Nonparametric Part Transfer for Fine-grained Recognition

Christoph Göring, Erik Rodner, Alexander Freytag, and Joachim Denzler*
Computer Vision Group, Friedrich Schiller University Jena
www.inf-cv.uni-jena.de

Abstract

In the following paper, we present an approach for fine-grained recognition based on a new part detection method. In particular, we propose a nonparametric label transfer technique which transfers part constellations from objects with similar global shapes. The possibility for transferring part annotations to unseen images allows for coping with a high degree of pose and view variations in scenarios where traditional detection models (such as deformable part models) fail. Our approach is especially valuable for fine-grained recognition scenarios where intraclass variations are extremely high, and precisely localized features need to be extracted. Furthermore, we show the importance of carefully designed visual extraction strategies, such as combination of complementary feature types and iterative image segmentation, and the resulting impact on the recognition performance. In experiments, our simple yet powerful approach achieves 35.9% *and* 57.8% *accuracy on the CUB-2010 and 2011 bird datasets, which is the current best performance for these benchmarks.*

1. Introduction

Within the last decade, research in visual object recognition mostly focused on category-level classification [12]. Although useful for coarse distinctions between object classes, common approaches fail in scenarios where the differentiation has to be done on a finer level, *i.e.*, only small interclass differences exist and specific details matter. This area of research, which is referred to as subordinate or *fine-grained recognition*, has become recently popular in our community (see e.g. [27, 22, 40, 39, 11]), because it offers a variety of challenging applications, such as bird recognition, tasks even difficult for human annotators [26].

A large majority of visual object recognition methods are solely based on global histogram features containing statistics of local features calculated in the whole image. An overwhelming number of publications justifies this approach for diverse classification problems, *e.g.*, differentiating cars from persons and bicycles [18]. However, for subordinate or fine-grained recognition, this approach is limited, since general appearances are highly similar for different classes and only small differences at certain positions allow for discrimination. Consequently, a suitable algorithm needs to detect important parts for a reliable distinction – ideally independent of the current object pose. Part-based approaches can be exploited to tackle this goal, since they allow for computing approximate pose invariant features [22, 39, 41]. For example, with given ground-truth part positions, the method of [3] is able to boost the performance up to 22% (14 classes subset, CUB-2011 dataset) compared to their method with automatic part detection. This highlights the importance of a highly accurate part detection method. Based on a part detection model, whether supervised [22, 41, 42] or unsupervised [39, 42, 38], the extracted information for every part can be combined for a final classification.

In this paper, we follow a part-based approach and show that standard parametric models for detection [14, 3] are not sufficient for tackling the large variations present in fine-grained recognition tasks. Given this analysis as a motivation, we propose a simple nonparametric part detection algorithm based on label transfer, which we refer to as nonparametric part transfer in the following. Our method allows for flexible poses, missing part annotations, and extreme changes in viewpoint. For fine-grained recognition, we use detected parts to compute localized features that focus on regions where discriminative visual elements are expected. Our running example throughout the paper will be the distinction of bird species, although our approach is not restricted to this application at all. In combination with a multiple feature extraction pipeline and linear classifiers, our approach achieves state-of-the-art performance on both CUB-200 datasets [37, 36], which are the standard datasets used in the field. In particular, we show the benefits of iterative segmentation-based masking and a combination of color and shape-based features for an off-the-shelf fine-grained recognition system. An outline is given in Figure 1.

*This work was partially supported by the Carl Zeiss AG through the "Pro-Excellence" initiative of the state of Thuringia, Germany.

