Discriminative Features for Bird Species Classification

Cheng Pang          Hongxun Yao          Xiaoshuai Sun
School of Computer Science & Technology, Harbin Institute of Technology, China
{pangcheng3, h.yao, xiaoshuaisun}@hit.edu.cn

ABSTRACT

Bird species classification has received more and more attention in the field of computer vision, for its promising applications in biology and environmental studies. Although methods derived from basic-level classification are introduced to bird species classification, most of them couldn’t get a satisfied result due to the absence of discriminative features and quantization errors. In this paper, we introduce discriminative features for bird species classification based on parts of birds. We first crop and align the images, obtaining some patches specifying the parts of a bird. The patches are collected, forming some codebooks to learn the intermediate-level features using sparse coding algorithm. We then learn a model which characterize the discrimination of each part of every species of birds. Finally, the learned features combined with the model are concatenated to form the final representation for training and classification. We show the effectiveness of the discriminative features on the CUB-200-2011 dataset.

Categories and Subject Descriptors
I.4.9 [Image Processing and Computer Vision]: Applications;
I.5.4 [Pattern Recognition]: Applications

General Terms
Algorithms, Design, Experimentation

Keywords
Discriminative Features, Bird Species Classification, Fine-grained Classification.

1. INTRODUCTION

Bird species identification is a scientific and technical challenge for computer vision and taxonomy. Birds are closely related with humans, they were loved and bred by people since ancient times. We keep watching birds to learn the changes of environ and maintain biological diversity, and there are plenty cues lie in the geographical distribution and sub-species of the birds. Suffered from great species variation, it is difficult for non-professionals to identify the sub-category of a bird only by its appearance. However, it is exhausting to annotate all the images by human beings with expert knowledge. Thus, an automatic classification system for bird species are needed, which will be great convenience for many practical applications. For researchers working outdoors, shoot photos can be classified and analyzed immediately by the system, heavy illustration books are no need. For the public, the system could provide much fun when combined with culture information like poems and legends. It will arouse people’s interest in birds and could benefit the protections of birds.

However, bird species classification is facing many difficulties, for the difference between species could be very subtle, and the large amount of different species makes classification even more difficult. The discriminative features are highly local, which lie in certain parts of a bird. In addition, the birds can vary dramatically in pose and orientation with various lighting and background, which lead to highly intra-class variance.

In recent years, fine-grained classification stood out from basic-level classification, bringing promising applications and new challenges to computer vision society. It aims to classify similar objects such as dogs, birds, flowers and planes [1][2][3][4]. Technologies widely used in basic-level classification have been introduced to species recognition. Angelova et al. proposed develop an efficient object detection and segmentation for recognition of birds and dogs [5]. Yao et al. proposed a random forest with discriminative decision trees algorithm, discovering image patches and pairs of patches that are highly discriminative for both subordinate categorization and activity recognition [6]. Template-based approaches are used to find the right alignment of local regions, and image response features are generated for classification [7]. Cognitive theory suggest that subordinate-level recognition is based on difference between appearance details and certain parts of the object [8], inspiring methods based on attributes and parts [9-11]. A recent method introduce a set of highly discriminative, part-based and one-vs-one features(POOF), which leads to a robust learning strategy with remarkably high accuracy [12]. Our method originate in such approaches and consider the need for both theoretical plausibility and practical performance. Some recent methods includes human in the loop, adapting the human’s cognitive processes to the classification system. The crowd offer hints that how humans identify different species of birds with similar appearance [13-15]. Some part detectors are developed to localize certain parts of a bird selected by experts, which could be comparable to manual annotation to some extend [16].
In this paper, we center on part-based recognition and develop an explicit framework, using our discriminative features. Alignment using parts is shown to be an ideal alignment method for bird species classification [12]. Firstly we choose two parts with care, one for feature extraction and another for alignment. The images will be aligned and cropped into patches, in which an intermediate-level features are computed using sparse coding method. Unlike bag-of-words approaches which encode image patches to visual words and often result in quantization errors, our features maintain the relationship between different categories. Then a set of weight will be learned, indicating which pairwise parts is more discriminative for every species of birds. Our discriminative features will be computed using the features and set of weight above, ensuring reliable performance of classification.

