

Acknowledgements

This thesis was borne out of my doctoral research that was privileged to receive the support of many directly and indirectly.

First, thanks to supervisors who have not only provided expert guidance and sound advice throughout the research process, but have also acted as mentors to train me to become an independent researcher. For this I am truly grateful to my primary supervisor Dr. Brendan McCane, co-supervisor Dr. Steven Mills and external supervisor Dr. Umapada Pal.

My research abode at the department of Computer Science at University of Otago has always been gracious in indulging me with every resource that I have asked for my research. The technical and administrative staff have always been kind to consider my requests.

Cheers to my "lab-mates" who became close friends and confidantes, Xiping Fu and Rassoul Mesbah, who started in the same research group a few months before me. They introduced me to the PhD life and were always there with helpful tips as well as fun chat sessions, that made the journey more enjoyable.

What started as a small research proposal turned out to be a multi-national and multi-institution collaboration. I would like to thank Dr. Stella Yu and her team, specially Peter Wang and Jeff Zhang, for their collaboration and hospitality during my 3 month stint at the International Computer Science Institute.

Lastly and above all, I would like to express my heartfelt gratitude to my parents without whose support and sacrifice, none of this would have been possible.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

In this introductory chapter, the research problem is presented first: fine-grained visual categorization (FGVC). Then the main genre of methodology used in the thesis is discussed: collaborative representation classifiers (CRC). The claim of this thesis is that CRC based methods should be good candidates to solve FGVC tasks. The contributions, papers and chapter flow is also summarised in this chapter.

1.1.1 Fine-grained Visual Categorization (FGVC)

Humans are naturally adept at the task of object detection and recognition from visual scenes, but to replicate this ability in intelligent machines is one of the core problems of artificial intelligence research. Computer vision or machine vision, as the research area is commonly referred, encompasses the related fields of digital image processing, pattern recognition and machine learning. Within the scope of computer vision, object recognition research has traditionally focused on solving the problem of detection (segmentation) and recognition (classification) of sufficiently different object classes from images.

The focused research over the past thirty years, along with the exponential increase in capacity and power of computing machines, have resulted in the development of sophisticated vision systems which can robustly detect and categorize objects with sufficient visual differences (base category classification), even from natural scene images (Russakovsky *et al.* (2015)). In fact, with rapid advancement in machine learning and neural networks (particularly deep convolutional nets in the last decade), state of the art vision systems have recently achieved near human accuracy in recognizing base categories (e.g. recognition of animal images as members of broad classes like dogs, cats, horses, etc) even in large challenging datasets like ImageNet.

In the past five years, a new and challenging area of research has gained popularity in machine vision, that of recognizing sub-categorical object classes (e.g. identification of type/species of birds/fish/insects from images) with fine grained differences in attributes. Fine Grained Visual Categorization (FGVC) is currently one of the open problems of computer vision as it poses certain interesting challenges (Chai (2015)).

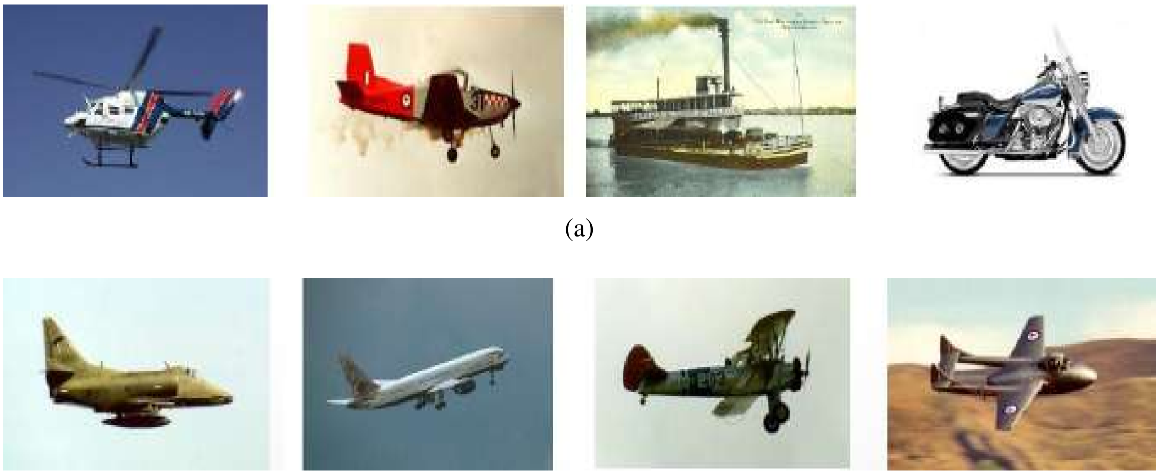
A case in point is automated species recognition, which has emerged as one of the representative problems of FGVC. In fine-grained species recognition, the variability in background and pose can be high compared to the subtle inter-class differences, thus making it a particularly challenging task (Rodner *et al.* (2015)). Furthermore, there can be considerable intra-class pose variation which may involve significant changes in object contour (for example, shape change of same bird species between flight vs. roosting images). The above statements are further illustrated in Fig 1.2. Four images each of the NZ endemic birds kaka and kea are shown. It can be readily observed that the visual differences between the classes are subtle, especially due to the strong confounding factors of background and pose variation.

These challenges may signify a possible necessity of new algorithmic approaches to tackle the problem of FGVC. There are interesting avenues of novel research, especially in cases of higher intra-class variability, brought about by changes in object appearance (strong variations in illumination, pose, deformation, etc.) and changes in background conditions (challenging natural scenes, detection in crowd, cluttered scenes, etc).

1.1.2 Collaborative Representation Classifiers (CRC)

Collaborative filters are popular in recommender systems to effectively encode user trends (Schafer *et al.* (2007)). Collaborative representation classification (CRC) represent the test image as an optimal weighted average of training images across all classes. The predicted label is the class having least residual. The process is explained with relevant formulae in the next chapter.

This inter-class collaboration for optimal feature representation is novel, considering the traditional purely discriminative approach. CRC has a closed



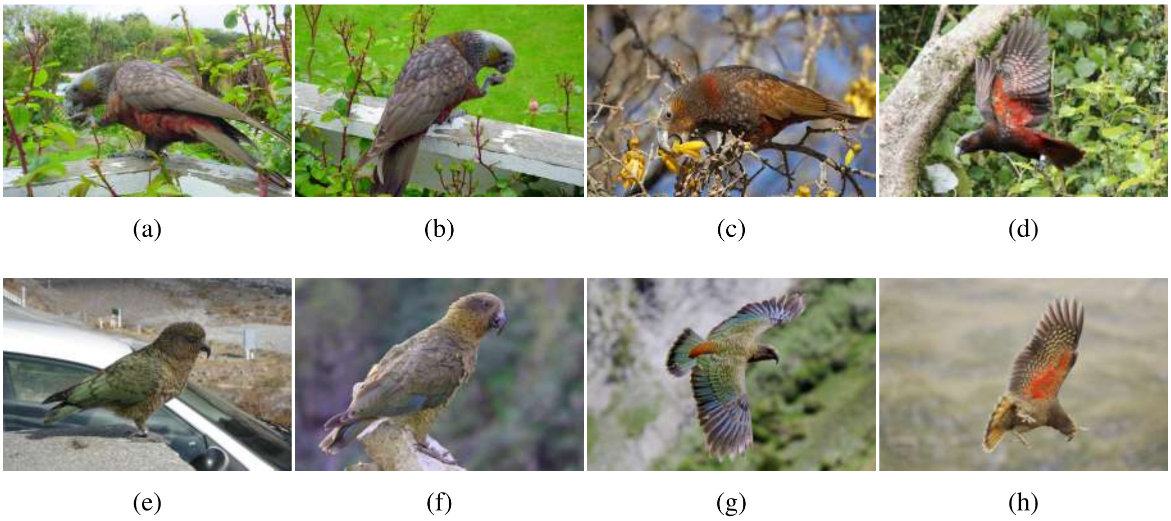
(a)

(b)

Figure 1.1: The first row has images of different types of vehicles, namely helicopter, airplane, ship and motorbike. These can be taken as examples of base object recognition or coarse categories. The second row gives images of different types of aircrafts, that is it presents sub-categories of the same type of vehicle. These may be taken as examples of fine-grained classes. Images have been taken from object classes of public datasets. eg. the aircraft images are from the Oxford FGVC-Aircraft dataset

form solution and does not need iterative or heuristic optimization; thus it is efficient and analytic. It is also a general feature representation-classification scheme and thus most off-the-shelf features and ensembles thereof are compatible with it.

In computer vision, CRC was first applied to the face recognition problem (Zhang *et al.* (2011)). This is because human faces have subtle inter-class differences and significant similarities across classes and CRC is effective in encoding these attributes across classes as mentioned before. However, most of the existing works on CRC based face recognition have reported results on benchmarks having well aligned and centered images with minimal background. Even the few works which have used face datasets in natural scene backgrounds have mostly employed pre-processing steps to align and crop the face region, thus removing the effect of the natural setting by manual intervention (Taigman *et al.* (2009)).


(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 1.2: Sample images of NZ endemic birds Kaka (top row) and Kea (bottom row) from the new NZBirds v1.0 dataset. The challenging nature of the bird species recognition problem is evident from the images, due to subtle inter-class differences and high variation in background and pose (in flight vs. roosting).

Face recognition, like FGVC, also involves the challenge of low inter-class variation since all human faces share many similar characteristics, leading to a sparsity in discriminative parts and features. Hence, given the high accuracies achieved by CRC in the face recognition problem, it seems logical to expect a certain level of applicability to the FGVC problem. Another major advantage of using the CRC framework is the fact that it is a feature representation and classification paradigm and hence can be used in conjunction with any state of the art features.

Thus it seems worthwhile to explore in depth the validity of the intuition that CRC may be particularly suitable for the FGVC problem. CRC based methods have been sporadically used in works that happen to involve some experiments on FGVC datasets among other problems (Cai *et al.* (2016)). But there has not been focused research to ascertain the appropriateness of CRC for FGVC.

In this work, first the robust performance of existing CRC methods is demonstrated in tackling FGVC tasks, taking species recognition as the representative problem. After establishing the alignment of CRC methods to the FGVC problem, a probabilistic collaborative representation of image patches is presented to address the problem of high randomised background variation between classes in FGVC, compared to subtle differences in foreground objects. Second, an improved local feature descriptor is introduced for better representation of repeated object patterns. Third, the collaborative loss function is embedded into an end-to-end deep convolutional network (CoCoNet). This helps in establishing a benchmark transfer learning protocol to learn small specialised fine-grained image datasets. Fourth, we also introduce a collaborative unsupervised learning technique as a generalisation of the standard K-means clustering (Lloyd (1982)). Here the cluster centers are updated using the colloborative weighted mean and hence K-means can be viewed as a special case. Lastly, we also show how distance metric learning may be used to enhance performance by improving discrimination of the collaborative cost function. As part of the research, four benchmark fine-grained species recognition dataset were compiled, those of NZ birds and butterflies as well as Indian birds and butterflies. These are summarised in Table 1.1.

1.2 Thesis Objectives

This thesis has 3 main objectives as follows.

1.2.1 Advance Collaborative Fine-grained Recognition

The first aim of the work is to demonstrate the applicability of collaborative representation classifiers (CRC) in fine-grained visual categorization (FGVC) and advance the state-of-the-art of the current CRC methods. For this benchmark existing methods are first evaluated and then improved.

1.2.2 Collaborative Supervised/ Unsupervised Learning

If it can be shown that CRC methods work well for FGVC problems, the next objective is to incorporate CRC into both supervised and unsupervised machine learning. For unsupervised learning, this would result in a generalised collaborative clustering method. For supervised learning, this would result in a collaborative convolutional network for improved transfer learning of fine-grained categories.

1.2.3 New Fine-grained Benchmark Image Datasets

The last objective as a natural by-product of the current research would be compilation of new benchmark fine-grained image datasets. Given the unique ecology of New Zealand, the aim is to compile species recognition datasets for endemic New Zealand species.

1.3 Thesis Overview

1.3.1 Thesis Structure and Chapter Linkages

The thesis consists of nine inter-related chapters presented below. The chapter linkages are illustrated in Fig. 1.3.

1. **Chapter 1** is the introductory chapter which outlines the research problem, the methodologies used, the project deliverables and summarises the outcomes.

2. **Chapter 2** serves as a literature review of the existing methods that address the research problem, explores the links between the genre of methods selected for this work and the earlier approaches, and gives some preliminary results to justify the scope and approach of this work based on the literature.

3. **Chapter 3** presents the new benchmark fine-grained image datasets for species recognition compiled as part of this research. We describe and give examples images from all classes of these datasets. The four datasets are of New Zealand Birds, New Zealand Lepidoptera (moths and butterflies), Indian Birds and Indian Butterflies.

4. **Chapter 4** presents a patch based probabilistic framework that incorporates background supression in the collaborative cost function itself. Thus it addresses a very important issue of FGVC problems, that of significant background variation compared to subtle differences in foreground objects.

5. **Chapter 5** presents a new binary descriptor that encodes repeated local patterns effectively. It is an improvement over the binary local patterns by incorporating rotational invariance. It retains a simple formulation compared to some recent rotation invariant methods.

6. **Chapter 6** presents an end-to-end collaborative convolutional network (CoCoNet) that improves transfer learning of fine-grained datasets by adding a collaborative layer into the ConvNet architecture.

7. **Chapter 7** presents a collaborative clustering method that maybe viewed as a generalised formulation of the classical K-means clustering. In each iteration, the cluster centers are updated as a weighted mean of the data points belonging to those clusters. The optimal weights are provided by the collaborative cost function; thus K-means would be a special case with all weight unity.

8. **Chapter 8** presents an algorithm to determine the best distance metric in the collaborative cost function for better discrimination and hence adds an additional layer of performance optimisation.

9. **Chapter 9** concludes the thesis, summarizes the work and the outcomes and outlines scope of future work.

1.3.2 Academic Papers

This doctoral thesis has resulted in seven academic papers that are in various stages of peer-reviewed publication process. In accordance with the University of Otago norms about including publications in thesis, the candidate would like to declare that he was the first author and main contributor (at least 90% of the research, analysis and writing) of the work presented in these papers. The contents of Chapters 4, 5, 6, 7 and 8 are mostly the compilation of the contents of these papers. There is, of course, some modifications, more detailed explanations, and reformatting as found appropriate by the candidate to suit the style of the thesis. Chapter 2, mainly focusing on the literature review, has been compiled for the thesis with only some parts taken from a publication as outlined in Table 1.1. Also Chapter 3 gives details of the four new datasets compiled during this work. The table presents information about the target outlets, the chapters that link to those papers, and current status at the time of submission of this thesis.

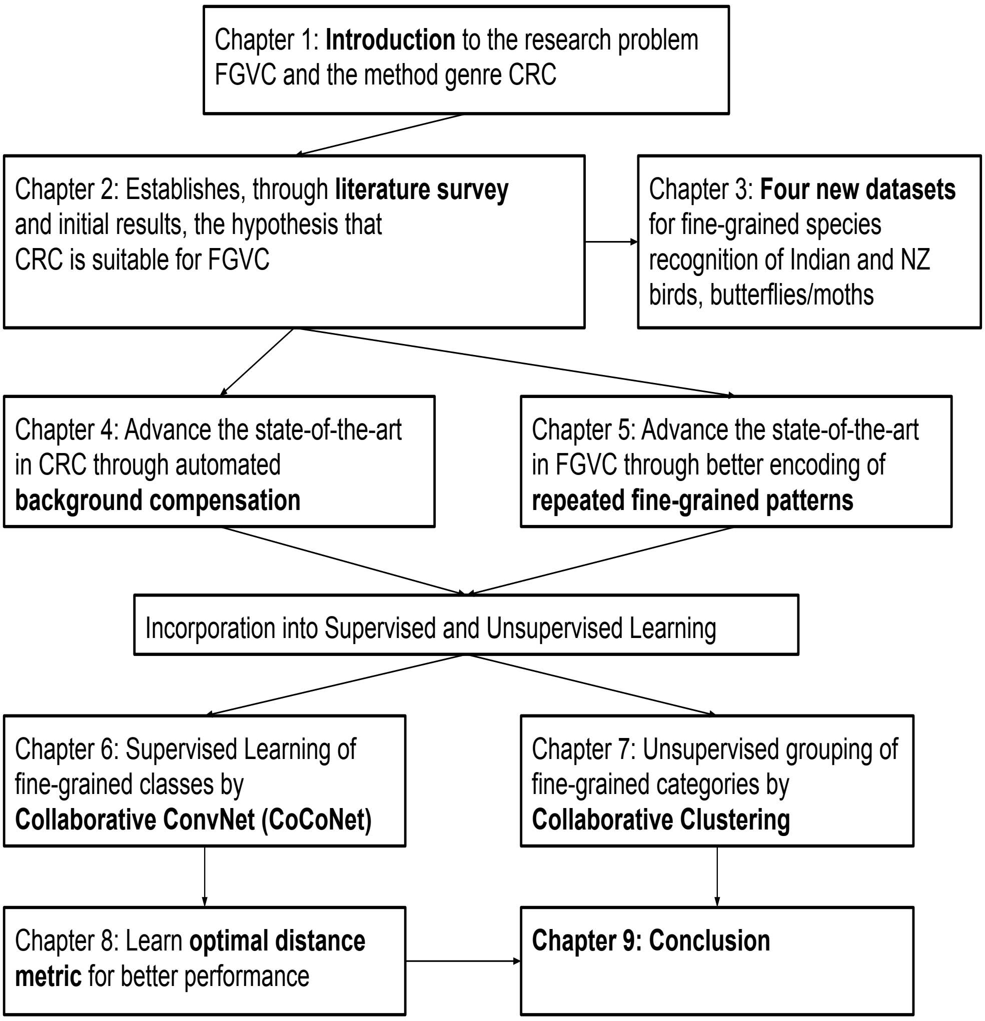

Chapter 1: **Introduction** to the research problem FGVC and the method genre CRC
Chapter 2: Establishes, through **literature survey** and initial results, the hypothesis that CRC is suitable for FGVC
Chapter 3: **Four new datasets** for fine-grained species recognition of Indian and NZ birds, butterflies/moths
Chapter 4: Advance the state-of-the-art in CRC through automated **background compensation**
Chapter 5: Advance the state-of-the-art in FGVC through better encoding of **repeated fine-grained patterns**
Incorporation into Supervised and Unsupervised Learning
Chapter 6: Supervised Learning of fine-grained classes by **Collaborative ConvNet (CoCoNet)**
Chapter 7: Unsupervised grouping of fine-grained categories by **Collaborative Clustering**
Chapter 8: Learn **optimal distance metric** for better performance
Chapter 9: Conclusion

Figure 1.3: Thesis flow and chapter linkages

Table 1.1: Research Outputs

Chap	Paper Name	Outlet	Status
2	Collaborative representation based fine-grained species recognition	Image and Vision Computing New Zealand (IVCNV), 2016	published
4	A Generalised Formulation for Collaborative Representation of Image Patches (GP-CRC)	British Machine Vision Conference (BMVC), London, 2017	published
4	PProCRC: Probabilistic Collaboration of Image Patches	IEEE Transactions in Image Processing	under review
5	LOOP Descriptor: Local Optimal-Oriented Pattern	IEEE Signal Processing Letters, 2018	published
6	CoCoNet: Collaborative ConvNet for deep transfer learning of fine-grained classes	Machine Vision and Applications	under review
7	Fine-grained Collaborative K-Means Clustering	Image and Vision Computing New Zealand (IVCNV), 2018	published
8	Distance Metric Learned Collaborative Representation Classifier	IEEE Transactions on Pattern Analysis and Machine Intelligence	under review

Chapter 2

Literature Review: Existing Methods, Benchmark Datasets and Preliminary Results

2.1 Chapter Summary

There are two main parts to this chapter on existing literature. The first part presents an account of existing research in fine-grained visual categorization (FGVC). It sets up briefly the research problem and then goes on to trace early research in the area. It then moves on to recount more recent trends, categorizing the current research approaches, while illuminating the gaps and scope for improvement. The second part focuses on collaborative representation classification (CRC) and the existing work in the topic, since that is the main methodology adopted in this work. The original CRC formulation is presented here along with some of its popular variants. The hypothesis is that CRC works in a way that should make it particularly suitable for FGVC problems. The claim is intuitively justified in this chapter based on earlier research in related problems. We also test out the hypothesis to obtain some preliminary benchmark results using a few existing CRC methods.

2.2 Fine-grained Visual Categorization

This section presents a brief history of the FGVC problem. First we recount the early work in the area and the formalization of the problem as well as the initial datasets. Then we move on to elaborate on the current methodologies, research groups and datasets in FGVC.

2.2.1 Early years of FGVC until 2010

Deep learning based vision systems have achieved human-like performance in recognizing base categories. Hence research has been shifting towards solving the more challenging task of recognizing finer sub-categories. Thus research interest in FGVC has grown rapidly in the current decade, however some early research can be traced back to more than 10 years back.

Oxford Visual Geometry Group. One of the earliest formal works in the topic is that on flower category detection between 2006 to 2009 at the Oxford Visual Geometry group by Maria-Elena Nilsback and Andrew Zissserman. Flower type recognition was taken as the representative fine-grained recognition problem, because it presents all the characteristic challenges of the field. Many flower species have high intra-species visual differences though they are of the same type. On the other hand, two different flower species may have quite similar appearances. Also, for images in the wild, significant variation in background can be a confounding factor across classes. Some flower species may be much more rare than the others, thus causing an imbalance in sample size of classes. These atrributes make flower species recognition a good benchmark FGVC task. A benchmark dataset encompassing 17 major flower types (with 80 images per type) common to the UK was developed in context of this work. This was later expanded to include 102 flower categories with 40 to 258 images per class.

The work addressed two major aspects of the problem: automated flower segmentation and flower category recognition. For automated flower segmentation, Nilsback *et al.* (2006) proposed a coupled model of two parts: a color model for the foreground and background of the image, and a viewpoint and deformation independent petal shape model. The final segmentation is achieved by using a binarized Markov random field based cost function using graph cuts. The work was later expanded by Nilsback *et al.* (2008), where a training set was used to estimate the parameters of the model for optimal performance. For flower type categorization, a visual bag of words was used (Nilsback *et al.* (2007)) to characterize several global features like colour, shape and texture. For classification, a nearest neighbour based architecture is optimised for vocabulary selection and combination, using validation datasets. The work was later extended in (Nilsback *et al.* (2009)) where a multiple kernel based support vector machine was employed for classification.

Ponce Research Group. Early contributions were made also by the Ponce Research Group at Beckman Institute, University of Illinois at Urbana-Champaign. Discriminating local parts were discovered using Laplacian blob detectors which were then represented by scale invariant feature transform (SIFT) descriptor and classification was achieved by a maximum entropy based framework. Two datasets were developed in this regard: Ponce birds dataset by Lazebnik *et al.* (2005) (6 classes with 100 samples per class) and Ponce butterfly dataset by Lazebnik *et al.* (2004) (619 images of 7 butterfly classes). However, the algorithms were developed to address the general object classification problem and hence were applied simultaneously to both base class datasets (Caltech256 dataset by Griffin *et al.* (2007)) as well as the above mentioned fine-grained classes and sub-categorical recognition was yet to be explored as a distinct research task.

Formalisation of FGVC. In 2007, Hillel and Weinshall published one of the pioneering works (Hillel *et al.* (2007)) in FGVC in which subordinate class recognition was addressed as a distinct research problem with unique nuances, that deserved a new solution framework. The importance of identifying distinguishing object parts was discussed. In the first stage, the representative parts of the base class are identified and modelled and those part models are represented by feature vectors. These features are then used for sub-categorical classification by a linear support vector machine (SVM). To evaluate the method, the work considers 6 base classes with 2 sub-categories each (collected from Caltech object datasets): Motorcycles (Cross, Sport), Faces (Male, Female), Guitars (Classical, Electric), Tables (Dining, Coffee), Chairs (Dining, Living), Pianos (Grand, Upright). Each class, say Motorcycle, is characterised by a set of P representative parts, each of which is then modelled to form a descriptor comprising of the first 15 DCT (discrete cosine transform) values of the image, location (x, y) co-ordinate and log-scale size of the part, mean and standard deviation of feature value and the log-likelihood of occurrence of that part model in the corresponding class. These P representative parts would be different for different base classes. The parts would be same for sub-categories, but the values are shown to be different enough to achieve sub-categorical classification when fed into the SVM classifier.

ImageNet and Caltech-UCSD Birds (CUB) datasets.

The ImageNet project (Russakovsky *et al.* (2015)) has had a huge impact in the general standardization and advancement of computer vision research particularly related to object recognition, segmentation, classification tasks. It has about 1.4 million image categories as of 2017. Research in FGVC has picked up momentum in the present decade particularly after the publication of the CUB birds datasets, which have now become one of the most popular benchmark datasets to evaluate FGVC algorithms. The CUB-200-2010 dataset (Welinder *et al.* (2010)) has 6033 images of 200 mostly American bird species and the extended CUB-200-2011 dataset (Wah *et al.* (2011)) has 11788 images with the same number of classes. Further details of the dataset are provided later in this chapter.

2.2.2 Current Trends in FGVC from 2010

Following the work by Hillel *et al.* (2007) and other similar works by early researchers, and particularly with the introduction of the CUB Birds benchmark datasets, FGVC now formed its own identity as a distinct research problem and a steady development of research methodology started particularly designed to address the area.

Parts localization and alignment

Because of the close similarity between fine-grained object classes, proper identification of discriminating object parts is an important step in FGVC. This part localization is either done manually through expert annotation or in an automated fashion by identifying keypoints. After localization, the next step is part alignment, that is compensating for pose variation. This becomes quite important for certain fine-grained recognition tasks like species recognition where the objects are deformable and can vary a lot with different poses. Consider the case of bird species recognition. The discriminating parts among different bird species may be the beak, head, wings, tail and feet. Depending on the camera angle, some parts may be partly visible and also depending on whether the bird is in flight or roosting, the wings can look quite different.

Manual annotation is not the option of choice for any computer vision system since it involves cost, time and possibility of human error. This is compounded in the case of specialised fine-grained recognition tasks like species recognition. In that case domain experts are needed to make accurate annotation and labeling. Unlike bird species recognition where the discriminating parts are more obvious and consistent, the case of flower species recognition might be even more challenging. Only an expert can possibly point out the exact features that separate two visually similar flowers. Thus the preferable route would be automated part localization. On the other hand, for automated part localization, one would be faced with precisely the same challenge: how to robustly design a part detector and whether one can trust the part detector in lieu of a human expert.

Part template detectors. One approach to discover parts in an automated manner is to create and train part templates, that then extract the part features from test images. One example is the work by Yang *et al.* (2012b). The templates are first inititalised with different sizes and locations. The objective function takes into account both the co-occurrence of the templates as well as the diversity between them. These trained detector templates then extract features from the test images and the features are concatenated into the final feature vector for the classification phase.

Deformable part detectors. The success of part templates led to the extensive use of the more powerful deformable part models (DPM) for the next five years. The first use of DPM in fine-grained recognition was by Zhang *et al.* (2013). In DPM, more than one detector may describe the same part unlike the previous part detectors. Each DPM is a combination of detectors. Each detector has a root histogram of gaussians (HoG) filter and a series of part HoG filters (Dalal *et al.* (2005)). This mixture helps cater to multi-modal objects, like different views of a bird (front/side, flight/roosting). The root filter is moved as a sliding window through the image to detect parts at different locations and at different scales. Parts are then placed optimally around the detected root locations.

Poselet detectors. Suppose keypoints are known for at least a part of the training set through either expert annotations or keypoint detectors. These keypoints then might be used to construct what are known as poselets (Bourdev *et al.* (2009)). Given keypoints in a training set, groups of patches with similar keypoints are first searched, after which a template-based detector (poselet) is configured for each of the patch groups. Several poselets may then go on to describe one part of the object.

