Collaborative Representation based Fine-grained Species Recognition

Tapabrata Chakraborti, Brendan McCane, Steven Mills
Department of Computer Science
University of Otago
Dunedin, New Zealand
Email: tapabrata;mccane;steven@cs.otago.ac.nz

Umapada Pal
Computer Vision and Pattern Recognition Unit
Indian Statistical Institute
Kolkata, India
Email: umapada@isical.ac.in

***Abstract*—Fine-grained Visual Categorization (FGVC) is an open problem in Computer Vision due to subtle differences between categories. The present paper demonstrates that Collaborative Representation based Classification (CRC) can address this problem successfully. Instead of the traditional discriminative approach of classification, CRC takes a co-operative approach by representing the query image as a weighted collaboration of training images across all classes in the feature space. The superior performance of CRC compared to some other modern classifiers including SVM is shown in this work using several popular descriptors like GIST+Color, SIFT and CNN features with Species Recognition chosen as the representative FGVC problem. Besides experiments on the Oxford 102 Flowers and CUB200-2011 Bird benchmarks, the present work also introduces a new challenging dataset NZBirds v1.0 with 600 images of 30 New Zealand endemic and native bird species.**

I. Introduction

A. Problem Background

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 computer vision. The research, primarily focussed on base category classification over the past couple of decades, along with the exponential increase in capacity and power of computing machines, has resulted in the development of sophisticated automated vision systems which can robustly detect and categorize objects with sufficient visual differences, even from natural scene images. In fact, as per the knowledge of the authors, with rapid advancement in neural network based machine learning (particularly deep convolutional nets), state of the art vision systems have recently achieved high accuracy in recognising base categories (Eg. recognition of animal images as members of broad classes like dogs, cats, horses, etc) even in large challenging datasets.

In the past 5 years, a new and challenging area of research has gained popularity in machine vision, that of recognizing sub-categorical object classes (Eg. identification of type/species of birds/fish/insects from images) with fine grained differences in attributes. Fine Grained Visual Categorization (FGVC) [1] is currently one of the open problems of computer vision as it poses certain interesting challenges.

978-1-5090-2748-4/16/$31.00

A case in point is automated species recognition [2], 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. Furthermore, there can be considerable intra-class pose variation which may involve significant changes in object contour (eg. Shape change of same bird species between flight vs. roosting images)

The above arguments are further illustrated in Fig 1. 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. Moreover, Kaka are a vulnerable species and the Kea is endangered. Thus for threatened species, the datasets available may have limited training images. The complexity of the problem is reflected in the experimental results where it is illustrated in further details.

B. Brief Literature Survey

Most current research in FGVC either employ the filter approach or the feature learning approach, which encompass the same methodological philosophy that is traditionally used to tackle any object classification problem in machine vision.

The filter approach in FGVC mainly focuses on representation of discriminative local parts to effectively model the fine inter-class differences. These parts are then modelled by local feature descriptors (SIFT, SURF, etc). The feature vectors thus obtained are finally used for classification with possible clustering to form bag of words and multi-scaling using spatial pyramids.

The second approach [4] in FGVC deals with automated feature learning along with classification in a single framework like the Convolutional Neural Network (CNN).

The above methods however have certain limitations which may limit their performance in some practical scenarios.

1) Part Localisation: The filter approach tries to identify discriminating local parts either by automated localisation or human annotation. Automated annotation of discriminating local parts is an involved problem in itself mainly due to the low inter-class differences in FGVC.

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

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

Fig. 1. Few sample images of NZ endemic birds Kaka [3(a)-3(d)] and Kea [3(e)-3(h)] 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).

Human annotation of suitable local parts in case of FGVC may require domain expertise especially in case of such specific problems like bird species recognition (ornithological expertise). In fact, recent research has focused on developing robust methods that do not require part identification.

2) Limited Data: For FGVC, the number of suitable parts are usually few due to low inter-class differences. This may lead to strong sparsity in feature description especially when number of training samples per class is low.

The feature learning approach involves state of the art deep architecture learning frameworks like CNN. These networks are data intensive and require large number of training samples per class for effective learning.

Current FGVC methods work well for datasets with sufficient training samples. However, availability of large image sets may be difficult for certain interesting FGVC problems like endangered/vulnerable species recognition for biodiversity analysis, where the number of quality images may be limited.

