Int J Comput Vis
DOI 10.1007/s11263-014-0741-5

Local Alignments for Fine-Grained Categorization

Efstratios Gavves · Basura Fernando · Cees G. M. Snoek · Arnold W. M. Smeulders · Tinne Tuytelaars

Received: 26 February 2014 / Accepted: 13 June 2014

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. Then, one may proceed to the classification by examining the corresponding regions of the alignments. More specifically, the alignments are used to transfer part annotations from training images to unseen images (supervised alignment), or to blindly yet consistently segment the object in a number of regions (unsupervised alignment). We further argue that for the distinction of subclasses, distribution-based features like color Fisher vectors are better suited for describing localized appearance of fine-grained categories than popular matching oriented shape-sensitive features, like HOG. They allow capturing the subtle local differences between subclasses, while at the same time being robust to misalignments between distinctive details. We evaluate the local alignments on the CUB-2011 and on the Stanford Dogs datasets, composed of 200 and 120, visually very hard to distinguish bird and dog species. In our experiments we study and show the benefit of the color Fisher vector parameterization, the influence of the alignment partitioning, and the significance of object segmentation on fine-grained categorization. We, furthermore, show that by using object detectors as voters to generate object confidence saliency maps, we arrive at fully unsupervised, yet highly accurate fine-grained categorization. The proposed local alignments set a new state-of-the-art on both the fine-grained birds and dogs datasets, even without any human intervention. What is more, the local alignments reveal what appearance details are most decisive per fine-grained object category.

Keywords Alignment · Image representation · Object classification

Communicated by Florent Perronnin.

E. Gavves (✉) · C. G. M. Snoek · A. W. M. Smeulders
University of Amsterdam, Amsterdam, Netherlands
e-mail: efstratios.gavves@gmail.com

B. Fernando · T. Tuytelaars
KU Leuven, ESAT PSI, iMinds, Leuven, Belgium

1 Introduction

According to cognitive psychology, fine-grained categorization of images, like the ones in Fig. 1, relies on identifying small differences in appearance of specific object parts (Rosch et al. 1976). Humans learn to distinguish different types of birds by addressing the differences in specific details. Recent works in computer vision have verified this mechanism (Farrell et al. 2011; Zhang et al. 2012; Chai et al. 2013; Zhang et al. 2013; Berg and Belhumeur 2013). The same holds for car types (Deng et al. 2009), aircraft types (Maji et al. 2013) and dog breeds (Khosla et al. 2011; Liu et al. 2012). Active learning methods have been proposed to extract attributes (Duan et al. 2012), volumetric models (Farrell et al. 2011) or part models (Branson et al. 2011). Such methods expect user input at runtime. In contrast, we aim for fine-grained image categorization from training example images, with no interaction other than the fine-grained label.

Various methods learn what details to focus on for fine-grained categorization. While good results have been obtained by relying on high dimensional template matching procedures (Yao et al. 2012), parts are adopted as the natural template (Zhang et al. 2013). Yet, it remains unclear *how important it is to be able to accurately localize*

Published online: 19 July 2014

corresponding locations over object instances even if that reduces the ability of capturing detailed information from raw visual data? While often these go hand in hand, *e.g.*, when using templates, we defend the view that actually it is the latter that matters most. Therefore, we argue that precise localization is not always necessary. Rough alignments suffice, as long as one manages to capture the distinctive details in the appearances.

Localizing consistent locations on instances of certain object categories is strongly related to part learning. Parts are divided in *intrinsic* parts, *i.e.* semantic parts that are shared by all (or at least most of) the sub-classes, as in (Branson et al. 2011; Liu et al. 2012), such as the *head* of a dog or the *body* of a bird, as opposed to *distinctive* parts, as in (Yao et al. 2012; Yang et al. 2012) specific to a few sub-classes. The large variability in poses and appearances renders the clean detection of intrinsic parts difficult. In contrast, distinctive parts are most likely to be found on few sub-classes only. They are more consistent in appearance, as the distinctive detail is better tailored to be detected on few sub-classes. Still, the number of sub-class specific parts soon becomes huge, each trained on a small number of examples. This limits the robust capturing of all viewpoints, poses and condition changes. Hence, detecting parts, be it intrinsic or distinctive, both have their difficulties in the learning phase.