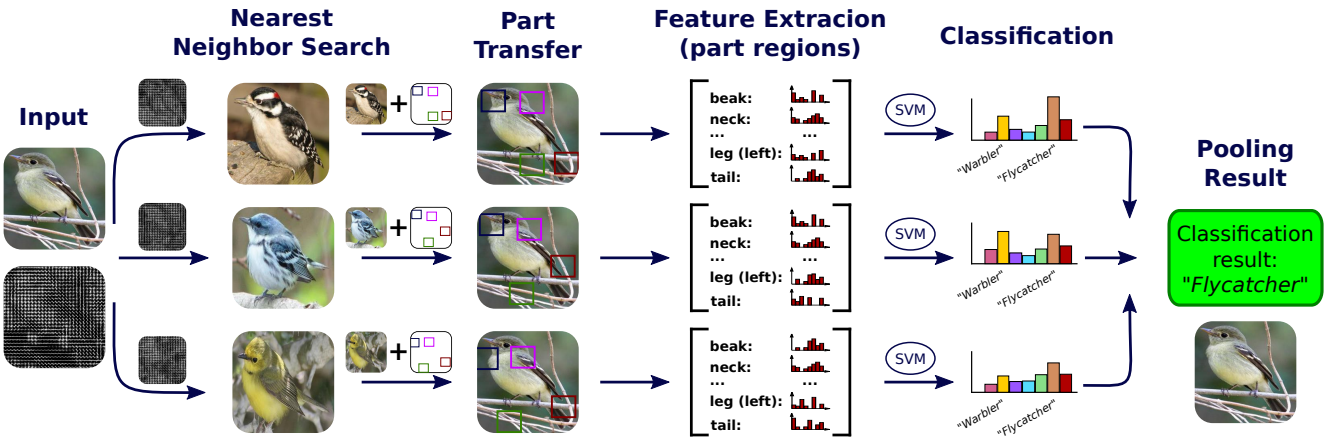

Figure 1. Visualization of our approach using part transfer for fine-grained visual categorization: training images visually similar to the test images are selected and existing part annotations are transferred to extract features from part locations only. Final results are obtained by pooling individual classification scores for every transferred part configuration. The figure is best viewed in color and by zooming in.

First, we review related work of fine-grained classification followed by an analysis of parametric part detection models and their ability to cope with large pose variations. In section 3, we present our nonparametric part transfer method to increase the flexibility of object part models. How to use part locations to focus fine grained feature extraction is depicted in section 4 together with important ingredients such as segmentation-based masking and classifier ensembles. Experiments are given in section 5.

2. Related work

Fine-grained recognition with global features One way to tackle fine-grained recognition is to directly apply visual bag-of-words approaches commonly used for standard object categorization. Due to the visual similarity between classes, there is a high likelihood for having a significant amount of common visual words shared between classes that do not help distinguishing the classes from each other. Therefore, Khan et al. [18] improve dictionary learning by fusing visual words from two different feature types. Chai et al. [7] replace the histogram with powerful Fisher vectors. Other approaches do not use dictionaries at all to avoid quantization errors [24, 39]. Reducing background clutter which might interfere classification is presented in [6, 25] using segmentation techniques. In contrast, we directly use detection results to restrict feature extraction to image regions that are likely to contain object parts. Following previous approaches, additional global features are computed on a mask obtained using GrabCut [28], which also eliminates background artifacts.

Another line of work uses active classification techniques, which require human interaction during testing to refine classification results [35, 5]. However, our main aim in this paper is to analyze fully automatic recognition approaches that do not require further human interaction, although such an approach could be combined with active classification methods to refine the results.

Part-based fine-grained recognition Global approaches discard the position of features, which are crucial in the fine-grained case. Part-based approaches avoid this information loss by extracting features on detected object parts only. Previous techniques extract unsupervised parts using an ellipsoid to model the bird pose [13] and fuse parts using specialized kernel functions [41]. In a recent paper, Zhang et al. [42] show how to include an increasing amount of supervision for training deformable part models which then allows for pose normalization by comparing corresponding parts. Parkhi et al. detect a single main part (in their case the head of a dog) to discover the rest of the body. The features used for classification are individually extracted from head and body [27]. For very specific applications, it is even possible to perform classification only based on a single main part without extending the area of feature extraction [22]. In contrast, our approach uses every part available, a necessity when tackling bird recognition.

Exemplar models and label transfer Transferring label information from training to test images has successfully been used in several computer vision applications. A prominent technique in this field are Exemplar-SVMs as introduced in [23]. The idea is to train a single SVM for every training image as positive sample and millions of negative samples, thereby bridging parametric and nonparametric modeling. Label information can then be transferred to new images from training samples with high detection scores of corresponding SVMs. Our idea is similar in spirit but avoids expensive training by using a nearest neighbor transfer technique.

Another line of research is label transfer for semantic segmentation and scene understanding proposed in [20]. Their idea is to transfer pixelwise labels from a set of K

nearest neighbors of the test images. The label transfer is based on a relaxed SIFT matching algorithm, called SIFT-Flow. In our method, we transfer specific part positions only, because the annotation is not available on a pixel level. Furthermore, instead of merging the part locations of the K most similar images in the training set to a more precise part detection result, we show how to build a multiple part feature representations to boost classification performance.

Parallel to us, the work of [21] and [16] very recently also demonstrated the power of part detections for fine-grained recognition. The paper of [21] builds on [2] for part detection. Their methodology is quite different from ours. We do not learn single part detectors and fuse them, instead we perform a simple but very powerful global matching (without any part position optimization) and a subsequent ensemble learning.