C. Proposed Approach

Collaborative Representation based Classification (CRC), first introduced in human face recognition by Zhang et al. [6], represents the query image as a weighted collaboration of features over all classes. Thus the CRC framework in face recognition exploits the fact that human faces for different individuals (classes) share some common features which may be exploited to overcome the problem of sparsity in case of few training samples.

Since then many interesting modifications and improvements of the CRC model have been proposed. These enhancements include Optimised CRC [7], Multi-scale Patch-based CRC [8], Relaxed CRC [9], and Probabilistic CRC [10]. A few of these that have been used in this work are discussed in brief in the following section.

The main philosophy of CRC is radically different to the traditional filter or feature learning approach. The traditional approaches in visual classification focus on discriminative categorization, whereas CRC uses similarities between classes to achieve a co-operative representation. The CRC framework has been thoroughly tested by many researchers on the face recognition problem and has been shown to perform consistently.

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. But as per the knowledge of the present authors, there has not been focussed research to ascertain the appropriateness of CRC for FGVC.

II. Methodology

A. Collaborative Representation based Classification (CRC)

The mathematical framework for CRC [6] is described in brief here. 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:

$$\hat{\alpha} = \arg\min_{\alpha} (\|y - X\alpha\|_2^2 + \lambda\|\alpha\|_2^2) \quad (1)$$

where $\hat{\alpha} = [\hat{\alpha}_1, \ldots, \hat{\alpha}_c] \in \Re^N \mid \hat{\alpha}_i \in \Re^{n_i}$ is the reconstruction matrix corresponding to class i.

A least-squares derivation yields the optimal solution as:

$$\hat{\alpha} = (X^T X + \lambda I)^{-1} X^T y \quad (2)$$

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

$$r_i(y) = \frac{\|y - X_i\hat{\alpha}_i\|_2^2}{\|\hat{\alpha}_i\|_2^2} \ \forall i \in 1, \ldots, c \quad (3)$$

The final class of test sample y is thus given by

$$C(y) = \arg\min_{i} r_i(y) \quad (4)$$

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,

$$X^{\#} = (X^T X + \lambda I)^{-1} X^T \quad (5)$$

Then the GCV cost function is given by:

$$G(\lambda) = \frac{\|y - X\alpha_\lambda\|_2^2}{trace(I - XX_\lambda^{\#})^2} \quad (6)$$

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.

1) Optimized Collaborative Representation (CROC): Chi and Porikli [7] 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:

$$r_i^{CR}(y) = \|y - X_i\alpha_i\|_2^2 \ \forall i \in 1, \ldots, c \quad (7)$$

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:

$$r_i(y) = r_i^{NS}(y) + \lambda r_i^{CR}(y) \quad (8)$$

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

$$r_i(y) = (1 - \lambda) r_i^{NS}(y) + r_i^{CR}(y) \quad (9)$$

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

2) Multi-scale Patch-based Collaborative Representation (MPCRC): Zhu et al. [8] 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 MPCRC becomes:

$$\hat{p}_j = \arg\min_{p_j} \|y_j - M_j p_j\|_2^2 + \lambda\|p_j\|_2^2 \quad (10)$$

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:

$$C(y_j) = \arg\min_{k} r_{jk}(y) \quad (11)$$

where

$$r_{jk} = \frac{\|y_j - M_{jk}\hat{p}_{jk}\|_2^2}{\|\hat{p}_{jk}\|_2^2} \ \forall i \in 1, \ldots, c \quad (12)$$

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

3) Relaxed Collaborative Representation (RCRC): Yang et al. [9] 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

$$\hat{\alpha} = \arg\min_{\alpha, w} \|y - X\alpha\|_2^2 + \lambda\|\alpha\|_2^2 + \tau w\|\alpha - \bar{\alpha}\|_2^2 \quad (13)$$

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.

4) Probabilistic Collaborative Representation (ProCRC): Cai et al. [10] 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):

$$\arg\max_{i} \text{Prob}[C(y)] = \arg\max_{i} e^{-\|X\hat{\alpha} - X_i\hat{\alpha}_i\|_2^2} \quad (14)$$

B. Datasets

Experiments have been performed on 3 separate datasets, two of which are among the well established benchmarks for FGVC for bird species (CUB200-2011 Birds) and flower species (Oxford 102) recognition. The present work also introduces a new NZ bird species dataset (NZBirds v1.0) developed by the authors. The images were used as is without bounding boxes or part annotations.

