Fine-Grained Categorization by Alignments

E. Gavves[1], B. Fernando[2], C.G.M. Snoek[1], A.W.M. Smeulders[1, 3], and T. Tuytelaars[2]

[1]University of Amsterdam, ISIS
[2]KU Leuven, ESAT-PSI, iMinds
[3]CWI Amsterdam

Abstract

The aim of this paper is fine-grained categorization without human interaction. Different from prior work, which relies on detectors for specific object parts, we propose to localize distinctive details by roughly aligning the objects using just the overall shape, since implicit to fine-grained categorization is the existence of a super-class shape shared among all classes. The alignments are then used to transfer part annotations from training images to test images (supervised alignment), or to blindly yet consistently segment the object in a number of regions (unsupervised alignment). We furthermore argue that in the distinction of fine-grained sub-categories, classification-oriented encodings like Fisher vectors are better suited for describing localized *information than popular matching oriented features like HOG. We evaluate the method on the CU-2011 Birds and Stanford Dogs fine-grained datasets, outperforming the state-of-the-art.*

1. Introduction

Fine-grained categorization relies on identifying the subtle differences in appearance of specific object parts. Research in cognitive psychology has suggested [24] and recent works in computer vision have confirmed [10, 31, 34] this mechanism. Humans learn to distinguish different types of birds by addressing the differences in specific details. The same holds for car types [8], sailing boat types, dog breeds [15, 16], but also when learning to discriminate different types of pathologies. For this purpose, active learning methods have been proposed to extract attributes [9], volumetric models [10] or part models [3]. They require expert-level knowledge at run time, which is often unavailable. In contrast, we aim for fine-grained categorization without human interaction.

Various methods have been proposed to learn in an unsupervised manner, what details to focus on for identifying fine-grained sub-categories, such as the recent works relying on templates [31, 32]. In [32] templates rely on high dimensionalities to arrive at good results, while in [31] they are designed to be precise, being effectively analogous to "parts" [11]. Yet, it remains unclear what is the most critical aspect of "parts" in a fine-grained categorization context: *is it the ability to accurately localize corresponding locations over object instances, or simply the ability to capture detailed information?* While often these go hand in hand, as indeed is the case for templates, we defend the view that actually it is the latter that matters. We argue that a very precise "part" localization is not necessary and rough alignments suffice, as long as one manages to capture the fine-grained details in the appearance.

Figure 1. The first image shows a *Hooded Warbler*, whereas the second image shows a *Kentucky Warbler*. Based on example images like these, fine-grained categorization tries to answer the question: *what fine-grained bird category do we have in the third image?* Rather than directly trying to localize parts (be it distinctive or intrinsic), we show in this paper that better results can be obtained if one first tries to align the birds based on their global shape, ignoring the actual bird categories.

Parts may be divided in *intrinsic* parts [3, 16] such as the *head* of a dog or the *body* of a bird, and *distinctive* parts [32, 31] specific to few sub-categories. Recovering intrinsic parts implies that such parts are seen throughout the whole dataset. However, the large variability that naturally arises for large number of classes complicates their detection. Distinctive parts, on the other hand, are destined to be found on few sub-categories only. They are more consistent in appearance, as the distinctive *details* are better tailored to be detected on few sub-categories. On the downside, however, the number of sub-category specific parts soon becomes huge for large number of classes, each trained on a

small number of examples. This limits their ability to robustly capture the viewpoints, pose and lighting condition changes. Hence, detecting parts, be it intrinsic or distinctive, seems to involve contradictory requirements.

Different from prior work, we propose not to learn detectors for individual parts, but instead localize distinctive details by first roughly aligning the objects. This alignment is rough and insensitive to vast appearance variations for large number of sub-categories. Furthermore, rough alignment is not sub-category specific, thus the object representation becomes independent of the number of classes or training images [33, 32]. For alignments we only use the overall shape.

A first novelty of our work is based on the observation that all sub-categories belonging to the same super-category share similar global characteristics regarding their shape and poses. Therefore, it is effective to align objects, as we will pursue. In the supervised case, annotated details are transferred from training images to test images. In the unsupervised case, we use alignments to delineate corresponding object regions that we will use in the differential classification.