3. Nonparametric part transfer

In the following, we first review parametric part detection models and analyze their underlying assumptions. This analysis motivates then our new nonparametric extension, which allows for flexible part detection models able to cope with a large variety of views and object poses.

3.1. Analysis of parametric part models

One of the most common approaches for object detection in 2D images is the deformable parts approach of [14] A deformable part model M is meant to be invariant to certain object deformations by allowing parts of the objects to move. It consists of a set of filters $\mathcal{F} = \{\mathbf{w}, \mathbf{w}_1, \ldots, \mathbf{w}_m\}$ and a model for their spatial layout expressed as a deformation model $\mathbf{d}$. The root filter $\mathbf{w}$ is intended to cover the whole object and the remaining filters cover parts of the object. The combined detection score is calculated by:

$$f_M(\mathbf{x}) = \max_{\mathbf{z}} \; f_{\mathbf{w}}(\mathbf{x}) - \mathcal{D}(\mathbf{z}; \mathbf{d}) + \sum_{k=1}^{m} f_{\mathbf{w}_k}(\mathbf{x} + \mathbf{z}_k) \quad (1)$$

where $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_m)$ are the latent part locations, $\mathcal{D}$ is the cost of $\mathbf{z}$ with respect to a learned deformation model $\mathbf{d}$, and b is a bias term. The entire model is usually learned with a latent SVM scheme and is described in detail in [14].

The deformation model of a DPM is given by the parameters $\mathbf{d} = (\mathbf{d}_1, \ldots, \mathbf{d}_m, \mathbf{v}_1, \ldots, \mathbf{v}_m)$ and the cost defined with part locations $\tilde{\mathbf{z}}_k$ relative to anchor positions $\mathbf{v}_k$ learned for each part[1]:

$$\mathcal{D}(\mathbf{z}; \mathbf{d}) = \sum_{k=1}^{m} \left[\tilde{z}_k^x, \tilde{z}_k^y, (\tilde{z}_k^x)^2, (\tilde{z}_k^y)^2\right] \cdot \mathbf{d}_k \quad (2)$$

[1]The deformation model given here ignores the fact that in the original DPM formulation a finer resolution is used for all parts. However, our analysis still holds and we only ignore this fact for the DPM review to simplify the presentation.

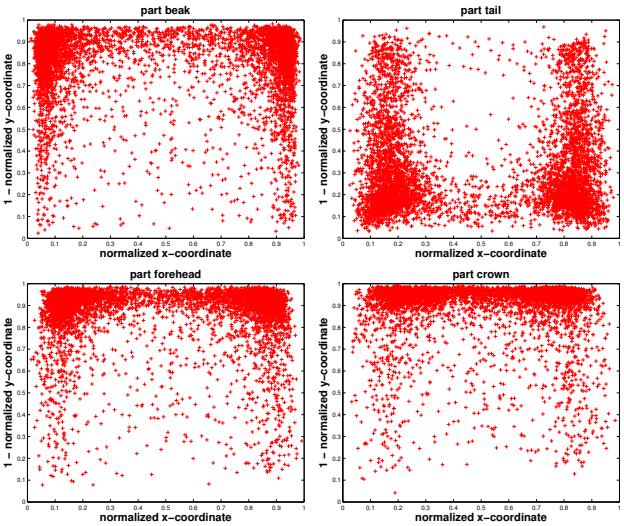

Figure 2. Visualization of part offsets for the CUB-2011 dataset [36]. The plots in the original publication where biased towards specific views and have been corrected after publication at CVPR.

This corresponds to a Gaussian model for the part locations, *i.e.*, covariance matrices $\mathbf{S}_k \in \mathbb{R}^2$ and mean vectors $\mu_k \in \mathbb{R}^2$ can be derived, such that the following holds:

$$\mathcal{D}(\mathbf{z}; \mathbf{d}) \propto -\log \prod_{k=1}^{m} \mathcal{N}((z_k^x, z_k^y) \mid \mu_k, \mathbf{S}_k) \; . \quad (3)$$

The question remains whether this model is complex enough to capture the high variability encountered in fine-grained recognition applications. Hence, we analyzed part offsets present in the CUB-2011 dataset [36] and visualized results for all bird species in Figure 2. The distribution is clearly non-Gaussian, therefore, a single DPM model would not be able to model the variation present in the training dataset. Note that in [14], parts and configurations are learned in an unsupervised fashion. However, the analysis of the "ground truth parts" provides a certain intuition for the complexity of the problem tackled.

