FINE-GRAINED BIRD SPECIES RECOGNITION VIA HIERARCHICAL SUBSET LEARNING

ZongYuan Ge[†‡], Chris McCool[†‡], Conrad Sanderson[◇], Alex Bewley[‡], Zetao Chen[†‡], Peter Corke[†‡]

[†]Australian Centre for Robotic Vision, Brisbane, Australia
[‡]Queensland University of Technology, Brisbane, QLD 4000, Australia
[◇]NICTA, PO Box 10522, Adelaide St, Brisbane, QLD 4001, Australia

ABSTRACT

We propose a novel method to improve fine-grained bird species classification based on hierarchical subset learning. We first form a similarity tree where classes with strong visual correlations are grouped into subsets. An expert local classifier with strong discriminative power to distinguish visually similar classes is then learnt for each subset. On the challenging Caltech200-2011 bird dataset we show that using the hierarchical approach with features derived from a deep convolutional neural network leads to the average accuracy improving from 64.5% to 72.7%, a relative improvement of 12.7%.

Index Terms— fine-grained classification, subset clustering

1. INTRODUCTION

Fine-grained image classification is a challenging computer vision problem. Distinct from general object classification which aims to find the correct overall category such as a bird or dog, fine-grained image classification aims to identify the particular sub-category of a given category [1, 13, 14]. As an example, for an overall category of *bird* we wish to discriminate between various sub-categories with similar appearance, as shown in Fig. 1. In fact, bird classification is an area of particular interest within fine-grained image classification [3, 5, 7, 8].

Recent work in bird classification has concentrated on the issues of pose and view-point variation by finding local parts or extracting normalised features. Several authors have examined ways in which locating the parts of the birds (and other animals) can be used to improve classification [4, 5, 14]. Extracting pose-normalised features has been another popular approach [18] and is the basis for the deep convolutional bird classification system of Donahue et al. [6].

Aside from the issue of pose and view-point changes, a major challenge for any fine-grained classification approach is how to distinguish between classes that have high visual correlations. In Fig. 1 it can be seen that the *hooded oriole* and *baltimore oriole* species are visually very similar, but can be easily differentiated from the *black throate* species. This visual similarity was exploited by Berg and Belhumeur [2] to build a similarity tree that divides visually similar classes into subsets, which in turn was used to help derive a visual field guide. However, the application of the similarity tree to automatic classification for bird images has not been explored.

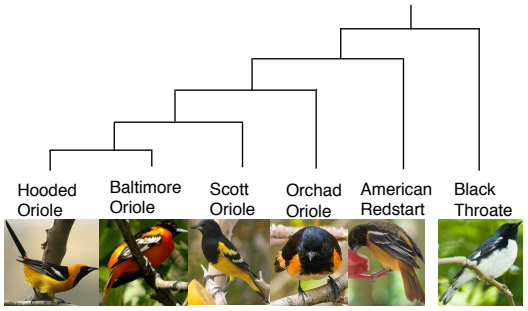

Fig. 1: One subset of the similarity tree of Berg and Belhumeur [2], built from the visual similarity matrix based on part-based one-vs-one features [3]. Species from the same node (eg. oriole) appear very similar to each other in terms of overall color and texture.

Inspired by the similarity tree of Berg and Belhumeur, we propose a hierarchical approach for fine-grained image classification. Our hierarchical approach begins by clustering visually similar classes before learning separate expert local classifiers which focus on discriminating the similar classes.

As a baseline for bird classification, we use the recently proposed deep convolutional feature approach of Donahue et al. [6]. This approach first performs part detection and pose normalisation, followed by extracting local features. The part detection and pose normalisation is achieved by using the deformable part descriptors model [18] on local parts which have been extracted using a pre-trained deep convolutional neural network (DCNN) learned from ImageNet [12]. Features obtained from the 6-th layer (fc-6) of the DCNN are used which are then classified using a linear regression approach.

The paper is continued as follows. In Section 2 we present our proposed hierarchical classification system in detail. Section 3 is devoted to a comparative evaluation with several recent methods on the task of fine-grained bird classification. Conclusions and possible future avenues of research are given in Section 4.