2. PROPOSED FRAMEWORK

2.1 Part-Based Feature Extraction

Our framework begin with basic-level feature extraction. Features are extracted from all images and then turned to intermediate-level features, maintaining the hints of ground truth species. First, flip the images if necessary, making all birds have the same orientation towards the left. Secondly, patches specified pairwise parts are extracted from the images using the method of POOF [12]. Alignment of the parts are finished at the same time. Not like POOF, we carefully choose 12 pairwise of parts that explicitly depict a bird for computational efficiency and generalization performance.

When align an images, two ground truth part locations are used every time. Make these two point in a horizontal line with 64 pixels between them. Crop the image into a rectangle region of 64 pixel tall and 128 pixel wide, centering at the midpoint of the two parts. So, we get 12 pairwise parts in a single image, which are denoted by \( p^i_k \), \( i \) is the id of the image and \( k \) is the id of pairwise parts, \( k \in \{1, 2, \ldots, 12\} \). If any part of pairwise \( k \) is invisible, \( p^i_k \) is missing. Then a low-level feature \( f^i_{k} \) is extracted from \( p^i_k \) using HOG and color histograms.

As shown in Figure 1, pairwise parts depict a bird in fine-grained and provide cues for discriminative local features of the birds. Moreover, the scale and pose dependencies can be removed to some extend and intra-class variation is largely reduced.

2.2 Discriminative Feature Learning

Since typical codebook-based methods encode a feature by quantifying the feature to its nearest neighborhood visual word, its connections to other visual words are also eliminated. The result can be seen as vector with only one non-zero entry. This could lead to inevitable quantization errors when quantify the features which is close to the boundary of adjacent visual words. We use sparse representation to reduce quantization errors and build discriminative model.

There are totally 200 species of birds in the dataset, we observed that the samples from a certain species of bird is sparse among the overall dataset. It can be inferred that we can represent a pairwise parts using a few bases acquired from the same pairwise parts of other samples. Since sparse representation is shown to be effective local image descriptors [17], which could capture subtle patterns and lead to less quantization errors.

A set of \( f^i_{k} \) extracted from patch \( p^i_k \), which is denoted by \( L^i = \{f^i_{k_1}, f^i_{k_2}, \ldots, f^i_{k_{12}}\} \in \mathbb{R}^{256} \). Let \( D^i = \{d^i_{c_1}, d^i_{c_2}, \ldots, d^i_{c_{200}}\} \in \mathbb{R}^{256^W} \) be an over complete codebook. \( D^i \) consists of all the low-level feature extracted from pairwise parts \( k \), we simply make \( d^i_{c_i} = f^i_{k} \). A \( M \)-dimensional descriptor will be computed using sparsity-constrained orthogonal matching pursuit, which is a new intermediate-level feature for \( p^i_{k} \). The optimization problem could be defined as:

\[
\arg\min_{c_i} \left\| L^i - D^i \ast C^i \right\|_2, \quad \text{s.t.} \quad \left\| C^i \right\|_0 \leq S, \quad \forall i
\]

where \( C^i = [c^i_{c_1}, c^i_{c_2}, \ldots, c^i_{c_{200}}] \in \mathbb{R}^{M} \) is the set of codes of \( L^i \). \( S \) is the number of non-zeros in each \( c^i_{c_j} \). We use \( S \) to control the sparsity of non-zeros in each \( c^i_{c_j} \) for better performance. Let \( c^i_{c_j} \) be the \( j \)-th element of \( c^i_{c_j} \), which is the coefficient of the sparse representation of \( f^i_{k} \). If the base \( d^i_{c_j} \) is similar to \( f^i_{k} \), \( \left| c^i_{c_j} \right| \) tends to be a large value. So \( c^i_{c_j} \) infers the similarity between them. Let \( Z \) be the total number of class, the intermediate-level feature \( f^i_{Z} = \left[ f^i_{c_1}, f^i_{c_2}, \ldots, f^i_{c_Z} \right] \) could be obtained:

