PART-BASED FINE-GRAINED BIRD IMAGE RETRIEVAL RESPECTING SPECIES CORRELATION

Cheng Pang[1,2] *Hongdong Li*[2,3] *Anoop Cherian*[3] *Hongxun Yao*[1]

[1]School of Computer Science & Technology, Harbin Institute of Technology {pangcheng3, h.yao}@hit.edu.cn

[2] College of Engineering & Computer Science, Australian National University, hongdong.li@anu.edu.au

[3] Australian Centre for Robotic Vision, Australian National University, anoop.cherian@anu.edu.au

ABSTRACT

Most of the existing works on fine-grained bird image categorization and retrieval focus on finding similar images from the same species and often give little importance to inter-species similarity. In this paper, we devise a new fine-grained retrieval task that searches similar instances from different species. To this end, we propose a two-step strategy. In the first step, we search for visually similar parts to a query image using a deep convolutional neural network (CNN). To improve the quality of the retrieved candidates, we incorporate structural cues into the CNN using a novel part-pooling layer. In the second step, we re-rank the retrieved candidates improving the species diversity. We achieve this by formulating a novel ranking function that balances between the similarity of the candidates to the queried parts, while decreasing the similarity to the query species. We provide experiments on the benchmark CUB200 dataset and demonstrate clear benefits of our schemes.

Index Terms— Fine-grained image categorization, image retrieval, part detection

1. MOTIVATION

Fine-grained image recognition and retrieval have received significant advancements in the recent years; thanks to the availability of highly successful deep learning architectures and efficient computational platforms. In general, the goal of fine-grained recognition is to learn subtle image cues that (i) distinguishes instances of one species against another [1, 2, 3, 4], (ii) localizes distinguishing parts [5, 6, 7], or (iii) segments images at the instance level [8, 9]. However, existing methods do not explore the correlations between features across different species; knowledge of which can be useful in a variety of scenarios, including the study of the evolution of animal body parts [10] and finding cross-species behavioral similarities [11]. In this paper, we present a novel task for the retrieval of fine-grained images that improves species diversity. Figure 1 illustrates our motivation and goal.

Our proposed fine-grained bird image retrieval framework consists of two stages. First, we use part instances from the query image to generate a candidate set of images containing these parts irrespective of the species of the retrieved images. We use a deep convolutional neural network (CNN) to implement this stage by generating part-specific binary hash codes [12, 13]. As our end goal is to retrieve images from diverse species, a retrieval of exact matches of the part instances may be inadequate (as the candidates may contain only instances from the same species as the query). To circumvent this issue, we propose a novel pooling layer in our CNN, dubbed *part-pooling*, that incorporates structural cues. Specifically, this pooling layer generates a part-specific spatial receptive field on the CNN features maps, where the size of this field is scaled proportional to the geometric structure of the queried part in relation to other parts. This scheme allows including nearby parts in the query, but avoids the influence of parts that are far. For example, searching for part *eye* could include the *beak*, but avoids including the *wings*. Our scheme allows candidates to have a degree of diversity, while avoiding retrievals that are structurally dissimilar (e.g., a *bird* versus a *dog*).

In the second stage of our algorithm, we re-rank the candidates by imposing species diversity. Precisely, we propose a novel ranking function that trades off between increasing the part similarity and decreasing the species similarity between the query and retrieved images. We present experiments on the standard CUB200 dataset consisting of bird images. Our results show that our two stage scheme demonstrates superior performance in retrieving diverse images.

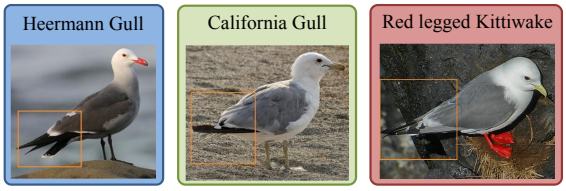

Fig. 1. Birds from different species with similar tails share interesting correlations: two of them are *gulls* and they are both sea birds. Our goal in this paper is to search for instances with visually similar parts but from different species.