Our second novelty is based on the observation that starting from rough alignments instead of precise part locations, noticeable appearance perturbations will appear even between very similar objects, due to common image deformations such as small translations, viewpoint variations and partial occlusions. Using as fine-grained representations [10, 32, 34, 1] raw descriptors such as [17, 2, 32], that are precise, yet sensitive to common image transformations, is therefore likely to be a sub-optimal choice, especially when part detection becomes challenging. We propose to use state-of-the-art feature encodings, like Fisher Vectors [23], typically used for image classification, as local descriptors. In contrast to the raw SIFT or template features preferred in the fine-grained literature [16, 31, 32], such localized feature encodings are less sensitive to misalignments. Indeed, as our experiments indicate, they are better suited than matching based features.

We present two methods for recovering alignments that require varying levels of part supervision during training. We evaluate our methods on the CU-2011 Birds and Stanford Dogs dataset [30]. The results vouch for unsupervised alignments, which outperform previous published results.

2. Related work

Fine-grained categorization has entered the stage in the computer vision literature only recently. Prior works have focused on various aspects of fine-grained categorization, such as the description of fine-grained objects, the detection of fine-grained objects and the use of human interaction to boost recognition.

Fine-grained description. For the description of fine-grained objects various proposals have been made in the literature. In [32] Yao *et al*. propose to use color and gradient pixel values, arriving at high-dimensional histograms. Farell *et al*. [10] use color SIFT features, whereas Yang *et al*. [31] propose to use shape, color and texture based kernel descriptors [2]. Different from the above works, we propose to use strong classification- and not matching-oriented, encodings to describe the alignment parts and regions. Sanchez *et al*. in [13] and Chai *et al*. in [6] rely on classification-oriented encodings, Fisher vectors specifically, to learn a global object level representations. Inspired by their work we also adopt Fisher vectors. However, we use Fisher vectors not only as global, object level representations, but also as localized appearance descriptors.

Fine-grained detection. The detection of objects in a fine-grained categorization setting ranges from the segmentation of the object of interest [19, 5, 6] to fitting ellipsoids [10] and detecting individual parts and templates [33, 34, 32, 31, 16]. In their seminal work [19] Nilsback and Zisserman show the importance of segmenting out background information for recognizing flowers. Furthermore, in [5, 6] Chai *et al*. demonstrate how co-segmentation may be employed to improve classification. In the current work we also use segmentation, but with the intention to acquire an impression of the object's shape and to recover interesting object regions.

Targeting more towards parts instead of segmentations, Yao *et al*. propose to either sample discriminative features using randomized trees [33] or convolute images with hundreds of thousands of randomly generated templates [32]. Since a huge feature space is generated, tree pruning is employed to discard the unnecessary dimensions and make the problem tractable. In [10, 34] Farrell *et al*. capture the poses of birds, whereas in [34] Zhang *et al*. furthermore propose to normalize such poses and extract warped features, arriving at impressive results. In [21] Parkhi *et al*. propose to use deformable part models to detect the head of cats and dogs and in [1] Berg and Belhumeur learn discriminative parts from pairwise comparisons between classes. Also, in [16] Liu *et al*. propose to share parts between classes to arrive at accurate part localization.

Different from the above works, we do not directly aim at localizing individual parts, but rather at aligning the object as a whole. Based on this alignment, we then derive a small number of predicted parts (supervised) or regions (unsupervised). Such regions are highly repeatable, while few in number, thus ensuring consistency across the dataset and a smaller parameter space to learn our fine-grained object descriptions.

Human interaction. In [20] Parikh and Grauman iteratively generate discriminative attributes. They then evaluate and retain the "nameable" ones, that is the ones that can be interpreted by humans. In [4] Branson *et al*. try to determine

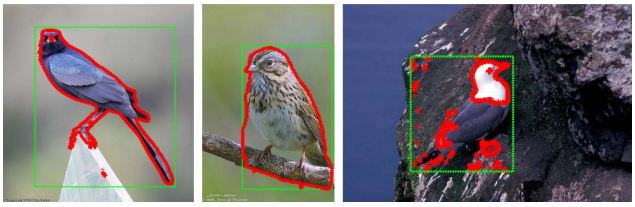

Figure 2. The computation of the segmentation mask can be accurate as in the left, ok as in the middle or completely fail as in the right image. Most times segmentations are somewhere in between the left and middle example, thus allowing us to obtain a rather good impression of the object's shape.