\[
f^i_{Z} = \sum_{j=1}^{Z} c^i_{c_j} d^i_{c_j}, \quad \text{s.t.} \quad \text{Class}(d^i_{c_j}) = t
\]
where \( \text{Class}(d_i^3) \) returns the ground truth class of \( d_i^3 \). \( f_i^k \) is the sum of coefficients from the same class, which reflects the probability that \( f_i^k \) (or sample \( i \)) comes from class \( t \). Dimension of \( f_i^k \) is much lower than that of \( t_i^1 \), which just depends on the number of classes. More importantly, the number of non-zeros in \( f_i^k \) can be controlled by the sparsity parameter \( S \). In other words, the connection between different classes can be maintained and the quantization errors will be reduced.

### 2.3 Discriminative Model Learning

Since all the intermediate-level features are obtained, an straightforward method is concatenating all \( f_i^k \) with different \( k \), forming the final representation of sample \( i \). We consider this, larger \( f_i^k \) usually suggests that the ground truth class of sample \( i \) is \( t \), but some \( f_i^k \) with many large \( f_i^k \) may be confused, because it share too many features with samples of other classes. In other words, pairwise parts \( k \) is not so discriminative, and concatenating all the parts doesn’t make sense. Hence, we introduce a model depicting the discrimination of different pairwise parts for each of the species.

Let \( W \in \mathbb{R}^{z \times K} \) be a matrix of discrimination. \( Z \) is the total number of species, \( K \) is total number of pairwise parts for one species. The element in the row \( z \) and column \( k \) is denoted by \( w_{z,k} \), which is the weight of \( k \)-th pairwise parts of species \( z \). Larger is the weight, more discriminative is the parts.

To learn \( W \), first use the method above to obtain a set of intermediate-level feature \( \left[ f_i^{1}, f_i^{2}, \ldots, f_i^{K} \right] \) for each parts \( k \). This progress is different from that we obtain \( f_i^k \), as \( S \) is set large enough to capture the relationship between more of the species. Then let \( Q^i = [q_i^1, q_i^2, \ldots, q_i^K] \in \mathbb{R}^{z \times 2} \) to depict the average of the features from each species:

\[
q_i^k = \text{average}(f_i^k), \quad s.t. \text{Class}(f_i^k) = i, \quad k \in \{1,2,\ldots,K\} \tag{3}
\]

Where \( q_i^k \) denotes the average of the feature of parts \( k \) from class \( i \). If parts \( k \) is more discriminative than other parts from class \( i \), parts \( k \) must share less features with those from other class, the variance of \( q_i^k \) tends to be larger. A vector consists of variance \( v_i^k = [v_i^1, v_i^2, \ldots, v_i^K]^T \) can be obtained:

\[
v_i^k = \text{var}(q_i^k) \tag{4}
\]

Where \( v_i^k \) is the variance of \( q_i^k \). Then all \( v_i^k \) are assembled to form a matrix:

\[
[v_i^1, v_i^2, \ldots, v_i^K]^T \in \mathbb{R} \tag{5}
\]

Each row vectors of this matrix is normalized to a range of 0-1 separately, forming a set of weights for the parts from one species. The result after normalization is the matrix of discrimination \( W \).

Now the final discriminative feature for training and test can be computed like this:

\[
x_i = \sum_{k=1}^{K} f_i^k \cdot w_{z,k}, \quad z = \text{Class}(f_i^k) \tag{5}
\]

Where \( x_i \) is the final feature for sample \( i \). For the training samples, the class of \( t_i^1 \) is already known. For the test ones, the class is estimated by the largest \( f_i^k \) in \( f_i^k \), for it suggests the most likely class. Though, it may suggest a wrong class, the wrong one must be highly related to the ground truth, and they share most of the discriminative features. So the weight of the parts will be enlarge as our wish. Figure 2 illustrate the framework we proposed. A linear SVM will be trained and used for species classification.