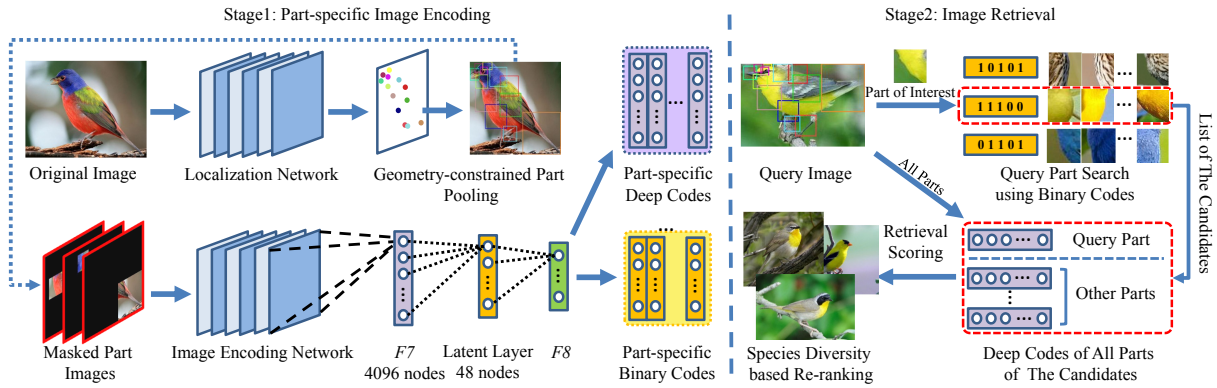

Fig. 2. Pipeline of the proposed part-based fine-grained image retrieval respecting species correlation. Key novelties of our method lie in the geometry-constrained part pooling, two-step retrieval strategy and species diversity based ranking function.

2. RELATED WORK

Fine-grained image categorization and retrieval are challenging tasks due to minor between-class differences and significant within-class variations. Deep learning has been applied to such tasks in the recent past. For example, Liu *et al*. [14] developed a deep model that learns features by jointly predicting attributes and landmarks of clothes for online shopping. Wang *et al*. [15] learns a similarity metric directly from images that can be used for exemplar-based object recognition and image de-duplication. However, these methods have not explored inter-species correlations and only focus on full objects. In contrast, our method uses intra-species correlations and focuses on similar parts. Another paper similar to ours is from Zhang *et al*. [16] that learns a fine-grained and structured feature representation that is able to locate similar images at different levels of relevance, while also including intra-species correlations. However, their method relies on hierarchical labels and shared attributes which need tedious annotations while our proposed method directly uses the consensus of visual similarity of parts. Similar to ours, a few other state-of-the-art methods use a deep CNN for image retrieval [13, 17]. However, given that a part usually contains much less visual information than a full object, these methods cannot deal with queries using only parts as they are designed to encode whole images without structural cues. Instead, we explicitly detect parts and obtain part-specific encodings.

3. THE PROPOSED METHOD

In this section, we detail our two-stage retrieval framework. Our overall pipeline is depicted in Figure 2, showing our image encoding stage (on the left) and the retrieval stage (on the right). We use two deep CNNs for the encoding phase, one for part detection (using the localization network with our part pooling) and another for encoding the image patches associated with the parts to generate binary codes for retrieval. In the retrieval stage, we first use the binary codes to generate a candidate set, which precedes re-ranking of these candidates using deep codes from all parts to include species diversity. Although our method could use several query parts, we assume to use only one part in the sequel for brevity.

3.1. Part-specific Image Encoding

The two CNNs, namely (i) our part-localization CNN [12] and (ii) our image encoding CNN (see Figure 2) are trained separately, but are fused by the geometry-constrained part pooling layer. This layer takes the output of the localization network to mask the parts, which are then used as the input to the image encoding network. This network generates part-specific codes for retrieval.

3.1.1. Part Pooling Layer

As noted above, our part-localization network is used to predict the locations and visibility of the parts together with the bounding box of the instance. The network is trained using ground-truth part locations on all training samples. We modified this network by designing our novel part-pooling layer. This layer filters image patches containing each part from the original images (by masking out other parts). To improve the retrieval accuracy, each of the patches (which are candidates to be retrieved) should center at one part and contain as few pixels of other parts as possible, reducing interferences from other parts. Prior work with part pooling layers [2, 5] directly pick part features from the feature maps using a fixed-size spatial window (163×163 for a 227×227 image) and thus may lead to incorporating several other parts in the query. While, including other parts may benefit the detection task by exploiting co-occurrences, it may also dilute the quality of the retrievals with parts that are irrelevant to the query. To resolve these contradicting requirements, we derive a novel geometry-based pooling layer that trades-off between being part specific and instance holistic.

