IEICE TRANS. INF. & SYST., VOL.E100–D, NO.8 AUGUST 2017

LETTER

Pre-Processing for Fine-Grained Image Classification

Hao GE[†a)], ***Nonmember***, **Feng YANG**[††], ***Member***, **Xiaoguang TU**[†], **Mei XIE**[†††], ***and*** **Zheng MA**[†], ***Nonmembers***

SUMMARY Recently, numerous methods have been proposed to tackle the problem of fine-grained image classification. However, rare of them focus on the pre-processing step of image alignment. In this paper, we propose a new pre-processing method with the aim of reducing the variance of objects among the same class. As a result, the variance of objects between different classes will be more significant. The proposed approach consists of four procedures. The "parts" of the objects are firstly located. After that, the rotation angle and the bounding box could be obtained based on the spatial relationship of the "parts". Finally, all the images are resized to similar sizes. The objects in the images possess the properties of translation, scale and rotation invariance after processed by the proposed method. Experiments on the CUB-200-2011 and CUB-200-2010 datasets have demonstrated that the proposed method could boost the recognition performance by serving as a pre-processing step of several popular classification algorithms.

key words: *fine-grained classification, object detection, neural network*

1. Introduction

Fine-grained image classification (FGIC) aims to find differences among subordinate classes, such as discriminating different models of cars, species of animals, and types of food. For instance, in fine-grained flower classification, it's very difficult for human to recognize whether the target flower is "pincushion", "sweet william" or "artichok" because of the small inter-class variance. However, it's possible to get a quick and precise classification result with the help of a well trained fine-grained classification model. That is to say, the FGIC is helpful.

As is well known, FGIC is a challenging problem because of the small inter-class variance and high intra-class variance, which may confuse the computer to make decisions. Meanwhile, the undertraining may be triggered due to the small size of training data. In generic image classification tasks, the IMAGENET, which contains millions of images, can be used. However, in Fine-grained classification tasks, we can only use CUB [9], [10], which contains only ten thousands of images. On account of the small magnitude and less annotations of this dataset, we have to make full use of these images and their annotations. That is the reason why we don't treat FGIC as a big data problem. In contrast, we construct our framework refer to the methods of small data problems. We believe that data mining and pre-processing will play a more important role than adjusting the network structure.

In most existing FGIC methods, there are two main steps, i.e., feature extraction and classification. It has been proven in many literatures [3], [6] that using the features automatically extracted from CNN and classifying with SVM is a good scheme for classification task. Recently, a new framework, which is called "end to end system" [1], has been proposed. In this system, a Neural Network is constructed and trained. The images of the dataset are directly feeded into the network, and the labels could be automatically obtained by a training strategy. However, compared with the face recognition on small datasets [5], it seems that an important step, the pre-processing step, has been forgotten. In most face recognition frameworks, the frontalization are implemented first, to normalize the faces to similar viewpoints. Inspired by this, we propose to use the spatial relationships of the "parts" to align the images in FGIC. Spatial relationships of the "parts" can be used to obtain more information. In this way, images in the dataset can be aligned so that the influence of different poses and viewpoints can be avoided.

In this paper, we propose a new pre-processing method to align the images in CUB, which is a famous bird classification dataset in the field of FGIC. Our alignment is independent to the other steps, and can be used as a pre-processing step for all the existing fine-grained bird classification frameworks. In addition, it may evolve into a general step for fine-grained classification frameworks in the future. The proposed pre-processing strategy is composed of four independent steps, i.e., location, rotation, cropping and resizing. With these four steps, objects in the images can be presented in similar fashion, such as locations, viewpoints and sizes. The proposed method could make full use of the information of the images as well as reducing the variation of objects among the same class. The experiments demonstrate that our method could achieve promising performance.