Rather, we propose to roughly localize distinctive details by first aligning the objects. This alignment is rough and insensitive to most appearance variations. Rough alignment is not sub-class specific, thus the object representation becomes independent of the number of classes or training images (Yao et al. 2011, 2012). In essence, rough alignment rests on the assumption that the sub-classes share a rough shape.

Within a fine-grained categorization setting sub-classes belonging to the same super-class often feature a similar pose and posture. As our first contribution we exploit this observation to predict the location of the interesting object parts in a top-down manner. We first align objects and then loosely define parts by their location on the object, either in a supervised or an unsupervised fashion. This contrasts to bottom-up models (Farrell et al. 2011; Yang et al. 2012; Chai et al. 2013; Zhang et al. 2013), where one explicitly learns appearance models for individual parts.

As our second contribution we propose to capture the *appearance* variations of the estimated fine-grained parts (using pooling-based encodings), rather than using *shape*-sensitive descriptors such as HOG (Dalal and Triggs 2005) or kernel descriptors (Bo et al. 2010) on deformable parts (Felzenszwalb et al. 2010) as used in (Farrell et al. 2011; Yang et al. 2012; Chai et al. 2013; Zhang et al. 2013). In particular, we use Fisher vectors (Perronnin et al. 2010) on color SIFT descriptors (van de Sande et al. 2010), which have shown to improve object detection accuracy in large datasets (Deng et al. 2009). To the best of our knowledge we are the first to evaluate whether appearance is more effective than shape when describing fine-grained parts, and with the exception of the concurrent work from (Chai et al. 2013; Zhang et al. 2013), the first to propose pooling-based encodings to describe fine-grained parts.

Thirdly, we assess the impact of segmentation on fine-grained categorization. We first quantify the relationship between segmentation accuracy and fine-grained categorization accuracy, having as a baseline perfect ground truth segmentation. This allows us to draw conclusions independent of the segmentation model unlike (Nilsback and Zisserman 2008; Chai et al. 2011, 2012). However, in a realistic recognition scenario no indication of the object location is provided, leading to difficulties in segmenting the foreground from the background (Chai et al. 2011). To overcome this we propose to compute object saliency maps by averaging object detector proposals and use them as priors for the subsequent segmentation. Last, to make the object part representation more robust to deformations and clutter, we refine square parts with segmentation.

The methodology we present allows for performing fine-grained categorization with minimal human interaction. Where user input is often required both during training and testing, either in providing the object location (Gavves et al. 2013; Chai et al. 2013; Zhang et al. 2013), its parts (Branson et al. 2011) or its attributes (Duan et al. 2012; Branson et al. 2014), we present a system that does not rely on any user input, not even during training, without sacrificing the fine-grained categorization accuracy. In fact, we show that competitive accuracy is obtained even when no bounding boxes are provided neither during training nor testing. Thus, we can limit the amount of human interaction required to just providing labelled training images.

To achieve a better understanding of the fine-grained categorization process and the limitations of visual features, we conclude by performing a qualitative analysis. Where visual features extracted from the fine-grained object fail to discern between species, possibly due to almost identical appearance, one could attempt to analyze the environment, as Darwin (1859) would argue. Moreover, we attempt to answer what makes a bobolink a bobolink. We find that advanced, orderless, features, such as Fisher vectors, operate as a spatial hashing function, that builds correspondences between spatial details and certain feature dimensions.