2. PROPOSED HIERARCHICAL CLASSIFICATION

Our proposed approach to hierarchical fine-grained image classification consists of two steps. First, the system performs a coarse classification to assign the test sample to the most likely subset k using a *subset selector*. Each subset consists of visually similar species; the subsets are automatically generated using a similarity tree. Secondly, if the confidence of the *subset selector* is sufficiently high, for each chosen subset k, fine-grained classification is performed using a local classifier *LocalSVM*$_k$. Each *LocalSVM*$_k$ has been trained to differentiate between the visually similar species belonging to this subset. If the confidence is low, a one-vs-all *GlobalSVM* classifier is used. An overview of the system can be seen in Fig. 2. The details of each component are explained in the following subsections.

2.1. Automatically Obtaining the Similarity Tree

There are two main issues with using the similarity tree of Berg and Belhumeur [2] to derive our hierarchical structure. First, it has a deep hierarchical structure of up to 17 layers and in this work we wish to explore the potential for a shallow structure of just 2 layers. Second, we want to generate the hierarchical structure in a fully automatic manner. In contrast, the similarity tree in [2] is learned from features obtained from manual part annotation which may not always be possible or desirable.

Our aim is to derive a similarity tree that groups all of the J_i samples of class i to the same subset (cluster), as well as grouping together similar classes. To do this we first obtain discriminant features by applying linear discriminant analysis (LDA) [15] to DCNN-based features (see Section 3 for more details). We use discriminant features as they will aid in having samples from the same class being assigned to the same subset (cluster). Using these discriminant features we then learn the similarity tree by performing k-means clustering.

An issue with this automatically derived similarity tree is that not all of the samples from a class are assigned to just one cluster (subset). To deal with this issue we use the result of k-means as an initial split of classes into subsets. We then determine the subset s_k which contains the majority of its samples for each class i and declare this as being the subset responsible for that class. Using this assignment of classes to subsets, we then learn a discriminative *subset selector* so that we can more accurately assign a sample to its correct subset.

2.2. Subset Selectors

We train a discriminative subset selector to minimise the number of mis-assignments of species to its subset. The k-th subset is assigned I_k classes, and so the subset selector *Selector*$_k$ is trained to correctly assign all the samples from these I_k classes. The positive samples to train the subset selector consist of all the training samples for the I_k classes and the negative samples are the remaining training samples.

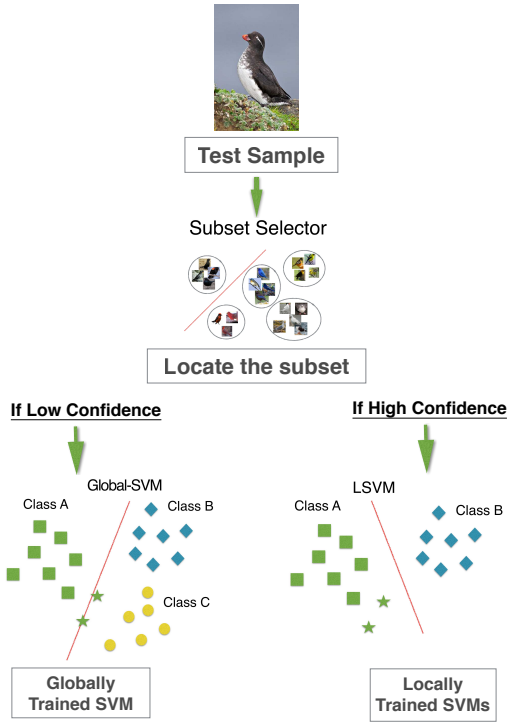

Fig. 2: An overview of the proposed hybrid system (the green stars are test samples for class A). A test image is first coarsely classified into a subset, and receives a confidence on the classification. If the confidence is higher than a predefined threshold, a local classifier *LocalSVM* specific to the chosen subset is used to make the final decision. Otherwise, a one-vs-all SVM (termed *GlobalSVM*) is used to make the decision.

In total, K subset selectors *Selector*$_{1..K}$ are trained, one for each subset of the hierarchical structure. These subset selectors are trained using a probabilistic SVM as this provides the probability that a sample belongs to a particular subset. This allows us to mitigate potential errors by incorporating this knowledge in the next step.

2.3. Local Expert Classifier Learning

Let $S = \{s_k\}_{k=1}^{K}$ denote the K subsets learned by the hierarchical clustering. An expert classifier (SVM) is then learned for each subset s_k which we term *LocalSVM*$_k$. Each *LocalSVM*$_k$ is a linear multi-class SVM. This is different to the classical one-versus-all approach because only the I_k classes assigned to the subset are used to train each *LocalSVM*.