2. Our Method

As discussed above, our pre-processing framework consists of four procedures, i.e., location, rotation, cropping and re-

Manuscript received April 7, 2017.
Manuscript publicized May 12, 2017.
[†]The authors are with the School of Communication and Information Engineering, University of Electronic Science and Technology of China, P.R. China.
[††]The author is with the School of Information and Engineering, Wenzhou Medical University, P.R. China.
[†††]The author is with School of Electronic Engineering, University of Electronic Science and Technology of China, P.R. China.
a) E-mail: ghzzxxcc19890929@live.cn
DOI: 10.1587/transinf.2017EDL8076

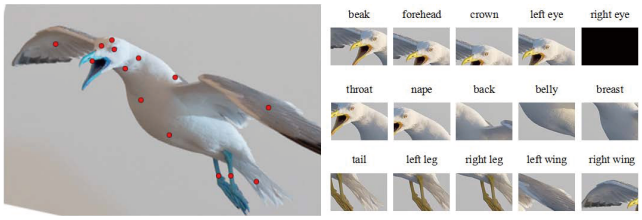

Fig. 1 Parts annotation (left: bird with part centers marked by red circles; right: $\frac{1}{4}W \times \frac{1}{4}H$ region for each part (W and H is the width and height of the left image). Black image means this "part" doesn't appear in the left image.)

sizing. The details of these four procedures will be described in this section.

2.1 Location

As center of the bird should be located first, we propose to take the average of the "part" as the "Center" of the bird. However, there remains some difficulties. As shown in Fig. 1, the "parts" around the "Head" are so dense. If we use the traditional $\frac{1}{4}Height \times \frac{1}{4}Width$ regions [8] to represent these "parts" (where Height and Width represents the height and width of the bounding box, respectively), these regions will be highly overlapping. Therefore, we integrate these 7 parts (beak, forehead, crown, nape, throat, left eyes and right eyes) into one "part", which is called "Head", and an identical label is assigned to the pair of legs and wings, respectively. We obtain 7 "parts", i.e., Head, Back, Belly, Breast, Leg, Wing and Tail, which is similar with [1]. We use the average of 6 "parts" (except the "Wings", as the result can be significantly influenced by the pose of the "Wings") as "Center" as the following function,

$$C_x = \frac{1}{N}\sum_{i=1}^{N} P(i)_x, \quad C_y = \frac{1}{N}\sum_{i=1}^{N} P(i)_y \tag{1}$$

where C_x and C_y indicates the coordinates of the "Center" of the bird, $P(i)_x$ and $P(i)_y$ indicates the X-coordinate and Y-coordinate of the "parts" respectively, and N is the number of the "parts" which appear in the image. On the other hand, as illustrated in the state-of-the-art methods [1], [3], [6], [7], bounding boxes are used in both training and testing phase, but part annotation is only used in training. In order to control variables, the part annotation should not be used in the testing step, so the "parts" should be detected first. After comparison with the other algorithms, we choose the algorithm proposed by H. Zhang [1] to implement the part detection. However, different "part" has different detection accuracy, so we can't treat them equally. We find the performance of the detection of "head" and "legs" are pretty good (more than 90%), so they are used independently in our work, and the other 5 "parts" are mainly used for locating the center of the bird. Therefore, our algorithm is insensitive to the location errors in the step of part detection.

2.2 Rotation

Rotation is the most important step in our framework, as it determines the viewpoints of the birds in the images. Our target is to align the images in both training and testing phase with the same rules. Observing from the images, we found that it is inappropriate to align all of the images with one rule, so we separate the images into three subsets first, images with only bird head, images with the front view of face and the remaining images.

2.2.1 Images with Only Bird Head

The first subset contains images with only bird head, which can be distinguished from the others by Eq. (2).

$$\sum_{i=2}^{7} P(i)_x = 0 \quad and \quad \sum_{i=2}^{7} P(i)_y = 0 \tag{2}$$

where $P(i)_x$ and $P(i)_y$ indicates the coordinates of the 6 "parts", i.e., "Back", "Belly", "Breast", "Leg", "Wing" and "Tail", which are calculated in the step of location. To this kind of images, we tried to align them, but we find that more detailed parts must be used. For example, we can use the spatial relationship between "Beak" and "Head" to align this kind of images. However, in testing phase, we can only detect 7 "parts", which means we don't have these more detailed "parts" to align the images in testing. The alignment should be performed simultaneously in both phases, so this kind of images should not be aligned in training either. Therefore, the rotation angle of this kind of images is zero, that is $\theta = 0$.