This paper is an extension of our previous work (Gavves et al. 2013). Compared to our earlier version, we present a richer related work section, and we enrich our methodology by *(i)* extending the types of fine partitionings and *(ii)* alleviating the bounding box requirement at runtime. Furthermore, we extend the experimental section by including seven more

experiments, qualitatively and quantitatively evaluating all extensions on the challenging *Birds* (Wah et al. 2011b) and *Dogs* (Khosla et al. 2011) datasets.

We proceed with presenting a list of related works on fine-grained categorization in Sect. 2. In Sect. 3 we describe the proposed method, including the localization, the extraction and the description of alignments. Experiments are presented in Sect. 4 and we conclude in Sect. 5.

2 Related Work

We organize our discussion on related fine-grained categorization works by the vision tasks involved: *localization*, *partitioning*, and *description*. Within each task we organize the papers by the amount of required human intervention.

2.1 Localization

Many works in fine-grained categorization assume that the (bounding box) location of the object is available, both at *training and test phase*, see (Yao et al. 2011, 2012; Yang et al. 2012; Jia et al. 2013; Berg and Belhumeur 2013; Chai et al. 2013; Donahue et al. 2013; Xie et al. 2013; Gavves et al. 2013). Knowing *a priori* the location of the fine-grained object allows to focus on the detection and description of the fine-grained details only. Hence, the above works report the highest recognition rates in the literature, although it was shown by Yao et al. (2012) that a bad bounding box can be more harmful than having no box at all. In the current work we localize fine-grained objects, without requiring a bounding box.

Others require annotations *only during training*. Inspired by the poselets of Bourdev and Malik (2009), Farrell et al. (2011) use volumetric primitives, the "birdlets", parameterized to reflect the 3-*D* geometry of the body and head of birds, resulting in pose normalized representations. Since birdlets require expensive 3-*D* ground truth annotations, they are limited to small datasets. Therefore, Zhang et al. (2012) propose to first employ simpler to detect 2-*D* poselets, which are then warped in order to arrive at a consistent, pose-normalized representation. Others require only bounding boxes for the location of the fine-grained objects during training. Wah et al. (2011a); Branson et al. (2011) employ deformable part models (Felzenszwalb et al. 2010) for detection, showing, however, that user feedback is necessary to improve accuracy. In contrast to the above works, we localize fine-grained objects without requiring anything but the class label for training.

Others working under such conditions proceed with fine-grained categorization, without expecting any information regarding the location of the fine-grained objects, *neither during training nor during testing*. While Sanchez et al. (2011) focus on image-level descriptions, purposefully ignoring the spatial aspect, the main focus has been to discover the object's location in an unsupervised manner, usually applying image-level segmentation like in Nilsback and Zisserman (2008), or co-segmentation methods like in Chai et al. (2011, 2012). We rely on segmentation as well.

We propose a multi-functional approach that performs accurate fine-grained categorization, when bounding boxes are *(i)* provided during training and testing, *(ii)* only during training, using supervised object detectors like (Felzenszwalb et al. 2010) at test time or *(iii)* not provided at all, using unsupervised object proposals like (Uijlings et al. 2013; Manén et al. 2013) both at training and test time. In the latter case we report competitive recognition rates that often outperform methods requiring bounding boxes. Last, we evaluate the importance of accurate segmentation.

2.2 Partitioning

When classifying different bird sub-classes, like telling the *Forster's Tern* apart from the *Least Tern*, see Fig. 1, one probably needs to discover details such as their beak color patterns. Since consistently localizing such details is assumed to be crucial, a large part of the fine-grained literature has put considerable effort in this task, see (Farrell et al. 2011; Zhang et al. 2012; Yao et al. 2011; Zhang et al. 2012; Yao et al. 2012; Yang et al. 2012; Liu et al. 2012; Berg and Belhumeur 2013; Xie et al. 2013; Chai et al. 2013).