the object's sub-category using visual properties that can be easily answered by a user, such as whether the object "has stripes". In [29] Wah *et al*. propose an active learning approach that considers user clicks on object part locations, so that the machine learns to select the most informative question to pose to the user. In [9] Duan *et al*. propose to use a latent conditional random field to generate localized attributes that are both machine and human friendly. A user then picks those attributes that are sensible. And in [3] Branson *et al*. show that part models designed for generic objects do not always perform equally well for fine-grained categories. They therefore propose online supervision to learn better part models. The above approaches require time-consuming user input and often expert-knowledge. Hence, their applicability is usually restricted to small datasets covering only a limited number of fine-grained categories [9]. In the current work we propose a fine-grained categorization method that does not require any human interaction.

3. Alignments

A local frame of reference serves to identify the spatial properties of an object. In the following we will employ both shape masks and ellipses as local frames of reference. We say an image is aligned with other images if we have identified a local frame of reference in the image that is consistent with (a subset of) the frames of reference found in other images. Consistent means that corresponding parts are found in similar locations, when expressed relative to this frame of reference.

As is common in fine-grained categorization [33, 32, 31], we have available both at training and at test time the bounding box locations of the object of interest. We focus exclusively on the classification problem, leaving the problem of object detection for another occasion. Ignoring the image content outside the bounding box is a reasonable thing to do, since context is unlikely to play any major role in recognition of sub-categories, *e.g*., all birds are usually either on trees or flying in the sky.

The rectangular bounding box around an object allows for extracting important information, such as the approximate shape of the object. More specifically, we use GrabCut [25] on the bounding box to compute an accurate figure-ground segmentation. Although GrabCut is not always as accurate and in rare cases fails to recover even a basic contour, in the vast majority of cases it is able to return a rather precise contour of the object, see Fig. 2.

3.1. Supervised alignments

In the supervised scenario the ground truth locations of basic object parts, such as the *beak* or the *tail* of the birds, are available in the training set. This is a typical scenario when the number of images is limited, so that human experts can provide information at such a level of granularity. In this setting, we aim at accurately aligning the test image with a small number of training images. Then, we can use the common frame of reference to predict the part locations in the test image.

Our first goal is to retrieve a small number of training pictures that have a similar shape as the object in the test image. Note that, at this stage, it does not matter whether these are images that belong to the same sub-category or not. To this end, we first obtain the segmentation mask of the object as described before. Since we are interested only in the outer shape of the object, we suppress all the interior shape information. This gives us a shape mask for the image, which we effectively summarize in the form of HOG features [7].

A HOG feature forms in theory a high-dimensional, dense space. In practice, however, all the sub-categories belong to the same super-category, hence the generated poses will mainly lie on a lower dimensional manifold. Therefore, we can expect that given an object, there are several others with similar shapes and, that due to the anatomical constraints of the super-category they belong to, are likely to be found in similar poses. Given the ℓ_2-normalized HOG feature of the image shape mask, we retrieve the nearest neighbor images from the training set using a query-by-example setting. As a result, we end up with a shortlist of other similarly posed objects, see Fig. 3.

Having retrieved the training images with the most similar poses, the bounding boxes can be used as frames of reference. We are now in position to use the ground truth locations of the parts in the training images and predict the corresponding locations in the test image. To calculate the positions of the same parts on the test image, one may apply several methods of varying sophistication, ranging from simple average pooling of part locations to local, independent optimization of parts based on HOG convolutions. We experimentally witnessed that averaging yields accurate results, accurate enough to recover rough alignments. To ensure maximum compatibility we repeat the above procedure for all training and testing images in the dataset, thus predicting part locations for *all* the objects in the dataset.