Birdlet detectors. DPM can be used for any object categorization tasks and poselets are a specialisation of DPM that were developed for fine-grained recognition. Farrell *et al.* (2011) further refined the concept of poselets to form Birdlets, which are specialised poselets for species recognition, applied to bird species categorization. The authors modeled the bird head and body with two ellipsoids. The part detectors (birdlets) were trained on groups of patches with similar pose using annotated keypoints. All detectors were applied to the test images to extract the ellipsoids of the head and the body and the resulting features were used for classification.

Handling sparse detectors. One general drawback of discriminating parts and deformable models, including specialised versions for FGVC like poselets and birdlets, is that since only a few of these are activated per image for fine-grained categories, this generates a sparse concatenated descriptor, which then results in low classification accuracy. There can be various ways to overcome this. One can be dimensionality reduction of the descriptor, through pooling poselets that describe the same part of the objects. A second general approach is collaboration between similar objects to make up for missing parts, by taking an optimal average description (this approach is explored in the present work using collaborative filters). The third approach is to replicate missing parts by plugging them with corresponding available poselets of similar objects, if pose relationship is known between poselets. Zhang *et al.* (2012) used this in their work by applying warp kernel operations on neighbouring poselets to aggregate activations with similar locations on the object, say, bird.

Foreground segmentation and background suppression. One major challenge of FGVC, as mentioned earlier, is that the foreground objects may have subtle differences while the background can have large variations randomised between classes. Thus backgrounds can become a confounding factor while learning the fine-grained classes and hence robust background compensation or suppression is an important part of FGVC. One way of achieving this compensation is by having a pre-processing step involving binary segmentation of the foreground using methods like graph cut and then use parts of the objects for classification. A weighted graph can be used for binary segmentation of the foreground-background. Each vertex of the graph has a prior probability of being background/foreground that serve as unary potentials for graph-cut (Parkhi *et al.* (2011)). Edge detectors are used to compute binary potentials which indicate whether two connected vertices will have same label.

Without Part Annotations. We have seen getting human labels and annotations is costly for fine-grained species recognition since domain expertise is required. Labels or annotations generated by citizen scientists may be considered unreliable, due to the subtle differences in classes like say bird species. The research by Jonathan Krause and Fei Fei Li at the Stanford AI Lab has deals with this issue in a couple of ways. In one of their works, they use crowdsourced annotations to limit the uncertainty (Deng *et al.* (2013)). They use an online game called Bubbles to make humans interact and generate parts and the general consensus is used as a pseudo-expert. In another work they try to achieve fine-grained recognition without part annotations (Krause *et al.* (2015)). They generate parts in a fully automated fashion using co-segmentation and alignment, that they combine in a discriminative mixture to achieve a competitive recognition rate.

Convolutional Networks for FGVC

Some of the most powerful vision systems now use deep convolutional neural network (CNN) for the majority of visual tasks. It is therefore not surprising that they are fast gaining popularity in fine-grained recognition tasks. However, fine-grained categorization problems pose certain specific challenges that warrants special considerations while using ConvNets to address them.

CNN Descriptors. Even with graphics processing unit (GPU) enabled modern deep vision systems, training a full CNN from scratch is time consuming and for a large scale problem, often taking days or even weeks. This is because deep convolutional networks by nature are data hungry, that is, they need a large amount of training data to achieve good performance. This might be a challenge in certain specialised FGVC tasks like endangered species recognition, where the number of training samples available will be limited. In such cases, one might choose to use a network that has been pre-trained on a large dataset like ImageNet on a generic object recognition task. The bottom/initial layers of a convolutional network (ConvNet) represent generic low level patterns of images. The first layer may just be simple edge detectors, and the patterns represented gradually increase in complexity over the layers. Hence if one trains a ConvNet on a large dataset for generic object recognition and then takes out the last few layers, it can be used as a feature extractor (Cai *et al.* (2016)). The output of the pre-trained ConvNet are used as feature descriptors for the fine-grained classification task. Of course, since the network is not trained on the target fine-grained smaller dataset, the performance will not be optimal, but in many cases will still be sufficient.

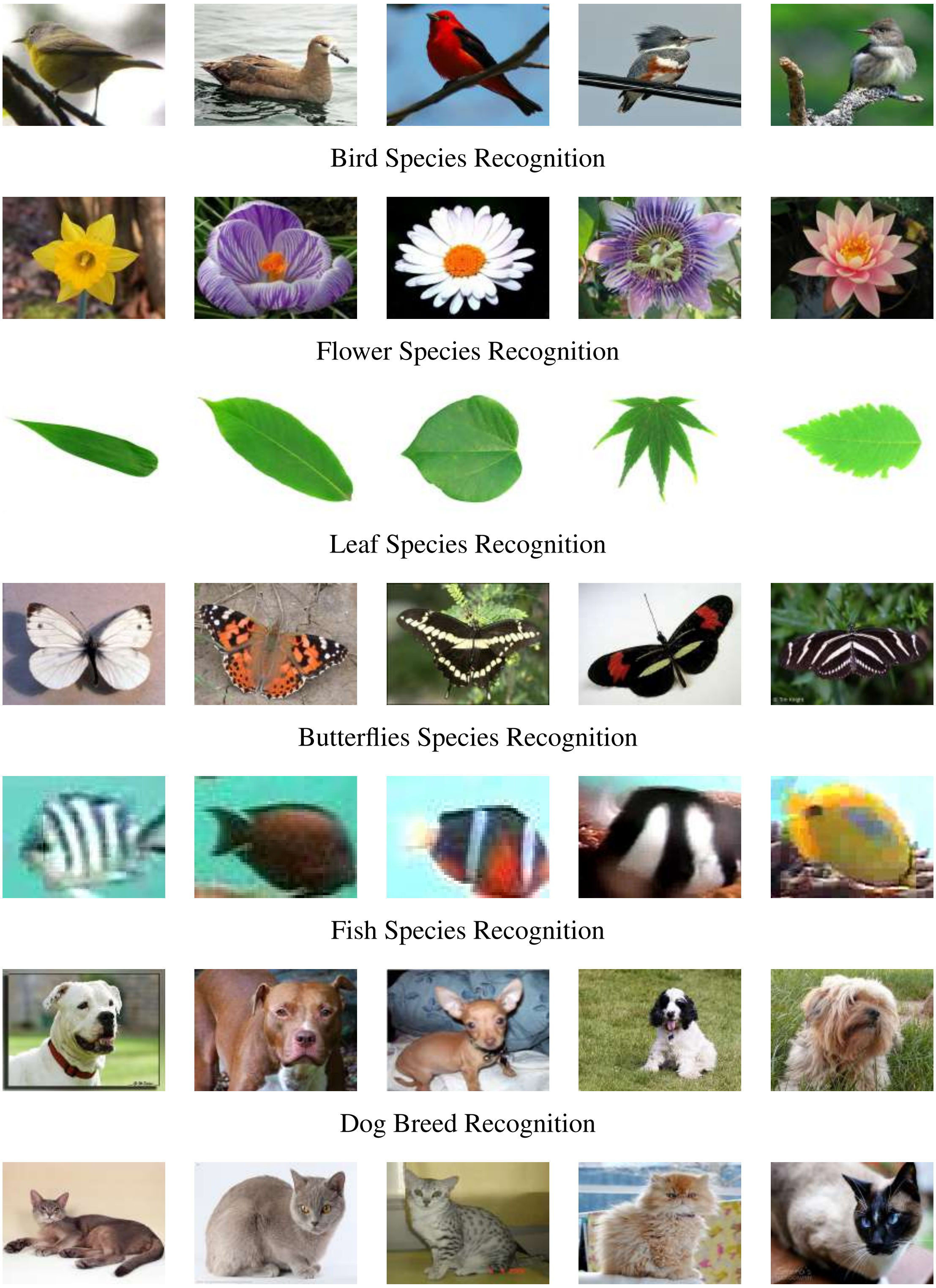
Fine-tuning and Transfer Learning. An obvious improvement in approach is to fine-tune the pre-trained network on the fine-grained dataset for the specific FGVC task. This will update the network weights, which were previously trained on the larger source dataset like ImageNet for a more generic object recognition task Simon *et al.* (2015). Then a validation set may be used to tune the hyper-prameters of the network as well, before the final testing phase. There has been recent research solely focusing on how to use ImageNet best for transfer learning like the work by Kornblith *et al.* (2019).

Specialised CNNs. Some research has been put into developing deep convolutional network (CNN) architectures with modifications to suit fine-grained recognition. One example in that genre would be the Bilinear CNN (BCNN) by Lin *et al.* (2015), where the authors present a CNN architecture that uses translationally invariant interactions between neighbouring fine-grained features. Another example is that of pose normalised deep convolutional networks by Branson *et al.* (2014). Here the pose normalisation is done first by patch wise alignment which are then used by a deep network. Thus in both cases there is the element of achieving pose alignment or translational invariant which is very important for such fine-grained recognition tasks like bird species recognition.

2.2.3 Major FGVC Applications and Datasets

The main task in FGVC is to recognize fine-grained sub-categories of objects with subtle differences. Species recognition is the main representative problem of FGVC as it presents most of its characteristic challenges and hence majority of the standard datasets pertain to that. We cover these in details here and a few non-species benchmark datasets too.

The Cross Language Evaluation Forum (CLEF) is an yearly program mostly based in Europe and ImageClef (Muller *et al.* (2010)) is a section of that dealing with benchmarking image retrieval and object recognition algorithms on standard datasets. As part of that every year since 2014, the LifeClef challenge has been organised which has fish, bird and plant species recognition datasets publicly available every year. Besides this, we discuss below some task specific species recognition datasets. A FGVC workshop dedicated to fine-grained visual problems is organised every year at the premier conference on computer vision CVPR. The first one was held at CVPR 2011 and the most recent sixth workshop on FGVC was held at CVPR 2019.



Bird Species Recognition

Flower Species Recognition

Leaf Species Recognition

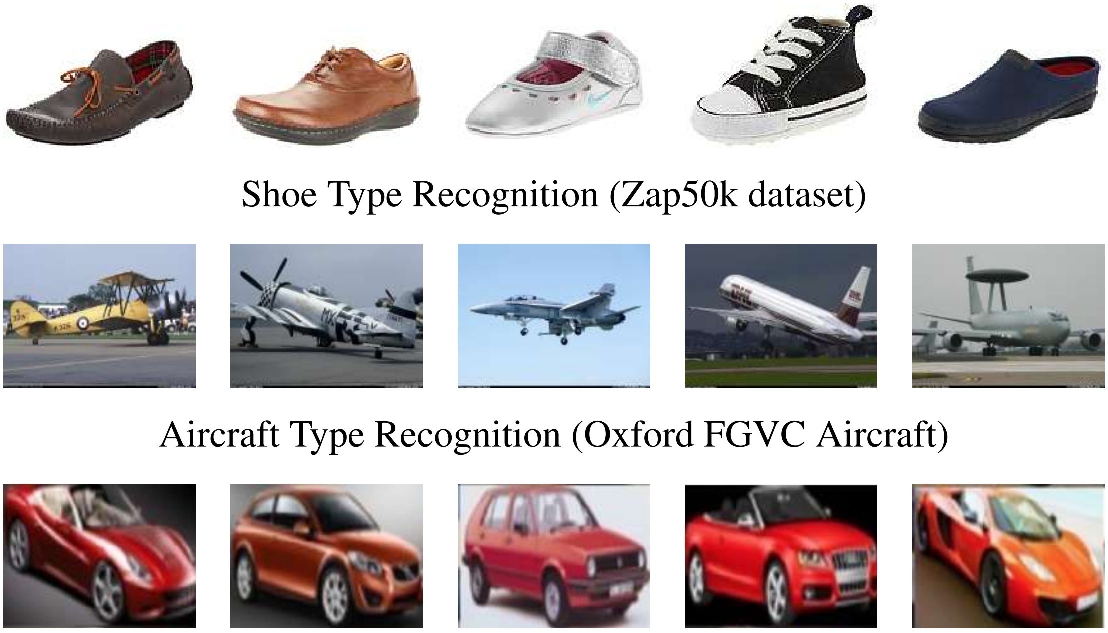
Butterflies Species Recognition

Fish Species Recognition

Dog Breed Recognition

Cat Breed Recognition

Figure 2.1: Fine-grained Species Recognition Tasks



Shoe Type Recognition (Zap50k dataset)

Aircraft Type Recognition (Oxford FGVC Aircraft)

Car Type Recognition (Stanford Cars)

Figure 2.2: Other FGVC tasks besides species recognition

Bird Species Identification:

- The most popular benchmark FGVC dataset is the Caltech-UCSD Birds database image dataset. The first version by (Welinder *et al.* (2010)) introduced in 2010, called CUB-200-2010, has 6033 images of 200 mostly North-American bird species. The CUB-200-2011 dataset by (Wah *et al.* (2011)) introduced in 2011 is the extended version with the same number of classes (200) but with a total of 11,788 images (almost double).

- Ponce Birds dataset, created at the Ponce Research Group at University of Illinois at Urbana-Champaign, consists of 600 images (100 samples each) of 6 different categories of birds (Lazebnik *et al.* (2005)).

- The North American Birds dataset (NABirds) has been developed through the joint collaboration of Cornell Tech, UCSD and CalTech. It has 70,000 images of 550 bird classes and hence, unlike most FGVC species datasets that are small in size, NABirds is large enough to train a CNN effectively from scratch (Horn *et al.* (2015)15).

Insect Species Identification:

- The University of Jena has developed 2 datasets for fine-grained biodiversity analysis (Rodner *et al.* (2015)). The first one, Costa Rican butterfly dataset, contains around 3000 images of a broad range of moth and butterfly taxa found in north western Costa Rica (female individuals with at least 5 samples per type). The second one, Ecuadorian moth dataset, consists of around 2000 images of one single family (taxon) of moths (Geometridae) found in southern Ecuador. One small drawback of this dataset is that the paper with the typed species name attached to the photograph is visible for some of the digitised images.

- The Leeds butterfly database (Wang *et al.* (2009)) consists of 832 images of 10 categories of butterflies with 55 to 100 images per category. Text descriptions and segmentation masks are also provided.

- The Ponce Group butterfly database (Lazebnik *et al.* (2004)) created at University of Illinois at Urbana-Champaign, consists of 619 images of 7 different classes of butterflies.

- Oregon State University fly database has 2 parts. The STONEFLY9 (Lytle *et al.* (2010)) dataset consists of 3826 images of 773 specimens of 9 taxa of Stoneflies. The EPT29 dataset (Larios *et al.* (2011)) consists of 4842 images of 1613 specimens of 29 taxa of EPT flies: Ephemeroptera (mayflies), Plecoptera (stoneflies), and Trichoptera (caddisflies). It is a very high resolution dataset of large volume, however, there is also a smaller curated front aligned cropped version (canonical pose).

Fish Species Identification:

- Croatian Fish species database (Jaeger *et al.* (2015)) developed by University of Jena, has around 800 images belonging to 12 classes.

- Fish4knowledge database (Boom *et al.* (2012)) is an EU funded international project involving several research labs and universities around the world. It focuses on fish recognition from underwater video, but also has a still images dataset for fine grained fish species recognition. The dataset has 27,370 fish images of 23 classes.

Flower/Leaf Species Identification:

- Leafsnap database (Kumar *et al.* (2012)) developed at Columbia University covers 185 tree species found in Northeastern united States. There are 23,147 oriented images taken in the lab as well as 7719 field images taken in the wild.

- Kingston 100 Leaves database (Mallah *et al.* (2013)) is a fairly small dataset developed at Kingston University, London. It has 100 classes of leaf types with 16 images per class.

- Oxford Flower database developed at the Oxford Robotics Lab, has 2 parts. The 17 category dataset has 80 images per class (Nilsback *et al.* (2006)), (Nilsback *et al.* (2007)). The number of images in the 102 category dataset varies between 40 and 258 per class. Both datasets have scale, pose and light variations (Nilsback *et al.* (2008)), (Nilsback *et al.* (2009)).

Cat/dog Species Identification:

- Oxford-IIIT Pets dataset (Parkhi *et al.* (2012)) compiled by the Oxford Robotics Group and IIIT Hyderabad, consists of 37 category of pet cats and dogs with around 200 images belonging to each class, with variations of pose, illumination and scale.

- Stanford Dogs dataset (Khosla *et al.* (2011)) has 20,580 images of 120 breeds of dogs from around the world.

Other non-species FGVC tasks:

- There may be other fine-grained recognition tasks not pertaining to species recognition. One example is recognition of shoe types using datasets like Zappos50k by University of Texas (Yu *et al.* (2014)). This dataset has 50,025 images of different types of shoes taken from Zappos.com (Yu *et al.* (2017)).

- There are similar datasets on vehicle type recognition. There is a fine-grained aircraft dataset by University of Oxford Robotics Group (Maji *et al.* (2013)). It is called the FGVC-Aircraft dataset and it has 100 images each of 102 aircraft models. There is a car model recognition dataset by Stanford AI group with 16,185 images of 196 classes of cars (Krause *et al.* (2013)).

2.3 Collaborative Representation Classifiers (CRC)

As mentioned in the previous chapter, the main hypothesis of the thesis is that collaborative representation classifiers should be good candidate methods to handle fine-grained recognition. In this section we present that overview of CRC, its variants as well as baseline results to support the claim and setup the rest of the thesis.

2.3.1 Overview of CRC and Formulation

The mathematical framework for CRC is described in brief below (Zhang *et al.* (2011)). Consider a training dataset with images in the feature space as $X = [X_1, \ldots, X_c] \in \Re^{d \times N}$ where N is the total number of samples over c classes and d is the feature dimension per sample. Thus $X_i \in \Re^{d \times n_i}$ is the feature space representation of class i with n_i samples such that $\sum_{i=1}^{c} n_i = N$.

The CRC model reconstructs a test image in the feature space $y \in \Re^d$ as an optimal collaboration of all training samples, while at the same time limiting the size of the reconstruction parameters, using the Lagrangian multiplier λ.

The CRC cost function is given as:

$$\hat{\alpha} = \arg\min_{\alpha} (||y - X\alpha||_2^2 + \lambda||\alpha||_2^2) \tag{2.1}$$

A least-squares derivation yields the optimal solution as:

$$\hat{\alpha} = (X^T X + \lambda I)^{-1} X^T y \tag{2.2}$$

The representation residual of class i for test sample y can be calculated as:

$$r_i(y) = \frac{||y - X_i \hat{\alpha}_i||_2^2}{||\hat{\alpha}_i||_2^2} \; \forall i \in 1, \ldots, c \tag{2.3}$$

The final class of test sample y is thus given by

$$C(y) = \arg\min_{i} \; r_i(y) \tag{2.4}$$

Optimal λ may further be chosen using *Generalized Cross Validation* (GCV) as follows:

We have $\hat{\alpha} = (X^T X + \lambda I)^{-1} X^T y$ from (2).
Let,

$$X^{\#} = (X^T X + \lambda I)^{-1} X^T \tag{2.5}$$

Then the GCV cost function is given by:

$$G(\lambda) = \frac{||y - X\alpha_\lambda||_2^2}{trace(I - XX_\lambda^\#)^2} \tag{2.6}$$

The optimal value of λ, for which $G(\lambda)$ is minimum, is graphically determined from the plot of $G(\lambda)$ vs (λ).

Some of the recent improvements and enhancements of the original CRC are listed below. There are many more in the existing literature, but only those that have been directly evaluated in the present work, have been mentioned here.

2.3.2 CRC Variants and use in Computer Vision

Optimized Collaborative Representation (CROC)

Chi *et al.* (2012) suggested a collaborative representation optimized classifier (CROC) to combine nearest subspace classifier (NSC) with either Collaborative Representation based Classification (CRC) or Sparse Representation based Classification (SRC) for multi-class classification. Nearest Subspace Classifier defines the residual for determining class prediction as follows, which is basically the nearest distance minimiser, but weighted across training samples across all classes:

$$r_i^{CR}(y) = ||y - X_i\alpha_i||_2^2 \ \forall i \in 1, \ldots, c \tag{2.7}$$

The final residual is defined as a combination of NSC with either CRC or SRC. CROC combining NSC and CRC would have the residual as:

$$r_i(y) = r_i^{NSC}(y) + \lambda r_i^{CRC}(y) \tag{2.8}$$

Likewise, a combination of NSC and SRC would be given by:

$$r_i(y) = (1 - \lambda)r_i^{NSC}(y) + r_i^{CRC}(y) \tag{2.9}$$

The optimal value of λ can then be solved following the Generalised Cross-Validation scheme explained before.

Multi-scale Patch-based Collaborative Representation (PCRC)

Zhu *et al.* (2012) introduced a patch-based framework to achieve multi-scale collaborative representation.

Let the query image y be divided into q overlapped patches $y = \{y_1, \ldots, y_q\}$ From the feature matrix X, local dictionary M_j is extracted corresponding to patch y_j. Thus the modified cost function for PCRC becomes:

$$\hat{p}_j = \arg\min_{p_j} ||y_j - M_j p_j||_2^2 + \lambda ||p_j||_2^2 \tag{2.10}$$

where $M_j = [M_{j1}, \ldots, M_{jc}]$ are the local dictionaries for the c classes and $\hat{p}_j = [\hat{p}_{j1}, \ldots, \hat{p}_{jc}]$ is the optimal reconstruction matrix for the j^{th} patch. The class of test sample is predicted as:

$$C(y_j) = \arg\min_k r_{jk}(y) \tag{2.11}$$

where

$$r_{jk} = \frac{||y_j - M_{jk}\hat{p}_{jk}||_2^2}{||\hat{p}_{jk}||_2^2} \; \forall i \in 1, \ldots, c \tag{2.12}$$

The classification of the entire test sample y is determined by majority voting of the classification labels of the patches y_j.

Relaxed Collaborative Representation (RCRC)

Yang *et al.* (2012a) developed an improved CRC method (RCRC) with relaxed constraints assigning adaptive weights to features for controlled contribution to final representation. The weights are so optimised that the variance of representative features from mean is controlled, to add stability to the representation.

Thus in the RCRC formulation, the cost function of CRC gets modified to

$$\hat{\alpha} = \arg\min_{\alpha} ||y - X\alpha||_2^2 + \lambda ||\alpha||_2^2 + \tau w ||\alpha - \bar{\alpha}||_2^2 \tag{2.13}$$

where τ is a positive constant and w is the weight vector such that $w = [w_1, \ldots, w_c] \mid w_i \in \Re$ and c is the number of classes.

All other symbols have usual meaning from the CRC formulation. The cost function is iteratively optimized.

Enhanced Collaborative Representation (ECRC)

Liu *et al.* (2014) enhanced the original CRC by incorporating the covariance matrix R of the training samples into the cost function:

$$\hat{\alpha} = \arg\min_{\alpha} \left((y - X\alpha)^T R^{-1} (y - X\alpha) + \lambda ||\alpha||_2^2 \right) \tag{2.14}$$

Kernel Collaborative Representation (KCRC)

Zhao *et al.* (2014) introduced the kernel trick into the CRC framework.

The cost function for KCRC becomes:

$$\hat{\alpha} = \arg\min_{\alpha} ||\alpha||_{l_p} \text{ subj. to } ||\phi(y) - \Phi\alpha||_{l_q} \leq \epsilon \tag{2.15}$$

Here the second term imposes the kernel condition in higher dimension.

Probabilistic Collaborative Representation (ProCRC)

Cai *et al.* (2016) recently proposed a probabilistic representation of the collaborative framework which jointly maximizes the likelihood that a test sample belongs to each of the multiple classes. The final classification is performed by checking which class has the maximum likelihood.

Thus the predicted class label for a test sample y is given by (symbols having usual meaning):

$$\arg\max_{i} \text{Prob}[C(y)] = \arg\max_{i} e^{-||X\hat{\alpha} - X_i \hat{\alpha}_i||_2^2} \tag{2.16}$$

Extended Probabilistic CRC (EProCRC):

Lan *et al.* (2017) recently extended the probabilistic CRC model by incorporating an additional prior information metric β_c into the cost function that measures the distance $||X - X_k||$ between the centroid of the training set from the centroid of the individual classes. Thus the predicted class label for a test sample y is given by (symbols having usual meaning):

$$\hat{\alpha} = \arg\min_{\alpha} \left(||y - X\alpha||_2^2 + \lambda||\alpha||_2^2 + \frac{\gamma}{K} \sum_{k=1}^{K} \beta_c ||X\alpha - X_k\alpha_k||_2^2 \right) \tag{2.17}$$

2.3.3 Initial results of applicability of CRC in FGVC

In this section we make some initital exploratory tests on our main hypothesis of this doctoral research, which is that collaborative filter classifiers should be suitable for fine-grained recognition and if so, the state of the art should be expanded and integrated into supervised and unsupervised learning frameworks. With this in mind, we do some initial benchmarking of results here using standard datasets, features and classifiers.

Datasets

Experiments are performed on two of the most popular and long-standing benchmark FGVC datasets: CUB200-2011 Birds and Oxford 102 Flower datasets.

- CUB200-2011 Birds dataset contains 11,788 images of 200 bird species with around 30 training samples for each species (Wah *et al.* (2011)).
- Oxford 102 Flowers contains 8,189 images from 102 categories, with each category having at least 40 images (Nilsback *et al.* (2008)).

Features

The effectiveness of CRC based classification has been tested using several popular descriptors namely GIST+Color, SIFT and CNN based features.

- *Global Invariant Scale Transform (here referred as GIST)* (Oliva *et al.* (2001)) is a low level global feature that describes the spatial envelope of the image using directional properties. Color descriptor (van de Weijer *et al.* (2007)) converts the image to color bag of words and extracts dense multi-scale overlapping patches. It finally forms a histogram of color words. The features are concatenated and fed into the Bag of Words and Spatial Pyramid pipeline.

- *Scale Invariant Feature Transform (SIFT)* features were proposed by Lowe (1999). A dense variant of that is used with patch size 16×16 with a stride of 8 pixels. After the extraction of the local key-points and the SIFT features, k-means clustering with a size of 1024 is used to generate the codebook or Bag of Words (BoW). A 2-level Spatial Pyramid representation is used for multi-scaling.

- *CNN features* pretrained on the VGG-19 architecture is used. The activations of the penultimate layer are used as features (Simonyan *et al.* (2014)).

Classifiers

Several classifiers have been adopted for comparative evaluation. They are mainly divided into three categories as cited below.

- *CRC based:* A family of Collaborative Representation based classifiers have been utilised including the original CRC implementation along with some of its recent enhancements CROC, PCRC, RCRC and Pro-CRC.

- *Softmax and SVM based:* Probabilistic regression based Softmax classifier has been used along with linear and χ^2 kernel based Support Vector Machines (SVM) Cortes *et al.* (1995). The binary SVM classifiers have been used in one-vs-all format to achieve multi-class categorization.

- *NSC and SRC based:* Sparse Representation based Classification (SRC) is similar to CRC but uses $\ell 1$ norm in the Lagrangian multiplier instead of $\ell 2$ while minimising the cost function. The Nearest Subspace Classifier (NSC) assigns a test sample to the class which has the minimum Euclidean distance to it in feature space.

2.4 Results and Discussion

Average recognition accuracies in percentage over 5-fold cross-validation for CUB200-2011 Birds dataset and Oxford 102 Flowers dataset are presented in Table 2.1. Several interesting observations may be made from the results. First, a gradual but consistent increase in accuracy can be observed as we transition from initial NSC/SRC based classifiers to CRC and optimized CRC (CRC) and then to more advanced modifications of CRC. Pro-CRC which is one of the most recent and state-of-the-art version of CRC, gives the best result in all of the cases among the CRC based classifiers. These trends are consistent across all the features.