Our solution is based on the intuition that the receptive field of a part, which is near to other parts, should be small than that of a part far from other parts. For example, the size of the tails is usually larger than that of the beaks. Thus, we first compute the distances between every part and find the k-nearest neighbor parts for every part. That is, if D denotes the average distance between a part and its top-k neighbours, and if $\overline{D}$ denotes the average distance to all parts, then we scale the default receptive field size S_d as follows:

$$S = (1 + \frac{D - \overline{D}}{\overline{D}}) \times S_d, \quad (1)$$

We use the updated part size S to generate a part mask; the original image is then pooled using this mask, and is fed into the encoding Alexnet network [13] to obtain 2 kinds of codes, namely (i) the binary codes and (ii) deep codes. To generate the part specific binary codes, we add a new layer between the fc7 and fc8 of the Alexnet CNN model; this layer acting as a binary latent layer on which the classification layer fc8 rely. We use sigmoid functions to generate binary codes from this layer, which are used to retrieve part-specific image candidates. As for the deep codes mentioned above, they are the CNN features extracted from the fc7 layer and is used in the re-ranking stage.

3.2. Retrieval Strategy

As alluded to above, given a query image with a query part of interest, we use a two-step retrieval strategy. In the first step, the binary codes specifying the part of interest is compared with the codes of the same part from all samples in the dataset, obtaining a ranked list for the query image measured by Hamming distance. In the second step, we use top n samples in the ranked list for species diversity based re-ranking. All parts of the query image will be compared to their corresponding parts of the candidates using the deep codes from the fc7 layer measured using the cosine distance. In this step, the compared parts of interest and the other parts play different roles; the former is expected to be similar, while the latter should be different for promoting diversity of the species. We therefore modified the loss function from adversarial network [18] for the measurement of the final similarity between a query image and the candidates.

Our re-ranking function is defined as follows. Suppose $\mathcal{X} = \{x^1, x^2, \cdots, x^m\}$ are deep codes for m body-parts and let $\mathcal{X}_q$ is such a set for the query image. Further let, $x_q^t \in \mathcal{X}_q$ is the deep code for part t which we use for the query. Then, the re-ranking function computes:

$$F(x_q^t, \mathcal{X}_q, \mathcal{X}) = \log P(x_q^t, x^t) + \frac{w}{m-1} \sum_{\substack{s=1 \\ s \neq t}}^{m} \log\left(1 - P(x_q^s, x^s)\right)$$

$$\text{where } P(x_i^s, x_j^s) = \frac{1}{1 + \text{dist}(x_i^s, x_j^s)}. \quad (2)$$

Here, dist is the distance between two deep codes x_i^s and x_j^s of corresponding parts (we use dist to be the cosine distance). The dissimilarity function P is scaled between 0 and 1. The first term in F is the logarithmic similarity between the parts of interest. The last term sums the logarithmic difference for other corresponding parts and is penalized by a constant weight w (which we set to be 0.2). The insight for using this objective is that it balances similarity of the query part and diversity of other parts.

4. EXPERIMENTS

4.1. Dataset, Benchmark and Settings

We test our method on CUB200-2011 bird dataset which is a challenging fine-grained dataset consisting of 11788 images for 200 species of birds. Although the dataset has been extensively studied, the dataset continues to be one of the most difficult ones in this domain due to the significant variations in the appearances and poses of the bird images it contains, concomitant with the large number of species.

To evaluate the performance of our scheme, we need to define a list of classes for each part based on the similarity between the corresponding parts. In this paper, we show evaluations on three parts: *chest*, *tail*, and *wing*. We manually generate 3 such class lists for these parts. In a part-class list, every species of bird is assigned to one part class according to the appearance of this part. We select 148 images as queries, with one query part of interest for each image. The IDs of the query images and parts of interest are available in the supplemental material, along with the 3 lists for the part classes.

We use the standard cumulative match characteristic (CMC) and Recall@K [19] for evaluating the performance. A correct retrieval is defined as one that has the same part instance, while is from a different bird species. All images in the dataset are used as instances to be retrieved, except for the query ones.

The part detection network is trained using all the training images. Our encoding network uses an Alexnet architecture pre-trained on ImageNet [20] and fine-tuned on the CUB200 bird dataset [21] after adding the new latent layer between layer fc7 and fc8. At the query stage, we use the detected part locations and bounding boxes for all instances to be retrieved. After training with the species labels, the image encoding net-

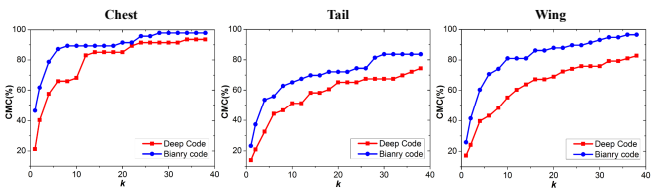

Fig. 3. CMC curves for three parts of interest using binary codes and deep codes.