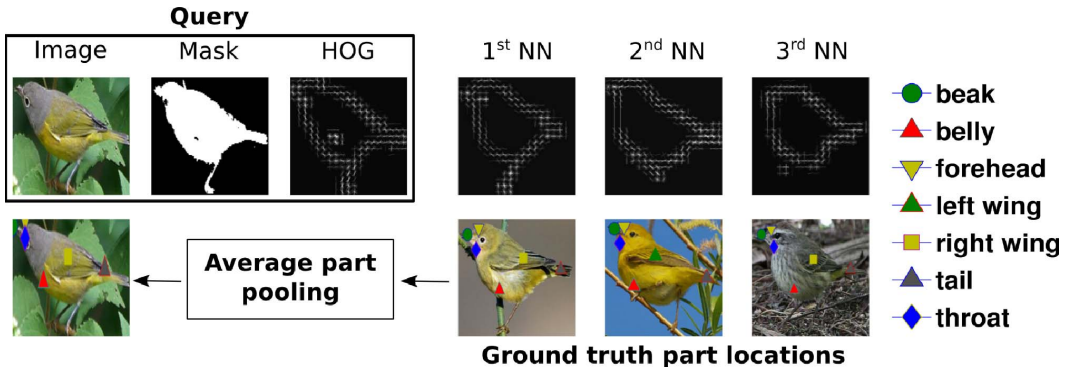

Figure 3. In the top left, we have a test image, for which we want to predict part locations. On the right, we have the nearest neighbor training images, their ground truth part locations and their HOG shape representations, based on which they were retrieved. Regressing the locations from the nearest neighbors to the test image we get the predicted parts, shown as the colorful symbols. The predicted part locations look quite consistent.

3.2. Unsupervised alignments

In the unsupervised scenario no ground truth information of the training part locations is available. However, we still have the bounding box that surrounds the object, based on which we can derive a shape mask per object.

Since no ground truth part locations are available, it does not make sense to align the test image to a small subset of training images. Instead, we derive a frame of reference based on the global object shape, inspired by local affine frames used for affine invariant keypoint description [18]. While not as accurate as the alignments in the previous subsection, this procedure allows us to obtain robust and consistent alignments over the entire database.

More specifically, we fit an ellipse to the pixels X of the segmentation mask and compute the local 2-d geometry in the form of the two principal axes

$$a_j = \bar{x} + \vec{e}_j \sqrt{\lambda_j} \tag{1}$$

In eq. (1) λ_j and $\vec{e}_j$ stand for the j-th eigenvalue and eigenvector of the covariance matrix $C = E[(X - \bar{x})(X - \bar{x})^T]$ and $\bar{x}$ is the average location of the mask pixels, see Fig. 4. GrabCut does not always return very accurate contours around the objects. Still, the centre of mass of the object is relatively stable to random fluctuations of the object contour. Thus, we let the ellipse axes meet each other at this point. To this end we extract the principal axes using *all* the foreground pixels of the shape mask.

For objects that have an elliptical shape the longer axis is usually the principal axis. Additionally, we follow the gravity vector assumption [22] and adopt the highest end point of the principal axis as its origin. Regarding the ancillary axis, we cannot easily define an origin in a consistent way. We therefore decide not to use the ancillary axis in the generation of consistent regions. This procedure fully defines the frame of reference, see Fig. 4.

Relative to this frame of reference, we can define different locations or regions at will. Here, we divide the principal axis equally from the origin to the end in a fixed number of segments, and define regions as the part of the foreground mask that falls within one such segment. Given accurate segmentation masks, the corresponding locations in different fine-grained objects are visited in the same order, thus resulting in pose-normalized representations, see Fig. 4. Small errors in the segmentations, as in the last row of picture of Fig. 4, have only a limited impact on the regions we obtain.

4. Final Image Representation

Our alignments are designed to be rough. Thus, using features that are precise, but sensitive to common image transformations, is likely to be suboptimal. Instead, we propose to use Fisher vectors [23] extracted in the predicted parts/regions. There are different ways one could sample from the alignment region to generate a Fisher vector. We turn our focus into two approaches, one that is more relevant to part based models and another one that is more relevant to consistent regions. For the first approach we sample in a $T \times T$ window around the center of the part, sampling descriptors every d pixels. Together with the object information this approach also captures some of the context that surrounds the object parts. For the second approach we sample densely every d pixels only on the intersection area of the segmentation mask and the region. This approach includes less context, as no descriptors centered to the background are extracted. Note that although the second approach is theoretically more accurate in capturing only the object appearance details, at the same time it might either include background pixels or omit foreground pixels, since segmentation masks are not perfect.