It can further be observed that Softmax does not perform at par with the other classifiers but SVM still holds up as a strong contender against CRC. However, the range of accuracy of SIFT and GIST+Color features for any classifier/dataset is insignificant compared to the performance of deep convolutional network features (VGG-19). Thus with the modern CNN features, CRC based classifiers, especially recent modifications like RCRC, PCRC and Pro-CRC consistently outperform SVM.

Table 2.1: Average accuracy (%) of CRC compared to other classifiers

	CUB Birds			Oxford Flowers		
	GIST	SIFT	Vgg-19	GIST	SIFT	Vgg-19
SoftMax	7.5	8.2	72.1	45.7	46.5	87.3
SVM	9.2	10.2	75.4	50.5	50.1	90.9
Kernel SVM	9.8	10.5	76.6	51.7	51.0	92.2
NSC	9.1	8.4	74.5	45.4	46.7	90.1
SRC	8.8	7.7	76.0	48.1	47.2	93.2
CRC	9.3	9.4	76.2	47.3	49.9	93.0
CROC	9.5	9.1	76.2	48.8	49.4	93.1
PCRC	9.9	9.7	76.9	49.7	50.3	94.3
RCRC	10.0	9.5	77.4	50.6	51.0	93.6
Pro-CRC	10.4	9.9	78.3	52.4	51.2	94.8

These initial results are encouraging for the hypothesis of this work that the workings of collaborative representation classifiers should be aligned for fine-grained object recognition. At least these preliminary results using benchmark classifiers, features and datsets, warrant further investigation which is carried out through the rest of this thesis, along with expansion of the state of the art.

2.5 Notes on statistical analysis in this thesis

Three types of statistical tests have been used in the thesis, and since the associated papers have already been published (or under review) in peer-reviewed international outlets, I have not changed the test and kept them as is. These are unpaired t-tests, signed binomial test and Wilcoxon signed rank test. Of these, unpaired t-tests consider accuracy and standard deviation, whereas Binomial test is frequency based and Wilcoxon test is rank based. Care must be taken while performing statistical analysis, particularly regarding experimental design and whether the underlying assumptions of these tests conform with the realities of the experimental data.

A good comparison of these three tests and relative merits/demerits and where and when to use them is found in the work by Demsar (2006). Among these tests, t-tests have the strongest/strictest assumption of normal distribution. T-tests also assume that the sample size is large enough and t-tests are known to be susceptible to outliers.

Wilcoxon signed rank test offers a safer alternative to t-tests as it is less prone to outliers and does not need the assumption of normal distribution. However, when the assumptions of t-tests are satisfied, t-tests offer the stronger indicator of statistical performance. Also like t-tests, Wilcoxon test needs the commensurability of differences, but only qualitatively: greater differences still count more, but the absolute magnitudes are ignored (different from t-tests).

Signed binomial tests is also recommended for use when comparing classifiers over several datasets/features. It has the advantage that since only success/failure is measured, it does not assume any commensurability of scores or differences nor does it assume normal distributions and is thus applicable to any data (as long as the observations, i.e. the data sets, are independent). On the other hand, it is much weaker than the Wilcoxon signed-ranks test, since the sign test will not reject the null-hypothesis unless one algorithm almost always outperforms the other.

A key point is rather to ensure is that for either binomial test or rank test, each trial of the experiment is independent of the other trials. The other thing to take care is the proper use of Bonferroni compensation when needed. Considering the relative safety of using Wilcoxon rank test and binomial sign test compared to other statistical tests, these methods have been used extensively in the present thesis, along with occasional use of the unpaired t-test.

Chapter 3

New Fine-grained Benchmark Datasets

Besides the development of new algorithms to deal with fine-grained recognition problems, this work has also resulted in the compilation of several benchmark image datasets for fine-grained species recognition. There are two bird species datasets for Indian and NZ endemic birds. There are two butterfly/moth species datasets for Indian and NZ Lepidoptera. These are available for research on request.

3.1 New Zealand Birds Dataset

The New Zealand birds dataset (NZBirds) is a small benchmark dataset of fine-grained images of NZ endemic birds, many of which are endangered. Currently it contains 600 images of 20 species of NZ birds and has been compiled by University of Otago in collaboration with The National Museum of NZ (Te Papa), the Department of Conservation (DOC) and the Ornithological Society of NZ (Birds NZ). Sample images of the NZ Birds dataset are presented in Figure 3.1.

Note that not all images in the dataset are of different birds. These images have been kindly made available for this study by individual birder photographers of New Zealand. So a photographer may have taken several shots of an individual bird from different angles. So, though on one hand, this causes repetition of data on the same bird, but on the other hand due to pose difference this acts as a sort of data augmentation and actually helps to avoid overfitting. Same is true for the other datasets.

3.2 New Zealand Lepidoptera Dataset

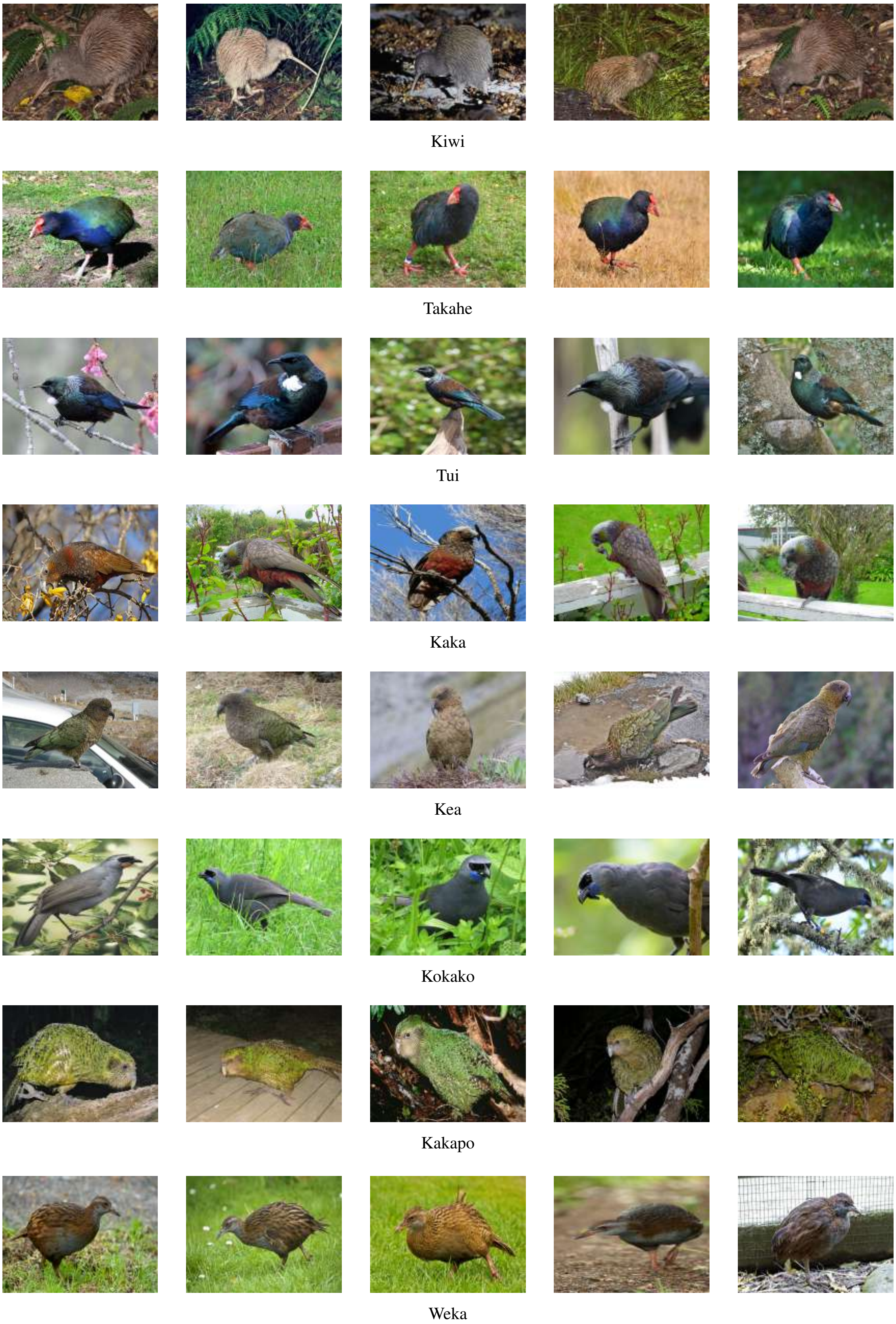
NZ Lepidoptera dataset is a new benchmark built during this work at the Department of Computer Science, University of Otago, NZ in collaboration with the CVPR Unit, Indian Statistical Institute. It has eight classes of NZ butterflies and moths, four categories each. The four butterfly classes are Admiral, Blue, Copper and Ringlet. The four moth classes are Erebidae, Geometridae, Hepialidae and Noctuidae. Currently it has 640 images with 80 images per class, subject to expansion in near future. Images of NZ moths have been partly compiled from the publicly available database of NZ Landcare Research. Sample images of the NZ Lepidoptera dataset are presented in Figure 3.2.

3.3 Indian Birds Dataset

The Indian Birds Dataset (IndBirds) was recently compiled at the Indian Statistical Institute in collaboration with the University of Otago, NZ. It has 14 classes of endemic Indian birds, with 100 images per species. Sample images of the Indian Birds dataset are presented in Figure 3.3.

3.4 Indian Butterflies Dataset

Indian Butterfly Dataset has been compiled as part of the present work in collaboration between the Indian Statistical Institute and the University of Otago, NZ. It is named Titli after the Hindi word for butterfly. The current version 1 has 6 classes with 60 images per class. The six butterfly classes are Papilionidae, Pieridae, Nymphalidae, Lycaenidae, Riodinidae and Hesperiidae. Sample images of the Indian Butterfly dataset are presented in Figure 3.4.



Kiwi

Takahe

Tui

Kaka

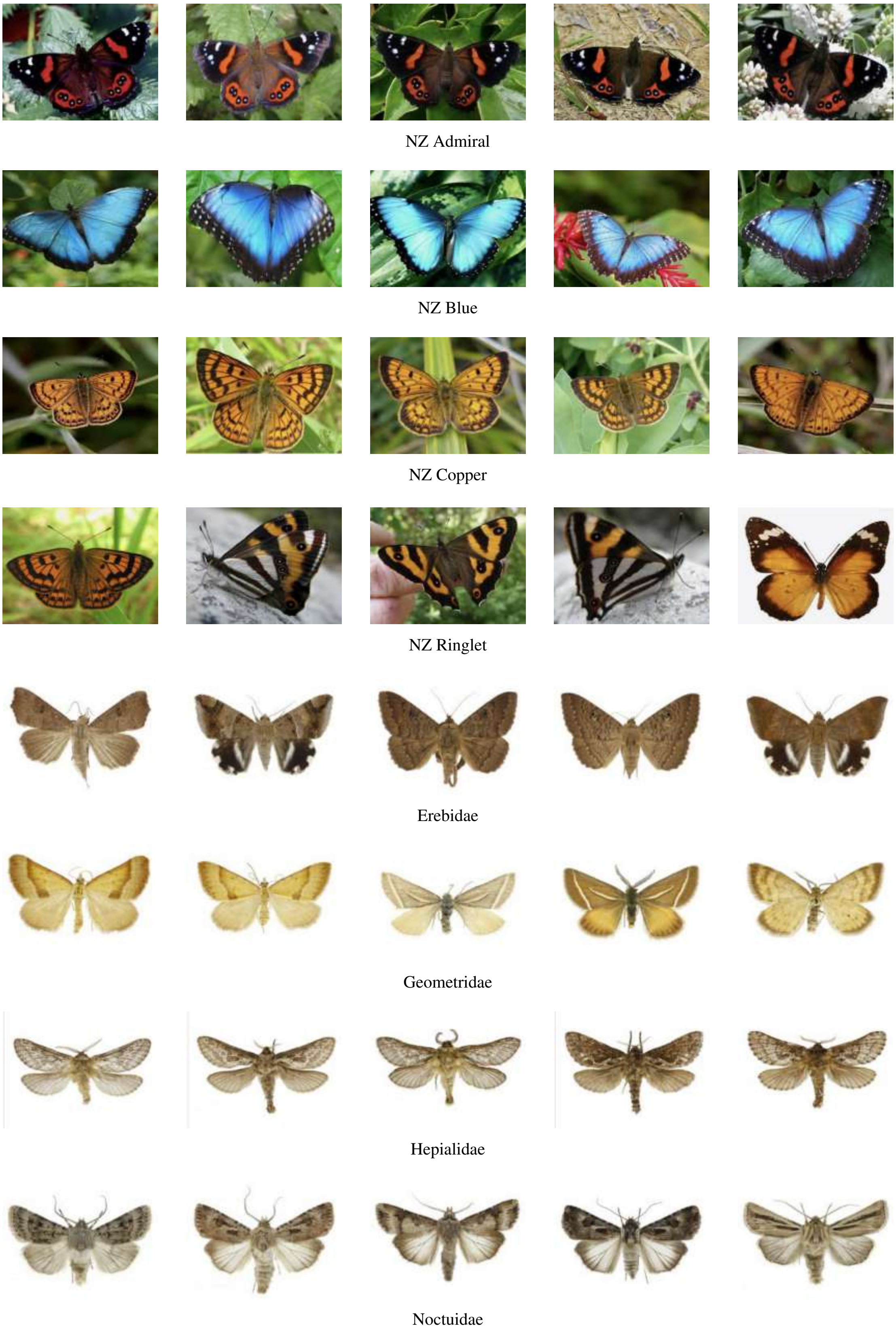
Kea

Kokako

Kakapo

Weka

Figure 3.1: NZ Birds dataset: 8 out of 20 classes; images vary per classs



NZ Admiral

NZ Blue

NZ Copper

NZ Ringlet

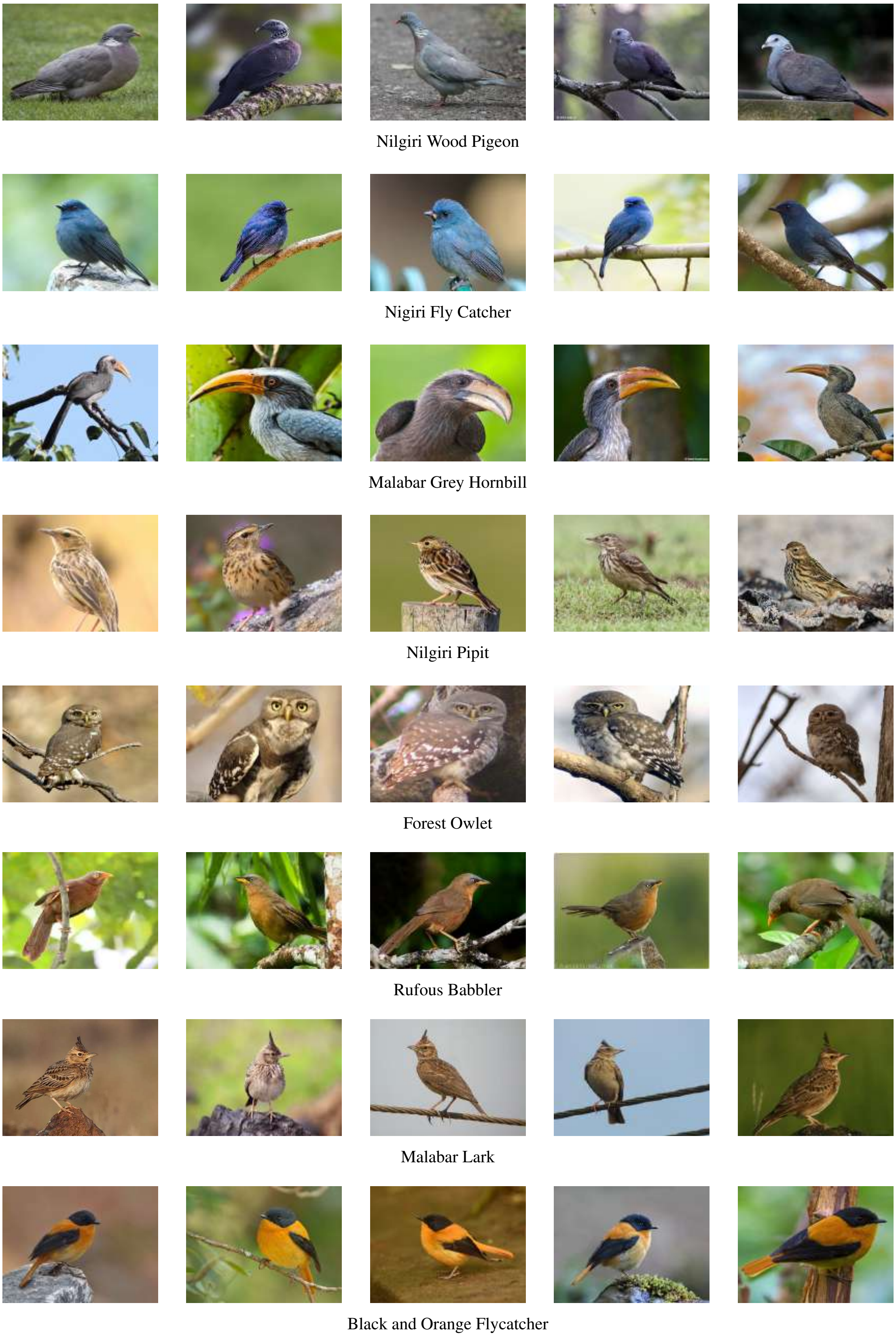
Erebidae

Geometridae

Hepialidae

Noctuidae

Figure 3.2: NZ Lepidoptera dataset:8 butterflies/moths;100 images per classs



Nilgiri Wood Pigeon

Nigiri Fly Catcher

Malabar Grey Hornbill

Nilgiri Pipit

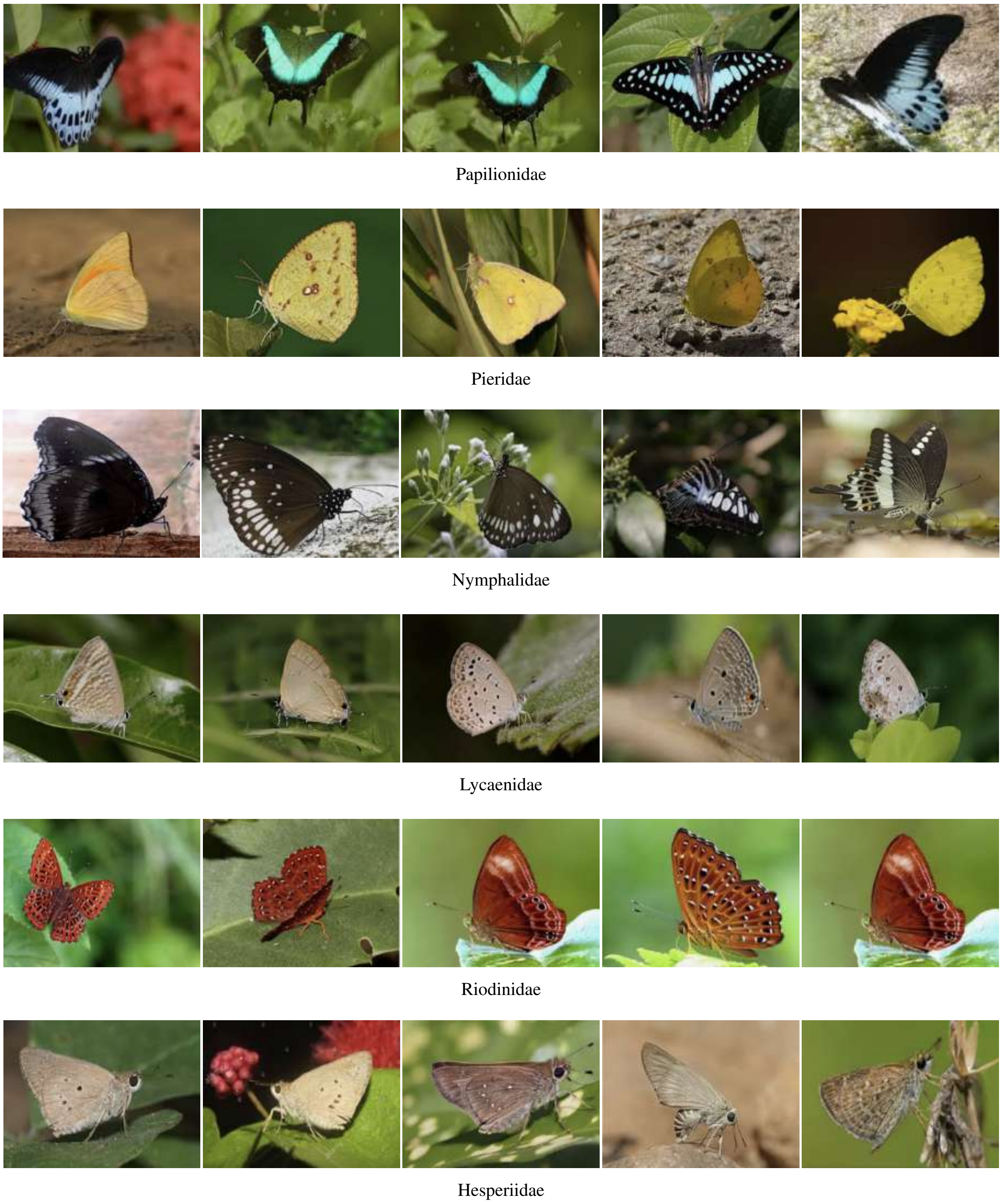
Forest Owlet

Rufous Babbler

Malabar Lark

Black and Orange Flycatcher

Figure 3.3: Indian Birds dataset: 8 out of 14 classes; 100 images per classs



Papilionidae

Pieridae

Nymphalidae

Lycaenidae

Riodinidae

Hesperiidae

Figure 3.4: Indian Butterflies dataset: 6 classes; 100 images per classs

Chapter 4

Background Suppression in Fine-grained Categories

4.1 Chapter Summary and Linkage

This thesis explores applications of collaborative filters to solve four major problems of fine-grained recognition: automated background compensation, encoding local repeated patterns for categorization, supervised collaborative classification, unsupervised collaborative clustering. The present chapter tackles the first of these problems, that of robust background compensation. Significant background variation across classes pose a significant confounding effect due to subtle differences in the foreground objects. Furthermore, if the objects of interest occupy a smaller non centrally aligned portion of the image compare to the background, then there is a tendency of the algorithms to learn the background, rather than the foreground. So robust background suppression is an important part of effective fine-grained recognition and this chapter covers that.

4.2 Introduction

In the previous chapter, it was shown, through some preliminary experiments, that collaborative representation classifiers (CRC) may be expected to effectively represent and categorize fine-grained image datasets. However it is known that the performance of collaborative representation classifiers degrades considerably when there is significant background which is randomised across classes (Chakraborti *et al.* (2016)). This may be found in such fine-grained recognition problems like species recognition with varying habitats. Many variations of CRC have been proposed but most, if not all, carry this drawback. One particular approach of overcoming this is to use majority voting by patches, where the background effect gets compensated if it is randomly distributed across classes (Zhu *et al.* (2012)). However, these methods still need to take into account several conditions like whether the test patch itself is an outlier, etc.

This chapter tries to overcome these challenges. Two new CRC methods are presented here which handle outlier background patches better than its predecessors. These are the generalised patch based CRC (GP-CRC) and the probabilistic patch based CRC (PProCRC). Background suppression is formulated into the main cost function, thus doing away with the need for initial pre-processing steps like detection/localisation (annotation, bounding box, cropping). A closed form analytic solution of the cost function is derived that is non-iterative and hence efficient.

4.3 Methods

The present work is an improvement on patch based CRC (PCRC) and probabilistic CRC (ProCRC), which have been discussed in Chapter 2. We point out the shortcomings of the existing methods and how the proposed modifications overcomes them. The two new CRC methods presented here are generalised patch based CRC (GP-CRC) and probabilistic patch based CRC (ProCRC).

4.3.1 Generalised Patch based CRC (GP-CRC)

GP-CRC addresses some of the inherent issues with the original patch based CRC formulation (PCRC). These have been touched upon in brief in the introduction, but here it is dealt in detail with concrete mathematical solutions. In essence, a more generalised and robust patch based CRC is developed.

Case 1: According to equation 4.1, in PCRC, the dictionary is extracted at the same corresponding position of the test patch, which means that the test patch is only represented as a weighted collaboration of the training patches at the exact same location in the images. This is quite a restricted formulation and in fact, unless the foreground object in the image is well centred, aligned and covers most of the image, this location matching of patches will rarely occur. This disadvantage can affect the recognition rate to quite an extent.

Solution: An augmented dictionary M is generated considering all patches across all classes. p_j is the weight matrix for the image patch y_j and hence Mp_j is the representation of y_j as a weighted collaboration of all training image patches.

Thus the modified cost function becomes:

$$\hat{p}_j = \arg\min_{p_j} (||y_j - Mp_j||_2^2 + \lambda||p_j||_2^2) \tag{4.1}$$

Other symbols have same meaning as in PCRC, with dimensions of corresponding matrices being adjusted appropriately.

Case 2: If the test sample(s) happens to contain background patches that are significantly different (new or rare) from the majority of patches in the training set, then there is a high probability of wrong classification.

Solution: To address this an additional term is added to the cost function. M_j is a sub-matrix of M consisting of all the training image patches (can be in a transformed domain) at the same location j of the incoming test patch y_j in the test image y. p_{jj} is a sub-matrix of p_j constructed similarly and hence $M_j p_{jj}$ is the representation of y_j as a weighted collaboration of all training image patches which are at the same position j with respect to their corresponding full images.

The representation of the current patch, $M_j p_{jj}$,is compared to the overall representation of the dataset Mp_j and the distance $||Mp_j - M_j p_{jj}||_2^2$ gives a cost metric of how different the patch is from the overall nature of the image dataset. This reduces the effect of outlier background patches on the system.

Thus the modified cost function becomes:

$$\hat{p}_j = \arg\min_{p_j} (||y_j - Mp_j||_2^2 + \lambda||p_j||_2^2 + \gamma||Mp_j - M_j p_{jj}||_2^2) \tag{4.2}$$

<u>**Case 3**</u>: If the images of the dataset are such that significant portions of the images are background, then many of the generated patches will only contain background, which may confound the system.

Solution: In general, the patch based collaborative representation framework is quite robust against this challenge, due to the very fact that since the test image is represented as a co-operation across all samples, it strengthens the larger number of recurring objects, that is the foreground, rather than the changing background which occurs in fewer patches. So it performs better than purely discriminative approaches. In our formulation, the term introduced for Case 2 also strengthens this aspect further.

Derivation of closed form solution: Let M'_j be a matrix of same dimension as that of M with only the values in M_j being the non-zero elements in M'_j. Then $M_j p_{jj}$ may be replaced by $M'_j p_j$, the two expressions being equivalent. Let $\bar{M}_j = M - M'_j$. So we have, $||Mp_j - M_j p_{jj}||_2^2 = ||Mp_j - M'_j p_j||_2^2 = ||(M - M'_j)p_j||_2^2 = ||\bar{M}_j p_j||_2^2 = \bar{M}_j^T \bar{M}_j ||p_j||_2^2$.

Thus equation 4.6 can be rewritten as

$$\hat{p}_j = \arg\min_{p_j} (||y_j - Mp_j||_2^2 + \lambda||p_j||_2^2 + \gamma\bar{M}_j^T \bar{M}_j ||p_j||_2^2) \tag{4.3}$$

Differentiating equation 4.7 with respect to p_j and equating to zero to find $\hat{p}_j$, we have

$$-M^T(y_j - Mp_j) + \lambda p_j + \gamma \bar{M}_j^T \bar{M}_j p_j = 0 \tag{4.4}$$

Hence,

$$\hat{p}_j = (M^T M + \gamma \bar{M}_j^T \bar{M}_j + \lambda I)^{-1} M^T y_j \tag{4.5}$$

The projection matrix $T_j = (M^T M + \gamma \bar{M}_j^T \bar{M}_j + \lambda I)^{-1} M^T$ can be computed offline and the solution $\hat{p}_j = T_j y_j$ can thus be computed efficiently in a closed form. The class of test sample is predicted similarly to PCRC, but M_{jk} being replaced by M_k in equation 4.3. Optimal values of γ and λ are found by gradient descent.