2.4. Hybrid Decision System

The accuracy of the proposed system is dependent on the accuracy of the assignment of a test sample to the correct subset of our hierarchy. If the wrong subset is chosen then we have no way to recover and a mis-classification will occur. To alleviate this issue, we present a hybrid decision system which makes use of the classical global classifier, *GlobalSVM*, as well as our local classifier, *LocalSVM*.

Our hybrid decision system makes use of the probability from the subset selector to combine *GlobalSVM* and the *LocalSVM*. It uses the locally trained classifier (*LocalSVM*$_k$) only when the confidence of the subset selector is greater than a pre-defined threshold τ. In all other cases the classical *GlobalSVM* trained with all birds species is used to make the classification decision.

3. EXPERIMENTS

We evaluate our approach on the Caltech birds dataset (CUB200-2011) [17]. It contains 11,788 images from 200 bird species in North America. Each species has approximately 30 images for training and 30 for testing. Each image comes with an annotated bounding box around the object of interest (the bird), as well as annotations for many constituent parts of the object.

The feature vectors that we use throughout our experiments are the DCNN features (DeCAF) trained from ImageNet [12]. We fine-tune these features, using Caffe [10], for the task of bird classification by replacing the final output layer (for the $1,000$ classes of ImageNet) with a 200 class layer for bird species. We then retrain the entire network using the training samples for the 200 bird classes with a learning rate of 0.01[1].

The experiments are divided into two parts: **(i)** performance of the proposed hierarchical approach for varying number of subsets, and **(ii)** performance comparison of the proposed system against several recent algorithms. Based on preliminary experiments, the threshold for confidence of the subset selector is set to $\tau = 0.98$ for all experiments.

We first evaluate the performance of the proposed system by varying the number of subsets $K = [2, 3, \ldots, 25]$. The results are presented in Fig. 3, along with the performance of the baseline system DPD-DeCAF [6]. The performance of the proposed system generally increases until $K = 8$, reaching 72.7%. For higher values of K (ie. more subsets), the performance tends to decrease in a non-monotonic manner, indicating that relatively large values of K are not necessarily helpful. A visualisation of the classes assigned to each subset is given in Fig. 4.

Comparisons against other methods are shown in Tables 1 and 2. In Table 1 parts annotations are exploited, while in Table 2 only bounding boxes are used. It can be seen that in Table 1 the proposed method (using the optimal $K = 8$) leads to a relative performance improvement of 12.7% over the baseline DPD-DeCAF system. When ground-truth labels are used for the subset selector, the proposed system can increase its performance from 72.7% to 78.6%. This indicates that if the performance of the subset selector can be improved, we can further improve the performance of the overall system.

In Table 2, where only bounding boxes are used instead of parts annotations, the best performance by the proposed method is obtained at $K = 14$. The proposed method achieves an accuracy of 69.2% compared to 61.8% obtained by a convolutional neural network method presented in [16], resulting in a relative performance improvement of 12.0%.

[1]This rate decreases by a factor of 10 every $5,000$ iterations for a total of $20,000$ iterations.

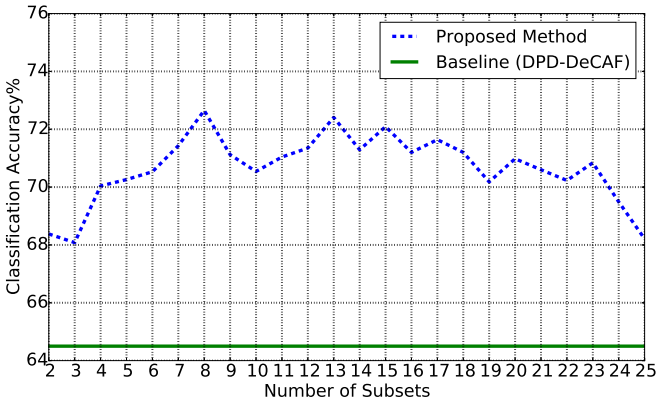

Fig. 3: Performance of the proposed method on the Caltech-UCSD CUB200-2011 bird dataset, while exploting part annotations. The number of subsets (K) is varied from 2 to 25. The subsets are selected automatically. Performance of the baseline system DPD-DeCAF [6] is also shown.