Some methods focus on an active learning approach for detecting locations. Wah et al. (2011a) consider user clicks, guiding the machine to pose the most informative question to the user, while Branson et al. (2011) propose online supervision to learn better part models. Given ground truth part annotations, part sharing between classes was shown by Liu et al. (2012) to result in accurate dog breed recognition. Going one step further, Xie et al. (2013) demonstrate excellent results for fine-grained categorization, assuming that ground truth part annotations are available also at runtime. We do not require part annotations at runtime.

The majority of works, however, targets towards automatic partitioning. Yao et al. (2011) use randomized trees to mine discriminative features. In (Yao et al. 2012) the same authors propose to randomly generate thousands of templates, which after being convolved with the unseen images lead to very high-dimensional representations. Extracting unsupervised templates, which take into account part appearance, co-occurrence and diversity, was shown by Yang et al. (2012) to deliver excellent results in several datasets. Inspired by the partial object model of Biederman (1987), Farrell et al. (2011) and Zhang et al. (2012) consider the head of a bird as most discriminative, using it to perform recognition. Moreover, Berg and Belhumeur (2013) showed that ground truth part annotations can be used for designing intricate features specific to certain sub-categories, arriv-

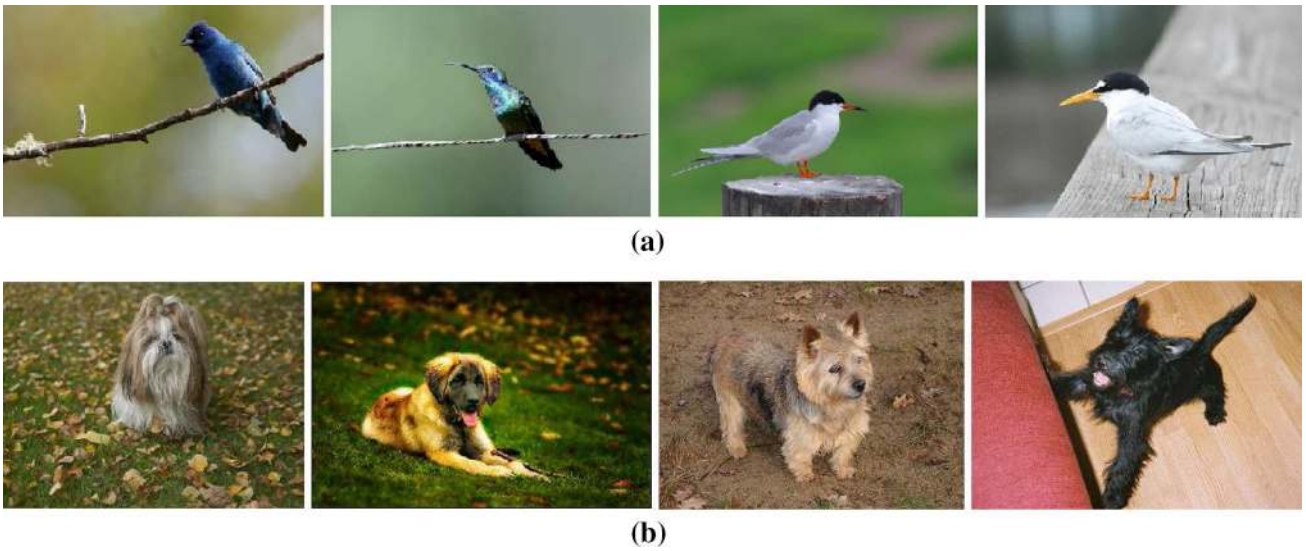

Fig. 1 Examples of fine-grained sub-classes for the Birds and Dogs datasets. Note the difficulty of recognizing these categories in a finer detail. **a** All four birds belong to different sub-classes, although some of them look very similar. **b** Dogs appear in all kinds of position, poses and scales. Based on example images like these, fine-grained categorization tries to discover *which fine-grained species each image belongs to*. Rather than directly trying to localize parts (be it distinctive or intrinsic, see text), we propose to first roughly align the objects based on their global shape, ignoring the actual fine-grained category. After aligning the object, we then proceed with consistent partitioning, arriving at successful classification

ing at excellent recognition rates. And recently, Chai et al. (2013) and Zhang et al. (2013) proposed to employ modified deformable part models (Felzenszwalb et al. 2010), to detect consistent fine-grained parts that allow for pose-normalized representations.