Summary. In earlier work only patches at same locations were compared, which only works well if the images are well aligned without much background variation. So, in the present work, instead of constructing M with patch features at same locations, an augmented M is constructed with features of all patches over all training images (in Case 1). This solution handles the case of misaligned foreground objects, but raises the chances of the representation learning the background. To compensate, the second term of the cost function (in Case 2) compares it to other location matched patches (Mj) in order to have a penalty if the query patch is too dissimilar to other patches at the same location. This is like regularization, trading off misaligned foreground objects with the risk of learning the background- all integrated in the cost function.

4.3.2 Probabilistic Patch based CRC (PProCRC)

In this section the new PProCRC method is described in detail along with some drawbacks of its predecessors, which it overcomes.

Drawbacks of earlier formulations. The proposed PProCRC method overcomes a couple of drawbacks of the PCRC and ProCRC methods, on which it is based. ProCRC gives a logical probabilistic framework to the CRC formulation, but suffers from the same drawback of most collaborative formulations, that of randomized background variation across fine-grained classes. Eg. in the case of sub-categorical species recognition, the collaborative filter helps to get a robust representation of the fine-grained classes, but these species classes often contain a wide range of background variation in habitat which may be repeated across classes, thus acting as a confounding factor for the inter-class collaborative representation.

PCRC and other patch based CRC methods tend to overcome the background challenge by having a majority voting based classification scheme as described before. This might compensate for the effect of background patches if they are in the minority or if the background patches are randomised across classes which is often the case. However, the patch based methods are prone to outliers that is if some images have rare backgrounds. Our patch based probabilistic formulation of collaborative representation overcomes these challenges as discussed below.

PProCRC Formulation. The PProCRC cost function is as a maximisation of the joint occurrence of three independent events that overcome the drawbacks of the earlier methods, while preserving the strengths of each. The main insight is that the predicted label of a patch (y_i) and the entire test image (Y) should be the same (that is equal to the label of one of the patches x of the training set to which it is the most similar in the collaborative space). This should be achieved under the condition that the patch y_i belongs to the test image Y (that is it is not an outlier patch) and that the training patch x also belongs to the training set and is not an outlier. An example of this can be a rare background patch which is not commonly repeated in the dataset, and hence is assigned low probability so as not to affect the voting outcome. These probabilities are modeled as Gaussians and separated into three independent events as follows.

1. Probability of a test patch having same label as one of the training patches, provided that training patch belongs to the training set (not an outlier) is given by:

$$P[l(y_i) = l(x) \mid x \in X].\ P[x \in X] = e^{-||y_i - X\alpha_i||} e^{-\lambda||\alpha_i||} \tag{4.6}$$

2. Probability of the test patch having the same label as the total test image, provided the test patch belongs to the test image (not an outlier) is given by:

$$P[l(y_i) = l(y) \mid y \in Y].\ P[y \in Y] = e^{-||y_i - Y\beta_i||} e^{-\gamma||\beta_i||} \tag{4.7}$$

3. Probability of the entire test image having the same label as the training patch (which has same label as test patch) is given by:

$$P[l(y) = l(x)] = e^{-||Y\beta_i - X\alpha_i||} \tag{4.8}$$

So the final cost function is given by the maximum of the joint occurrence of these 3 events as:

$$\max_{\alpha_i,\beta_i} [\exp(-||y_i - X\alpha_i|| - ||y_i - Y\beta_i|| - \lambda||\alpha_i|| - \gamma||\beta_i|| - ||Y\beta_i - X\alpha_i||)]$$

$$= \min_{\alpha_i,\beta_i} [-||y_i - X\alpha_i|| - ||y_i - Y\beta_i|| - \lambda||\alpha_i|| - \gamma||\beta_i|| - ||Y\beta_i - X\alpha_i||] \tag{4.9}$$

Next a closed form solution of the cost function is obtained as follows:

- Differentiating cost function with respect to α_i:

$$(2X^T X + \lambda I)\hat{\alpha}_i - X^T Y\hat{\beta}_i = X^T y_i \tag{4.10}$$

- Differentiating cost function with respect to β_i:

$$(2Y^T Y + \gamma I)\hat{\beta}_i - Y^T X\hat{\alpha}_i = Y^T y_i \tag{4.11}$$

Solving the simultaneous equations 13 and 14, the optimal values of $\hat{\alpha}_i$ and $\hat{\beta}_i$ are calculated, which are then used for the classification phase through patch majority voting as in the PCRC scheme.

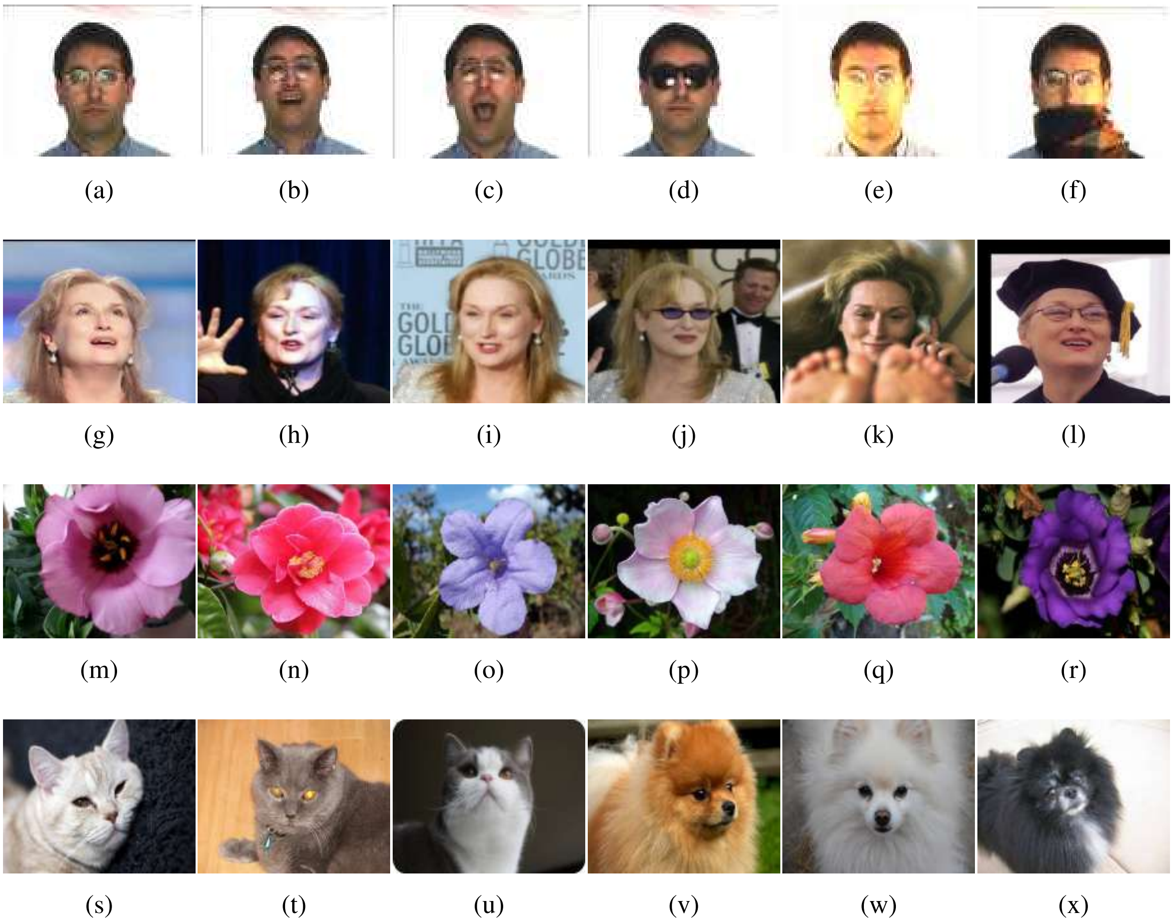


Figure 4.1: Sample images from the face recognition and species recognition datasets. (a)-(f) and (g)-(l) are images of one individual from AR dataset and LFW dataset respectively, showing variation in clothing, eyewear, headgear, other faces in image, varying background and partial covering of face. (m)-(r) are six images of different flowers from the Oxford 102 dataset showcasing the challenge of low inter-class variation; (s) to (x) are three images each of one breed/class of dogs and that of cats, showcasing the challenge of high intra-class variation.

4.4 Experiments and Results

4.4.1 Experimental Setup

Benchmark Datasets. The proposed method and its competitors have been evaluated on four fine-grained image datasets: two face recognition benchmarks (AR and LFW) and two species recognition benchmarks (Oxford Flowers and Oxford-IIIT Pets). Results on the new fine-grained species recognition dataset IndBirds, compiled as part of this doctoral project, is also presented here.

Face recognition benchmark datasets have been chosen due to ready availability of performance data of collaborative representation based classifiers in the existing literature. However, the major limitation of the published results of CRC methods applied to the face recognition problem is that the benchmark datasets used are early ones. Hence these have the foreground object (the human face) as the focus and covering most of the image, and hence are not representative of real life scenarios. The proposed CRC methods are seen to be more robust in performance compared to earlier variants when there is such background variation.

• *AR Face Dataset*: It was developed at the Ohio State University and contains more than 400 color face images of 126 people with changes in illumination, emotion and occlusion (Martinez *et al.* (1998)). For fair comparison to reported results using CRC methods, a subset of 50 male and 50 female subjects were chosen and the images are resized to 32×32 for the experiments.

• *LFW Face Dataset:* The Labeled Faces in the Wild (LFW) dataset (Huang *et al.* (2007)), compiled by the University of Massachusetts Amherst (UMass), contains unconstrained images of 5749 individuals in a natural setting. The original dataset presents challenging backgrounds, along with pose variation and partial occlusions. LFW-a (Taigman *et al.* (2009)) is a front aligned subset of it, which is used in some related works with other CRC methods. In LFW-a, 158 subjects with at least 10 sample images each are chosen and are resized to 121×121 pixels.

The AR dataset has front aligned faces with minimal background and pose variation. The LFW dataset contains human images from natural settings, but the experiments were performed on aligned and cropped version (LFW-a) of the original dataset. Further experiments on the original LFW dataset with images in the wild are performed, which demonstrate that the performance of most CRC methods degrade considerably in a natural setting. The decrease in accuracy for the proposed method is much less in comparison, which shows that it is more robust to background variation.

The problem of random background variation across classes is more significant in the case of fine-grained sub-categorical object recognition. Here the objects in different classes are quite similar visually and have only subtle differences. Collaborative representation may help to better utilise the similar foregrounds, but the diverse background can have a confounding effect. Species Recognition has been chosen as the representative problem of fine-grained classification, to showcase the superior performance of the proposed PProCRC under these conditions.

For species recognition, Oxford flowers dataset, Oxford-IIIT pets dataset and the Indian birds dataset (compiled as part of this thesis) have been used. Descriptions of these datasets are provided in Chapter 2 and 3.

Competing Classifiers:

Non-CRC classifiers. The performance of the proposed PProCRC method is compared with that of several competing classifiers, both CRC based as well as non-CRC based. Three popular modern non-CRC classifiers are chosen, namely support vector machines (SVM), adaptive boosting (AdaBoost) and random decision forests (RDF).

- *Support Vector Machines:* Multiclass categorization is performed with the binary SVM (Cortes *et al.* (1995)) classifier with χ^2 kernel in a one-versus-all fashion.

- *AdaBoost:* The weak classifier generates an optimised threshold in order to classify the data into two classes. The boosting part calls the classifier iteratively, and updates the weights of misclassified examples after each iteration. This creates a cascade of "weak classifiers" which act as a "strong classifier" in a repeated one-versus-all strategy (Freund *et al.* (1999)).

- *Random Decision Forest:* RDF (Ho (1999)) is an ensemble of Decision Trees. Since individual decision trees are prone to over-fitting, bootstrap aggregated (bagged) Classification and Regression Trees (CART) are used to achieve better generalisation.

CRC based classifiers. First those CRC classifiers are taken that are directly related to the formulation of the present method. As has been described in Section 4.2, these are the original CRC, patch based CRC (PCRC), generalized patch based CRC (GP-CRC) and probabilistic CRC (ProCRC). Besides these several other recent variations of CRC have been used like Enhanced CRC (ECRC) by (Liu *et al.* (2014)), Relaxed CRC (RCRC) by (Yang *et al.* (2012a)), Kernel CRC (CRC) (Zhao *et al.* (2014)), and the state-of-the-art Extended Probabilistic CRC (EProCRC) by (Lan *et al.* (2017)). These are described briefly below. Details of these may be found in Chapter 2.

Feature Descriptors. Two popular feature descriptors have been used: Dense SIFT and ensemble of GIST+HOG. But it should be noted that the proposed algorithm is general and is agnostic to feature choice.

- *Scale Invariant Feature Transform (SIFT):* is extracted and a patch size of 10 × 10 is chosen with overlap (Lowe (1999)).

- *Global Invariant Scale Transform (GIST):* is a global feature that describes the spatial envelope of the image using directional properties. It extracts dense multi-scale overlapping patches (Oliva *et al.* (2001)).
- *Histogram of Oriented Gradients (HOG):* features are extracted in a dense grid fashion in 3×3 cells which are concatenated at each grid location to generate the descriptor (Dalal *et al.* (2005)).

  The Dense SIFT features are used separately, while the GIST and HOG features are used as a concatenated ensemble.

4.4.2 Hyper-parameter Tuning

The hyper-parameters are tuned using stock implementations of optimization algorithms like genetic algorithms and simulated annealing during cross-validation.

- For SVM, the kernel parameter gamma and the regularization parameter C are tuned.
- For Adaboost, the key is to strike a good balance between learning rate and number of weak classifiers used. Lesser learning rate is better but more the number of classifiers needed. We started with a learning rate 0.01 at first and then settled with 0.001 for the final model.
- For Random Forest, the main approach was to set the number of trees high (10000) and then run a series of values for maximum depth parameter in an array to find the optimal value.
- For CRC based methods, the regularization parameter lambda is tuned starting with a random number between 0 and 1.

Patch sizes of 16x16 were used. For example, in the case of SIFT features, 128-dimensional SIFT descriptors were computed over 16x16 pixel patches, sampled densely over a grid with a regular spacing of 8 pixels in both the horizontal and vertical directions. Of course computation time would be n times that speed per image, if the number of patches is n per image.

4.4.3 Results and Analysis

Results. For the AR dataset, images have been resized to 32×32, while for the LFW-a benchmark the aligned face images are cropped to 121×121 to eliminate background and then resized to 32×32. These conditions are maintained to have parity with reported results in earlier articles on CRC. Experiments are also carried out on the original LFW images with just resizing to 121×121. For all the three species recognition input images are resized to 121×121, preserving the original background. A patch size of 10×10 is used for both the face recognition and species recognition datasets. For each dataset, experiments are conducted with 5 fold cross validation and percentage classification accuracies along with standard deviation are presented in Table 4.1 (face recognition) and Table 4.2 (species recognition) with the highest accuracy in each column highlighted in bold.

Among the CRC-based methods, the basic CRC has the least accuracy and then there is a consistent increase in the performance of the CRC variants. The proposed Probabilistic patch based CRC (PProCRC) comfortably outperforms all the competing CRC methods including the two that it is based on, that is the original patch based CRC (PCRC) and the probabilistic CRC (ProCRC). It also has some improvement in performance over the state-of-the-art enhanced probabilistic CRC (EProCRC). Compared to the non-CRC methods, PProCRC has significantly better results than all three, SVM, AdaBoost and RDF. These inferences are consistent for both tasks (face recognition and species recognition) and across the 6 datasets and 2 features. The other proposed method GP-CRC also outperfoms the competitors except the very recent EProCRC.

It may be further observed that there is a significant degradation in performance on the original LFW compared to the less challenging LFW-a. The results on the original LFW images preserve background, with only resizing of the entire image. LFW-a has the images from LFW, but cropped to exclude background and the faces are aligned to front, and in grayscale. It is seen that the proposed PProCRC has the least deterioration in accuracy between LFW and LFW-a among all the competing classifiers. This demonstrates the robustness of the proposed methods to changes in background and pose.

Evaluating Robustness of PProCRC on IndBirds dataset.

Effect of Feature Choice: Fine-grained species recognition datasets can have significant variances in foreground and background color compared to similar object contour (if pose is same). The effect of color as a confounding attribute is illustrated in Fig. 4.2 using examples from the new IndBirds dataset. Fig. 4.2(a) and 4.2(b) are of Nilgiri flycatcher (blue colored) and Black-yellow flycatcher. Note that the shape of the birds are almost identical, only major difference being the color. Experiments on the IndBirds dataset are run using Dense SIFT and GIST+HOG features for PProCRC, EProCRC and ProCRC classifiers with color images and grayscale images.

The results are presented in Table 4.3 and it may be observed that in presence of color information, the performance improves for both features and for all classifiers. However, the percentage degradation in performance for PProCRC is less than that of EProCRC and ProCRC. This example brings to light the importance of feature choice for classifier performance and thus strengthens the case of integrating the collaborative representation into a CNN based feature learning framework. This has been presented in Chapter 6.

Effect of Image Quality: Image quality can also play an important role in feature-classifier performance. This is pertinent for fine-grained visual categorization problems involving species recognition, because in many cases the images acquired in natural settings can be of poor quality. There may be motion blur, poor illumination, partial obscuring in the wild, etc. The effect

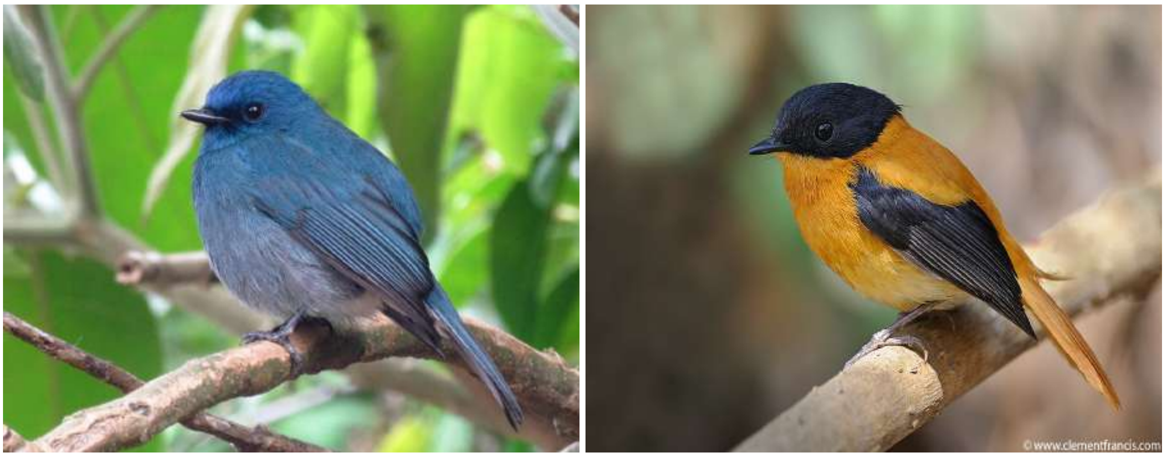
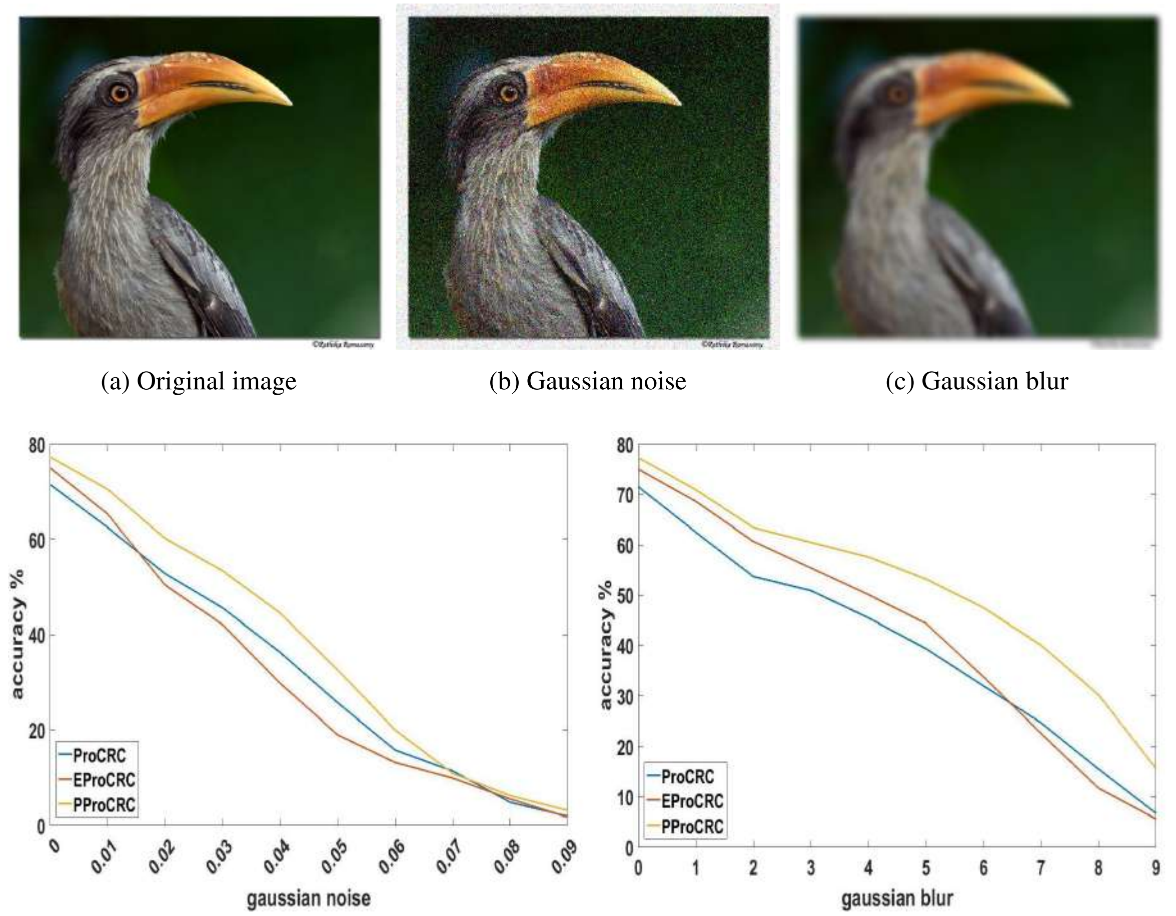


Figure 4.2: Example from IndBirds dataset with similar appearance with only color difference: (a) Nilgiri flycatcher (Blue), (b) Black-yellow flycatcher

of deterioration in image quality is evaluated on the IndBirds dataset as follows. *Gaussian Noise* is added to each color component of each image pixel separately. The standard deviation is varied from 0 to 0.09 in steps of 0.01. *Blur* is imposed with a Gaussian kernel of standard deviation ranging from 0 to 9 in steps of 1. The size of the filter window is set to 4 times the standard deviation.

The effects of these on the performance of ProCRC, EProCRC and PProCRC is graphically represented in Fig. 4.3. Fig. 4.3(a) is the original image of a Malabar Gray Hornbill and Fig.4.3 (b) and (c) are respectively the same image with added Gaussian noise (standard deviation = 0.02) and Gaussian blur (standard deviation = 4). Fig. 4.3(d) and (e) trace the fall of accuracy of ProCRC, EProCRC and PProcRC with increase in Gaussian noise and blur respectively using Dense SIFT features. It is seen that performance of all 3 classifiers falls off as expected, but PProCRC gives the most robust performance among the three. Similar trends are noticed for GIST+HOG features.

Statistical Analysis. To test the statistical significance of the improvement in performance of the proposed PProCRC over its competitors, statistical tests are conducted.



(a) Original image

(b) Gaussian noise

(c) Gaussian blur

(d) accuracy vs. noise

(e) accuracy vs. blur

Figure 4.3: Effect of noise and blur on classification accuracy in NZBirds.v2 dataset

Wilcoxon signed rank test is performed to compare the performance between PProCRC and EProCRC. Wilcoxon test is more generalised, since it does not assume a normal distribution of data unlike the paired/unpaired t-tests. This is more applicable to compare results across several classifiers, features and datasets simultaneously, but for one particular task at a time (in our case face recognition and species recognition separately). First the test is conducted on the accuracy results for the face recognition tasks as presented in Table 4.4. The ranks (*R*) are allocated according to the magnitude of difference in accuracy between the two methods. If there is a tie in the absolute difference, then the rank is split between the two. Eg. if there is tie for the values for 3rd and 4th rank, then both are given 3.5 rank. The corresponding signs (S) are allocated depending on which method outperforms for that particular experimental setting. The ones for which PProCRC is better have sign 1; the rest have sign -1. The Wilcoxon parameter $W = \sum SR$ is calculated. For the face recognition task $W = 17$. The same calculations are performed for the species recognition task and presented in Table 4.5 and $W = 19$. For each task, maximum possible rank value for $n = 6$ experiments (combination of 2 features and 3 datasets) is $n(n + 1)/2 = 21$. The Wilcoxon signed rank test states that the null hypothesis (PProCRC and EProCRC are equally good) may be rejected in one-direction (PProCRC better than EProCRC) at 5% level of significance if $W \geq 17$. Hence for both tasks (face and species recognition) it may be concluded that the proposed PProCRC performs significantly better than the state-of-the-art EProCRC.

Signed binomial test is next carried out between PProCRC and EProCRC, since it can be used across different tasks simultaneously because it considers frequency of success in the calculations rather than the accuracy values. Again the null hypothesis is that the two are equally good, that is there is 50% chance of each beating the other on any particular trial. Now over the 2 tasks (face recognition and species recognition) there are 6 datasets, 2 descriptors each (Dense SIFT and GIST+HOG) and 10-fold cross-validated results. Thus in total we have 100 experiments, and out of these PProCRC outperformed EProCRC 70 times (that is 70% of the trials). The signed binomial test yields that given the assumption that both methods are equally good, then the probability of PProCRC outperforming EProCRC in 70% of the trials is 0.13% (one-tail p-value of 0.0013). The two-tail p-value is 0.0027. Considering a level of significance of $\alpha = 0.05$, one has to apply the Bonferroni adjustment. Here there are 2 descriptors and 6 datasets, hence 12 combinations of experimental condition. So the 5% level of significance is divided by 12 to get adjusted $\alpha = 0.0042$. Since the one-tail and two-tail p-values obtained are both less than 0.0042, it may be concluded that though improvement in mean accuracy of GP-CRC over ProCRC is marginal, it is still statistically significant considering the frequency of out-performance.

4.5 Discussion

Two new collaborative representation classifiers were presented in this chapter. The first one (GP-CRC) is a generalised patch based CRC that achieves better background suppression. This is then extended into a probabilistic patch based CRC (PProCRC) that outperforms the state-of-the-art (EProCRC). Experiments have been performed on the species recognition problem and a new benchmark fine-grained bird image dataset (IndBirds) has been introduced.