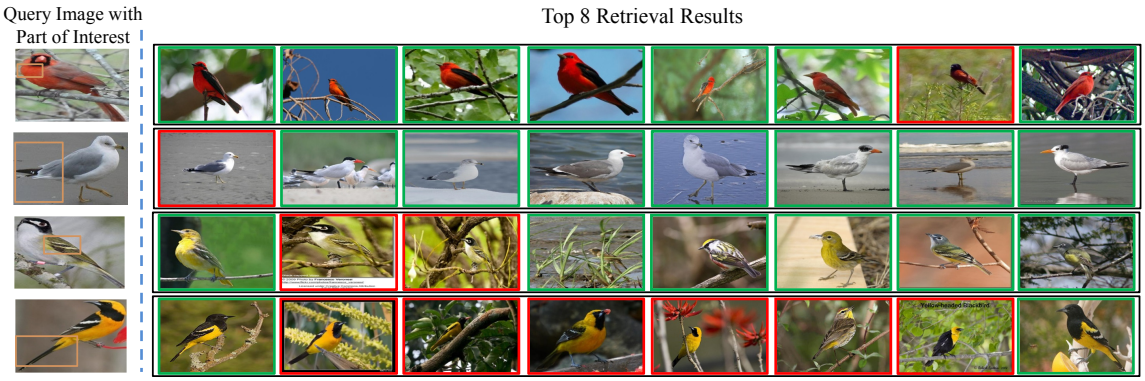

Fig. 4. Top k ranked results for different parts-of-interest with one row per query. We show correctly retrieved instances in green and incorrect ones in red.

Table 1. Recall@K Scores on CUB200-2011.

part	Recall(%)@K	1	10	40	60	80	100	120	140	160	180	200	220	240	260	280	300
Chest	Retrieval w/o part pooling layer and with part diversity	0	0.2	4.8	8.4	12.8	16.3	21.2	26.0	28.2	30.4	32.6	36.6	39.2	40.5	43.2	46.7
	Retrieval with part pooling layer and w/o part diversity	0	1.3	4.9	10.6	13.2	18.1	21.2	26.0	29.5	32.6	35.7	40.5	43.6	47.1	49.3	51.1
	Retrieval with part pooling layer and with part diversity	0.4	3.1	7.5	12.3	16.7	20.7	25.1	29.1	32.6	34.8	37.4	42.7	46.3	49.3	52.0	53.7
Tail	Retrieval w/o part pooling layer and with part diversity	0	0.8	1.3	2.8	4.4	5.2	7.0	8.0	8.8	10.1	10.8	12.1	12.1	12.4	12.4	12.6
	Retrieval with part pooling layer and w/o part diversity	0	0.9	6.2	10.6	11.5	12.4	15.9	17.7	18.6	21.2	21.2	22.1	23.9	24.8	24.8	25.7
	Retrieval with part pooling layer and with part diversity	0	0.9	7.1	11.5	11.5	15.0	16.8	19.5	19.5	21.2	22.1	22.1	23.9	24.8	25.7	26.6
Wing	Retrieval w/o part pooling layer and with part diversity	0	1.8	5.5	10.9	15.5	15.5	16.4	17.3	17.3	17.3	21.8	25.5	27.3	29.1	29.1	30.9
	Retrieval with part pooling layer and w/o part diversity	0.9	4.6	10.9	14.5	18.2	20.9	20.9	21.8	24.6	25.5	26.4	26.4	27.3	33.6	33.6	33.6
	Retrieval with part pooling layer and with part diversity	0.9	5.5	10.9	16.4	18.2	22.7	24.6	24.6	26.4	26.4	27.3	27.3	28.2	33.6	33.6	33.6

work is able to produce discriminative features when images of parts are input. The number of nodes in the latent layer is set to be 48, thereby generating a 48-dimensional code for each part.

4.2. Results and Discussion

In this section, we evaluate the performance of every novel module in our scheme bringing out their benefits. In Figure 3, we provide the retrieval results using binary codes and deep codes. This figure clearly shows that our deep codes using novel species diversity ranking function, wins by a large margin (especially for a small k) over binary codes (that are directly from the latent CNN layer). These plots demonstrate the effectiveness of our two-step ranking strategy.