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

Fig. 2. Sample images of the critically endangered NZ endemic brid Kakapo from the new NZBirds v1.0 benchmark dataset

1) Caltech-UCSD (CUB200-2011) Birds: It contains 11,788 images of 200 bird species with around 30 training samples for each species [11]. The main challenge of this dataset is considerable variation and confounding features in background information compared to subtle inter-class differences in birds.

2) Oxford 102 Flowers: It contains 8,189 images from 102 categories, with each category having at least 40 images [12]. The main challenge of this dataset is considerable variation in scale, pose and lighting conditions within each category.

3) NZBirds (v1.0): It has been developed by the Vision Group at the Department of Computer Science, University of Otago, NZ in close collaboration with Te Papa (National Museum of New Zealand) and Birds NZ (Ornithological Society of NZ). This is also an international research project with the CVPR Unit, Indian Statistical Institute.

Besides being an additional benchmark FGVC database, the NZBirds dataset is envisioned to act as a unique evaluation resource since it comprises of mostly endemic and native NZ bird species. It currently contains 600 images from 30 species and is subject to expansion in the near future.

The images in Fig 1 are of the NZ endemic bird species Kaka and Kea, taken from the NZBirds dataset. A few additional sample images of Kakapo are shown in Fig 2. New Zealand, due to its unique geo-climatic niche, is home to many endemic and native species, especially birds. Unfortunately, many of these species are considered threatened or endangered.

Thus the new NZBirds dataset adds a valuable resource to the FGVC community subject to further expansion and compilation. It is envisioned to address new interesting FGVC problems like bio-diversity analysis of endangered species using few training data.

C. Features

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

1) GIST+Color: Global Invariant Scale Transform (here referred as GIST) is a low level global feature that describes the spatial envelope of the image using directional properties [13]. Color descriptor [14] converts the image to color names 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.

2) SIFT: VLFeat has been used to extract SIFT features for comparison [15]. The chosen patch size is 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 and the final feature dimension for each image is 5120. The MATLAB API has been used for this.

3) CNN features: CNN features have been extracted using the VGG-19 framework [16]. The activations of the penultimate layer are used as local features, extracted over five scales of $\{2^{-1}, 2^{-0.5}, 2^{0}, 2^{0.5}, 2^{1}\}$. All local features are concatenated irrespective of scale and location. The final feature dimension is 4096 for all datasets with $\ell 2$ normalisation. The CNN features thus obtained for each dataset are then arranged as mat files and used for different collaborative representation based classifiers on the MATLAB platform. Again the MATLAB API has been used to extract the CNN features.

D. Classifiers

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

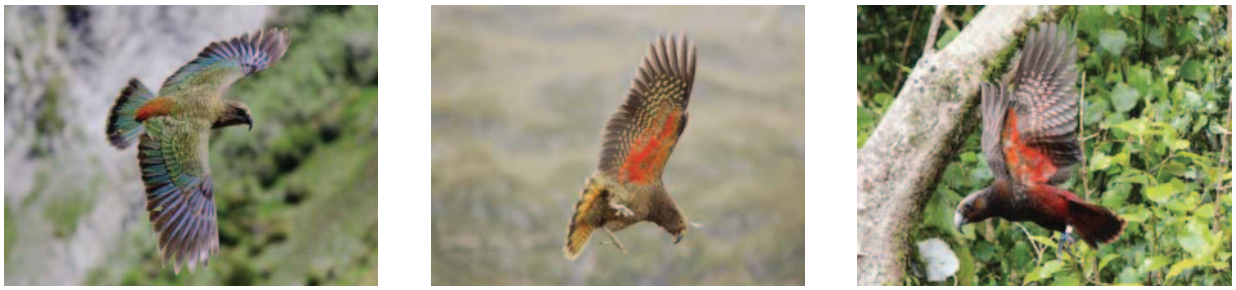
1) CRC based: A family of Collaborative Representation based classifiers have been utilised including the original CRC implementation along with some of its recent enhancements CROC, MPCRC, RCRC and Pro-CRC.

2) Softmax and SVM based: Probabilistic regression based Softmax classifier has been used along with linear and χ^2 kernel based Support Vector Machines (SVM), with parameter optimization as in [10]. The binary classifiers have been used in one-vs-all format to achieve mult-class categorization.