Table 4.1: Face Recognition Accuracy (%)

	AR Face Dataset		LFW Face Dataset		LFW-a Face Data	
	DenseSIFT	GIST/HoG	DenseSIFT	GIST/HoG	DenseSIFT	GIST/HoG
SVM	90.4 ± 5.1	88.5 ± 6.4	31.4 ± 3.0	25.0 ± 3.6	45.3 ± 2.6	41.7 ± 2.9
AdaBoost	91.8 ± 5.6	91.1 ± 5.9	33.1 ± 3.3	26.6 ± 3.1	47.7 ± 2.8	42.2 ± 2.7
RDF	91.5 ± 5.3	90.6 ± 6.0	33.6 ± 3.4	26.2 ± 3.2	48.1 ± 2.5	42.5 ± 2.8
CRC	91.2 ± 5.2	90.0 ± 6.1	31.5 ± 3.1	25.6 ± 3.8	44.9 ± 2.5	42.1 ± 3.0
ECRC	92.4 ± 5.0	91.6 ± 6.5	33.4 ± 3.5	26.3 ± 3.0	47.6 ± 2.7	42.5 ± 2.8
PCRC	93.5 ± 5.5	92.3 ± 6.3	33.6 ± 3.3	27.5 ± 3.4	48.3 ± 2.6	42.9 ± 2.5
RCRC	94.1 ± 5.3	92.8 ± 6.4	35.0 ± 3.2	29.9 ± 3.7	49.9 ± 2.4	43.2 ± 2.7
KCRC	94.6 ± 5.7	93.3 ± 6.6	35.3 ± 3.1	30.4 ± 3.3	50.7 ± 2.5	43.4 ± 3.0
ProCRC	95.0 ± 5.1	93.9 ± 5.8	38.7 ± 3.0	35.0 ± 3.5	52.6 ± 2.9	44.4 ± 3.0
GP-CRC	95.8 ± 5.6	94.2 ± 5.3	39.9 ± 3.2	38.5 ± 3.3	53.1 ± 2.7	45.0 ± 2.6
EProCRC	96.7 ± 3.0	95.7 ± 3.5	48.5 ± 3.6	43.2 ± 3.5	**56.7** ± 2.4	47.1 ± 2.8
PProCRC	**97.5** ± 2.3	**96.2** ± 3.8	**50.2** ± 3.4	**44.3** ± 3.7	56.0 ± 2.6	**47.9** ± 2.5

Table 4.2: Species Recognition Accuracy (%)

	Oxford Flowers		Oxford-IIIT Pets		Indian Birds	
	DenseSIFT	GIST/HoG	DenseSIFT	GIST/HoG	DenseSIFT	GIST/HoG
SVM	64.4 ± 5.3	61.2 ± 5.5	60.2 ± 4.6	56.4 ± 4.3	66.7 ± 4.8	65.5 ± 5.0
AdaBoost	64.0 ± 5.7	63.1 ± 5.2	61.7 ± 4.2	58.5 ± 4.5	67.0 ± 4.7	65.9 ± 4.9
RDF	64.9 ± 5.4	63.6 ± 5.6	62.0 ± 4.2	58.1 ± 4.6	67.1 ± 5.1	66.3 ± 5.2
CRC	64.6 ± 5.9	60.5 ± 5.7	61.5 ± 4.8	56.3 ± 4.4	67.3 ± 4.9	66.0 ± 4.8
ECRC	66.1 ± 5.5	63.3 ± 5.4	62.2 ± 4.1	58.6 ± 4.7	67.9 ± 5.0	66.8 ± 5.1
PCRC	66.3 ± 5.2	65.8 ± 5.1	62.8 ± 4.5	60.5 ± 4.6	68.1 ± 5.2	67.2 ± 4.9
RCRC	68.8 ± 5.1	67.4 ± 5.3	64.1 ± 4.6	61.9 ± 4.5	69.4 ± 4.7	67.9 ± 5.0
KCRC	69.0 ± 5.3	67.7 ± 5.5	64.5 ± 4.5	62.2 ± 4.9	69.8 ± 4.9	68.0 ± 4.8
ProCRC	72.3 ± 5.7	69.4 ± 5.9	67.7 ± 4.9	63.6 ± 4.3	71.5 ± 5.0	69.7 ± 5.1
GP-CRC	73.7 ± 5.8	71.7 ± 5.8	68.9 ± 4.7	64.2 ± 4.8	72.3 ± 5.1	70.1 ± 4.9
EProCRC	78.1 ± 5.6	**77.7** ± 5.4	73.6 ± 4.5	68.4 ± 4.7	75.0 ± 5.2	72.4 ± 4.8
PProCRC	**79.9** ± 5.5	77.0 ± 5.2	**75.1** ± 4.4	**69.6** ± 4.3	**77.3** ± 5.0	**74.8** ± 4.7

Table 4.3: Effect of Color (IndBirds dataset)

	Color Images		Grayscale Images		Percentage Reduction	
	DenseSIFT	GIST+HOG	DenseSIFT	GIST+HOG	DenseSIFT	GIST+HOG
ProCRC	71.5	69.7	68.1	67.4	4.8	3.3
EProCRC	75.0	72.4	72.9	70.5	2.8	2.6
PProCRC	77.3	74.8	76.6	74.2	0.9	0.8

Table 4.4: Face Recognition Wilcoxon Test

	Oxford Flowers		Oxford-IIIT Pets		Indian Birds	
	DenseSIFT	GIST+HOG	DenseSIFT	GIST+HOG	DenseSIFT	GIST+HOG
EProCRC	96.7	95.7	48.5	43.2	56.7	47.1
PProCRC	97.5	96.2	50.2	44.3	56.0	47.9
$\mid$ Difference $\mid$	0.8	0.5	1.7	1.1	0.7	0.8
Rank (R)	3.5	1	6	5	2	3.5
Sign (S)	1	1	1	1	-1	1

Table 4.5: Species Recognition Wilcoxon Test

	Oxford Flowers		Oxford-IIIT Pets		Indian Birds	
	DenseSIFT	GIST+HOG	DenseSIFT	GIST+HOG	DenseSIFT	GIST+HOG
EProCRC	78.1	77.7	73.6	68.4	75.0	72.4
PProCRC	79.9	77.0	75.1	69.6	77.3	74.8
| Difference |	1.8	0.7	1.5	1.2	2.3	2.4
Rank (R)	4	1	3	2	5	6
Sign (S)	1	-1	1	1	1	1

Chapter 5

Encoding Repeated Local Fine-grained Patterns

5.1 Chapter Summary and Linkage

This thesis explores the suitability of collaborative representation classifiers (CRC) for fine-grained recognition. The previous chapter overcame one major problem of existing CRC methods, that of lack of robustness in case of background variation. This was a significant step since many fine-grained visual categorization (FGVC) problems suffer from significant randomized background variation. The next step is to focus on representation of fine-grained objects themselves, before feeding them into the CRC for classification. To do this, we develop in this chapter an improved way to represent repeated local patterns which are characteristic of many fine-grained recognition problems, specially species recognition like those of flowers, butterflies, etc.

5.2 Introduction

Local binary descriptors have been shown to be effective encoders of repeated local patterns for robust discrimination in several visual recognition tasks (Huang *et al.* (2011)). He *et al.* (1990) in their seminal paper on the subject, introduced the concept of textons. These encode localized textures/patterns in an image into binary words, and the frequency histogram of these words describes the image. The first popular implementation was local binary pattern (LBP) (Ojala *et al.* (1994)). Since then, many interesting modifications and improvements of these descriptors have been developed. A few of these are modified census transform (MCT) (Frobaand *et al.* (2004)), local gradient pattern (LGP) (Jun *et al.* (2012)), local directional pattern (LDP) (Jabid *et al.* (2010a)), uniform local binary pattern (ULBP) (Ming *et al.* (2015)), etc. LBP encodes the local intensity variation in the neighborhood of each image pixel into a binary word, the decimal equivalent of which then acts as a representative feature encapsulating the pattern of local intensity variation in that neighborhood. The histogram of the LBP values over all pixels of an image serves as the descriptor for that image. MCT (Frobaand *et al.* (2004)) is a modified version of LBP, where instead of taking the central pixel intensity of the neighborhood as the threshold, the average intensity over the entire neighborhood is treated as the threshold while forming the binary word. LGP (Jun *et al.* (2012)) considers the gradient of neighboring pixel intensities with respect to the central pixel intensity to generate the binary word. Both MCT and LGP were demonstrated to be more robust to noise than the original LBP operator. LDP was proposed (Jabid *et al.* (2010a)) as a more noise invariant alternative to LBP. It incorporates the outputs of the 3×3 Kirsch masks in 8 directions, for each pixel, into a binary word and its corresponding decimal value is the desired LDP value, corresponding to that pixel.

An inherent shortcoming of both LBP and LDP, as well as many of the other related descriptors like MCT, LGP, etc., is that each assigns an ad hoc bias while sequencing the generated binary digits to form the binary word. This is because the choice of the initial pixel of this circular sequence of bits determines the binary-to-decimal conversion weight assigned to each of the binary digits, as explained later in the chapter. This adds an inherent rotation variance. A few rotation invariant versions have also been suggested in existing literature, like LDP^{ri} (Jabid *et al.* (2010b)) for LDP. But these have their own set of limitations and the present work presents an alternative rotation invariant scheme.

The main contributions of this chapter is a new binary local pattern, the LOOP descriptor. Rotation invariance is added into the main formulation of LOOP, thus overcoming a drawback of most existing descriptors of the genre. In the process we decrease post processing time complexity and increase accuracy of classification. Note that although the efficacy of the proposed method is demonstrated here on butterfly and moth species recognition, the methodology may be applied to other similar applications exhibiting repeated local patterns/textures.

5.3 Methods

In this section we give a brief description of two popular binary descriptors: local binary patterns (LBP) and local directional patterns (LDP). These two are the methods on which the proposed LOOP descriptor is directly based. Then we move on to describe the LOOP descriptor in detail along with how it overcomes the drawbacks of LBP and LDP, both mathematically and visually through illustrative examples. We also briefly mention the rotation invariance of LOOP.

5.3.1 Local Binary Pattern (LBP)

LBP (Ojala *et al.* (1994)) is a popular descriptor which captures the local intensity variation patterns of an image and has good discrimination characteristics.

Let i_c be the intensity of an image I at pixel (x_c, y_c) and i_n $(n = 0, \ldots, 7)$ be the intensity of a pixel in the 3×3 neighborhood of (x_c, y_c) excluding the center pixel i_c.

Then the LBP value for the pixel (x_c, y_c) is given by

$$LBP(x_c, y_c) = \sum_{n=0}^{7} s(i_n - i_c).2^n \tag{5.1}$$

where

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

A major disadvantage of LBP is the arbitrary sequence of binarization weights. Depending on the chosen starting pixel of the sequence of binary weights $(2^n, n = 0, \ldots, 7)$, the 8 neighbors of the output 3×3 grid are allocated subsequent weightage n sequentially. There is no clear logic behind the proper assignment of weight and the result obtained is susceptible to rotation variance. The same pattern rotated between images of the same class or even within different parts of the same image will generate a different binary word, thereby confounding the classification process. In fact, this bias has persisted over to other related descriptors as well, like LDP, MCT, LGP, etc.

5.3.2 Local Directional Pattern (LDP)

LDP is an improved local pattern descriptor which incorporates a directional component by using Kirsch compass kernels. It was shown to be less susceptible to noise than the traditional LBP operator (Jabid *et al.* (2010a)).

Let i_c be the intensity of an image I at pixel (x_c, y_c) and i_n , $n = 0, 1, \ldots, 7$ be the intensity of a pixel in the 3×3 neighborhood of (x_c, y_c) excluding the center pixel i_c. 3×3 Kirsch edge detectors centered at (x_c, y_c) in eight possible directions are given in Fig. 5.1.

The 8 responses of the Kirsch masks are m_n, $n = 0, \ldots, 7$ corresponding to pixels with intensity i_n, $n = 0, \ldots, 7$ and let m_k be the k^{th} highest Kirsch activation. Then all the neighboring pixels having Kirsch response higher than m_k are assigned 1, and others 0.

But the empirically assigned value of k is *ad hoc*. This fixes the possible number of ones to $k - 1$ and number of zeros to $(n + 1 - (k - 1) = n - k + 2$ where n is as defined above by the neighborhood pixel number. Hence the possible number of binary words is reduced from $2^{(n+1)}$ to C^{n+1}_{k-1}.

Then the LDP value for the pixel (x_c, y_c) is given by

$$LDP_k(x_c, y_c) = \sum_{n=0}^{7} s(m_n - m_k).2^n \tag{5.3}$$

where

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.4}$$

Rotation Invariant LDP (LDP^{ri}): A rotation invariant version was introduced in (Jabid *et al.* (2010b)). Here the neighbor pixel with highest Kirsch mask output is assigned the highest order in the binary word, and then the other bits are taken sequentially as in previous formulations. Thus it assigns an empirical rule to the starting point of the binary word construction. However, it suffers from the self-imposed restriction of always having a leading 1, which immediately reduces the number of available combinations in the binary word by half. The problem of fixed number of 1s and 0s also persists from the original LDP, depending on the value of the threshold k.

5.3.3 Local Optimal Oriented Pattern (LOOP)

As discussed earlier, the major disadvantage of LBP and LDP is the arbitrary sequence of binarization weights that adds dependancy to orientation. LDP also suffers from the empirical assignment of value to the threshold variable, which puts an *ad hoc* restriction on the number of bits allowed to be 1, thus reducing the number of possible words, as discussed before. LOOP presents a non-linear amalgamation of LBP and LDP that overcomes these drawbacks while preserving the strengths of each.

Let i_c be the intensity of an image I at pixel (x_c, y_c) and i_n $(n = 0, 1, \ldots, 7)$ be the intensity of a pixel in the 3×3 neighborhood of (x_c, y_c) excluding the center pixel i_c. The 8 Kirsch masks, as used in LDP previously, are oriented in the direction of these 8 neighboring pixels i_n $(n = 0, 1, \ldots, 7)$ thus giving a measure of the strength of intensity variation in those directions, respectively (refer to Fig. 5.1).

This information is incorporated by assigning the binarization weight to each neighboring pixel corresponding to the strength of Kirsch output in the direction of that pixel. The underlying rationale behind this approach is that the Kirsch mask output in a particular direction provides an indication of the probability of occurrence of an edge in that direction. Since the LBP indicates the intensity variation over the neighboring pixels in the same directions, the value of the Kirsch output is employed to assign the decimal-to-binary weights.

As discussed earlier, the 8 responses of the Kirsch masks are m_n corresponding to pixels with intensity i_n, $n = 0, \ldots, 7$. Each of these pixels are assigned an exponential w_n (a digit between 0 and 7) according to the rank of the magnitude of m_n among the 8 Kirsch mask outputs.

Then the LOOP value for the pixel (x_c, y_c) is given by

$$LOOP(x_c, y_c) = \sum_{n=0}^{7} s(i_n - i_c).2^{w_n} \tag{5.5}$$

where

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

Thus the LOOP descriptor encodes rotation invariance into the main formulation. Moreover, the proposed LOOP algorithm also negates the empirical assignment of the value of the parameter k in the traditional LDP method (eqn. 5.3).

5.3.4 Scale and Rotation Invariance

A multi-scaled amalgamated histogram is constructed to achieve scale independence. This is done by forming a spatial gaussian pyramid and then concatenating the histograms of LOOP values obtained at each scale to form the final histogram which acts as the descriptor for the image.

Figure 5.1 illustrates the rotation invariance property of LOOP descriptor, compared to the lack thereof in LBP and LDP. As demonstrated in Figure 5.1, binary words are formed according to LBP rule and the weights are assigned according to the LDP mask activations.

Tie break in weight assignment: Referring to Figure 5.1, the one with more differing nearest neighbour is assigned higher weight. Eg. For the two -2155 for pattern 2, nearest neighbours for one are -2155 and -275 (difference is 1880) while nearest neighbours for the other are -2155 and -2035 (difference is 120). So the former is assigned higher weight than latter (2^1 vs. 2^0).

5.4 Experiments and Results

5.4.1 Experimental Setup

Datasets. Three image datasets of moth and butterfly species have been used to showcase the performance of the methods. Among these the Leeds and Ponce butterfly datasets are pre-existing and have been discussed in Chapter 2 (few sample images provided here in Fig. 5.2). The NZ lepidoptera dataset

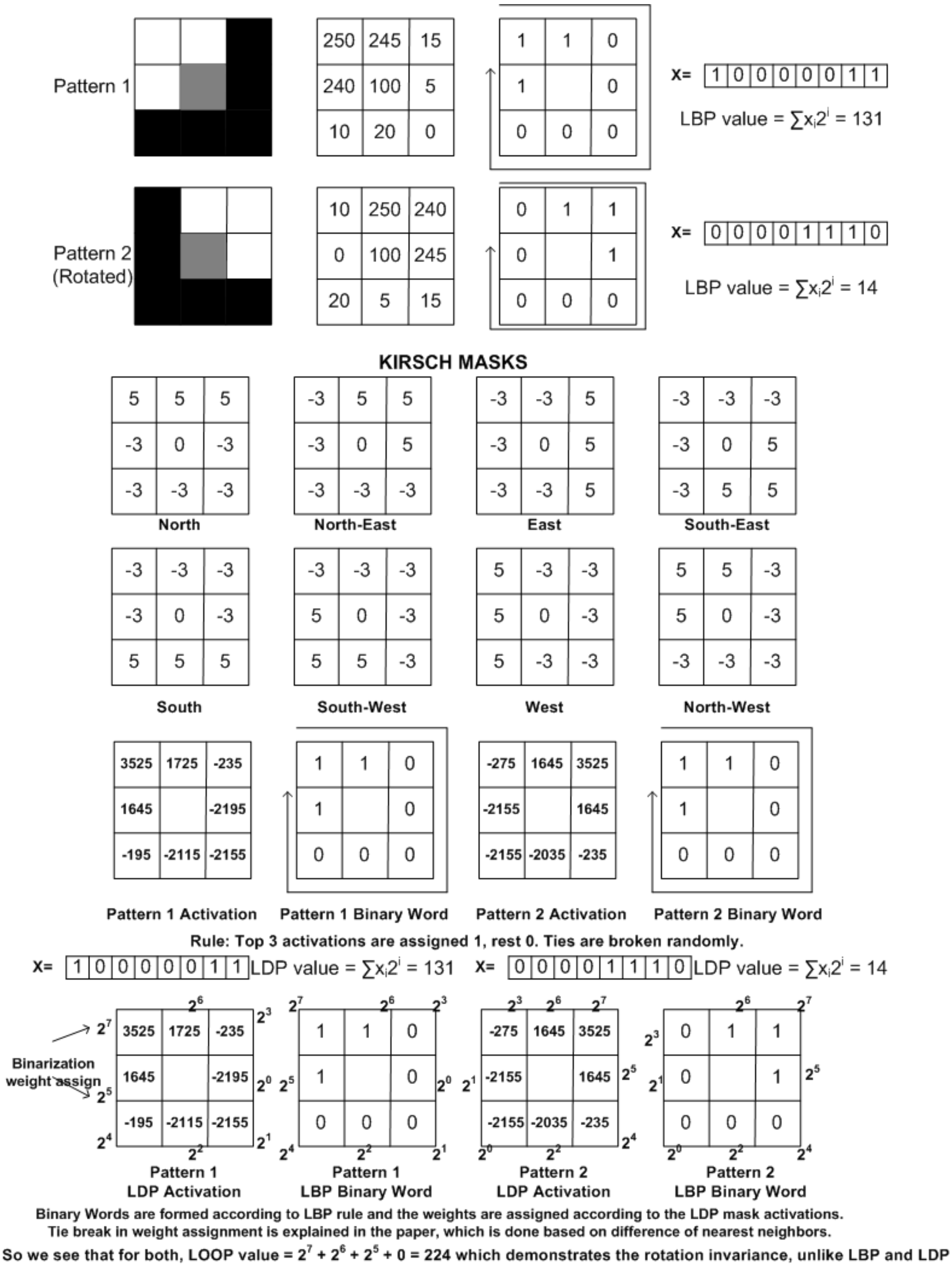

Pattern 1
Pattern 2
(Rotated)
X= 1 0 0 0 0 0 1 1
LBP value = ∑xᵢ2ⁱ = 131
X= 0 0 0 0 1 1 1 0
LBP value = ∑xᵢ2ⁱ = 14
KIRSCH MASKS
North
North-East
East
South-East
South
South-West
West
North-West
Pattern 1 Activation
Pattern 1 Binary Word
Pattern 2 Activation
Pattern 2 Binary Word
Rule: Top 3 activations are assigned 1, rest 0. Ties are broken randomly.
X= 1 0 0 0 0 0 1 1 LDP value = ∑xᵢ2ⁱ = 131
X= 0 0 0 0 1 1 1 0 LDP value = ∑xᵢ2ⁱ = 14
Binarization weight assign
Pattern 1 LDP Activation
Pattern 1 LBP Binary Word
Pattern 2 LDP Activation
Pattern 2 LBP Binary Word
Binary Words are formed according to LBP rule and the weights are assigned according to the LDP mask activations.
Tie break in weight assignment is explained in the paper, which is done based on difference of nearest neighbors.
So we see that for both, LOOP value = 2⁷ + 2⁶ + 2⁵ + 0 = 224 which demonstrates the rotation invariance, unlike LBP and LDP

Figure 5.1: Numerical Example to calculate LBP, LDP and LOOP.

has been collected as part of this thesis and has been discussed in details in Chapter 3.

Classifiers. The collaborative representation classifier (CRC) (Zhang *et al.* (2011)) has been shown recently to be effective in handling small fine-

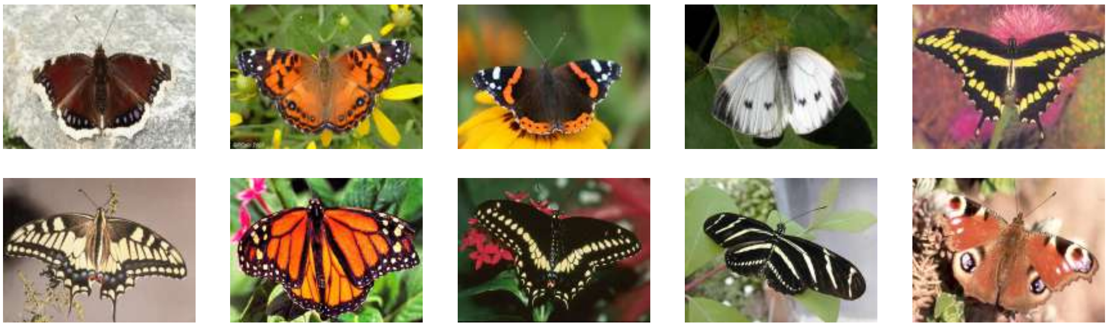
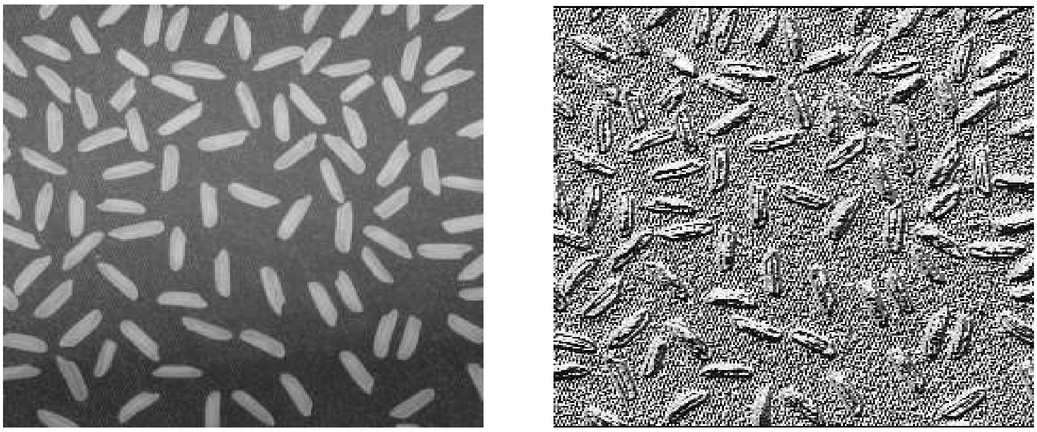


Figure 5.2: Sample images from the Leeds butterfly dataset (top row) and the Ponce butterfly dataset (bottom row). Images from the new NZ Lepidoptera dataset are provided in more details in Chapter 3.

grained datasets where the differences in objects between classes is subtle compared to randomized significant background variation within and between classes (Chakraborti *et al.* (2016)). CRC represents the test image as an optimal weighted average of training images of all classes and the subsequent residual per class is used to calculate the predicted category. A recent Probabilistic formulation of CRC called ProCRC (Cai *et al.* (2016)) is used here.

For comparison with a popular off-the-shelf classifier, a support vector machine (SVM) (Cortes *et al.* (1995)) with a χ^2 kernel is chosen with settings as in the ProCRC paper (Cai *et al.* (2016)) for fair comparison. Multiclass categorization is performed with the binary SVM classifier in a one-versus-all fashion in turns.

Descriptors. The performance of several local binary pattern encoders like LBP, MCT, LDP, LDP^{ri}, LGP is compared with the proposed LOOP descriptor. Among these, LOOP is built influenced by LBP and LDP, while MCT and LGP are chosen as relevant modifications of these methods. LDP^{ri} is a rotation invariant modification of LDP. It is also compared with three popular modern binary descriptors: BRIEF (Binary Robust Independent Elementary Features) (Calonder *et al.* (2010)), BRISK (Binary Robust Invariant



(a) Sample image

(b) LOOP output

Figure 5.3: Standard test image (rice.png) and the LOOP output.

Scalable Keypoints) (Leutenegger *et al.* (2011)), and ORB (Oriented FAST and rotated BRIEF) (Rublee *et al.* (2011)). BRIEF was the first of these and presents a simple configuration similar to LBP, without rotation invariance. ORB descriptor is rotation invariant and uses an optimal sampling pair. BRISK has both of these attributes and also has the additional characteristic of a more advanced hand-crafted sampling pattern composed of concentric rings.

5.4.2 Results and Analysis

For each dataset, experiments are conducted with 5 fold cross validation and percentage classification accuracies are presented in Table 1 with the highest accuracy in each column highlighted in bold. The LOOP output on a standard test image 'rice.png' is presented in Figure 5.3 as an illustrative example.

It is observed that LOOP performs much better than LBP and LDP, the two descriptors on which it is based. Performance of LOOP is also better than the competing LBP variants: MCT, LGP and LDP^{ri}. LOOP successfully outperforms BRIEF, but has comparable results with BRISK and ORB, both of which are rotation invariant. However, LOOP has a simpler formulation and lower execution time than ORB and BRISK. Using the standard tic-toc functionalities of MATLAB, it is observed that ORB and BRISK have 21% and 27% higher computation time respectively than LOOP.

Table 5.1: Classification Accuracy (%)

	Leeds		**Ponce**		**NZ**	
	SVM	**ProCRC**	**SVM**	**ProCRC**	**SVM**	**ProCRC**
LBP	62.1	64.7	68.4	70.5	55.6	59.3
MCT	63.9	65.8	69.9	72.3	58.7	61.1
LDP	66.6	68.5	71.7	74.1	60.9	64.5
LDP^{ri}	69.2	72.5	75.1	77.9	64.8	68.3
LGP	69.4	72.9	75.0	77.6	64.2	68.6
BRIEF	65.5	67.1	70.4	73.6	59.9	63.0
BRISK	69.8	73.5	77.7	79.3	**66.2**	69.5
ORB	71.0	73.8	**78.4**	79.9	65.8	70.1
LOOP	**71.5**	**74.4**	78.3	**80.4**	66.0	**70.6**

Among the competing classifiers listed in Table 5.1, only LDP^{ri}, BRISK and ORB are rotation invariant. LOOP yields comparable results to BRISK and ORB, but has lower run time due to simpler formulation. LOOP yields only marginal improvement in performance compared to LDP^{ri}, but has similar complexity in formulation and comparable run time. Next it is determined whether the increase in average accuracy of LOOP over LDP^{ri} is statistically significant.

Sign Binomial Test. For each descriptor, 2 classifiers and 3 datasets are used, hence 6 combinations per descriptor. Also there are 5 fold cross-validation per combination. So for each descriptor, there are 30 sets of accuracy results. Assuming the null hypothesis to be that the two competing methods (LOOP and LDP^{ri}) are equally good, then there is 50% chance of each beating the other.