Similar to the above works, we detect interesting object locations for discriminating between sub-species. Different from the above works, we do not aim at directly localizing individual parts. Instead, we propose to first align the object as a whole. Based on this alignment, we then derive a small number of partitionings. Although our alignments and the subsequent partitionings can benefit from supervision during training, we show that obtaining them in an unsupervised manner is feasible, leading to high recognition in fine-grained categorization that outperforms the state-of-the-art.

2.3 Description

For the description of features several possibilities have been explored in the literature, some of them requiring user assistance, while the majority is fully unsupervised.

Methods that propose user-assisted features mainly focus on interpretable attributes. Discovering discriminative, user-accredited attributes, *e.g.* whether a bird has spots or not, has been repeatedly explored by Parikh and Grauman (2011); Branson et al. (2010). In a similar manner, Duan et al. (2012) detect mid-level attributes, which are, however, location and not image-level specific. Since attributes need to be interpretable to make sense, human labor and often expert knowledge is required, rendering these approaches useful for small datasets only as in Duan et al. (2012). In our work we do not attempt to represent fine-grained objects in terms of mid-level features or attributes.

Most works in the fine-grained categorization literature do not require human-interpretable features. Raw features, such as intensity SIFT proposed by Lowe (2004) or kernel based descriptors proposed by Bo et al. (2010) have shown to be good choices in describing the distributions of low level appearance details, such as edges or color (Farrell et al. 2011; Yang et al. 2012). However, being sensitive to misalignments renders them less suited for objects that are distorted in the presence of common image deformations. To cope with such misalignments, feature encodings have also been proposed. Locality-constrained linear coding in Yao et al. (2011), bag-of-words in Zhang et al. (2012) and Fisher vectors in Sanchez et al. (2011); Chai et al. (2012, 2013) were shown to describe fine-grained categories accurately. For an excellent review on how to adapt Fisher vector for fine-grained categorization we refer to Gosselin et al. (2013). Furthermore, Berg and Belhumeur (2013) showed that supervised features trained to be discriminative for pairs of classes achieve state-of-the-art results. And Donahue et al. (2013) showed that employing a deep learning architecture specialized to fine-grained subcategories arrives at remarkable recognition rates, at the expense of requiring additional images for feature learning. Here, we also propose to use unsupervised features, more specifically Fisher vectors (Perronnin et al. 2010). Different from most previous works, we extend Fisher vectors to operate not only