A common strategy to cope with different views is to train a mixture of components, $\mathcal{M} = \{M_0, M_1, ...\}$, in which case z is augmented to add the latent component label that the example belongs to. However, as can be seen in our analysis, the distribution has a large number of different components and the data can not be easily clustered. Therefore, learning proper mixtures of deformable part models is an ill-posed problem and good generalization abilities can not be expected a priori. Furthermore, not only is the deformation model too restrictive, but also a single linear appearance model is not flexible enough to capture the large variability of birds.

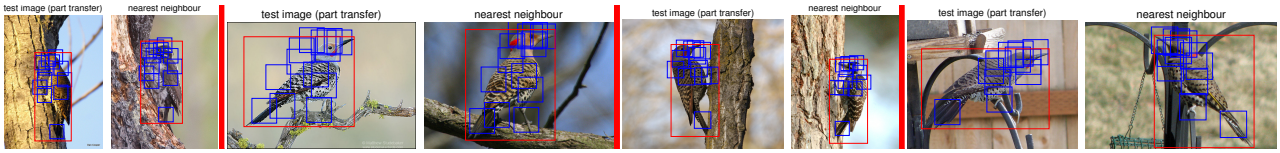

Figure 4. Example detections of our nonparametric part transfer technique together with the used nearest neighbor match.

Figure 3. Different bird poses present in CUB-2011.

3.2. Nearest neighbor part transfer

Our previous analysis motivates the development of more complex models able to tackle large intraclass variances, which are usually encountered in fine-grained recognition scenarios. In our running example, the distinction of different bird species, the high variation in part positions arises from the large number of different poses of birds in images, which are visualized for a single species in Figure 3. For fine-grained recognition, part localization can be a crucial step. This is demonstrated by the results of [3], whose approach was able to boost recognition accuracy up to 22% (14 classes subset of the CUB-2011 dataset) when being provided with the ground-truth part positions.

To overcome the limitations of linear detection models, [33] proposed to use non-linear models to increase the model complexity. However, these methods are costly during learning and prediction, despite the speed-up with explicit feature transformations [30]. Therefore, we are following an orthogonal line of research and learn not only a small number of parametric models from the given training data, but make use of a simple part transfer technique. Our method searches for training images with an object shape similar to the current test image and then transfers the part annotations from them directly. Exemplary visual results obtained from the CUB-2011 dataset are given in Figure 4.

For training, the only thing that additionally needs to be done is to compute a suitable feature representation, which mainly focuses on the global shape of the object rather than on color information for small part details. First, we use the given bounding box of the object, which is available for the CUB-2011 dataset, to re-scale the image to a fixed size of 256×256 pixels. After that, the overall layout is represented with histograms of oriented gradients (HOG), which have been shown to adequately describe the rough shape of an object [8].

For part detection during testing, we again compute a global HOG feature from the image region defined by the provided bounding box.[2] The K nearest neighbors with respect to the HOG feature are then obtained from the training set and part positions of the nearest neighbors are scaled proportionally to the bounding box of the test image. Similar to deformable part models, the size of the parts is fixed to a constant scale. In our case, we use squared parts with a length proportional to the diagonal d of the bounding box in the test image. The transferred part positions together with the scaled part sizes are the final result of our nonparametric part detection method.

Following established techniques for data augmentation [19, 14], we also add the flipped version of each training image to the training set to allow for matches between images of different orientation. In the case of a test image matching a flipped image, part correspondences are handled properly, *e.g.*, the flipped left wing position will be the right wing position of the flipped image.

We want to point out that our part detection technique is indeed related to the idea of Exemplar-SVM [23]. However, instead of learning a single nonparametric HOG detector, the authors propose to use a trained detector for each positive example. A main motivation of their technique is to allow for transferring segmentation masks to test images. In contrast, we transfer part annotations that are important to facilitate fine-grained recognition later on. Furthermore, the expensive learning stage of [23] is not necessary in our case, since we do not need to perform sliding-window detection and hard negative mining for tackling the large number of possible negative examples.

4. Part feature representations

In the following, we show how to compute multiple features that capture shape as well as color information for each of the parts. Furthermore, we show how to build ensembles of part feature representations to allow for more robust classification decisions.

4.1. Feature extraction for single parts

One of the main reasons for the difficulty of fine-grained recognition is that although small details in appearance matter, a method still has to cope with a large intraclass variance, *e.g.*, different poses and view points for the task

[2] The CUB datasets also provide bounding box annotations for test images and consequently our approach should be able to exploit this source of information. Apart from this, bounding boxes are usually cheap to obtain and their usage is therefore no strong requirement.