2.2.2 Images with the Front View of Face

The second subset contains images with the front view of face. We distinguish this kind of images from the others with the following functions.

$$\theta_H = \arctan(-(H_y - C_y), H_x - C_x) \tag{3}$$

$$\theta_{P(i)} = \arctan(-(P(i)_y - C_y), P(i)_x - C_x) \tag{4}$$

where H_x and H_y indicates the coordinates of the "Head", $P(i)_x$ and $P(i)_y$ indicates the coordinates of the 5 "parts", i.e., "Back", "Belly", "Breast", "Leg" and "Tail". The ranges of θ_H and $\theta_{P(i)}$ are both from $-\pi$ to π. In this kind of images, all $P(i)$ should be in a line, which means all $P(i)$ should follow either Eq. (5) or Eq. (6).

$$|\theta_H - \theta_{P(i)}| < M_0 \quad or \quad 2\pi - M_0 < |\theta_H - \theta_{P(i)}| < 2\pi \tag{5}$$

$$\left|\left(-\frac{\theta_H}{|\theta_H|}(\pi - |\theta_H|)\right) - \theta_{P(i)}\right| < M_0 \tag{6}$$

where M_0 is a threshold variable. The experiments show that best result is achieved when $M_0 = \frac{1}{16}\pi$. Equations (5) and (6) ensure the "parts" in this kind of images are in the same

or reverse direction with "Head", respectively. As a special case, if $\theta_H = 0$, Eq. (6) should be $|\pi - |\theta_P(i)|| < M_0$. In most of this kind of images, birds are standing in the ground and facing us, so $\theta_H = \frac{\pi}{2}$ is quite common. Therefore, we rotate all this kind of images to make sure $\theta_H = \frac{\pi}{2}$, so the rotation angle of this kind of image is calculated by the following function.

$$\theta = \frac{\pi}{2} - \theta_H \tag{7}$$

2.2.3 The Remaining Images

The remaining images are in the third subset. Different from the previous two kinds of images, birds in this subset all contain head and body, and are pictured from the side. Therefore, we can rotate this kind of images to the similar viewpoints. We calculate the angle of rotation base on the spatial relationship of "Head" and "Center" according to the following function.

$$\theta = \frac{3}{4}\pi - \theta_H \tag{8}$$

2.2.4 Flipping

With the previous steps, all images have been rotated. However, we find that some rotated images look unnatural, as is illustrated in the left picture of Fig. 2. Images in this case will enhance the intra-class invariance, so a discriminant conditions should be proposed to flip those mis-rotated images. Through some comparisons, we realize that whether an image should be flipped depends on the spatial relationships between the "Head", "Center" and "Legs". If an image looks natural after the rotation step, the vector from "Center" to "Legs" should be within the range of certain angle, which is described as Eq. (9).

$$M_3 < \arctan(-(L'_y - C'_y), L'_x - C'_x) < M_4 \tag{9}$$

where L'_x and L'_y indicate the coordinates of the "Legs" after the rotation step, C'_x and C'_y indicate the coordinates of the "Center", M_3 and M_4 are the threshold variables. After a lot of experiments and comparisons, we find that $M_3 = -\pi$ and $M_4 = -\frac{1}{12}\pi$ have the best performance. If the spatial relation between "Legs" and "Center" isn't satisfied with Eq. (9) for an image, this image should be flipped.

However, "Legs" may not included in an image. In this condition, the spatial relationships between "Head" and "Center" can be used to determine whether an image should be flipped according to the following function:

$$M_5 < \theta_H < M_6 \tag{10}$$

θ_H can be calculated by Eq. (3). $M_5 = -\frac{1}{2}\pi$ and $M_6 = \frac{1}{2}\pi$ have the best performance. Therefore, if "Legs" is not included in an image and the spatial relationship between "Head" and "Center" isn't satisfied with Eq. (10) for this image, it should be flipped. The right picture in Fig. 2 is flipped and rotated, which looks much more natural than the left one.