3) 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.

III. Results and Discussion

Average recognition accuracies in percentage are presented over 5-fold cross-validation for CUB200-2011 Birds dataset (Table I), Oxford 102 Flowers dataset (Table II) and NZBirds dataset (Table III).



(a) (b) (c)

Fig. 3. presents images of Kea [3(a)-(b)] and Kaka [3(c)] in flight to illustrate one of the challenges of fine-grained species recognition problem. The image of Kea (b) was misclassified as Kaka (c) due to similarity of ventral side of wings between Kea and Kaka as well as the dissimilarity between the dorsal of and ventral sides of the wing of Kea as seen from (a) and (b).

TABLE I
RECOGNITION ACCURACY FOR CUB200-2011 BIRDS DATASET

	GIST+Color	SIFT	VGG-19
Softmax	7.5	8.2	72.1
SVM	9.2	10.2	75.4
Kernel SVM	9.8	10.5	76.6
NSC	9.1	8.4	74.5
SRC	8.8	7.7	76.0
CRC	9.3	9.4	76.2
CROC	9.5	9.1	76.2
MPCRC	9.9	9.7	76.9
RCRC	10.0	9.5	77.4
Pro-CRC	10.4	9.9	78.3

TABLE II
RECOGNITION ACCURACY FOR OXFORD102 FLOWERS DATASET

	GIST+Color	SIFT	VGG-19
Softmax	45.7	46.5	87.3
SVM	50.5	50.1	90.9
Kernel SVM	51.7	51.0	92.2
NSC	45.4	46.7	90.1
SRC	48.1	47.2	93.2
CRC	47.3	49.9	93.0
CROC	48.8	49.4	93.1
MPCRC	49.7	50.3	94.3
RCRC	50.6	51.0	93.6
Pro-CRC	52.4	51.2	94.8

Several interesting observations may be made from the results reported in Table I. 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 the 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 for any classifier 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, MPCRC and Pro-CRC consistently outperform SVM.

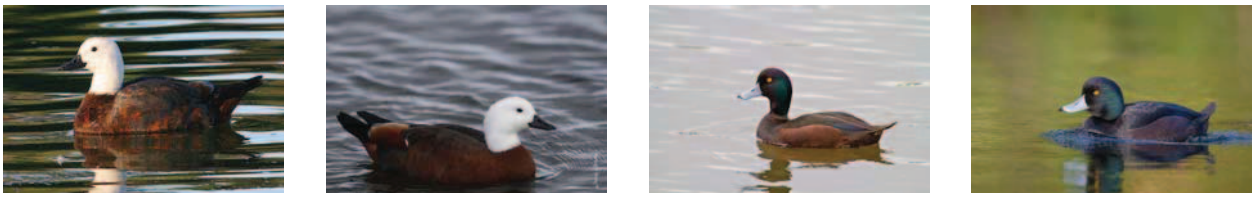
The Flower Species Recognition results, as reported in Table II, follow mostly the trends established for Bird Species Recognition in Table I, thus further supporting earlier observations. Like before, here also it may be observed that there is a consistent increase in accuracy for all features with advancement in CRC based classifiers, with Probabilistic CRC once again giving the best results, outperforming SVM in each case. CNN based features again significantly outperform SIFT and GIST+Color descriptors, thus once again demonstrating that deep neural networks are particularly effective.

TABLE III
RECOGNITION ACCURACY FOR NZ BIRDS DATASET

	GIST+Color	SIFT	VGG-19
Softmax	60.2	61.5	79.7
SVM	63.4	63.8	81.4
Kernel SVM	65.7	65.4	82.5
NSC	59.9	60.2	82.0
SRC	60.6	60.9	83.6
CRC	60.3	60.5	82.8
CROC	62.0	62.1	83.7
MPCRC	63.9	63.6	85.2
RCRC	63.4	63.7	86.4
Pro-CRC	66.1	66.9	89.8

Similar trends are also reflected in Table III which tabulates results for the newly presented NZBirds dataset, thus further supporting the claim of this paper.