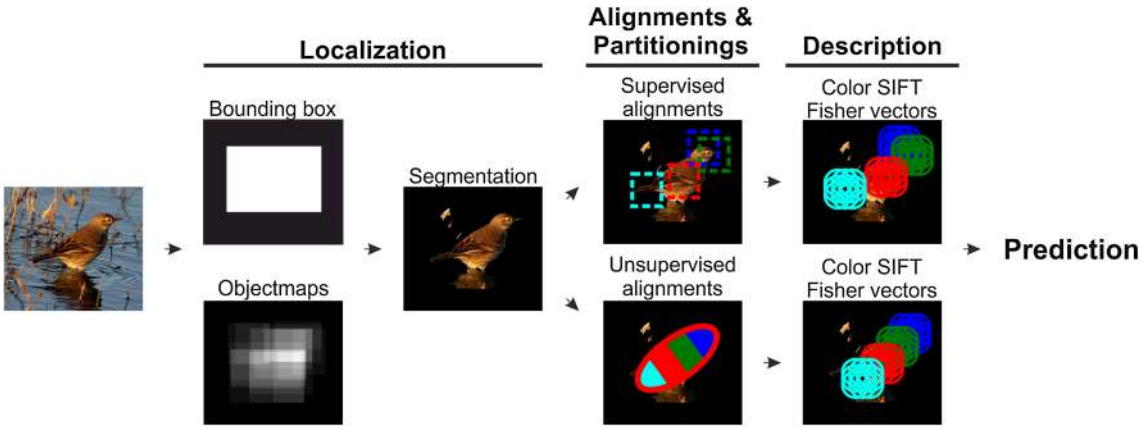

Fig. 2 Block diagram of the system. The proposed system results in competitive recognition rates, even when no user input (bounding boxes, part locations) is to be expected, not even during training. Naturally, additional annotations allow for even higher recognition accuracy. The individual blocks are detailed in Sect. 3

as global, object-level representations, but also to encode the *localized* appearance of object parts.

Another interesting aspect of the description of object locations is the exploitation of domain specific, low-level appearance, such as color. Intuitively, in fine-grained sub-categories of the natural world, such as birds species, see (Wah et al. 2011b) or dogs breeds, see (Khosla et al. 2011), color is bound to have a great impact in telling sub-categories apart. Surprisingly enough, the recent fine-grained literature (Farrell et al. 2011; Yao et al. 2012) often focuses on more traditional color based descriptors as found in (Swain and Ballard 1991) rather than state-of-the-art solutions, see (van de Sande et al. 2010). We evaluate and highlight the potential of color in fine-grained categorization, when advanced color descriptors are considered.

3 System

Within a fine-grained categorization setting we assume an image I contains an object belonging to one of the $1, \ldots, K$ sub-categories of interest. Naturally, there might be several other objects present in the image and not just the fine-grained object. Furthermore, we do not restrict the location and scale of the fine-grained object. Although in a fine-grained categorization setting these problems are often evaded by assuming that bounding boxes are provided by humans at query time like in (Yao et al. 2012; Berg and Belhumeur 2013; Chai et al. 2013; Gavves et al. 2013), in real world scenarios it is not always realistic to expect such user input. Therefore, localization of the object of interest needs to precede any further fine-grained analysis regarding the specific sub-category that is depicted. For localization, we propose to use object detection as a soft prior for segmentation, to avoid important details to be missed.

The localization provides a local frame of reference that serves to identify the spatial properties of the object. When we identify a local frame of reference in an image, consistent with other local frames of reference in other images, then we call the image aligned. Consistent means that corresponding parts are found in corresponding locations, when expressed with respect to their frame of reference.

By design we opt for finding the parts consistently, at the cost of less precise detections, accepting the small drift in part appearance that might occur. To avoid being oversensitive to such drifts, we choose our supervised and unsupervised alignments to be rough but consistent, rather than precise but unstable. Given the rough nature of our alignments, we show that orderless, powerful features are the preferred choice.

An overview of the system is illustrated in Fig. 2.

3.1 Localization

3.1.1 Why Not an Object Detector?

In order to discover the spatial support of an object the apparent choice is to employ an object detection algorithm, see (Uijlings et al. 2013; Felzenszwalb et al. 2010; Manén et al. 2013; Vedaldi et al. 2009). In that case, we predict the best possible bounding box that surrounds the object of interest as tightly as possible. A successful detection D is evaluated with respect to the amount of *overlap* between the predicted bounding box and the ground truth bounding box G

$$overlap = \frac{D \cap G}{D \cup G}. \quad (1)$$

The overlap penalizes both inclusion of extra background and the exclusion of foreground. Since detection is difficult by nature, usually some error margin is allowed. This

error margin is expressed as a minimum overlap threshold, above which detection is considered to be correct. State-of-the-art challenges (Everingham et al. 2007) set this threshold to 50 %. The design of the overlap measure in Eq. (1), therefore, suggests that detections should minimize the amount of the background in the detection D, even if some foreground is missed.