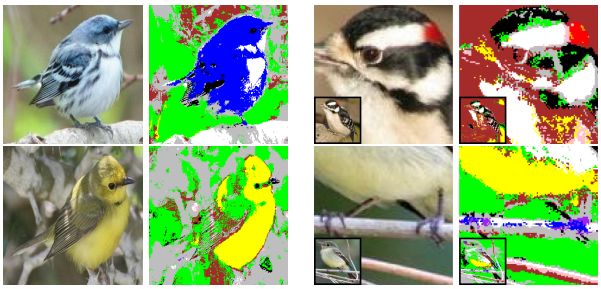

Figure 5. Example images of the CUB-2011 dataset [36] with corresponding color name features as proposed by [32].

of bird recognition. This trade-off is especially important when carefully designing the feature extraction step.

By using the part-based approach introduced in section 3, we already obtained a large degree of invariance with respect to different poses and view points: (i) features are extracted on relevant positions only and (ii) a mapping of corresponding parts is directly possible. However, we still have to deal with a large appearance diversity even within a single part. This fact is visualized in Figure 3 by showing image regions of parts for different images of a single bird species. Therefore, we follow established approaches by using an unordered bag-of-visual-words model based on complimentary types of local features. As argued by [18], the combination of color and texture information is important for the task of bird recognition, which is again our running example in this paper to analyze our part-based approach. Some bird species, for example the red-bellied woodpecker, have a characteristic head color (a fact which is contrary to what the name would suggest). Furthermore, texture or shape information can help to recognize birds with characteristic eyes or beaks. In detail, we use two types of local features to capture color and texture/shape information.

Local shape and color descriptors Instead of directly using the RGB values within a local pixel neighborhood, Weijer et al. [32] proposed to map colors to a space spanned by a basis of L colors. The colors have a correspondence in the English language, such as red, orange, blue, and yellow ($L = 11$ was used in their paper). In detail, a color $\mathbf{c}$ is mapped to the probability vector $[p(t_\ell|\mathbf{c})]_{\ell=1}^{L}$ stating how likely $\mathbf{c}$ would be described with the color name t_ℓ. Probabilities are estimated using images obtained by Google image search. Figure 5 shows some results of the color name representation, where only the most probable color prototype is shown. In particular, it is interesting to see how the dominant color of the bird is captured, *e.g.*, the blue colored body in the first example and the red colored back of the head in the second example.

In addition, we capture coarse shape information with *OpponentSIFT* descriptors presented in [31] by using gradient orientation histograms similar to the common SIFT descriptor but obtained from the images converted in the opponent color space.

Part-specific codebooks Codebooks for the bag-of-visual-word-models are learned for each part individually. The intuition is that this allows for learning prototypical local elements specific for each part, *e.g.*, elements of a beak. Part-specific codebooks are created using k-means clustering of local features extracted from the corresponding part only. Consequently, each part is represented by two histograms, one for the color features and one for OpponentSIFT descriptors, which are both normalized and finally concatenated.

4.2. Classification with part feature ensembles

As introduced before, parts detected with our nonparametric technique (see section 3) are represented with combined features. For a transferred part configuration, we thereby obtain m combined histograms for m annotated parts. However, not all parts have to be visible in an image due to the relative position of bird and camera (see Figure 7 for some examples). In absence of further knowledge, we use zero imputation [29]. We also studied other more sophisticated imputation methods, but did not observe a significant difference in terms of resulting recognition performance. To finally fuse information of all parts, several quite general possibilities exist: (i) late binding: every part is modeled by its own classifier and classification scores of all parts are combined via (weighted) pooling, or (ii) early binding: features are concatenated and a single classifier is trained using all information simultaneously. Due to the advantage of early binding to exploit dependencies between the different parts, our final part representation is the concatenation of all part features.

Although this sounds promising so far, we are still not done yet: features based on part detections transferred from only a single nearest neighbor will usually not be robust enough with respect to wrong matches. Therefore, we build a part feature ensemble by transferring part locations from each of the k nearest neighbors. For every part configuration transferred from the nearest neighbors of a test image, parts are represented by computing histograms over color and texture as introduced before. Final classification results are obtained by pooling scores over all transferred part configurations using average pooling and the category decision is done by taking the category with the maximum score. For classification, we use a SVM with the χ^2-kernel and the explicit feature transformation technique presented in [34].

4.3. Additional global feature extraction

Instead of relying only on part feature representations, we also compute a global representation using the whole bounding box. Feature types are the same as used for describing parts (bag-of-visual words with OpponentSIFT and color names) but with additional spatial pyramid pooling.

Figure 6. GrabCut segmentation results provide a mask for global feature extraction.