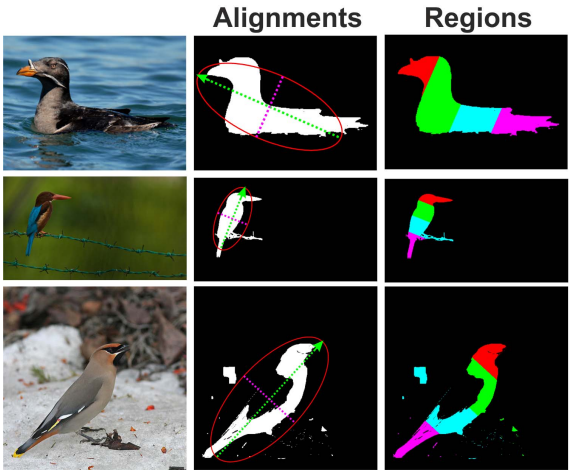

Figure 4. In the left column we see random birds, for which we have already extracted a segmentation mask. After fitting an ellipse, we obtain the two axes in the middle column pictures, the principal green and the ancillary magenta ones. After the gravity vector assumption [22] we assume the origin of the principal axis to be the highest point in the direction of the green arrow. Based on this frame of reference, we split equally in the right column pictures the principal axis to obtain consistent regions.

5. Experiments

5.1. Experimental setup

Datasets. We first run our experiments on the CU-2011 Birds dataset [30], one of the most extensive datasets in the fine-grained literature. The CU-2011 Birds dataset is composed of 200 sub-species of birds, several of whom bear tremendous similarities, especially under common image transformations, see Fig. 1. We use the standard training/test split provided by the authors. Following the standard evaluation protocol [33, 32, 31], we mirror the train images to double the size of the training set and use the bounding boxes to normalize the images. In total, we have 11,788 training and 5,894 testing images. We use the ground truth part annotations only during learning, unless stated otherwise. We also include results for Stanford Dogs [15], using similar experimental settings as for CU-2011 Birds. For both datasets we use their standard evaluation metrics, that is the category normalized mean accuracy.

Implementation details. We extract SIFT descriptors [28]. We sample densely every 3 pixels and at multiple scales (16x16, 24x24, 32x32 and 40x40 windows) for all experiments, unless stated otherwise. After extracting the SIFT descriptors we reduce dimensionality to 64 by applying a PCA transformation. For Fisher vectors we use a Gaussian mixture model with 256 components. We extract HOG [28] features on an 8 pixel spaced grid. We apply power- and ℓ_2- normalization on Fisher vectors [23] and ℓ_2 normalization on HOG. We use a linear SVM classifier [26] with a fixed parameter $C = 10$.

	Descriptors	
Part selection	*HOG*	*Fisher vectors*
Oracle	31.8	**52.5**

Table 1. Comparison of Matching vs Classification Descriptors based on accuracy. Fisher vectors are better equipped in describing part appearance than HOG for fine-grained categorization.

5.2. Matching vs Classification Descriptors

In this first experiment we evaluate what are good descriptors for describing parts in a fine-grained categorization setting. In order to ensure a fair comparison, as well as to test the maximum recognition capacity of parts for such a task, we use the ground truth part annotations both in training and in testing, as if an oracle algorithm for the part locations was available. If Fisher vectors outperform HOG on perfectly aligned ground truth parts, then we expect this to be the case even more for less accurate parts. The CU-2011 Birds contains 15 part annotations per bird, many of which are spatially very close to each other. In order to avoid a too strong correlation between the parts and also control the dimensionality of the final feature vector we use only the following 7 parts, which cover the bird silhouette: *beak, belly, forehead, left wing, right wing, tail* and *throat*. In a square window of $T \times T$ pixels centering the part location we extract HOG and Fisher vectors. We set $T = 100$ pixels, a value that seemed to work well in practice. The Fisher vectors from the 7 parts are concatenated with a Fisher vector from the whole bounding box to arrive at the final object representation. Similarly, for the HOG object descriptors we also compute a HOG vector using the bounding box, rescaled to 100×100 pixels.