It is observed that of the 30 experimental runs, LOOP outperforms LDP^{ri} 22 times. The one-tail P value at 5% level of significance is 0.0081. Now using Bonferroni correction, at 5% level of significance, the corrected α for the 6 combinations (2 classifiers and 3 datasets per descriptor) is $0.05/6 = 0.0083$. Since the calculated chance is 0.0081 (less than the corrected α), one can reject the null hypothesis and conclude that LOOP has a statistically significant better performance than LDP^{ri}.

Also as explained earlier, LDP^{ri} has the constraint of always having a leading 1 thus halving the number of possible words, along with the restriction of having a fixed number of 1s and 0s in the binary word. LOOP is free from these limitations in design.

5.5 Discussion

A novel binary local pattern descriptor, LOOP, which overcomes some disadvantages of its predecessors LBP and LDP, is presented in this chapter. It is tested on Lepidoptera species recognition with encouraging initial results that warrant further exploration. It outperforms the descriptors on which it is based, along with a few other variants. It has comparable results with popular binary descriptors like BRISK and ORB, but gains in time complexity. This chapter has only presented experiments and results on one representative problem, that of Lepidoptera classification. But LOOP is a generalized binary descriptor and may be used in further research for other small fine-grained datasets.

Chapter 6

Collaborative Fine-grained Supervised Deep Learning

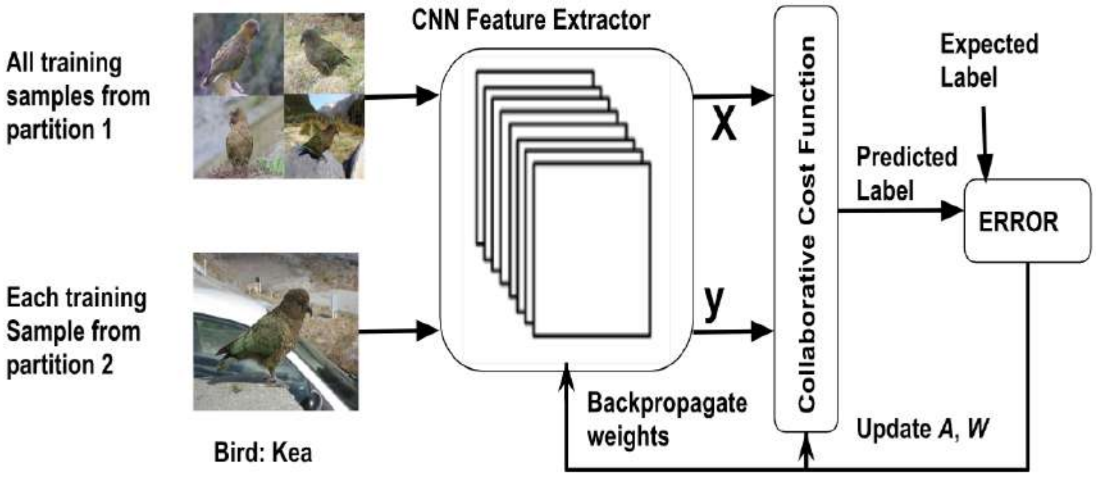
6.1 Chapter Summary and Linkage

Based on the available literature, it was indicated by preliminary results that collaborative representation classifiers (CRC) are suitable for fine-grained visual categorization (FGVC). Then this work progressed the state-of-the-art in both FGVC (by proposing a new way to represent fine-grained local repeated patterns) as well as CRC (by automating background compensation into the formulation). Thus the next step is to integrate CRC-like representations into modern machine learning methods like convolutional neural networks. The present chapter introduces a Collaborative ConvNet (CoCoNet) for better transfer learning of small fine-grained datasets with limited samples.

6.2 Introduction

Deep convolutional networks have proven to be proficient in classifying base image categories with sufficient generalization when trained with a large dataset. However, many real life applications of significance (Chai (2015)) may be characterized by fine-grained classes and limited availability of data, like endangered species recognition (Rodner *et al.* (2015)) or analysis of biomedical images of a rare pathology. In such specialized problems, it is challenging to effectively train deep networks that are data hungry. Transfer learning is a popular approach to train on small fine-grained image datasets with limited samples (Simon *et al.* (2015)). The ConvNet architecture is trained first on a large benchmark image dataset (eg. ImageNet) for the task of base object recognition. The network is then fine-tuned on the target smaller dataset for fine-grained recognition. Since the target dataset is small, there is an increased chance of overtraining. On the other hand, if the dataset has fine-grained objects with varying backgrounds, this can cause difficulty in training convergence. This makes the optimal training of the dataset challenging (Horn *et al.* (2017)). In case of small datasets with imbalanced classes (Horn *et al.* (2017)), the problem is compounded by the probability of training bias in favour of larger classes. A few specialized deep learning methods have been proposed in recent times to cater to these issues, like low-shot/zero-shot learning (Li *et al.* (2017)) for small datasets and multi-staged transfer learning (Simon *et al.* (2015)) for fine-grained classes. In spite of these advances, deep learning of small fine-grained datasets remains one of the open popular challenges of machine vision (Krause *et al.* (2014))(Krause *et al.* (2015)).

These challenges are addressed in the current chapter through a Collaborative ConvNet (CoCoNet). It fine-tunes a pre-trained deep network through a novel collaborative representation layer in an end-to-end fashion. This establishes a protocol for multi-stage transfer learning of fine-grained data with limited samples. The intuition is that we have already seen that CRC methods tend to work well for fine-grained recognition, so the logical next step becomes to incorporate the power of collaborative filtering into the current best approach available for classification, that is, convolutional networks. This is further strengthened by the findings of Chapter 2 and the works of Cai *et al.*


All training samples from partition 1
Each training Sample from partition 2
Bird: Kea
CNN Feature Extractor
X
y
Backpropagate weights
Collaborative Cost Function
Expected Label
Predicted Label
ERROR
Update A, W

(a) BCNN

Figure 6.1: Schematic Diagram of Collaborative ConvNet (CoCoNet).

(2016), who have recently shown that some modern versions of CRC give better performance with CNN learned features from a pre-trained ConvNet compared to a fully-connected softmax based classification layer.

6.3 Methods

6.3.1 The CoCoNet model

CoCoNet gives a collaborative loss which is back propagated through an end-to-end model. The training set is divided into 2 sections $p1$ and $p2$. One possible split can be 2:1 between $p1$ and $p2$, having m and n images respectively randomly selected with equal representation across classes.

Let y be the $d \times 1$ feature vector of one image in $p2$, such that the feature matrix for $p2$ is Y of dimension $d \times n$. Let x be the $d \times 1$ feature vector of one image in $p1$, such that the feature matrix for $p1$ is X of dimension $d \times m$.

The collaborative cost function is given by:

$$P\,(A, W, X) = ||(Y - XA)W||_2^2 + \lambda||A||_2^2 + \gamma||W||_2^2 \tag{6.1}$$

The collaborative reconstruction matrix A is thus of dimension $m \times n$. The goal is to find an optimal feature representation of each sample in $p2$ with respect to the “training" images in $p1$ via a representation vector $\vec{a}_i \in A$.

The weight matrix W is used to compensate for imbalance of classes and each of its elements is initialised with a weight proportional to the size of the class to which the corresponding feature vector in Y belongs. W counteracts the imbalance in classes as a penalty term for larger classes by increasing the cost. W is of dimension $n \times 1$.

After finding the initial optimal A through least squares, the weight matrix W, reconstruction matrix A and the feature representation matrix X are updated through partial derivatives for each backward pass of the CoCoNet as presented in Algorithm 1.

Least squares minimization gives the initial optimal value of A as:

$$\hat{A} = \left[X^T X W^T W + \lambda I\right]^{-1} X^T Y W W^T \tag{6.2}$$

Fix $\{A, X\}$, update W:

$$\frac{\partial P}{\partial W} = -(Y - XA)^T (Y - XA)W + \gamma W \tag{6.3}$$

Fix $\{W, X\}$, update A:

$$\frac{\partial P}{\partial A} = -X^T (Y - XA) W W^T + \lambda A \tag{6.4}$$

Fix $\{W, A\}$, update X:

$$\frac{\partial P}{\partial X} = -(Y - XA) W W^T A^T \tag{6.5}$$

Once all the partial derivatives are obtained, CNN weights are updated by standard back-propagation of gradients for each batch in $P1$ and $P2$. A schematic is presented in Fig. 6.1..

Algorithm 1: Training with CoCoNet

```
Initiate weight matrix $W$ with elements proportional to class size ;
Split the training set into two parts $P1$ and $P2$ ;
Extract feature matrix $X$ of $P1$ through CNN section of CoCoNet. ;
Find initial optimal reconstruction matrix $A$ by eqn. 6. ;
for each batch in $P2$ do
    Fix $\{A, X\}$, update $W$ by eqn. 7 ;
    Fix $\{W, X\}$, update $A$ by eqn. 8 ;
    Fix $\{W, A\}$, update $X$ by eqn. 9 ;
    for each batch in $P1$ do
        back-propagation of gradients to update weights of CoCoNet ;
    end
end
```

6.3.2 Reducing computation cost through SVD.

The optimal representation weight matrix $\hat{A}$ from eqn 6.2 has the term $(X^TXW^TW+\lambda I)^{-1}$, where X is of dimension $d \times m$. Here d is the dimension of the descriptor and m is the total number of data points in the partition $P1$ of training data. This poses the problem of high computation cost for large datasets (m is large). So singular value decomposition (SVD) is used to reduce the matrix inverse computation to dimension $d \times d$, so as to make it independent of dataset size. This is a crucial modification needed for applications like image retrieval from large unlabeled or weakly labeled image repositories.

Taking the singular value decomposition (SVD) of X^T, one can factor $X^T X$ as:

$$X^T X = (USV^T)^T USV^T = VS^T U^T USV^T = V(S^T S)V^T \tag{6.6}$$

Since S only has d non-zero singular values, one can truncate $S^T S$ and V to be smaller matrices. So V is $N \times d$, S is $d \times d$ and V^T is $d \times N$. Also note that since W is of dimension $n \times 1$. Thus $W^T W$ comes out as a scalar value w in eqn. 6.6. w is absorbed in S to have $\hat{S}$.

Using the Woodbury matrix inverse identity (Woodbury (1950)), the inverse term then becomes:

$$(V\hat{S}^T SV^T + \lambda I)^{-1} = \frac{1}{\lambda} + \frac{1}{\lambda^2} V(\hat{S}^{-1} + \frac{1}{\lambda} V^T V)^{-1} V^T = \frac{1}{\lambda} + \frac{1}{\lambda^2} V(\hat{S}^{-1} + \frac{1}{\lambda} I)^{-1} V^T \tag{6.7}$$

Note that the inverse term $(\hat{S}^{-1} + \frac{1}{\lambda} I)^{-1}$ is only $d \times d$, so it scales to many data points.

6.3.3 Enhanced Learning by CoCoNet

CoCoNet uses the collaborative cost function in an end-to-end manner. So one does not have the fully final classification layers. The CNN extracts features and feeds it to the collaborative layer. The collaborative cost function estimates error, updates its own parameters as well as feeds it back to the CNN. The error and gradients are then back propagated through the CNN to update the features. So CoCoNet is different from just cascading a CNN based feature learner with a collaborative filter, because the parameters are not updated in latter in an end-to-end fashion. For the same dataset and same number of given samples, the collaborative layer represents all samples together as an augmented feature vector. Thus after error is found out, the error gradient may be found with respect to each of the image vectors in the feature matrix and the gradients are then back propagated. This collaborative representation is not just the augmented feature matrix with all samples, it is also optimised by the collaborative filter. This adds an additional level of optimisation besides the CNN learned features and tuned parameters.

6.4 Experiments and Results

6.4.1 Experimental Setup

Datasets. Five benchmark image datasets are used in this work for pre-training and fine-tuning in total. *ImageNet* (Russakovsky *et al.* (2015)) has about 1.4 million image categories as of 2017 and has been used for pre-training the networks as base category classifiers. Then for transfer learning, four bird species recognition datasets have been used: two pre-existing NABirds and CUB datasets (described in Chapter 2) and two new NZbirds and IndBirds datasets (described in Chapter 3 and compiled as part of this thesis).

Competing Classifiers. The performance of CoCoNet is evaluated against two popular recent methods both among collaborative representation classifiers (CRC) and deep convolutional neural networks (CNN), besides testing against constituent components as ablation study. Among current CRC methods, comparison is made against the state-of-the-art Probabilistic CRC (ProCRC) (Cai *et al.* (2016)). Among recent deep CNN models, the popular Bilinear CNN is chosen (Lin *et al.* (2018)), (Lin *et al.* (2015)) as the benchmark competitor. Of course, there are a few more recent variants of ProCRC, like enhanced ProcCRC (EProCRC) (Lan *et al.* (2017)), as well as of BCNN, like improved BCNN (Lin *et al.* (2017)). But the vanilla versions have been deliberately chosen here because the aim is to establish a benchmark evaluation in this work. For the same reason comparisons have also been made with the original CRC formulation plus two well-known benchmark CNN architectures: AlexNet (Krizhevsky *et al.* (2012)) and the more recent VGG16 and

Table 6.1: CUB 200-2011 Test Accuracy (%)

	ImageNet → CUB **(1 stage)**	NABirds → CUB **(1 stage)**	ImageNet → NABirds → CUB **(2 stage)**
AlexNet	52.2 ± 5.4	55.4 ± 5.3	58.6 ± 5.5
AlexNet+CRC	57.5 ± 5.1	59.9 ± 5.0	61.3 ± 5.6
AlexNet+ProCRC	60.8 ± 5.3	63.5 ± 5.2	65.5 ± 5.6
CoCoNet1	**64.4 ± 5.2**	**67.0 ± 5.2**	**69.4 ± 5.5**
Vgg16	60.1 ± 5.8	63.9 ± 5.9	66.4 ± 5.7
Vgg16+CRC	66.3 ± 5.7	70.2 ± 5.9	72.9 ± 5.8
Vgg16+ProCRC	69.4 ± 5.9	72.6 ± 5.8	77.7 ± 5.4
CoCoNet2	**73.7 ± 5.7**	**75.8 ± 5.7**	**81.5 ± 5.6**
Vgg19	71.9 ± 5.5	74.1 ± 5.7	77.5 ± 5.9
Vgg19+CRC	76.2 ± 5.6	79.0 ± 5.5	80.2 ± 5.9
Vgg19+ProCRC	79.3 ± 5.4	82.5 ± 5.5	83.8 ± 5.8
CoCoNet3	**83.6 ± 5.5**	**87.4 ± 5.6**	**89.1 ± 5.6**
Bilinear-CNN	84.0 ± 5.3	85.7 ± 5.8	87.2 ± 5.5

VGG19 (Simonyan *et al.* (2014)). These help to conduct the ablation study. The VGG and AlexNet architectures are presented in Fig 6.2. A description

Table 6.2: IndBirds Test Accuracy (%)

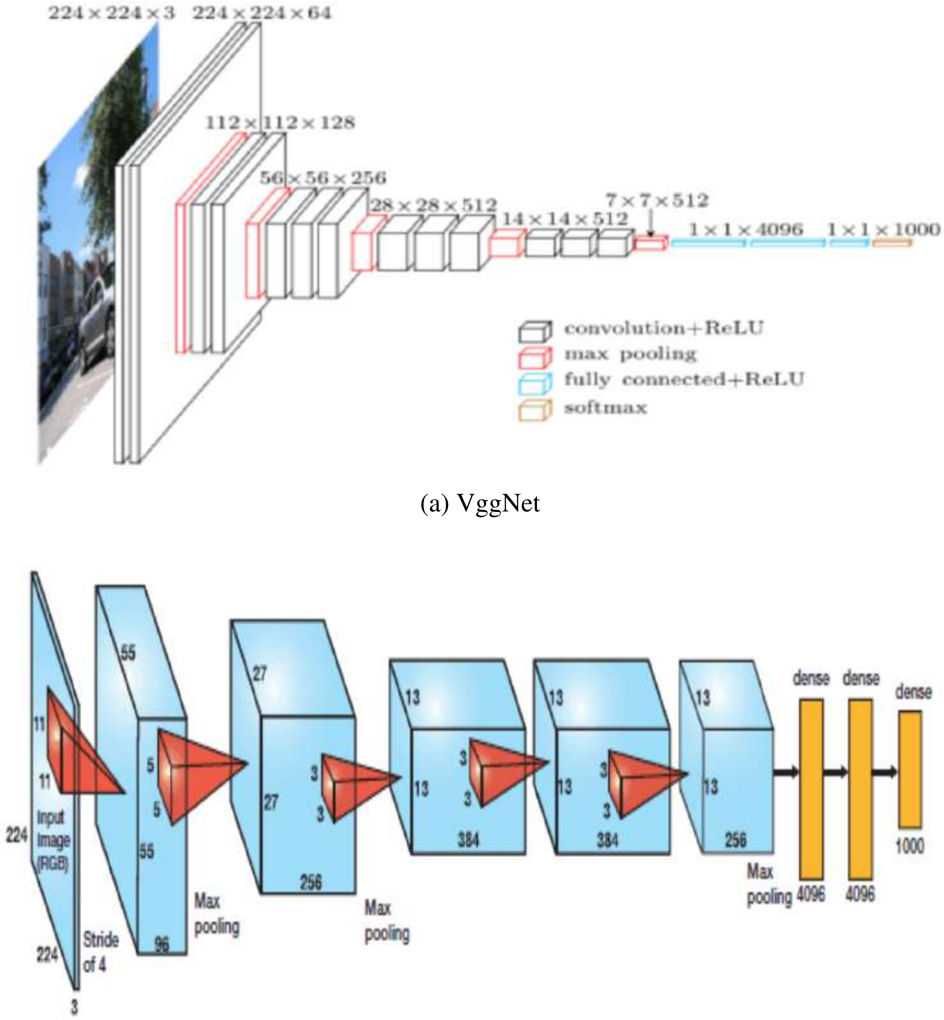
	ImageNet → IndBirds **(1 stage)**	NABirds → IndBirds **(1 stage)**	ImageNet → NABirds → IndBirds **(2 stage)**
AlexNet	60.1 ± 4.4	63.2 ± 4.7	66.6 ± 4.5
AlexNet+CRC	65.8 ± 4.8	68.5 ± 4.5	71.7 ± 4.8
AlexNet+ProCRC	70.5 ± 4.5	73.9 ± 4.9	75.4 ± 4.7
CoCoNet1	**73.3 ± 4.6**	**77.0 ± 4.7**	**80.4 ± 4.4**
Vgg16	69.7 ± 4.8	74.2 ± 4.1	77.7 ± 4.5
Vgg16+CRC	74.5 ± 4.7	79.3 ± 4.4	83.0 ± 4.6
Vgg16+ProCRC	78.6 ± 4.1	82.8 ± 4.3	85.7 ± 4.1
CoCoNet2	**81.9 ± 4.7**	**86.5 ± 4.4**	**89.9 ± 4.3**
Vgg19	76.2 ± 4.2	82.5 ± 4.7	84.8 ± 4.2
Vgg19+CRC	80.6 ± 4.4	86.3 ± 4.0	87.4 ± 4.4
Vgg19+ProCRC	84.0 ± 4.9	89.1 ± 4.1	91.0 ± 4.2
CoCoNet3	**87.4 ± 4.3**	**92.9 ± 4.4**	**94.7 ± 4.5**
Bilinear-CNN	85.1 ± 4.7	88.6 ± 4.2	91.5 ± 4.3

of ProCRC may be found in Chapter 2. Maji *et al.* introduced the BCNN architecture for fine-grained visual recognition (Lin *et al.* (2018))(Lin *et al.*

Table 6.3: NZBirds Test Accuracy (%)

	ImageNet → NZBirds **(1 stage)**	NABirds → NZBirds **(1 stage)**	ImageNet → NABirds → NZBirds **(2 stage)**
AlexNet	49.9 ± 5.6	52.3 ± 5.4	55.0 ± 5.8
AlexNet+CRC	54.4 ± 5.5	58.1 ± 5.7	60.4 ± 5.8
AlexNet+ProCRC	57.5 ± 5.8	61.9 ± 5.9	65.8 ± 5.5
CoCoNet1	**62.2 ± 5.2**	**66.6 ± 5.6**	**69.7 ± 5.7**
Vgg16	55.7 ± 5.9	57.9 ± 5.6	59.8 ± 5.3
Vgg16+CRC	59.6 ± 5.6	60.4 ± 5.5	62.7 ± 5.1
Vgg16+ProCRC	63.1 ± 5.7	66.5 ± 5.8	68.1 ± 5.5
CoCoNet2	**68.3 ± 5.1**	**69.8 ± 5.3**	**71.6 ± 5.4**
Vgg19	61.5 ± 5.0	63.7 ± 5.1	65.6 ± 5.7
Vgg19+CRC	63.9 ± 5.3	66.1 ± 5.5	68.7 ± 5.6
Vgg19+ProCRC	66.2 ± 5.5	71.3 ± 5.1	72.9 ± 5.8
CoCoNet3	**71.8 ± 5.2**	**74.4 ± 5.2**	**77.2 ± 5.6**
Bilinear-CNN	69.4 ± 5.6	71.8 ± 5.5	73.3 ± 5.0

(2015)). These networks represent an image as a pooled outer product of features learned from two CNNs and encode localized feature interactions that



(a) VggNet

(b) AlexNet

Figure 6.2: Architectures of Benchmark ConvNets: AlexNet and VggNet

are translationally invariant. BCNN is a type of orderless texture representation that can be trained in an end-to-end manner.

Experiments. Each of the three target datasets (CUB, NZBirds, IndBirds) is trained through a combination of one stage and two stage transfer learning. For one stage transfer learning, two separate configurations have been used: 1) the network is pre-trained for general object recognition on ImageNet and then fine-tuned on the target dataset; 2) the network is pre-trained for bird recognition on the large North American bird dataset (NABirds) and then fine-tuned on the target dataset. For 2 stage training, the network is trained successively on ImageNet, NABirds and then the target dataset. Note that for pre-training, always the original architecture (AlexNet/VggNet) is used, CoCoNet only comes into play during fine-tuning. Note that the configurations with AlexNet, VGG16 and VGG19 are named CoCoNet1, CoCoNet2 and CoCoNet3 respectively. During both pre-training and fine-tuning, 0.001 learning rate is taken at the start, but shifts to 0.0001 once there is no change in loss anymore, keeping the total number of iterations/epochs constant at 1000. The Adam (Kingma *et al.* (2014)) optimiser is chosen and single-stage fine-tuning of a pre-trained network took on an average 10 hrs, 4 hrs and 4.5 hrs respectively for the CUB, NZBirds and IndBirds dataset respectively on a iMac machine with 8GB RAM, 2GB AMD Radeon GPU, 3.2GHz Intel i5 processor. Next it is investigated how the end-to-end formulation of CoCoNet fares in controlled experiments with competing configurations. The same experiments are repeated using the original architecture (AlexNet/VggNet), and then the change in accuracy is observed with cascaded CNN+CRC and the end-to-end CoCoNet. The results with cascaded CNN+ProCRC as well as Bilinear CNN are then tabulated. For each dataset, images are resized to 128×128 and experiments are conducted with 5 fold cross validation and percentage classification accuracies along with standard deviation are presented in Table 6.1 (CUB), Table 6.2 (IndBirds) and Table 6.3 (NZBirds) with the highest accuracy in each column highlighted in bold. Before fine-tuning, the pre-trained network was truncated before the output layer in order to change the number of classes according to the target fine-grained dataset.

6.4.2 Results and Analysis

Results. It may be readily observed from the tabulated results, that the proposed method overall easily outperforms the vanilla architectures (AlexNet/ VggNet) as well as the cascaded configurations (AlexNet/ VggNet+CRC).

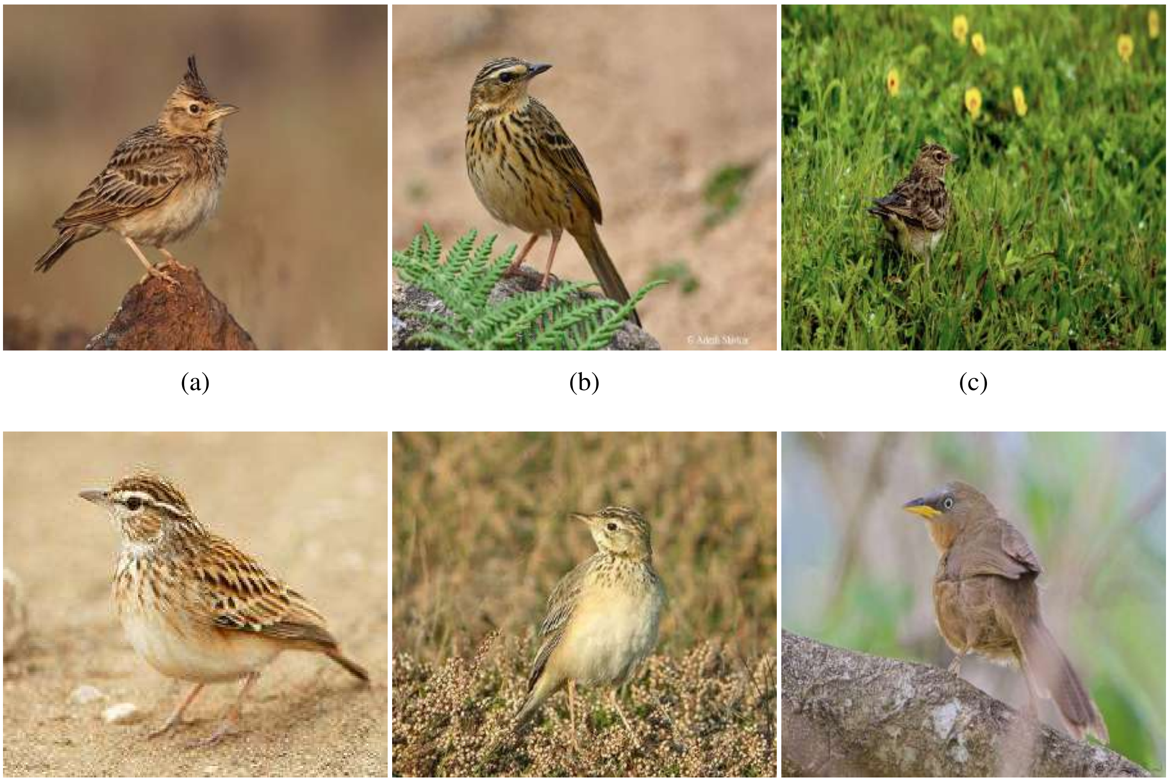
CoCoNet also yields higher accuracy than the recent collaborative method ProCRC, when the latter is cascaded with a ConvNet. CoCoNet also gives better results than the popular Bilinear CNN, which is considered a benchmark for fine-grained recognition. This trend in the performance is reflected across the three datasets and for both one stage and two stage transfer learning. It may also be noted that there is an increase in accuracy for two stage learning over single stage learning (over all datasets and architectures). The ablation study in the tables also reveals that the proposed method outperforms its constituent parts considerably and consistently.

Statistical Analysis. The Signed Binomial Test is performed to investigate the statistical significance of the improvement in performance of CoCoNet (best among the configurations) vs. BCNN. This can be used across different datasets and methods simultaneously because it considers frequency of success in the calculations rather than assuming a gaussian distribution of accuracy values (like t-tests). CoCoNet3, the configuration using VGG-19 network, is chosen as that is the best performing. The null hypothesis is that the two are equally good, that is there is 50% chance of each beating the other on any particular trial. For each of the three datasets (CUB, NZBirds and Indbirds), there are three transfer learning configurations (two single stage and a double stage) and five-fold cross-validated results. Thus over the three datasets, in total we have 45 experiments of CoCoNet vs. BCNN, and out of these CoCoNet outperformed the latter 33 times (that is 73.33% of the trials). The signed binomial test yields that given the assumption that both methods are equally good, then the probability of CoCoNet outperforming BCNN in 73.33% of the trials is 0.12% (one-tail p-value of 0.0012). Considering a level of significance of $\alpha = 0.05$, the Bonferroni adjustment is applied. There are 3 transfer learning protocols and 3 datasets: hence 9 combinations of experimental condition. So we divide the 5% level of significance by 9 to get adjusted $\alpha = 0.0055$. Since the one-tail p-value obtained is less than 0.0055, it may be concluded that the improvement in accuracy is statistically significant considering the frequency of out-performance.