In Table 1, we further show results evaluating the advantages of our part-pooling layer in combination with the re-ranking function. As is clear, using this layer increases the Recall@K performance significantly for all values of K. Further, comparing rows 2 and 3, we see that including species diversity enhances our overall results. In Figure 4, we provide qualitative results showing the top-K retrievals for different parts of interest. As is seen, the proposed method can find instances with similar patterns in the corresponding parts such as red chests, black, and white strips on the wing and tails with black tip, and yellow base. Further, interestingly our scheme could also find correlations between species. For example, for the query and the correct instances in the second row belong to the *gull* class and they all live by the sea. As for the third row, birds with similar wing patterns are usually native to eastern America.

Failure Modes: We mainly observed two modes of failure. The first one is caused by variations in the appearances within a species. For example, the variations in the bird pose leads to the first four incorrect instances in the last row of Figure 4 as our scheme considers these instances as from a different species. A related issue is when the appearance changes with the age of a bird or changes between individuals (the first incorrect instance in the second row). The second failure mode is caused by the noise from background which enlarges the difference between corresponding parts and makes instances from the same species to be dissimilar (the third row).

5. CONCLUSION

In this paper, we proposed a novel fine-grained bird image retrieval task that searches for instances having similar body parts, but from different species. We proposed a novel two-step strategy to solve this task consisting of first generating a list of candidate retrievals using a novel CNN architecture with part-pooling layers, and then re-ranking these candidates using a function that promotes species diversity. Experiments on CUB200 dataset have shown the effectiveness of the proposed method. Our method successfully discovered interesting correlations among distinct species. Although, in this paper we focused on bird images only, it can be straightforwardly extended to other fine-grained retrieval applications.

6. ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China under Project No. 61472103.

7. REFERENCES

[1] Feng Zhou and Yuanqing Lin, "Fine-grained image classification by exploring bipartite-graph labels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1124–1133.

[2] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang, "Part-stacked cnn for fine-grained visual categorization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1173–1182.

[3] Jakub Sochor, Adam Herout, and Jiri Havel, "Boxcars: 3D boxes as CNN input for improved fine-grained vehicle recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3006–3015.

[4] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1449–1457.

[5] Han Zhang, Tao Xu, Mohamed Elhoseiny, Xiaolei Huang, Shaoting Zhang, Ahmed Elgammal, and Dimitris Metaxas, "Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1143–1152.

[6] Di Lin, Xiaoyong Shen, Cewu Lu, and Jiaya Jia, "Deep lac: Deep localization, alignment and classification for fine-grained recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1666–1674.

[7] Jiongxin Liu, Yinxiao Li, and Peter N Belhumeur, "Part-pair representation for part localization," in *European Conference on Computer Vision*. Springer, 2014, pp. 456–471.

[8] Anelia Angelova and Shenghuo Zhu, "Efficient object detection and segmentation for fine-grained recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 811–818.

[9] Yuning Chai, Victor Lempitsky, and Andrew Zisserman, "Symbiotic segmentation and part localization for fine-grained categorization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 321–328.

[10] Sean B Carroll, "Endless forms: the evolution of gene regulation and morphological diversity," *Cell*, vol. 101, no. 6, pp. 577–580, 2000.

[11] Brian K Hall, *Homology: The hierarchial basis of comparative biology*, Academic Press, 2012.

[12] Kevin J Shih, Arun Mallya, Saurabh Singh, and Derek Hoiem, "Part localization using multi-proposal consensus for fine-grained categorization," *arXiv preprint arXiv:1507.06332*, 2015.

[13] Kevin Lin, Huei-Fang Yang, Jen-Hao Hsiao, and Chu-Song Chen, "Deep learning of binary hash codes for fast image retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 27–35.

[14] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1096–1104.

[15] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu, "Learning fine-grained image similarity with deep ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1386–1393.

[16] Xiaofan Zhang, Feng Zhou, Yuanqing Lin, and Shaoting Zhang, "Embedding label structures for fine-grained feature representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1114–1123.

[17] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese, "Deep metric learning via lifted structured feature embedding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4004–4012.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[19] Herve Jegou, Matthijs Douze, and Cordelia Schmid, "Product quantization for nearest neighbor search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 117–128, 2011.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[21] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona, "Caltech-ucsd birds 200," 2010.