### 3. EXPERIMENTAL RESULT AND DISCUSSION

Our method is test on the full CUB-200-2011 dataset, which contains over 10000 pictures for 200 species of birds. Some annotation are provided: segmentations, part locations, attributes and bounding boxes. We use ground truth part locations for alignment and generating the pairwise parts. There are at most 12 patches of pairwise parts generated from a single image if all of the parts are visible. Basic-level features are extracted from the patches, where we use pyramid HOG and HSV histogram with two size: 8*8 and 16*16. Once the certain pairwise parts are absent, the corresponding features are set to be a vector of zeros. The code books for feature learning consist of all the patches generated from all training samples, then the intermediate-level features are learned. Specifically, when learning the feature for a training sample, we exclude the sample itself from the codebook as a base, because it doesn’t make sense to represent a sample using itself. And the sparsity parameter \( S \) is set to be 45, just a little more than the number of samples of the same class. The discriminative model is learned using the same codebooks with a larger \( S \), which is set to be 100 to consider more classes and find discriminative pairwise parts. At last, a linear SVM is used for classification. The recommended data split is used.
The results of classification using the full dataset are shown in Table 1. The CUB-200-2011 dataset is introduced in [1], and it gives the first classification method for this dataset which is selected as a baseline by many researchers. The baseline method reported an accuracy of 17.31% using ground truth part locations. The second method is based-on segmentation, improving the baseline slightly. Yao et al. propose a random forest with discriminative decision tree algorithm, which combining randomization and discrimination for fine-grained categorization, and achieve an accuracy of 19.20%. Yang et al. use a template-based method to find finding the right alignment of image regions that contain the same object parts. However, the method is highly automatic and gets a better results than previous ones, the alignment is a little coarse. Catherine et al. include humans in the loop to detect unfamiliar classes based on attribute learning, they get an accuracy of 51.40%. Deng et al. use a “bubble bank” containing discriminative features for fine-grained classification. Notice that the POOF [12] have managed a rather high accuracy of 68.7%, but it is complex and extremely time consuming. We run POOF, and the executive time is about a week, which is 7 times of that of our method.

Table 1: Mean average precision of bird species classification using different methods. Ours outperforms other methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catherine et al. [2]</td>
<td>17.3</td>
</tr>
<tr>
<td>Anelia and Zhu [5]</td>
<td>17.5</td>
</tr>
<tr>
<td>Yao et al. [6]</td>
<td>19.2</td>
</tr>
<tr>
<td>Yang et al. [9]</td>
<td>28.2</td>
</tr>
<tr>
<td>Catherine et al. [14]</td>
<td>49.4</td>
</tr>
<tr>
<td>Deng et al. [13]</td>
<td>58.4</td>
</tr>
<tr>
<td>Ours</td>
<td>64.6</td>
</tr>
</tbody>
</table>

Figure 3: Some species that have been improved the most and their confused species.

We have found that the discriminative model improve 1% of the accuracy when compared to our method without using the model. Consider the large amount of the total number of the species, this is also a meaningful improvement. For some of species, the model improves the classification dramatically. Figure 3 illustrates some species of birds that have been improved the most. For example, the accuracy of class#84 is increased by 39%. Class#63 and class#65 are the confused classes for class#84, for the samples of class#84 are often misclassified into these classes. But now the error of classification is decreased by 80%.

CONCLUSIONS

We proposed a new discriminative feature for bird species classification. Our framework is based-on pairwise parts which could capture subtle differences between different birds and reduce intra-class variance simultaneously. Then sparse coding algorithm are used for intermediate-level feature extraction and discriminative model learning. By introducing the model, the learned features are finally organized in an appropriate way, forming a kind of discriminative features characterizing subtle features of different species. The process not only lead to less quantization errors, but also highlight the discriminative components of a bird for better performance. Test on the CUB-200-2011 dataset shows that our method outperforms most of the proposed methods. When compared to the method with higher accuracy, ours shows more efficiency in computation and transparence in classification.

4. REFERENCES