Qualitative Results. Fig. 6.3(a) is that of a Malabar Lark, which is one of the species in the IndBirds dataset. One of the key distinguishing parts of the Malabar Lark is its head crest. Fig. 6.3(b) is an image of the Nilgiri Pipit from the same dataset. Fig. 6.3(c) presents a test image of the Malabar Lark that was **misclassified** as a Nilgiri Pipit by the proposed CoCoNet as well as its nearest competitors: Bilinear CNN as well as cascaded CNN+ProCRC. It can be seen that in that image, due to the pose of the bird, the head crest is not clearly visible. Fig. 6.3(d) and 6.3(e) are those of Nilgiri Pipit, while 6.3(f) is that of Rufous Babbler. It can be seen from 6.3(d) that the Nilgiri Pipit is characterised by distinct dark patterns on its back, which is not clearly visible from the front, as shown in Fig. 6.3(e). The image in Fig. 6.3(e) was **correctly classified** by the proposed CocoNet as Nilgiri Pipit but was misclassified by its competitors (cascaded CNN+ProCRC and BCNN) as Rufous Babbler.

6.5 Discussion

An end-to-end collaborative convolutional network (CoCoNet) architecture is presented in this chapter for fine-grained visual recognition with limited samples. The new architecture adds a collaborative layer which adds an additional level of optimization based on collaboration of images across classes, the information is then back-propagated to update CNN weights in an end-to-end fashion. This collaborative representation exploits the fine-grained nature of the data better with fewer training images. The proposed network is evaluated for the task of fine-grained bird species recognition, but the method is general enough to be used in other fine-grained classification tasks like detection of rare pathology from medical images. The other major advantage is that most existing CNN architectures can be easily restructured into the proposed configuration. Results indicate that the proposed algorithm per-



(a) (b) (c)

(d) (e) (f)

Figure 6.3: Classification and Misclassification Examples from the new IndBirds dataset: (a) Malabar Lark, (b) Nilgiri Pipit, (c) Malabar Lark, misclassified as Nilgiri Pipit by both proposed CoCoNet and competitors, due to obfuscation of the discriminating head crest. (d) Nilgiri Pipit with characteristic dark pattern on back (e) Front-facing image of Nilgiri Pipit with back patterns not visible. Correctly classified by proposed CoCoNet but misclassified by competitors as Rufous Babbler (f).

forms much better than its constiutent parts, a recent CRC method (probabilistic CRC: ProCRC) and a benchmark deep network method (Bilinear-CNN: BCNN).

Chapter 7

Collaborative Fine-grained Unsupervised Clustering

7.1 Chapter Summary and Linkage

The previous chapter incorporated collaborative filters into a generic CNN architecture for better transfer learning of fine-grained datasets with limited samples. So it looked at supervised learning through a novel Collaborative ConvNet (CoCoNet). The present chapter takes the next logical direction of investigating the incorporation of collaborative filters into unsupervised learning. So here a new collaborative clustering algorithm is presented where the cluster centers are updated iteratively as a weighted average of the assigned data points (generalisation of k-means) and these weights are optimally obtained from the collaborative cost function.

7.2 Introduction

Supervised deep learning based vision systems are considered the state-of-the-art and they have achieved near human accuracy in recent years. However, these methods work best when a large amount of well labeled and annotated data is available and there are many applications of practical significance where these prerequisites are not adequately met (Chai (2015)). Automated recognition of endangered species in the wild (Rodner *et al.* (2015)) and detection of a rare pathology from medical images are examples of problems that may be characterized by scarcity of training images and imbalance of classes. Furthermore, these are highly specialized applications requiring labeling/annotations by domain experts. This might not be readily available and would be cost prohibitive to acquire in adequate quantity for deep learning. Thus unsupervised (weakly labeled or unlabeled) clustering methods are still pertinent for such applications.

In this chapter we present a collaborative clustering algorithm for fine-grained data, as an optimal weighted generalization of the classic k-means. The collaborative clustering method encodes the distances of each data point to the cluster centres and this function is optimised to find the representation weights. These weights are then used to update the cluster centers in each iteration. k-means is a special case (Lloyd (1982)) of this, where the representation weights are unity and hence the distance between data points are Euclidean and the cluster centers are updated by a simple mean. There are other methods referred to as "collaborative clustering" in existing literature, but these refer to a collaboration or ensemble of clustering methods (Cornuejols *et al.* (2018)), rather than using the collaborative filter analytically.

The proposed collaborative clustering algorithm is compared with k-means and several of its major variants: k-modes (Chaturvedi *et al.* (2001)), k-medians (Jain *et al.* (1998)) and k-medoids (Park *et al.* (2009)). Performance is also compared against DBSCAN (Ester *et al.* (1996)), which is currently the most cited clustering method, as well as against a state-of-the-art variation of DBSCAN called DSets-DBSCAN (Hou *et al.* (2016)), where the authors present a non-parametric formulation based on dominant sets using similarity matrix of input data. Fine-grained species recognition is chosen as the representative problem like in the rest of this thesis. The tasks are bird species recognition and butterfly species recognition. A dense variant of SIFT (Lowe (1999)) is taken as the first feature descriptor, as well as an ensemble of GIST (Oliva *et al.* (2001)) and HoG (Dalal *et al.* (2005)) features as the second descriptor. It is seen that the proposed collaborative clustering easily outperforms k-means and its variants, and also gives overall improvement against DBSCAN.

7.3 Methods

In this section we present the proposed collaborative clustering method in details, its formulation and derivation. We also present the competing methods in brief.

7.3.1 Collaborative k-means Clustering

Collaborative filters represent the query sample as a weighted average of available data points across all categories of the dataset. The representation weights are then optimised via the collaborative cost function and the final categorization is assigned according to the sample with least residual. Collaborative filters should be well suited to represent fine-grained clusters with subtle differences and limited samples, since it finds optimal representation of data across clusters. The intuition is to incorporate this co-operative approach within the k-means clustering framework in this work. For clustering, this would translate to the cluster centers being represented as weighted mean of data points, where these weights are optimised via the collaborative cost function as analysed below.

Let the number of required clusters be K. Consider a dataset with N images in the feature space of d dimensions each, such that the feature matrix is $X \in \mathbb{R}^{d \times N}$. Choose K samples out of the N samples as a random initialisation of the cluster centres as $Y \in \mathbb{R}^{d \times K}$. Each cluster center is y_k, where

$k = 1, \ldots, K$. α_k is the representation weight vector of dimension d for the cluster k.

The collaborative cost function is given by:

$$P(\alpha_k) = ||y_k - X\alpha_k||_2^2 + \lambda||\alpha_k||_2^2 \tag{7.1}$$

The optimal value of α_k for each cluster center y_k are given by:

$$\hat{\alpha_k} = (X^T X + \lambda I)^{-1} X^T y_k \tag{7.2}$$

The residual for sample i with respect to y_k (k_{th} cluster) is given by:

$$r_i(y_k) = \frac{||y_k - X_i\hat{\alpha}_{ki}||_2^2}{\hat{\alpha}_{ki}^2} \tag{7.3}$$

Calculate $r_i(y_k)\forall i = 1, \ldots, N$ and $k = 1, \ldots, K$.

The sample i is allocated to the cluster center with lowest residual as follows:

$$C(X_i) = \text{arg} \min_k \; r_i(y_k) \tag{7.4}$$

This concludes the first pass.

Let X^k be n_k number of columns of $X \in k_{th}$ cluster, $k = 1, \ldots, n$. $X^k = [X_1^k, \ldots, X_{n_k}^k] \in \mathbb{R}^{d \times n_k}$ where $\sum_{k=1}^{K} n_k = N$. Let $\hat{\alpha}_{kj}$ be the representation weight corresponding to $X_j^k \in X^k, k = 1, \ldots, n_k$.

In the next iteration, the new cluster centres are computed through:

$$y_k = \frac{1}{n_k} \sum_{j=1}^{n_k} X_j^k \hat{\alpha}_{jk} \tag{7.5}$$

Same steps are repeated until the termination condition is reached.

Reducing computation cost through SVD. The optimal representation weight matrix $\hat{\alpha}$ from eqn 7.2 has the term $(X^T X + \lambda I)^{-1}$, where X is of dimension $d \times N$. Here d is the dimension of the descriptor and N is the total number of data points in the dataset. This poses the problem of high computation cost for large datasets (N is large). So singular value decomposition (SVD) is used to reduce the matrix inverse computation to dimension $d \times d$, so as to make it independent of dataset size. This is a crucial modification needed for applications like image retrieval from large unlabeled or weakly labeled image repositories.

The singular value decomposition (SVD) of X^T is used as:

$$X^T X = (USV^T)^T USV^T = VS^T U^T USV^T = V(S^T S)V^T \tag{7.6}$$

Since S only has d non-zero singular values, $S^T S$ and V are truncated to be smaller matrices. So V is $N \times d$, S is $d \times d$ and V^T is $d \times N$.

Using the Woodbury matrix inverse identity (Woodbury (1950)), the inverse term then becomes:

$$VS^T SV^T + \lambda I)^{-1} = \frac{1}{\lambda} + \frac{1}{\lambda^2} V(S^{-1} + \frac{1}{\lambda} V^T V)^{-1} V^T = \frac{1}{\lambda} + \frac{1}{\lambda^2} V(S^{-1} + \frac{1}{\lambda} I)^{-1} V^T \tag{7.7}$$

Note that the inverse term $(S^{-1} + \frac{1}{\lambda} I)^{-1}$ is only $d \times d$, so it will scale to many data points.

7.3.2 Competing Clustering Methods

k-means and its variants. Collaborative clustering may be looked upon as a generalization of the k-means algorithm (Lloyd (1982)). Collaborative

Algorithm 2: Collaborative k-Means Clustering

```
1  Choose number of clusters $K$
2  Initiate the cluster centers randomly from the data points
3  Form the feature matrix $X$ and the cluster center matrix $Y$
4  Find initial reconstruction vector $\alpha$ by eqn. 7.2.
5  while Termination condition is not reached do
6      for each cluster center $y_k \in Y$ do
7          Find the collaborative weights $\alpha$ by eqn. 7.2.
8          for each image $x \in X$ do
9              Find the distances of $x$ from cluster center $y_k$ using eqn.
                 7.3 and 7.4.
10         end
11     end
12     Update cluster centers by eqn. 7.5.
13     Continue from Step 5 till termination condition reached.
14 end
```

clustering represents cluster centers as optimal weighted sums of data points. Thus, k-means is a specialized case where these weights are all unity and only the Euclidean distance from cluster centers is hence considered. The performance of collaborative clustering is evaluated against k-means and three of its major variants: k-modes (Chaturvedi *et al.* (2001)), k-medians (Jain *et al.* (1998)), k-medoids (Park *et al.* (2009)). k-modes and k-medians, as the names suggest, utilise the cluster modes and medians instead of the means during the updates. The k-medoids algorithm chooses datapoints as centers (medoids or exemplars) and uses a generalization of the Manhattan Norm instead of the Euclidean distance.

DBSCAN. Density-based spatial clustering of applications with noise (DBSCAN) (Ester *et al.* (1996)) is currently the most cited clustering algorithm. It groups together densely packed data points (with many nearby neighbours) and marks points in low density areas as outliers. Thus DBSCAN is somewhat robust to noise and unlike the k-means algorithms, does not require apriori knowledge of required number of cluster centers. A recent state-of-the-art variation of DBSCAN, named D-Sets DBSCAN (Hou *et al.* (2016)), is also taken for comparison. D-Sets DBSCAN is a non-parametric formulation based on dominant sets using similarity matrix of input data.

7.4 Experiments and Results

In this section, the experimental setup is presented: the datasets, the feature descriptors, and the competing clustering methods for comparison.

7.4.1 Experimental Setup

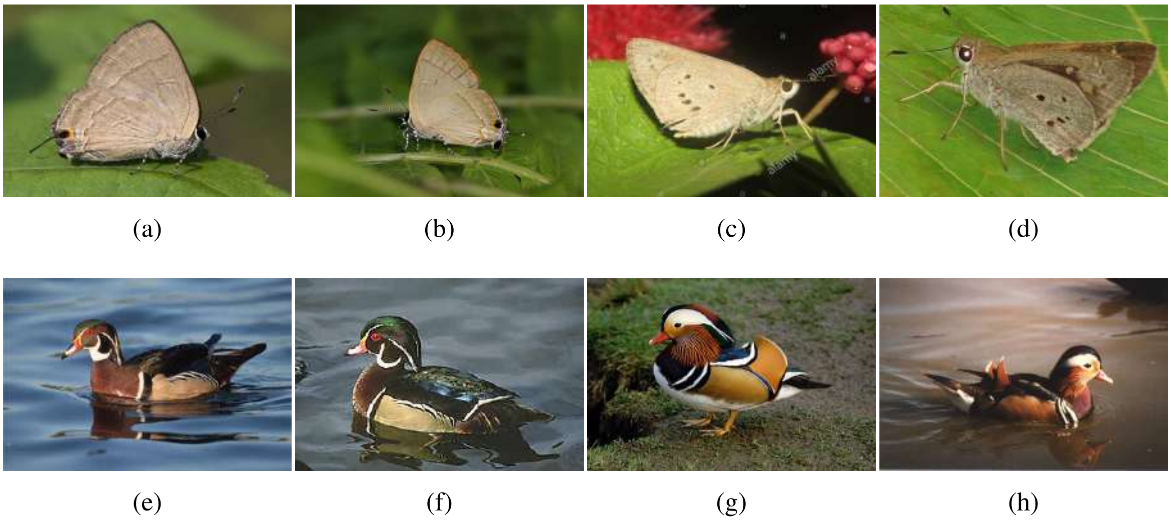
Benchmark Datasets. The proposed clustering algorithm is tested on four species recognition datasets. Of these two are for bird species recognition and the other two for butterfly species recognition. The Ponce birds and butterfly datasets are pre-existing benchmark datasets (described in Chapter 2). The Indian birds and butterfly datasets are new and compiled as part of this thesis (described in Chapter 3).

Feature Descriptors. 2 popular feature descriptors are used: Dense SIFT and ensemble of GIST+HoG. But it should be noted that the proposed algorithm is general and is agnostic to feature choice. A dense variant of scale invariant feature transform (SIFT) (Lowe (1999)) is extracted and a patch size of 10×10 is chosen with overlap. Global invariant scale transform, here referred to as GIST (Oliva *et al.* (2001)), is a global feature that describes the spatial envelope of the image using directional properties. It extracts dense multi-scale overlapping patches. Histogram of oriented gradients (HoG) features (Dalal *et al.* (2005)) are extracted in a dense grid fashion in 3×3 cells which are concatenated at each grid location to generate the descriptor. The Dense SIFT features are used separately, while the GIST and HoG features are used as a concatanated ensemble.

7.4.2 Results and Analysis

Experimental Results. Clustering is performed on each combination of descriptor, dataset and algorithm. The average percentage accuracy is presented in Table 7.1 for bird and butterfly species recognition. The highest results in each column are highlighted in bold. It is observed from both tables that collaborative clustering significantly outperforms its direct competitors, that is the baseline k-means algorithm and its major variants. Also for both tasks, for the majortiy of algorithms, DenseSIFT based features yields better results than GIST+HoG. It is also noticed that collaborative clustering outperforms the original DBSCAN and also gives slight improvement overall over the recent variant DSets-DBSCAN (Hou *et al.* (2016)). Though the improvements are marginal, it should be noted that the proposed algorithm has a much more lightweight formulation and implementation than DSets-DBSCAN. Moreover, Wilcoxon rank test is also performed to further explore the performance of collaborative clustering vs. DSets-DBSCAN.

Statistical Analysis. Wilcoxon signed rank test is performed across both tasks (bird and butterfly recognition) between collaborative clustering and DSets-DBSCAN and presented through Table 7.2. The ranks (R) are allocated according to the magnitude of difference in accuracy between the two methods. If there is a tie in the absolute difference, then the rank is split between the two. For example, if there is a tie for the values for 3rd and 4th rank, then both are given 3.5 rank. The corresponding signs (S) are allocated depending on which method outperforms for that particular experimental setting. The ones for which collaborative clustering is better have sign 1. the



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 7.1: Mis-clustering examples: (a)-(b) are Lycaenidae and (c)-(d) are Hesperiidae from the new Indian Butterfly dataset; these are fine-grained classes. Both DSets-DBSCAN and collaborative clustering wrongly assigned 3(c) to the Lycaenidae cluster. (e)-(f) are of Wood Duck and (g)-(h) are of Mandarin from the Ponce Duck Dataset. DSets-DBSCAN wrongly assigned 3(h) to Wood Duck cluster, but collaborative clustering correctly identified it as Mandarin.

Table 7.1: Clustering Accuracy %

datasets →	**Ponce Birds**		**Indian Birds**		**Ponce Butterflies**		**Indian Butterflies**	
methods ↓	**Gist/HoG**	**SIFT**	**Gist/HoG**	**SIFT**	**Gist/HoG**	**SIFT**	**Gist/HoG**	**SIFT**
k-Means	72.6	73.5	70.0	72.7	67.4	69.2	65.3	68.4
k-Medians	77.7	78.4	74.3	76.6	70.8	73.9	69.9	73.5
k-Modes	77.1	78.8	74.6	76.2	71.1	73.7	70.3	73.1
k-Medoids	79.0	80.3	76.9	79.5	74.0	75.3	73.6	75.0
DBSCAN	83.6	84.1	80.5	84.8	79.4	81.8	78.5	80.8
DSet-DBSCAN	87.5	**88.6**	85.0	89.2	74.5	85.4	83.7	**85.5**
Collab. Clust.	**88.8**	88.1	**86.3**	**89.9**	**75.3**	**87.2**	**84.9**	85.2

Table 7.2: Wilcoxon Signed Rank Test

datasets →	**Ponce Birds**		**Indian Birds**		**Ponce Butterflies**		**Indian Butterflies**	
methods ↓	**Gist/HoG**	**SIFT**	**Gist/HoG**	**SIFT**	**Gist/HoG**	**SIFT**	**Gist/HoG**	**SIFT**
DSet-DBSCAN	87.5	88.6	85.0	89.2	74.5	85.4	83.7	85.5
Collab. Clust.	88.8	88.1	86.3	89.9	75.3	87.2	84.9	85.2
| Difference |	1.3	0.5	1.3	0.7	0.8	1.8	1.2	0.3
Rank (R)	6.5	2	6.5	3	4	8	5	1
Sign (S)	+1	-1	+1	+1	+1	+1	+1	-1

rest have sign -1. The Wilcoxon parameter $W = \sum SR$ is calculated for the 8 pairs of values and $W = 30$. Maximum possible rank value for $n = 8$ experiments is $n(n + 1)/2 = 36$. The Wilcoxon signed rank test states that the null hypothesis (collaborative clustering and DSets-DBSCAN are equally good) may be rejected (collaborative clustering better than DSets-DBSCAN) at 5% level of significance if $W \geq 30$. Hence, it may be concluded that the proposed collaborative clustering performs significantly better than DSets-DBSCAN.

Normalised Mutual Information (NMI). NMI is considered to be a standard procedure to investigate the performance of two closely performing clustering methods. It is given by:

$$NMI(Y, C) = \frac{2 \times I(Y, C)}{[H(Y) + H(C)]} \tag{7.8}$$

Here Y are expected/class labels and C are estimated/cluster labels. H and I are entropy and mutual information functions respectively. The entropy function $H(Y)$ is given by

$$H(Y) = -\sum_{y} P(Y = y) \times log[P(Y = y)] \tag{7.9}$$

The function takes the similar corresponding form for $H(C)$. The mutual information is given by

$$I(Y, C) = H(Y) - H(Y|C) \tag{7.10}$$

where $H(Y|C)$ is the entropy of expected labels within each cluster. Following the calculations described in (Vinh *et al.* (2010)), the NMI between the proposed Collaborative Clustering against the closest competitor DSet-DBSCAN is calculated. It is observed that

$$\frac{NMI(Y, DsetsDBSCAN)}{NMI(Y, Collab.Clust.)} < 1 \tag{7.11}$$

This signifies that Collaborative Clustering outperforms DSets-DBSCAN by normalised mutual information.

Qualitative Example. One of the challenges of fine-grained image categorization is utilising discriminating parts which may be obfuscated due to pose variation, bad illumination, partial obstruction by surrounding objects, etc. In Fig. 7.2, examples of correct and wrong performance of collaborative clustering are provided. Fig 7.2. (a)-(b) are Lycaenidae and (c)-(d) are Hesperiidae from the new Indian Butterfly dataset (Titli.v1); these are fine-grained classes. Both DSets-DBSCAN and collaborative clustering wrongly assigned 7.2(c) to the Lycaenidae cluster. (e)-(f) are of Wood Duck and (g)-(h) are of Mandarin from the Ponce Duck Dataset. DSets-DBSCAN wrongly assigned 7.2(h) to Wood Duck cluster, but collaborative clustering correctly identified it as Mandarin.

7.5 Discussion

Collaborative clustering is presented as a generalization of the benchmark k-means algorithm. The contribution is to find out cluster centroids in each iteration as weighted mean of data points, where the weights are optimized using a collaborative filter. The data points are given this weighted representation with respect to the cluster centers. Thus k-means may be considered as a specialized case where the weights are unity and hence the distance from the cluster centers are Euclidean. Recent research has shown that colloborative filters are well suited in representing fine-grained image data and give good results even with limited labels/annotations. So in this work, the proposed collaborative clustering is used to categorize fine-grained species images (birds and butterflies) and compare results with k-means and its variants as well as the highly cited DBSCAN algorithm, along with its recent variant DSets-DBSCAN.

The application of collaborative clustering is demonstrated to the task of categorizing fine-grained image data for species recognition. However, collaborative clustering is a generalized algorithm which may be applied to any similar tasks and with non-image data. Also it would be interesting to investigate incorporating automated evaluation of number of cluster centers into the framework. As expansion of this work, one can explore the case of retrieval of poorly labeled images from large datasets. Consider the case of a specialised problem like endangered species recognition requiring labeled images from domain experts. Large image repositories may have unlabeled data or unreliably labeled data from citizen scientists. In those cases the collaborative clustering algorithm may provide a more robust representation to retrieve similar images.

Chapter 8

Collaborative Distance Metric Learning

8.1 Chapter Summary and Linkage

The earlier chapters have demonstrated the use of CRC methods in the FGVC problem. In particular, CRC has been used for deep supervised learning and unsupervised clustering. CRC methods have also been developed for effective background compensation and representation of repeated local patterns. But in all these CRC formulations in previous chapters, always $l2$ norm was used. In fact in most of CRC literature, mostly Euclidean distance and sometimes Frobenius distantce have been invariably used. In this chapter, we replace the Eucledian distance by the Mahalanabis distance and learn an optimal distance metric from there. This adds an additional layer of abstraction and improves results.

8.2 Introduction

Any neural network inspired machine learning algorithm basically fits a function to given data using many parameters so as to learn discriminatory features from the input in an end-to-end manner. These features are then used to do the final discrimination operation using a standard distance metric. Though the network tries to learn the optimal feature space, it seldom tries to learn an optimal distance metric in the cost function, and hence misses out on an additional layer of abstraction (Mensink *et al.* (2013)).

The intuition for this work is that if the deep learned features are fed into a cost function with a distance metric which is also learned in tandem in an end-to-end manner, then it might help to further maximize the inter-class distance and help for such advanced classification tasks like fine-grained visual categorization. Deep convolutional networks are already proficient at recognizing base classes with sufficient data, but robust classification of sub-classes with fine-grained differences is still an open problem (Chai (2015)). Thus as the representative problem to demonstrate the method, fine-grained species recognition is chosen (Rodner *et al.* (2015). As the cost function, a collaborative representation classifier (CRC) is used as a natural progression from earlier chapters.

The main contribution is to learn a generic distance metric in the cost function of a deep network in tandem with the learned features in an end-to-end manner. An analytical derivation of the partial derivatives is provided. It is needed to optimise the distance metric and then back-propagate the gradients. The resulting system has wide generalisation since it is agnostic of the deep architecture and so can be used for any classification task. The method achieves state-of-the art results on three benchmark fine-grained species recognition datasets with the standard VGG-19 (Simonyan *et al.* (2014)) deep network. Standard publicly available models pre-trained on ImageNet (Russakovsky *et al.* (2015)) are used and fine-tuned on the three datasets, CUB Birds (Wah *et al.* (2011)), Oxford Flowers (Nilsback *et al.* (2007)) and Oxford-IIIT Pets (Parkhi *et al.* (2012)), for fair comparison and ready reproducibility.

8.3 Distance Metric Learned CRC (DML-CRC)

Most CRC methods, if not all, use the Eucledian l_2 norm or the Frobenius norm in the cost function. It is replaced here by a general Mahalanobis distance metric Σ which can be optimised analytically, giving:

$$J(\alpha, \Sigma) = (y - X\alpha)^T \Sigma^{-1} (y - X\alpha) + \lambda ||\alpha||_2^2 + \gamma ||\Sigma||_2^2 \tag{8.1}$$

Let X be the training set in some feature domain using the pre-trained deep model. Now, y is each incoming image in the same feature domain, being used to fine-tune the network. The aim is to find optimal Σ, α so as to minimize the cost function during the fine-tuning process.

Differentiating J with respect to α, keeping Σ constant:

$$\frac{\partial J}{\partial \alpha} = -2X^T \Sigma^{-1} (y - X\alpha) + 2\lambda\alpha = 0 \tag{8.2}$$

Differentiating J with respect to Σ, keeping α constant:

$$\frac{\partial J}{\partial \Sigma} = -\Sigma^{-1} (y - X\alpha)(y - X\alpha)^T \Sigma^{-1} + 2\gamma\Sigma = 0 \tag{8.3}$$

Solving the simultaneous equations 8.2 and 8.3, the new values of Σ and α are:

$$\Sigma = \frac{\Sigma^{-1} (y - X\alpha)(y - X\alpha)^T \Sigma^{-1}}{2\gamma} \tag{8.4}$$

$$\alpha = (X^T \Sigma^{-1} X + \lambda I)^{-1} X^T \Sigma^{-1} y \tag{8.5}$$

During a specific round of back-propagation, once the new Σ and α are set, the weights are then propagated back using the partial derivative with X as follows.

$$\frac{\partial J}{\partial X} = -2\Sigma^{-1} (y - X\alpha)\alpha^{-1} \tag{8.6}$$

For further details on similar back-propagation schemes, the reader may refer to the work by Wang *et al.* (2016).

8.4 Experimental Setup

In this section, the experimental setup is described: the datasets, chosen deep network and competing classifiers.

8.4.1 Benchmark Datasets

Three benchmark fine-grained species recognition datasets have been used. These are CUB birds, Oxford flowers and Oxford pets. These have been described in Chapter 2.

8.4.2 Training on VGG-19 Deep Convolutional Network

The standard VGG-19 deep convolutional network has been used from the Oxford Robotics group (Simonyan *et al.* (2014)). It has 19 layers, is trained on more than one million images from the ImageNet (Russakovsky *et al.* (2015)) dataset, and can classify up to 1000 object categories. The pre-trained VGG-19 model has been fine-tuned on the target datasets. For details of the training protocol, please directly refer to the benchmark work by Simon *et al.* (2015) on neural constellation activations. For fair comparison, the baseline models provided by Simon *et al.* (2015) in their GitHub repository have been used: pre-trained VGG-19 models on ImageNet and well as fine-tuned models on CUB Birds, Oxford Flowers and Oxford-IIIT Pets dataset using the CAFFE deep learning framework.