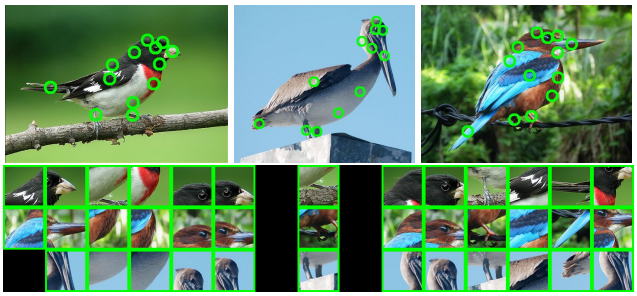

Figure 7. Example images of the CUB-2011 dataset [36] showing exploited part annotations (*top*) and corresponding regions for feature extraction (*bottom*). Occluded parts are displayed in black.

Furthermore, the influence of background clutter will be much higher here. Therefore, we apply GrabCut segmentation [28] to estimate the foreground. The algorithm of [28] performs iterative segmentation with a conditional Markov random field, where unary potentials are modeled with a Gaussian mixture model re-estimated in each iteration, and pairwise potentials are added to favor strong image edges. Some example results of this segmentation technique can be seen in Figure 6.

5. Experiments

5.1. Experimental setup

We evaluate our approach on the CUB-2010 [37] and the CUB-2011 dataset [36], which are the common benchmarks for fine-grained recognition algorithms. Both datasets consist of 200 different bird species and provide a bounding box for each image. In contrast to the CUB-2010 dataset, the CUB-2011 dataset also contains part annotations for 15 parts, *e.g.*, left and right eye, or beak (see Figure 7 for example images), which is necessary for our part transfer algorithm and also the information exploited in [3, 42]. We present results computed on the whole CUB-2011 dataset as well as on the commonly used 14 class subset [3]. Our results on CUB-2010 are based on estimating part positions for test as well as training images by using our nonparametric part detection approach to transfer part locations from the CUB-2011 dataset. Note that using a few additional annotations during training is a common strategy for this dataset, for example [42] uses manual head and body part annotations and [9] utilizes discriminative parts obtained with an annotation game.

We compare our approach with the recent state-of-the-art in fine-grained recognition: deformable part descriptors (DPD) [42], part-based one-vs-one features (POOF) [3], pooling-invariant image feature learning (PDL) [17], the template learning approach of [38], the segmentation technique of [1], and the Bubblebank method of [9]. Furthermore, we also compare our approach with a DPM baseline with five components, where part locations are initialized in a supervised manner during training using the provided part annotations. After part detection, part and global features are calculated as described before.

Approach	**CUB-2011/14**	**CUB-2011/200**
PDL [17]	-	38.91%
Template learning [38]	-	43.67%
DPD [42]	-	50.98%
POOF [3]	70.10%	56.78%
Ours, DPM	67.59%	39.18%
Ours, part transfer	69.85%	54.76%
Ours, part transfer ensemble with $k = 5$	**73.86%**	**57.84%**

Table 1. Mean accuracy results on CUB-2011.

5.2. Evaluation

CUB-2011 Recognition results for the CUB-2011 dataset are given in Table 1. As can be seen, our method outperforms previous methods for the 14 class subset as well as for the whole 200 classes dataset. Furthermore, as expected we gain by using our part transfer instead of a deformable part model (part transfer vs. DPM with over 15% gain for the whole dataset), and also by using a part transfer ensemble instead of only a single part representation (part transfer ensemble vs. part transfer with at least 3% gain for both settings). A qualitative comparison is given in Figure 8. In contrast to the DPM baseline, our approach is able to properly transfer parts and thereby to avoid misclassifications in nearly all cases. An example where our method fails is given in the lower left image, where the background is heavily cluttered and thereby misleads our HOG-based part transfer technique.

It should be noted that the best but still unpublished (although available on arXiv) approach on the CUB-2011 dataset is currently a deep learning version [10] of the DPD approach [42]. This approach achieved a performance of 64.96% on the whole dataset by just exchanging the feature representation inside of DPD with features calculated with a pre-trained deep learning architecture. Therefore, we believe that the performance of our approach can also be improved significantly by incorporating deep learning features and we consider this as interesting future work.

Compared to [21], we achieved an mAP performance of 76.94% ([21]: 62.42%) using 14 classes and 55.36% ([21]: 44.13%) using 200 classes.

CUB-2010 The results on the CUB-2010 dataset are given in Table 2 and show that we are consistently able to

Figure 8. Qualitative evaluation on CUB-2011. We displayed cases where our approach and the DPM baseline show significant differences.

Approach	CUB-2010
Template learning [38]	28.20%
Segmentation [1]	30.20%
Bubblebank [9]	32.50%
DPD [42]	34.50%
Ours, part transfer	**35.94%**

Table 2. Mean accuracy results on CUB-2010.