An example of misclassification is shown in Fig 3. It is evident from Fig 3(a) and 3(b) that the dorsal (bluish) and ventral (reddish) wingspans of the Kea are quite different. 3(b) is much closer to 3(c) though the latter image is that of a Kaka. In fact, in our experiments all the combinations, except VGG-



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

Fig. 4. (a)-(b) are of Paradise Shelduck and (c)-(d) are of NZ Scaup. Both have similar Duck like features, main difference being body color. Experiments show GIST+Color features perform better than SIFT for these two classes, whereas on average over all classes, SIFT performs better. This illustrates the significant effect of choice of features on performance of CRC and the other classifiers

19 features with Pro-CRC based classification, misclassified 3(b) to be a Kaka, where in fact, it is a Kea.

IV. Conclusion

Though the superior performance of the CRC based methods have been experimentally demonstrated, there remains the possibility of premature generalization. This is mainly due to the presence of a myriad of other possible features (like HOG, BRIEF/BRISK, SURF, etc) as well as alternative classifiers like ANN, RBF, etc. Thus it is impossible to predict the absolute superiority or effectiveness of CRC for FGVC problems, though the results obtained are quite encouraging.

Fig 4. emphasizes the necessity of future investigation into the effect of the adopted features on the performance of CRC. Fig 4(a) and 4(b) are those of the Paradise Shelduck. Fig 4(c) and 4(d) are those of the NZ Scaup. Both have similar almost indistinguishable duck features, except for difference in color. For these 2 classes, GIST+Color performed better than SIFT, maybe due to the presence of the color features. This is in contrast to the overall better performance of SIFT over GIST+Color features on average over all classes.

This shows that the choice of features has a significant effect on CRC performance. In addition to the experimental validation provided in this work, a mathematical insight into the possible alignment between CRC and FGVC would help in generalisation.

Lastly, since the VGG based features have yielded superior results than the other descriptors for all classifiers, a natural extension of work would be a complete VGG based CNN architecture for training. The present work has not included this since the approach has been to show the strength of CRC to boost filter based methods rather than in feature learning schemes.

Acknowledgment

The authors acknowledge the support of Te Papa (National Museum of New Zealand) and Birds NZ (Ornithological Society of New Zealand) who kindly provided access to their repository of images of NZ endemic birds. We aim to publish the NZBirds dataset along with the details of the individual contributing photographers in our project webpage. Finally, thanks to Mr. Prabir Mondal for his help in consolidating and housekeeping the NZBirds dataset.

References

[1] Y. Chai. Advances in Fine-grained Visual Categorization. University of Oxford, 2015.

[2] E. Rodner, M. Simon, G. Brehm, S. Pietsch, J.-W.Wgele, and J. Denzler. Fine-grained Recognition Datasets for Biodiversity Analysis. In *Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[3] J. Deng, J. Krause, and F.-F.Li. Fine-Grained Crowdsourcing for Fine-Grained Recognition. In *Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[4] J. Krause, T. Gebru, J. Deng, L.-J.Li,and F.-F. Li. Learning Features and Parts for Fine-Grained Recognition. In *Proc. IEEE Intl. Conf. on Pattern Recognition (ICPR)*, 2014.

[5] J. Krause, H. Jin, J. Yang, and F.-F.Li. Fine-grained recognition without part annotations. In *Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[6] L. Zhang, M. Yang, and X. Feng. Sparse representation or collaborative representation: Which helps face recognition? In *Proc. IEEE Intl. Conf. on Computer Vision (ICCV)*, 2011.

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

[8] P. Zhu, L. Zhang, Q. Hu, and Simon C.K. Shiu. Multi-scale patch based collaborative representation for face recognition with margin distribution optimization. In *Proc. European Conf. on Computer Vision (ECCV)*, 2012.

[9] M. Yang, L. Zhang, D. Zhang, and S. Wang. Relaxed collaborative representation for pattern classification. In *Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[10] S. Cai, L. Zhang, W. Zuo, and X. Feng. A probabilistic collaborative representation based approach for pattern classification. In *Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[11] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. http://www.vision.caltech.edu/visipedia/CUB-200-2011.html.

[12] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *IEEE Sixth Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, page 722729, 2008.

[13] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision (IJCV)*, 2001.

[14] J. van de Weijer, C. Schmid, J. Verbeek, and D. Larlus. Learning color names from real-world images. In *Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[15] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *Proc. of Intl. Conf. on Multimedia (ACM)*, 2010.

[16] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. of Intl. Conf. on Learning Representations (ICLR)*, 2014.