Table 1: Accuracy of various systems on the Caltech-UCSD CUB200-2011 bird dataset, exploiting part annotations.

Method	**Accuracy**
Pooling feature learning [11]	38.9%
Symbiotic Model [5]	59.4%
POOF [3]	56.9%
Part transfer [9]	57.8%
DPD-DeCAF [6]	64.5%
Proposed method (automatic subsets, K=8)	**72.7%**
Proposed method (ground truth subsets, K=8)	78.6%

Table 2: As per Table 1, but instead of using part annotations, only bounding box information is used.

Method	**Accuracy**
Bounding Box [16]	53.3%
Bounding Box-aug [16]	61.8%
Proposed method (automatic subsets, K=14)	**68.6%**

Fig. 4: Example images of 10 classes for each of the subsets for the best performing system ($K = 8$). It can be seen that the classes assigned to each subset are visually similar.

4. CONCLUSION

In this paper, we have introduced a novel direction to tackle the problem of fine-grained classification. We have proposed the use of a hierarchical classifier so that classes that have high visual correlations are grouped together into the same subsets. An expert classifier is then learnt for each subset.

The novel hybrid hierarchical classification system yields performance improvements over the recent deep convolutional neural network system proposed in [6]. This hybrid approach combines the classical *GlobalSVM* classification approach with a novel *LocalSVM* classification approach. Evaluations on the challenging CUB200-2011 dataset [17] show that classification accuracy for a fully automatic system can be increased from 64.5% to 72.7%, a relative improvement of 12.7%.

Future work will examine ways to close the gap between the performance of the automatic system and the performance of the ground truth system. The ground truth (assigning all test samples to their correct subset) achieves a classification accuracy of 78.6%, which is considerably better than the 72.7% of the fully automatic system. This implies that performing more accurate assignment of a sample to its subset can yield considerable performance improvements. One possible approach to obtain more accurate assignment would be to learn visual features that best differentiate the subsets rather than all of the classes.

Acknowledgments

The Australian Centre for Robotic Vision is supported by the Australian Research Council via the Centre of Excellence program. NICTA is funded by the Australian Government through the Department of Communications, as well as the Australian Research Council through the ICT Centre of Excellence program.

5. REFERENCES

[1] K. Anantharajah, Z. Ge, C. McCool, S. Denman, C. B. Fookes, P. Corke, D. W. Tjondronegoro, and S. Sridharan. Local inter-session variability modelling for object classification. *WACV*, 2014.

[2] T. Berg and P. N. Belhumeur. How do you tell a blackbird from a crow? In *ICCV*, 2013.

[3] T. Berg and P. N. Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *CVPR*, 2013.

[4] S. Branson, P. Perona, and S. Belongie. Strong supervision from weak annotation: Interactive training of deformable part models. In *ICCV*, 2011.

[5] Y. Chai, V. Lempitsky, and A. Zisserman. Symbiotic segmentation and part localization for fine-grained categorization. In *ICCV*, 2013.

[6] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *ICML*, 2014.

[7] R. Farrell, O. Oza, N. Zhang, V. I. Morariu, T. Darrell, and L. S. Davis. Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. In *ICCV*, 2011.

[8] E. Gavves, B. Fernando, C. G. Snoek, A. W. Smeulders, and T. Tuytelaars. Fine-grained categorization by alignments. In *ICCV*, 2013.

[9] C. Goring, E. Rodner, A. Freytag, and J. Denzler. Nonparametric part transfer for fine-grained recognition. In *CVPR*, 2013.

[10] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[11] Y. Jia, O. Vinyals, and T. Darrell. Pooling-invariant image feature learning. *arXiv:1302.5056*, 2013.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[13] N. Kumar, P. N. Belhumeur, A. Biswas, D. W. Jacobs, W. J. Kress, I. C. Lopez, and J. V. Soares. Leafsnap: A computer vision system for automatic plant species identification. In *ECCV*. 2012.

[14] J. Liu, A. Kanazawa, D. Jacobs, and P. Belhumeur. Dog breed classification using part localization. In *ECCV*. 2012.

[15] C. R. Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):159–203, 1948.

[16] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. *CVPR Workshop on Deep Vision*, 2014.

[17] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. In *Computation & Neural Systems Technical Report, California Institute of Technology*, number CNS-TR-2011-001, 2011.

[18] N. Zhang, R. Farrell, F. Iandola, and T. Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *ICCV*, 2013.