As we see in Table 1, Fisher vectors are much better in describing parts for fine-grained categorization than matching based descriptors like HOG. Where HOG scores an accuracy of 31.8 the Fisher vectors result in a final score of 52.5. The reason is that HOG descriptors capture only zero order spatial gradient information of the window around the part. However, for fine-grained classes the gradients are often quite similar, since they belong to the same superclass. Hence, Fisher vectors are able to better describe the little nuances in the gradients, since they are specifically designed to capture also first and second order statistics of the gradient information. We plot in the left image of Fig. 5 the individual accuracies per class for Fisher vectors and for HOG, noticing that Fisher vectors outperform for 184 out of the 200 sub-categories. In the following experiments we report results using only Fisher vectors for describing the appearance of parts and alignments.

5.3. Supervised alignments

In the second experiment we test whether supervised alignments actually benefit the recognition of fine-grained categories, as compared to a standard classification pipeline.

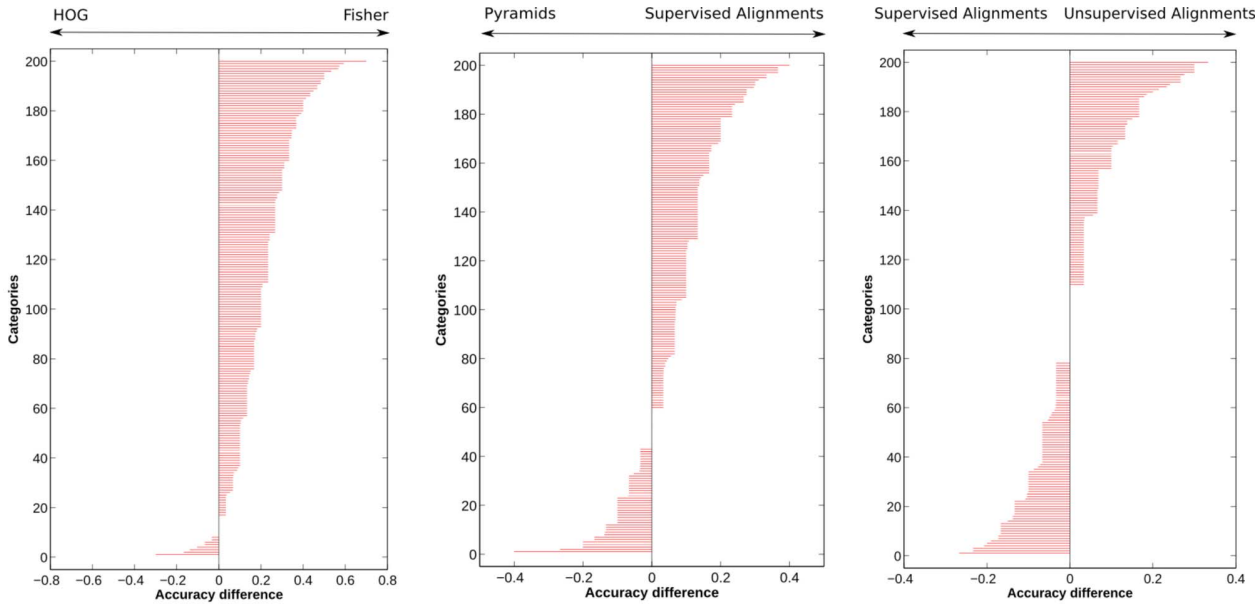

Figure 5. A fine-grained category-by-category comparison. We report results on the 200 concepts in CU-2011 Birds, measured in terms of accuracy. Falling at the right side of the reference line x=0 means that for oracle parts the Fisher vector is better than HOG (left picture), Fisher vector on parts is more accurate than a 2×2 spatial pyramid kernel (middle picture), and Fisher vectors on unsupervised alignments are more accurate than Fisher vectors on parts derived from supervised alignments (right picture). The difference is not that big (2%), but note that for Fisher vector unsupervised alignments no ground truth part locations are required.

Part selection	**Fisher vectors**
$[2 \times 2]$ spatial pyramid kernel	39.8
Supervised alignment on beak only	37.8
Supervised alignments	**47.1**

Table 2. Supervised alignments are more accurate than a spatial pyramid kernel and an alignment based on the beak of a bird only, while being rather close to the theoretical accuracy of the oracle parts in Table 1.