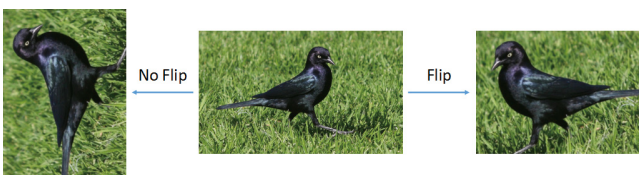

Fig. 2 Middle: original image. Left: processed image without the step of flip. Right: processed image.

2.3 Cropping

After location and rotation, birds are presented in similar viewpoints. The next step is to locate the bounding box. To the images which contain only bird head, we adopt the bounding box provided by the dataset directly. To the remaining images, the center of the "parts" should be $\frac{1}{8}$width and $\frac{1}{8}$height far from the bounding box, as most experiments use $\frac{1}{4}$width $\times$ $\frac{1}{4}$height as the part region [8]. Therefore, the bounding box is located by finding a smallest possible rectangle to contain all of the "parts", and then symmetrically enlarging the rectangle to its four thirds. However, this may cause mistakes to some of the images with the front view of face. To solve this problem, we set the short edge of the bounding box to 1/2 of the long edge if it's smaller than that.

This bounding box is different with the bounding box provided by the dataset, for example, we can see that not all of the birds are contained in the bounding box in Fig. 3 (some of the wings or tail has been cut). However, considering that the size of the "part" region is only $\frac{1}{4}$height $\times$ $\frac{1}{4}$width of the bounding box, so all the "part" regions can be included by this bounding box. Meanwhile, "part" regions contain almost all of the information of an image, so this bounding box also contains almost all of the information which can be used.

2.4 Resizing

The images got translation invariance in the step of location, and rotation invariance in the step of rotation. To obtain the scale invariance, we keep the aspect ratio of the images unchanged and resize the image to make sure the long edge

Fig. 3 Processed images with the long edge of 800 pixels.

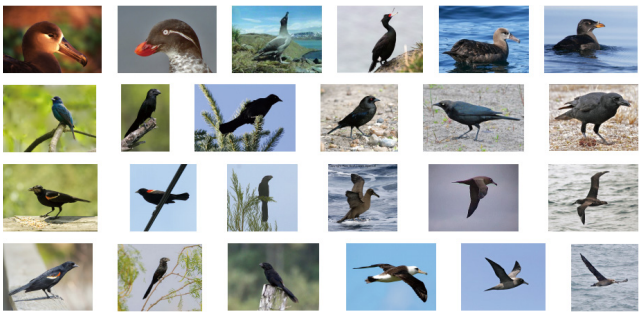

Fig. 4 Original images.

Table 1 Comparison with the original method on CUB-200-2011.

Method	Original Accuracy	After Alignment
Lee et al. [4]	41.01%	51.73%
Göering et al. [8]	57.84%	64.59%
Gravves et al. [2]	62.70%	65.32%
Zhang et al. [7]	64.96%	68.40%
Zhang et al. [6]	76.37%	78.29%
Lin et al. [3]	80.26%	81.23%

is 800. As is shown in Fig. 4, the birds in the original images are facing different directions, and the sizes of them are also different. By comparison, the processed images in Fig. 3 have similar viewpoints and sizes.

3. Experiments

3.1 Caltech-UCSD Bird-200-2011 Dataset

We test our method on the well known fine-grained dataset, CUB-200-2011 [10]. CUB-200-2011 contains 200 bird species, each species with about 60 images. The default training/test split is used, therefore we can obtain about 30 training images with its species label and the bounding box per category. Meanwhile, it also provides the coordinates of 15 "parts". As mentioned in Sect. 2, we replace the original 15 "parts" with 7 "parts", namely "Head", "Back", "Belly", "Breast", "Legs", "Wings" and "Tail".