Method	CUB-2011/14
NN part transfer	69.85%
... without GrabCut masking only	68.09%
... without part features only	66.08%
... without global features only	61.56%

Table 3. Detailed feature analysis on CUB-2011 with mean accuracies. Only specific single aspects are excluded.

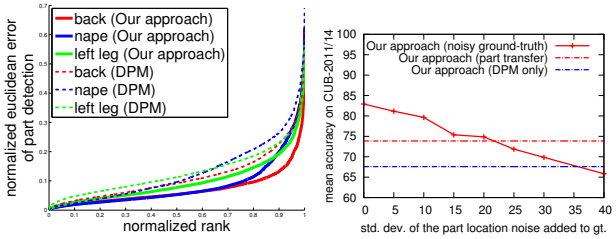

Figure 9. *Left*: Percentile plot of the error in part positions for three parts. *Right*: Dependency of part detection and classification accuracy (CUB-2011/14 dataset).

outperform previous methods. Note that for this dataset, we only estimate part locations by part transfer from CUB-2011, but do not use additional training examples from the CUB-2011 dataset for feature extraction. The most interesting result here is that we are able to outperform the Bubblebank approach of [9] despite of their heavy use of crowdsourced relevant part annotations.

Analysis of individual part detection In the following, we analyze the detection error of individual parts and compare our part transfer method with the DPM baseline, where semantic parts are used for initialization. We plotted the error distribution for three parts (back, nape, and leg) as a percentile plot in Figure 9 (left figure), where the error is defined as the normalized Euclidean distance between estimated and ground-truth part positions. First, it can be seen that the error distribution varies significantly across parts, which is consistent with the findings of [36]. Furthermore, and more importantly, our simple part transfer approach leads to significantly lower errors then the DPM. An analysis of the dependency between part detection and classification accuracy is finally visualized in the right part of Figure 9. In particular, we use the ground-truth part locations of the test images for our fine-grained recognition approach but added zero-mean Gaussian noise to it with a varying standard deviation. The plot reveals the part detection precision has a significant impact on the resulting classification performance, another motivation for nonparametric detection methods such as the one presented here.

Analysis of the feature representation Apart from our part detection method, we also proposed feature combinations of part and global features as well as using GrabCut [28] to reduce the influence of background clutter. Table 3 gives results of our method with and without *one* of these aspects. As can be seen in the accuracy numbers, a combination of local and global features is important and GrabCut masking can additionally help in general.

6. Conclusions

In this paper, we tackled the problem of fine-grained recognition, which is highly challenging particularly due to severe variations in object poses and viewpoints. We therefore introduced a nonparametric approach for part detection which is based on transferring part annotations from related training images to an unseen test image. This allows for a feature extraction step that focuses on those parts of images where discriminative features are likely to be located. In our experiments, we observed a significant gain over established part detection techniques like DPMs and also provided a theoretical analysis why DPMs suffer from high pose variations. Additionally, we showed how well-known techniques for object recognition, such as the combination of complementary feature types and image masking, can easily be added to obtain a simple yet powerful recognition

system for fine-grained classification scenarios. Despite of the simplicity of our approach, we were able to outperform previous approaches on the standard benchmark datasets CUB-2010 and 2011. Furthermore, our paper clearly motivates the use of nonparametric part and label transfer techniques, and might also help to bridge the gap between the branches of category-based recognition and instance-level matching. For future work, we are interested in exploitating part transfer for lifelong learning settings, *e.g.* during the process of active learning [15]. In addition, supervised feature transformations [4] might further help reducing the intra-class variations still apparant in correctly aligned parts of similar bird species.

References

[1] A. Angelova and S. Zhu. Efficient object detection and segmentation for fine-grained recognition. In *CVPR*, 2013. 6, 7

[2] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. *PAMI*, 35(12):2930–2940, 2013. 3

[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. 1, 4, 6

[4] P. Bodesheim, A. Freytag, E. Rodner, M. Kemmler, and J. Denzler. Kernel null space methods for novelty detection. In *CVPR*, pages 3374–3381, 2013. 8

[5] S. Branson, P. Perona, and S. Belongie. Strong supervision from weak annotation: Interactive training of deformable part models. In *ICCV*, 2011. 2

[6] Y. Chai, V. S. Lempitsky, and A. Zisserman. Bicos: A bi-level co-segmentation method for image classification. In *ICCV*, pages 2579–2586, 2011. 2