Here our supervised alignments use ground truth part annotations *only in training*. We use the same 7 parts as in the previous experiment plus a Fisher vector extracted from the whole bounding box. We predict their location by averaging the locations of the parts in the top 20 nearest neighbors. If a part is not present for the majority of the top 20 nearest neighbors, we consider this part absent and set the corresponding Fisher vector to zero. We compare our proposed supervised alignment method against a 2×2 spatial pyramid using Fisher vectors computed from all SIFT descriptors in the bounding box. Also, inspired by [16], we repeat the same experiment using only the predicted location of the *beak*, whose window captures most of the information around the head. We extract the Fisher vector on $T \times T$ windows around the predicted part locations, where T is again set to 100 pixels. We show the results in Table 2.

As we observe in Table 2, parts bring an 17% accuracy improvement over a standard spatial pyramid classification approach, since they better capture the little nuances that differentiate sub-classes that are otherwise visually very similar. Furthermore, we note that extracting Fisher vectors on the supervised alignments is 47.1% accurate, which is rather close to the 52.5% obtained when extracting Fisher vectors on the parts provided by the ground truth. This indicates that we capture the part locations well enough for an appearance descriptor like the Fisher vector. In fact, the mean squared error between our estimated parts and the ground truth ones is 12%, after normalizing the respective locations with respect to the bounding box geometry. We plot in the middle picture of Fig. 5 the individual accuracies per class for our part prediction method and the spatial pyramids. Our supervised alignments perform consistently better for 141 out of the 200 classes. We conclude that extracting localized information in the form of alignments or parts matters in a fine-grained categorization setting.

5.4. Unsupervised Alignments

In this experiment we compare the unsupervised alignments with the supervised ones. After extracting the principal axis we split the bird mask into four regions, starting from the highest point, considering only the pixels within the segmentation mask. We furthermore compare our method against a horizontally split $[4 \times 1]$ spatial pyramid. We show the results in Table 3.

We observe that describing the object based on the unsupervised alignments results in more accurate predictions compared to the supervised case (49.4% vs 47.1%). When

Figure 6. Best recognized categories, that is *Pied billed Grebe, Heermann Gull, Bobolink* and *European Goldfinch*. We observe that birds in these sub-classes have consistent appearance.

Part selection	Fisher vectors
Supervised alignments	47.1
$[4 \times 1]$ spatial pyramid kernel	39.4
Fisher vector from the foreground mask only	42.6
Unsupervised alignments	**49.4**

Table 3. Unsupervised alignments are more accurate than supervised ones, while at the same time requiring no supervision at all.

computing a single Fisher vector only from the foreground mask we obtain an accuracy of 42.6%. Note that unsupervised alignments use no ground truth part annotations, neither in training nor in testing. We repeat the experiment considering different number of regions. For 2 regions the accuracy decreases from 49.4% to 46.2%, whereas for 6 regions we obtain 49.3%. In the subsequent experiments we always use 4 regions.

We, furthermore, plot the individual accuracy differences per class for supervised and unsupervised alignments in the right picture in Fig. 5. The distribution of classes is split roughly equally for supervised and unsupervised alignments, with unsupervised alignments having slightly larger accuracy differences. We conclude that compared to supervised parts, unsupervised alignments describe the localized appearance of fine-grained objects at least as good, often better.

Birds	Accuracy
Pose pooling kernels [34]	28.2
Pooling feature learning [12]	38.9
POOF [1]	56.9
This paper: Unsupervised alignments	**62.7**

Table 4. State-of-the-art comparison in CU-2011 Birds [30]. Unsupervised alignments with Fisher vectors outperform the state-of-the-art considerably.

Dogs	Accuracy
Discriminative Color Descriptors [14]	28.1
Edge templates [31]	38.9
This paper: Unsupervised alignments	**50.1**

Table 5. State-of-the-art comparison in Stanford Dogs [15]. Unsupervised alignments with Fisher vectors outperform the state-of-the-art considerably.

5.5. State-of-the-art comparison

In experiment 4, we compare our unsupervised alignments with state-of-the-art methods reported on CU-2011 Birds and Stanford Dogs. We add color by sampling SIFT descriptors from the opponent color spaces [27]. Results for birds are shown in Table 4. Compared to the very recently published POOF features [1], unsupervised color alignments are 10% more accurate, while not requiring ground truth part annotations. Compared to the pose pooling kernels, unsupervised alignments recognize bird subcategories 84% more accurately. And compared to learned features proposed in [12] unsupervised alignments perform 36.5% better. Also for Stanford Dogs we outperform the state-of-the-art, in spite of the larger shape and pose variation among the dogs compared to the birds, see Table 5.