Just like some previous methods [1]–[3], bounding box is used in both training and testing phase, but the part annotations are only used in training. Therefore, we follow the detection network proposed in [1] to do the part detection in testing, and the detected "part" information helps us doing the pre-processing. Our pre-processing framework is tested in several FGIC algorithms, and the results are listed in Table 1.

3.2 Caltech-UCSD Bird-200-2010 Dataset

We also evaluate our algorithm on CUB-200-2010 [9]. CUB-200-2010 contains 200 bird species, each species with about 30 images, which is smaller than CUB-200-2011. The main difference is that there's no part annotation in this dataset, which means we can't locate the position of the 7 "parts". So the part detection network [1], which we used to align the images in the testing phase of the previous experiment, is applied in this dataset. We train this part detection network in CUB-200-2011, and use it in CUB-200-2010 to locate the 7 "parts". With the coordinates of the 7 "parts", the pre-processing could be conducted. The performance of our pre-processing framework is evaluated in several FGIC algorithms, and the comparison results are listed in Table 2, which indicates that our pre-processing framework can be well generalized to other datasets.

Table 2 Comparison with the original method on CUB-200-2010.

Method	Original Accuracy	After Alignment
Zhang et al. [7]	34.50%	42.13%
Göering et al. [8]	35.94%	40.68%
Chai et al. [11]	47.30%	51.04%
Lin et al. [3]	65.25%	66.89%

4. Conclusion

In this paper, we propose a new and efficient pre-processing method for fine-grained classification to reduce the intra-class variance for fine-grained classification. The proposed method can extract the objects from the original images and normalize them to similar viewpoints and sizes. We propose to use the spatial relationships of the part annotations to align the images, which can make full use of the annotations. As far as we know, this is the first attempt to focus on the issue of pre-processing for fine-grained image classification. Experimental results on CUB-200-2011 (Table 1) and CUB-200-2010 (Table 2) have demonstrated that the proposed pre-processing algorithm could achieve promising performance by serving as the pre-processing step of several popular classification frameworks.

References

[1] H. Zhang, T. Xu, M. Elhoseiny, X. Huang, S. Zhang, A. Elgammal, and D. Metaxas, "SPDA-CNN: Unifying semantic part detection and abstraction for fine-grained recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp.1143–1152, 2016.

[2] E. Gavves, B. Fernando, C.G.M. Snoek, A.W.M. Smeulders, and T. Tuytelaars, "Fine-grained categorization by alignments," 2013 IEEE International Conference on Computer Vision, pp.1713–1720, 2013.

[3] D. Lin, X. Shen, C. Lu, and J. Jia, "Deep LAC: Deep localization, alignment and classification for fine-grained recognition," 2015 IEEE Conference on Computer Vision and Pattern Recognition, pp.1666–1674, 2015.

[4] Y.J. Lee, A.A. Efros, and M. Hebert, "Style-aware mid-level representation for discovering visual connections in space and time," 2013 IEEE International Conference on Computer Vision, pp.1857–1864, 2013.

[5] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," 2015 IEEE Conference on Computer Vision and Pattern Recognition, pp.4295–4304, 2015.

[6] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," Computer Vision – ECCV 2014, Lecture Notes in Computer Science, vol.8689, pp.834–849, Springer International Publishing, Cham, 2014.

[7] N. Zhang, R. Farrell, F. Iandola, and T. Darrell, "Deformable part descriptors for fine-grained recognition and attribute prediction," International Conference on Computer Vision, pp.729–736, 2013.

[8] C. Göering, E. Rodner, A. Freytag, and J. Denzler, "Nonparametric part transfer for fine-grained recognition," 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp.2489–2496, 2014.
[9] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, P. Perona, "Caltech-UCSD birds 200," CNS-TR-2010-001, California Institute of Technology, 2010.
[10] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-UCSD birds-200-2011 dataset," CNS-TR-2011-001, California Institute of Technology, 2011.
[11] Y. Chai, V. Lempitsky, and A. Zisserman, "Symbiotic segmentation and part localization for fine-grained categorization," 2013 IEEE International Conference on Computer Vision, pp.321–328, 2013.