8.4.3 Competing Classifiers

Two CRC based and two non-CRC based methods have been used here for comparison. Note that all the methods have been used with VGG-19 features, but can be applied with any learned features.

CRC based deep network classifiers

There are many variants of CRC available; patch based CRC (PCRC) (Zhu *et al.* (2012)) as a major sub-class and probabilistic CRC (ProCRC) (Cai *et al.* (2016)) as a recent variant are chosen. Further details of these two methods may be found in Chapter 2.

Non-CRC based classifiers used with VGG-Net

Constellation models are chosen due to the popularity of the paper in fine-grained recognition and also because their pre-trained models have been used here directly for fair comparison. The other choice is the very recent paper on part attention models to compare against the state-of-the-art.

Constellation Neural Activations by Simon *et al.* (2015) finds activation patterns with the help of convolutional networks in a completely unsupervised manner (no annotation or bounding box) to identify discriminatory parts for fine-grained classification. This is one of the popular baseline works in fine-grained classification and also provides the pre-trained models used in the current work.

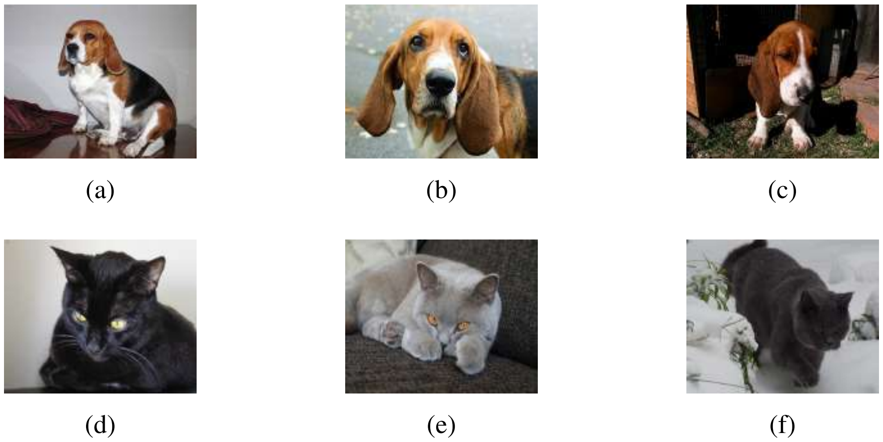
Object Part Attention Models by Peng *et al.* (2018) is a very recently published work in fine-grained recognition and can be considered state-of-the-art. It reports results on the same datasets used in this work with VGG-19 features. This work combines an object level and a part level attention models with a spatial constraint that preserves spatial patterns.

Table 8.1: Classification results of proposed DML-CRC versus competitors.

	CUB Birds	**Oxford Flowers**	**Oxford-IIIT Pets**
CRC [4]	75.24	91.83	83.30
PCRC [12]	76.95	93.06	84.88
ProCRC [13]	78.33	94.87	86.92
Constellation [11]	81.01	95.34	91.60
OPAM [14]	85.83	97.10	93.81
DML-CRC	**88.49**	**98.65**	**95.12**
DML-ProCRC	**89.95**	**99.33**	**96.58**

8.5 Experimental Results

For each dataset, experiments are conducted with five fold cross validation and percentage classification accuracies are presented in Table 8.1 with the accuracy of our method highlighted in bold. Among the CRC-based methods, basic CRC has the least accuracy and then there is an increase in the performance of the CRC variants. The proposed DML-CRC outperforms the original CRC and its variants comfortably. DML-CRC is also compared against two deep learning based methods, Constellation Model (Simon *et al.* (2015)) and OPAM (Peng *et al.* (2018)). The rationale of choosing these two



(a) (b) (c)

(d) (e) (f)

Figure 8.1: Qualitative results from the Oxford-IIIT Pets dataset. It has fine-grained image classes of cats and dogs. Row 1 gives a misclassification example: (a) Basset Hound, (b) Beagle (characterised by longer ear) (c) Beagle, misclassified as Basset Hound by the proposed DML-CRC and its competitors, due to partial obfuscation of the discriminating longer ear of Beagle (both dogs have similar colored patchy skin). Row 2 gives an example of correct classification: (d) Bombay Cat (e) British Shorthair (f) British shorthair correctly classified by proposed DML-CRC but misclassified by its competitors due to outlier black color of the cat (Bombay cat is generally black with narrower mouth while British Shorthair is gray with broader mouth).

particular methods, have been discussed in previous section. The proposed DML-CRC gives better results than both of these methods, thus establishing a new state-of-the-art. Fig. 8.1 presents qualitative results from the Oxford-IIIT Pets dataset.

It is important to note here that the original CRC cost function has been used deliberately, to emphasize the contribution of the distance metric learning. This is demonstrated by the fact that even with vanilla CRC, the state-of-the-art is outperformed albeit marginally in few cases. So it might be expected, that if a more recent version of CRC is used (like ProCRC), the margin of outperformance might increase. So the ProCRC cost function is plugged in, in place of the original CRC and the results are reported in Table 8.1, and as expected the performance improves further.

8.6 Conclusion

It has been shown that learning the distance metric for final discrimination of a convolutional network in an end-to-end manner enhances the performance of the system, keeping other factors like network architecture, data and training protocol constant. State-of-the-art results are achieved on several fine-grained recognition datsets, but the method is architecture agnostic and can be adopted for most visual classification tasks.

Chapter 9

Conclusion

9.1 Summary of Research

In this doctoral research, my main aim has been to advance the state-of-the-art in fine-grained visual categorization (FGVC) especially for those applications where number of training samples are limited. Recent developments in deep learning based vision systems have almost solved the task of base category recognition if ample data is available. But recognition of fine-grained categories with limited data and/or imbalanced classes remains an open problem. The challenge is compounded by the fact that in many cases the subtle differences in foreground objects are confounded by significant background variation across classes.

As the main genre of methods to tackle these challenges, I chose collaborative representation classifiers (CRC). Collaborative filters have been popular in recommender systems, where for example, they may be used to track user trends to suggest transactions in online shopping. They are found to be good at representing corroborative data for intelligent decision making. In computer vision, they have mainly been used in face recognition problems before this work. Human faces have distinguishing features but also have lot of similarities with each other. Thus it also poses the problem of high inter-class similarities and subtle inter-class differences. This inspired me to look into CRC methods as possible candidates for improving fine-grained recognition.

So in this work, I first demonstrated the robust performance of existing CRC methods in tackling FGVC tasks, taking species recognition as the representative problem. The reason for choosing species recognition was mainly that my base of research, New Zealand, due to its unique geographical location, has many endemic species due to evolution in isolation, quite a few of which are endangered. Thus the problem has real life significance in ecological conservation and bio-diversity of New Zealand. The project is directly aligned with the NZ Government national bio-diversity strategies 2020, particularly for building a Terrestrial and Freshwater Biodiversity Information System (TFBIS). However, the techniques developed in this work are also directly applicable to other similar FGVC tasks, like say, biomedical image analysis of rare pathologies.

9.2 Outcomes and Contributions

The present work has several major contributions. After establishing the alignment of CRC methods to the FGVC problem, I advanced the state-of-the art in CRC methods by proposing a probabilistic collaborative representation of image patches for robust background suppression. This helps in accounting for the problem of high randomised background variation between classes in FGVC, compared to subtle differences in foreground objects. Secondly, we incorporate the collaborative loss function into an end-to-end deep convolutional network. This collaborative convnet (CoCoNet) helps in establishing transfer learning protocol to learn small specialised fine-grained image datasets. We also introduce a collaborative unsupervised learning technique as a generalisation of the standard K-means clustering. Here the cluster centers are updated using the colloborative weighted mean and hence K-means can be viewed as a special case. We also show how learning an optimal distance metric can improve the performance of the collaborative cost function.

As part of the research, I also compiled several benchmark fine-grained image datasets for automated species recognition. The New Zealand Birds dataset was compiled at University of Otago, with contributions from the Department of Conservation, the Ornithological Society of New Zealand (Birds NZ), and the National Museum (Te Papa). The New Zealand lepidoptera (butterflies and moths) dataset was built with help from NZ Landcare Research. The Indian Birds dataset and Indian butterflies dataset were compiled through ongoing collaboration with the Indian Statistical Institute.

The work has resulted in several peer-reviewed publications through well known outlets in the field. The project won the Diane Campbell Hunt Memorial Award 2017. A part of the field work was conducted through funded visits to the International Computer Science Institute (ICSI), University of California, Berkeley, through the Kevin Novins Travelling Scholarship. Another part was concluded at the Computer Vision and Pattern Recognition (CVPR) Unit, Indian Statistical Institute.

References

Boom, B. J. *et al.* (2012). *Supporting Ground-Truth annotation of image datasets using clustering.* In Proc. Int. Conf. on Pattern Recognition (ICPR.

Bourdev, L. D. *et al.* (2009). *Poselets: Body part detectors trained using 3d human pose annotations.* In Proc. Intl. Conf. on Computer Vision (ICCV).

Branson, S. *et al.* (2014). *Improved Bird Species Recognition Using Pose Normalized Deep Convolutional Nets.* In Proc. British Machine Vision Conference (BMVC).

Cai, S. *et al.* (2016). *A Probabilistic Collaborative Representation based Approach for Pattern Classification.* In Proc. IEEE Intl Conf. on Computer Vision and Pattern Recognition (CVPR).

Calonder, M. *et al.* (2010). *Brief: Binary robust independent elementary features.* In Proc. European Conference on Computer Vision (ECCV)).

Chai, Y. (2015). *Advances in Fine-grained Visual Categorization.* Thesis: University of Oxford.

Chakraborti, T. *et al.* (2016). *Collaborative representation based fine-grained species recognition.* In Proc. IEEE Intl. Conf. on Imaging and Vision Computing New Zealand (IVCNZ).

Chaturvedi, A. *et al.* (2001). *K-modes Clustering*, Volume 18 (1): 35–55. Journal of Classification.

Chi, Y. *et al.* (2012). *Connecting the dots in multi-class classification: From nearest subspace to collaborative representation.* Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR).

Cornuejols, A. *et al.* (2018). *Collaborative clustering: Why, when, what and how*, Volume 39 (2): 81–95. Information Fusion.

Cortes, C. *et al.* (1995). *A support vector networks*, Volume 20(3):273–297. Machine Learning).

Dalal, N. *et al.* (2005). *Histograms of oriented gradients for human detection.* In Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR).

Demsar, J. (2006). *Statistical Comparisons of Classifiers over Multiple Data Sets*, Volume 7, 1-30. Journal of Machine Learning Research (JMLR).

Deng, J. *et al.* (2013). *Fine-Grained Crowdsourcing for Fine-Grained Recognition.* In Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR).

Ester, M. *et al.* (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise.* In Proc. Intl. Conf. on Knowledge Discovery and Data Mining.

Farrell, R. *et al.* (2011). *Birdlets: Subordinate categorization using volumetric primitives and posenormalized appearance.* In Proc. Intl. Conf. on Computer Vision (ICCV).

Freund, Y. *et al.* (1999). *A Short Introduction to Boosting*, Volume 14(5):771–780. Journal of Japanese Society for Artificial Intelligence.

Frobaand, B. *et al.* (2004). *Face Detection with the Modified Census Transform.* In Proc. IEEE Conf. on Face and Gesture Recognition.

Griffin, G. *et al.* (2007). *Caltech-256 object category dataset.* Technical Report 7694, California Institute of Technology.

He, D. *et al.* (1990). *Texture Unit, Texture Spectrum, And Texture Analysis*, Volume 28:509–512. IEEE Trans. on Geoscience and Remote Sensing.

Hillel, A. B. *et al.* (2007). *Subordinate class recognition using relational object models.* In Proc. Intl. Conf. on Neural Information Processing Systems (NIPS).

Ho, T. K. (1999). *Random Decision Forests.* In Proc. Intl. Conf. on Document Analysis and Recognition.

Horn, G. V. *et al.* (2015). *Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection.* In Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR).

Horn, G. V. *et al.* (2017). *The Devil is in the Tails: Fine-grained Classification in the Wild.* arXiv:1709.01450 [cs.CV]).

Hou, J. *et al.* (2016). *DSets-DBSCAN: A Parameter-Free Clustering Algorithm*, Volume 25 (7): 3182–3193. IEEE Trans. on Image Processing.

Huang, D. *et al.* (2011). *Local Binary Patterns and Its Application to Facial Image Analysis: A Survey*, Volume 41(6):765–781. IEEE Trans. on Systems, Man and Cybernetics- Part C.

Huang, G. B. *et al.* (2007). *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.* University of Massachusetts, Amherst, Technical Report.

Jabid, T. *et al.* (2010a). *Gender Classification using Local Directional Pattern (LDP).* In Proc. Intl. Conf. Pattern Recognition.

Jabid, T. *et al.* (2010b). *Local directional pattern (LDP)–A robust image descriptor for object recognition.* Proc. IEEE Intl. Conf. on Advanced Video and Signal Based Surveillance (AVSS).

Jaeger, J. *et al.* (2015). *Croatian Fish Dataset: Fine-grained classification of fish species in their natural habitat.* In Proc. of British Machine Vision Conference (BMVC).

Jain, A. K. *et al.* (1998). *Algorithms for Clustering Data.* Prentice-Hall.

Jun, B. *et al.* (2012). *Robust face detection using local gradient patterns and evidence accumulation*, Volume 45(9):3304–3316. Pattern Recognition.

Khosla, A. *et al.* (2011). *Novel dataset for Fine-Grained Image Categorization. First Workshop on Fine-Grained Visual Categorization (FGVC).* In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).

Kingma, D. P. *et al.* (2014). *Adam: A Method for Stochastic Optimization.* In Proc. Intl. Conf. on Learning Representations (ICLR).

Kornblith, S. *et al.* (2019). *Do better Imagenet models transfer better?* In Proc. Intl. Conf. on Computer Vision and Pattern Recognition (CVPR).

Krause, J. *et al.* (2013). *3D Object Representations for Fine-Grained Categorization.* IEEE Workshop on 3D Representation and Recognition, at ICCV (3dRR-13).

Krause, J. *et al.* (2014). *Learning Features and Parts for Fine-Grained Recognition.* In Proc. Intl. Conf. on Pattern Recognition.

Krause, J. *et al.* (2015). *Fine-grained recognition without part annotations.* In Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR).

Krizhevsky, A. *et al.* (2012). *ImageNet Classification with Deep Convolutional Neural Networks.* In Proc. Neural Information Processing Systems (NIPS).

Kumar, N. *et al.* (2012). *Leafsnap: A Computer Vision System for Automatic Plant Species Identification.* In Proc. European Conf. on Computer Vision (ECCV).

Lan, R. *et al.* (2017). *An extended probabilistic collaborative represen-tation based classifier for image classification.* In Proc. IEEE Intl. Conf.on Multimedia and Expo (ICME).

Larios, N. *et al.* (2011). *Stacked Spatial-Pyramid Kernel: An Object-Class Recognition Method to Combine Scores from Random Trees.* IEEE Workshop on Applications of Computer Vision.

Lazebnik, S. *et al.* (2004). *Semi-Local Affine Parts for Object Recognitionn.* In Proc. British Machine Vision Conference (BMVC).

Lazebnik, S. *et al.* (2005). *A Maximum Entropy Framework for Part-Based Texture and Object Recognition.* In Proc. IEEE Intl. Conf. on Computer Vision (ICCV).

Leutenegger, S. *et al.* (2011). *BRISK: Binary robust invariant scalable keypoints.* In Proc. IEEE International Conference on Computer Vision (ICCV).

Li, A. *et al.* (2017). *Zero-Shot Fine-Grained Classification by Deep Feature Learning with Semantics.* arXiv:1707.00785 [cs.CV].

Lin, T.-Y. *et al.* (2015). *Bilinear CNN Models for Fine-Grained Visual Recognition*. In Proc. IEEE Intl. Conf. on Computer Vision (ICCV).

Lin, T.-Y. *et al.* (2017). *Improved Bilinear Pooling with CNNs*. In Proc. British Machine Vision Conference (BMVC).

Lin, T.-Y. *et al.* (2018). *Bilinear Convolutional Neural Networks for Fine-Grained Visual Recognition*, Volume 40(6):1309–1322). IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI).

Liu, Z. *et al.* (2014). *Enhanced collaborative representation based classification*. In Proc. IEEE Intl. Conf. on Information and Automation (ICIA).

Lloyd, S. (1982). *Least squares quantization in PCM*, Volume 28 (2): 129–137. IEEE Trans. in Information Thoery.

Lowe, D. G. (1999). *Object recognition from local scale-invariant features*. In Proc. IEEE Intl. Conf. on Computer Vision (ICCV).

Lytle, D. *et al.* (2010). *Automated processing and identification of benthic invertebrate samples*, Volume 29(3), 867-874. Journal of the North American Benthological Society.

Maji, S. *et al.* (2013). *Fine-Grained Visual Classification of Aircraft*. arXiv:1306.5151 [cs.CV]).

Mallah, C. *et al.* (2013). *Plant Leaf Classification Using Probabilistic Integration of Shape, Texture and Margin Features*. In Proc. SPPRA.

Martinez, A. *et al.* (1998). *The AR Face Database*, Volume 24. CVC Technical Report.

Mensink, T. *et al.* (2013). *Distance-Based Image Classification: Generalizing to New Classes at Near-Zero Cost*, Volume 35(11), 2624-2637. IEEE Trans. Pattern Analysis and Machine Learning.

Ming, Y. *et al.* (2015). *Uniform Local Binary Pattern Based Texture-Edge Feature for 3D Human Behavior Recognition*, Volume 10(5):1–15. Plos One.

Muller, H. *et al.* (2010). *ImageCLEF: Experimental Evaluation in Visual Information Retrieval*. Springer.

Nilsback, M.-E. *et al.* (2006). *A Visual Vocabulary for Flower Classification*. In Proc. Computer Vision and Pattern Recognition (CVPR).

Nilsback, M.-E. *et al.* (2007). *Delving into the whorl of flower segmentation*. In Proc. British Machine Vision Conference (BMVC).

Nilsback, M.-E. *et al.* (2008). *Automated flower classification over a large number of classes*. In Proc. Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP).

Nilsback, M.-E. *et al.* (2009). *Delving Deeper into the Whorl of Flower Segmentation*, Volume 28(6):1049–1062. Image and Vision Computing.

Ojala, T. *et al.* (1994). *Performance evaluation of texture measures with classification based on Kullback discrimination of distributions*. In Proc. Intl Conf. on Pattern Recognition (ICPR).

Oliva, A. *et al.* (2001). *Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope*, Volume 42 (3): 145–175. Ineternational Journal of Computer Vision (IJCV).

Park, H. *et al.* (2009). *A simple and fast algorithm for K-medoids clustering*, Volume 18 (1): 35–55. Expert Systems with Applications.

Parkhi, O. M. *et al.* (2011). *The truth about cats and dogs*. In Proc. IEEE Intl. Conf. on Computer Vision (ICCV).

Parkhi, O. M. *et al.* (2012). *Cats and Dogs*. In Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR).

Peng, Y. *et al.* (2018). *Object-Part Attention Model for Fine-Grained Image Classification*, Volume 27(3), 1487-1500. IEEE Trans. Image Processing.

Rodner, E. *et al.* (2015). *Fine-grained Recognition Datasets for Biodiversity Analysis*. In Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR).

Rublee, E. *et al.* (2011). *ORB: an efficient alternative to SIFT or SURF*. In Proc. IEEE International Conference on Computer Vision (ICCV).

Russakovsky, O. *et al.* (2015). *ImageNet Large Scale Visual Recognition Challenge*. International Journal of Computer Vision (IJCV).

Schafer, J. B. *et al.* (2007). *Collaborative Filtering Recommender Systems*, Volume 4321 (291): 291–234. The Adaptive Web, Lecture Notes in Computer Science, Springer.

Simon, M. *et al.* (2015). *Neural Activation Constellations: Unsupervised Part Model Discovery with Convolutional Networks*. In Proc. IEEE Intl. Conf. on Computer Vision (ICCV).

Simonyan, K. *et al.* (2014). *Very deep convolutional networks for large-scale image recognition.* arXiv:1409.1556 [cs.CV].

Taigman, Y. *et al.* (2009). *Multiple One-Shots for Utilizing Class Label Information.* In Proc. British Machine Vision Conference (BMVC).

van de Weijer, J. *et al.* (2007). *Learning color names from real-world images*. Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR).

Vinh, N. X. *et al.* (2010). *Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance*, Volume 11: 2837–2854. The Journal of Machine Learning Research.

Wah, C. *et al.* (2011). *The caltech-ucsd birds-200-2011 dataset*. Computation and Neural Systems Technical Report, CNS-TR-2011-001.

Wang, J. *et al.* (2009). *Learning Models for Object Recognition from Natural Language Descriptions*. In Proc. British Machine Vision Conference (BMVC).

Wang, K. *et al.* (2016). *Dictionary Pair Classifier Driven Convolutional Neural Networks for Object Detection*. In Proc. Intl. Conf. on Computer Vision and Pattern Recognition (CVPR).

Welinder, P. *et al.* (2010). *Caltech-UCSD Birds 200*. California Institute of Technology. CNS-TR-2010-001. 2010.

Woodbury, M. A. (1950). *Inverting modified matrices*, Volume 42 (106): 336. Memorandum report.

Yang, M. *et al.* (2012a). *Relaxed collaborative representation for pattern classification*. In Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR).

Yang, S. *et al.* (2012b). *Unsupervised template learning for fine-grained object recognition*. In Proc. Intl. Conf. on Neural Information Processing Systems (NIPS).

Yu, A. *et al.* (2014). *Fine-Grained Visual Comparisons with Local Learning*. In Proc. Intl. Conf. on Computer Vision and Pattern Recognition (CVPR).

Yu, A. *et al.* (2017). *Semantic Jitter: Dense Supervision for Visual Comparisons via Synthetic Images.* In Proc. Intl. Conf. on Computer Vision (ICCV).

Zhang, L. *et al.* (2011). *Sparse Representation or Collaborative Representation: Which Helps Face Recognition?* In Proc. IEEE Intl. Conf. on Computer Vision (ICCV).

Zhang, N. *et al.* (2012). *Pose pooling kernels for sub-category recognition.* In Proc. Intl. Conf. on Computer Vision and Pattern Recognition.

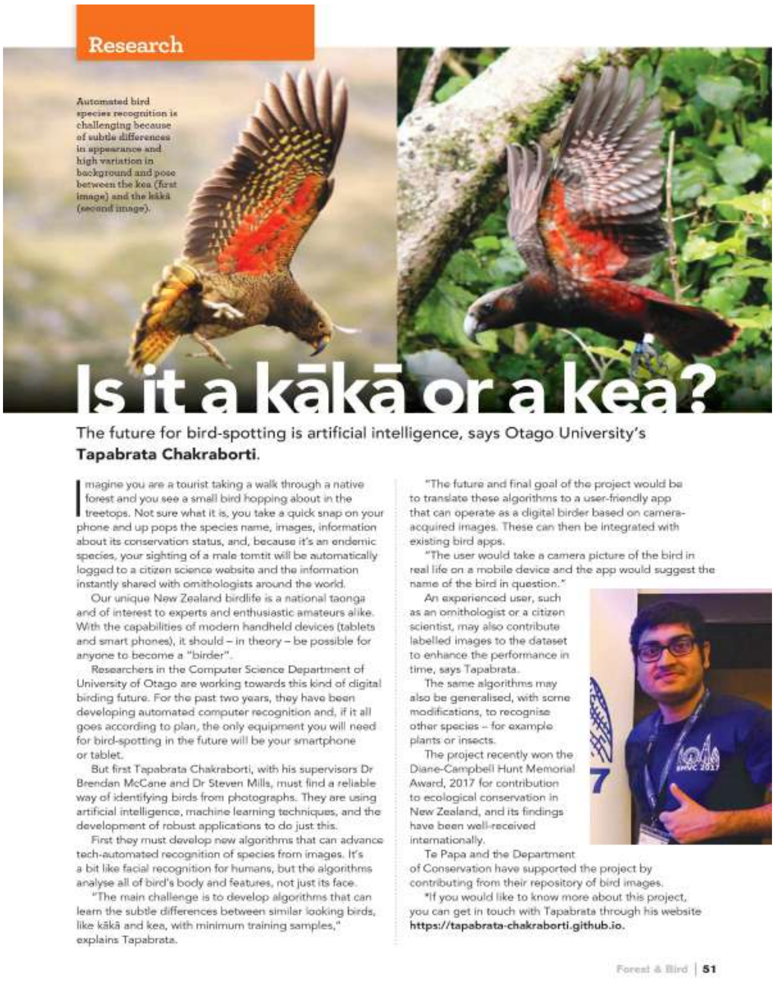
Zhang, N. *et al.* (2013). *Deformable part descriptors for fine-grained recognition and attribute prediction.* In Proc. Intl. Conf. on Computer Vision.

Zhao, J. *et al.* (2014). *Kernel collaborative representation for face recognition.* In Proc. Intl. Conf. on Signal Processing (ICSP).

Zhu, P. *et al.* (2012). *Multi-scale patch based collaborative representation for face recognition with margin distribution optimization.* In Proc. European Conf. on Computer Vision (ECCV).

Appendix A

End Plate: Magazine Article



Research

Automated bird species recognition is challenging because of subtle differences in appearance and high variation in background and pose between the kea (first image) and the kākā (second image).

Is it a kākā or a kea?

The future for bird-spotting is artificial intelligence, says Otago University's **Tapabrata Chakraborti**.

Imagine you are a tourist taking a walk through a native forest and you see a small bird hopping about in the treetops. Not sure what it is, you take a quick snap on your phone and up pops the species name, images, information about its conservation status, and, because it's an endemic species, your sighting of a male tomtit will be automatically logged to a citizen science website and the information instantly shared with ornithologists around the world.

Our unique New Zealand birdlife is a national taonga and of interest to experts and enthusiastic amateurs alike. With the capabilities of modern handheld devices (tablets and smart phones), it should – in theory – be possible for anyone to become a "birder".

Researchers in the Computer Science Department of University of Otago are working towards this kind of digital birding future. For the past two years, they have been developing automated computer recognition and, if it all goes according to plan, the only equipment you will need for bird-spotting in the future will be your smartphone or tablet.

But first Tapabrata Chakraborti, with his supervisors Dr Brendan McCane and Dr Steven Mills, must find a reliable way of identifying birds from photographs. They are using artificial intelligence, machine learning techniques, and the development of robust applications to do just this.

First they must develop new algorithms that can advance tech-automated recognition of species from images. It's a bit like facial recognition for humans, but the algorithms analyse all of bird's body and features, not just its face.

"The main challenge is to develop algorithms that can learn the subtle differences between similar looking birds, like kākā and kea, with minimum training samples," explains Tapabrata.

"The future and final goal of the project would be to translate these algorithms to a user-friendly app that can operate as a digital birder based on camera-acquired images. These can then be integrated with existing bird apps.

"The user would take a camera picture of the bird in real life on a mobile device and the app would suggest the name of the bird in question."

An experienced user, such as an ornithologist or a citizen scientist, may also contribute labelled images to the dataset to enhance the performance in time, says Tapabrata.

The same algorithms may also be generalised, with some modifications, to recognise other species – for example plants or insects.

The project recently won the Diane-Campbell Hunt Memorial Award, 2017 for contribution to ecological conservation in New Zealand, and its findings have been well-received internationally.

Te Papa and the Department of Conservation have supported the project by contributing from their repository of bird images.

"If you would like to know more about this project, you can get in touch with Tapabrata through his website **https://tapabrata-chakraborti.github.io.**

Forest & Bird | 51

Figure A.1: Featured in Spring 2018 edition of Forest and Bird Magazine.