Although no direct comparison can be made, we report also some numbers from prior works on CU-2010 Birds, which is the previous version of CU-2011 Birds. The highest recorded accuracy is 28.2% for templates and kernel descriptors [32]. Using co-segmentation, [6] reports an accuracy of 25.5%, whereas randomized features [33] perform 19.2% accurately. On a subset of 14 out of 200 bird species the codebook-free approach of [32] is 44.7% accurate. It is interesting to note that interactive approaches on CU-2010 Birds report approximately 50% accuracy either within 25 seconds of human interaction [29] or after asking 15 questions to the user [4]. Our approach requires no supervision to reach a similar accuracy, albeit on a bigger dataset.

In Fig. 6 we plot pictures from four categories for which alignments reach high accuracy, *i.e. Pied billed Grebe, Heermann Gull, Bobolink* and *European Goldfinch*. The primary reason for the good recognition performance of these classes appears to be their consistent appearance both in training and testing sets.

In Fig. 7 we show images of the two categories most

Figure 7. The two most confused categories, that is *Loggerhead Shrikes* in the left column and *Great Grey Shrikes* in the right column. These two classes have very similar appearance, thus often resulting in confusion also for alignments.

confused to each other: *Loggerhead Shrike* and *Great Grey Shrike*. These two sub-species belong to the same family and have very similar appearance even when color is added, resulting to high confusion.

6. Conclusions

In this paper we aim for fine-grained categorization without human interaction. Different from prior work, we show that localizing distinctive details by roughly aligning the objects allows for successful recognition of fine-grained subclasses. We show that for rough alignments, classification-oriented encodings, such as Fisher vectors, are a better choice than matching based features, such as HOG. We present two methods for extracting alignments, requiring different levels of supervision. We evaluate on the CU-2011 Birds and Stanford Dogs dataset, outperforming the state-of-the-art. We conclude that rough alignments lead to accurate fine-grained categorization.

7. Acknowledgements

The projects IMPact BeeldCanon, AXES, STW STORY and the Dutch national program COMMIT support this research.

References

[1] 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.

[2] L. Bo, X. Ren, and D. Fox. Kernel descriptors for visual recognition. In *NIPS*, 2010.

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

[4] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *ECCV*, 2010.

[5] Y. Chai, V. Lempitsky, and A. Zisserman. Bicos: A bi-level co-segmentation method for image classification. In *ICCV*, 2011.

[6] Y. Chai, E. Rahtu, V. Lempitsky, L. Van Gool, and A. Zisserman. Tricos: a tri-level class-discriminative co-segmentation method for image classification. In *ECCV*, 2012.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[9] K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *CVPR*, 2012.

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

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

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

[13] F. P. Jorge Sanchez and Z. Akata. Fisher vectors for fine-grained visual categorization. In *CVPR*, 2011.

[14] R. Khan, J. Van de Weijer, F. S. Khan, D. Muselet, C. Ducottet, and C. Barat. Discriminative color descriptors. In *CVPR*, 2013.

[15] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei. Novel dataset for fine-grained image categorization. In *CVPR, FGVC workshop*, 2011.

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

[17] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[18] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *IJCV*, 65(1-2):43–72, 2005.

[19] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008.

[20] D. Parikh and K. Grauman. Interactive discovery of task-specific nameable attributes. In *CVPR, FGVC workshop*, 2011.

[21] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar. Cats and dogs. In *CVPR*, 2012.

[22] M. Perdóch, O. Chum, and J. Matas. Efficient representation of local geometry for large scale object retrieval. In *CVPR*, 2009.

[23] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010.

[24] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cogn. psych.*, 8(3):382–439, 1976.

[25] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM TOG*, 23(3):309–314, 2004.

[26] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *ICML*, 2007.

[27] K. Van De Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *PAMI*, 2010.

[28] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *MM*, 2010.

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

[30] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, 2011.

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

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

[33] B. Yao, A. Khosla, and L. Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *CVPR*, 2011.

[34] N. Zhang, R. Farrell, and T. Darrell. Pose pooling kernels for subcategory recognition. In *CVPR*, 2012.