This setup, however reasonable for object detection, may cause problems to the subsequent segmentation required for fine-grained categorization, see (Wah et al. 2011a; Branson et al. 2011). To illustrate with an example, having a box overlapping 50 % with the object of interest suffices for an object detector. However, 50 % of overlap also implies that a large chunk of the object's body may be missed, thus potentially losing the crucial details that make the difference between, *e.g.* the "Magnolia Warbler" and the "Myrtle Warbler". Furthermore, performing segmentation for all the bounding box candidates returned by state-of-the-art object detectors, like (Uijlings et al. 2013; Cinbis et al. 2013), would be computationally challenging. To this end we propose to alter the way traditionally object detectors are employed and use them as soft priors for segmentation.

3.1.2 Objectmaps

We start from an object (proposal) detector. In order to remain agnostic to the type of detector, we make no assumptions other than the detector should return a sizable number of bounding boxes $\{D_i\}$ that indicate potential existence of the object in a particular image region. While some detectors, *e.g.* (Uijlings et al. 2013), are designed to return several box candidates, others, *e.g.* (Felzenszwalb et al. 2010), are parameterized to return only few. For the latter ones we set their reliability threshold sufficiently low, thus acquiring several promising candidates as well.

As explained above, we do not consider these bounding boxes to be accurate enough to be trusted for as is. However, we do consider them accurate enough as soft voters, that collectively return the confidence that the pixel p lies on an object, that is

$$o(p) = \frac{\sum_i D_i(p)}{Z}, \tag{2}$$

where $D_i(p) = 1$ when the i-th bounding box contains the pixel p and Z is a normalization constant such that $\max o(p) = 1$. We will refer to the spatial prior $o(p)$ as *objectmap*.

Not all bounding boxes returned by object detectors are relevant. We therefore employ filter functions to prune the ones that are unlikely to cover part of the object. The first filter relates to the size of the bounding boxes. As observed by Uijlings et al. (2013); Carreira and Sminchisescu (2012), the size of the relevant bounding boxes strongly depends on the specific dataset at hand and a minimum bounding box size is usually enforced. We discard the bounding boxes with unlikely geometries according to the training images, *e.g.*, too extreme width-to-height aspect ratios. Although some boxes will incorrectly be discarded, the rough location estimation depends on the collective power of several bounding boxes. Hence, missing a few relevant ones is not critical, as long as the majority concentrates around the object of interest.

The second filter relates to the tendency of object detector algorithms to maximize recall of returned boxes. For example, to avoid any missed detections, the selective search of Uijlings et al. (2013) generates on average 1,000–3,000 candidate boxes per image, whereas a DPM detector of Felzenszwalb et al. (2010) visits more than 100,000 locations for a normal sized image, a number of visits that is feasible because of the dynamic programming involved. We compute a saliency map (Itti et al. 1998) of the image to discard the detections D_i that occur in regions less likely to contain the actual object. The saliency score is helpful when the image is not cluttered with too many objects. Empirically, we have observed that this is often the case with certain fine-grained categories such as birds, as taking a picture of a fine-grained object, *e.g.*, a bird, implies a special interest to the particular sub-category and often results in a clear photo of the object.

After having obtained the objectmap for the fine-grained object in the image, we proceed with the segmentation. The segmentation component of our approach is based on GrabCut, see Rother et al. (2004). GrabCut uses a gaussian mixture model, which groups pixels with similar appearance together, such that the foreground is separated from the background. The gaussian mixture model is trained iteratively in an alternate fashion. During the first step the foreground and background probability density functions are updated, based on the current pixel foreground/background labels. During the second step, the pixel labels are re-estimated via graph-cut inference, using the updated foreground and background probability density functions to calculate the unary terms and the image gradients for the binary terms.