[7] Y. Chai, E. Rahtu, V. S. Lempitsky, L. J. V. Gool, and A. Zisserman. Tricos: A tri-level class-discriminative co-segmentation method for image classification. In *ECCV*, pages 794–807, 2012. 2

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005. 4

[9] J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *IEEE CVPR*, June 2013. 6, 7

[10] 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. *ArXiv e-prints*, Oct. 2013. 6

[11] K. Duan, D. Parikh, D. J. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *CVPR*, pages 3474–3481, 2012. 1

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, June 2010. 1

[13] 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*, 2009. 2

[14] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010. 1, 3, 4

[15] A. Freytag, E. Rodner, P. Bodesheim, and J. Denzler. Labeling examples that matter: Relevance-based active learning with gaussian processes. In *GCPR*, pages 282–291, 2013. 8

[16] E. Gavves, B. Fernando, C. Snoek, A. Smeulders, and T. Tuytelaars. Fine-grained categorization by alignments. In *ICCV*, 2013. 3

[17] Y. Jia, O. Vinyals, and T. Darrell. Pooling-Invariant Image Feature Learning. *ArXiv e-prints*, Jan. 2013. 6

[18] F. S. Khan, J. van de Weijer, A. D. Bagdanov, and M. Vanrell. Portmanteau vocabularies for multi-cue image representation. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *NIPS*, pages 1323–1331, 2011. 1, 2, 5

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 4

[20] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *PAMI*, 33(12):2368–2382, 2011. 2

[21] J. Liu and P. N. Belhumeur. Bird part localization using exemplar-based models with enforced pose and subcategory consistency. In *ICCV*, 2013. 3, 6

[22] J. Liu, A. Kanazawa, D. W. Jacobs, and P. N. Belhumeur. Dog breed classification using part localization. In *ECCV*, pages 172–185, 2012. 1, 2

[23] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011. 2, 4

[24] G. Martínez-Muñoz, N. L. Delgado, E. N. Mortensen, W. Zhang, A. Yamamuro, R. Paasch, N. Payet, D. A. Lytle, L. G. Shapiro, S. Todorovic, A. Moldenke, and T. G. Dietterich. Dictionary-free categorization of very similar objects via stacked evidence trees. In *CVPR*, pages 549–556, 2009. 2

[25] M.-E. Nilsback and A. Zisserman. Delving deeper into the whorl of flower segmentation. *Image Vision Comput.*, 28(6):1049–1062, 2010. 2

[26] A. of European Records and R. Committees. Taxonomic recommendations, 2003. 1

[27] O. M. Parkhi, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Cats and dogs. In *CVPR*, 2012. 1, 2

[28] C. Rother, V. Kolmogorov, and A. Blake. "grabcut": interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, 2004. 2, 6, 7

[29] M. Saar-Tsechansky and F. Provost. Handling missing values when applying classification models. *JMLR*, 8:1625–1657, 2007. 5

[30] V. Sreekanth, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Generalized RBF feature maps for efficient detection. In *Proceedings of the BMVC*, 2010. 4

[31] K. E. A. van de Sande, T. Gevers, and C. G. Snoek. Evaluating color descriptors for object and scene recognition. *PAMI*, 32(9):1582–1596, 2010. 5

[32] J. Van De Weijer, C. Schmid, J. Verbeek, and D. Larlus. Learning color names for real-world applications. *IEEE Transactions on Image Processing (TIP)*, 18(7):1512–1523, 2009. 5

[33] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *ICCV*, pages 606–613, 2009. 4

[34] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *PAMI*, 34(3):480–493, 2012. 5

[35] C. Wah, S. Branson, P. Perona, and S. Belongie. Multiclass recognition and part localization with humans in the loop. In *ICCV*, 2011. 2

[36] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, Caltech, 2011. 1, 3, 5, 6, 7

[37] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, Caltech, 2010. 1, 6

[38] S. Yang, L. Bo, J. Wang, and L. Shapiro. Unsupervised template learning for fine-grained object recognition. In *NIPS*, 2012. 1, 6, 7

[39] B. Yao, G. Bradski, and L. Fei-Fei. A codebook-free and annotation-free approach for fine-grained image categorization. In *CVPR*, 2012. 1, 2

[40] B. Yao, A. Khosla, and F.-F. Li. Combining randomization and discrimination for fine-grained image categorization. In *CVPR*, pages 1577–1584, 2011. 1

[41] N. Zhang, R. Farrell, and T. Darrell. Pose pooling kernels for subcategory recognition. In *CVPR*, pages 3665–3672, 2012. 1, 2

[42] N. Zhang, R. Farrell, F. Iandola, and T. Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *ICCV*, 2013. 1, 2, 6, 7