Using objectmaps we end up with figure-ground segmentations, as shown in Fig. 3. While the segmentation masks are not perfect, we recover sufficient spatial support for the object for most of the images.

3.2 Alignments and Partitionings

3.2.1 Supervised Alignments

In a supervised setting the ground truth locations of basic object parts, such as the *beak* or the *tail* of birds, are available in the training set. This is a typical scenario when the number of images is limited, so that human experts can pro-

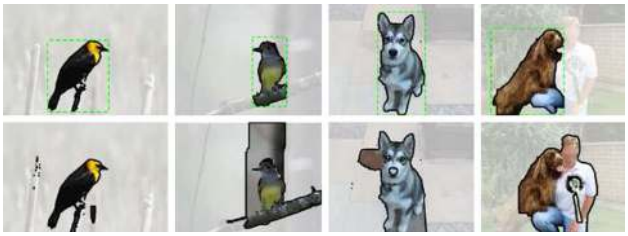

Fig. 3 Figure-ground segmentations by objectmap localizations. The result of the GrabCut segmentation algorithm is shown in the first row, when a bounding box is provided by the user, a common methodology in the fine-grained literature, see Yao et al. (2012); Berg and Belhumeur (2013); Chai et al. (2013). The objectmaps computed with an object detector, here selective search of Uijlings et al. (2013), are shown in the second row. For these objectmaps no user input of any form is required. Naturally, having no bounding box usually results in a less accurate segmentation, especially when other salient objects appear in the image as well. However, objectmaps still tend to concentrate on the fine-grained object

vide annotations at such a fine level of granularity. In the supervised alignment 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.

Different from general object categories that are often visually quite dissimilar from one another, fine-grained sub-categories typically share a great deal of similarities, mainly regarding their shape, their appearance and their poses. Hence, if the exterior shape of a fine-grained object is accurately captured, one can compare it with similar shapes in the training set and align the respective fine-grained objects. Note that, at this stage, it does not matter whether these are images that belong to the same sub-category or not. Having computed the figure-ground segmentation, we proceed with the description of the object shape. However, the segmentation mask is usually not perfect and often background is included or foreground is omitted, see Fig. 3. What is more, the interior of the object may contain inner edges, *e.g.* due to the intricate color patterns of a bird. As we are interested in the object silhouette and not the inner edges we extract HOG features from the binary segmentation mask and *not* from the segmented object. As a result, the gradients of HOG will focus on and accentuate the outer boundaries, while suppressing the interior shape edges.

After having extracted the segmentation mask, we encode the object shape by computing a HOG feature, that is $h_i = H(S_i)$. A HOG descriptor forms a high-dimensional space, which in theory may be populated by all shapes possible. Fine-grained objects, however, tend to have similar shapes and are seen in a limited repertoire of poses. More specifically, the observed exterior shapes reside on a lower dimensional manifold. Given an unseen fine-grained object, we can expect that its shape will probably be located in a specific region on this manifold. The fine-grained objects on this part of the manifold will have similar exterior shapes and, due to the anatomical constraints of the super-category they belong to, also similar poses on average. We take advantage of this principle to retrieve the N training exemplar images I_N from the training set D_t which have the most similar exterior shapes using a query-by-example setting. For the comparisons we employ the ℓ_2-distance metric on the unit-length normalized HOG vectors. In the end we have a shortlist of exemplar objects with similar poses, although no supervision was required regarding object poses or geometry. Examples of pose retrieval given an object of interest are shown in the upper row of Fig. 4.

Having retrieved the exemplar images with the most similar poses, we are in a position to transfer information from the training set to the test images. For the training exemplars I_N we know the ground truth part locations $\mathbf{x}$, as well as the appearance of the image regions that surround the parts $V_{\mathbf{x}}$. In order to calculate the locations of the part of interest in the test image I_q, we employ a geometric part aggregation function $f(\cdot)$, that is

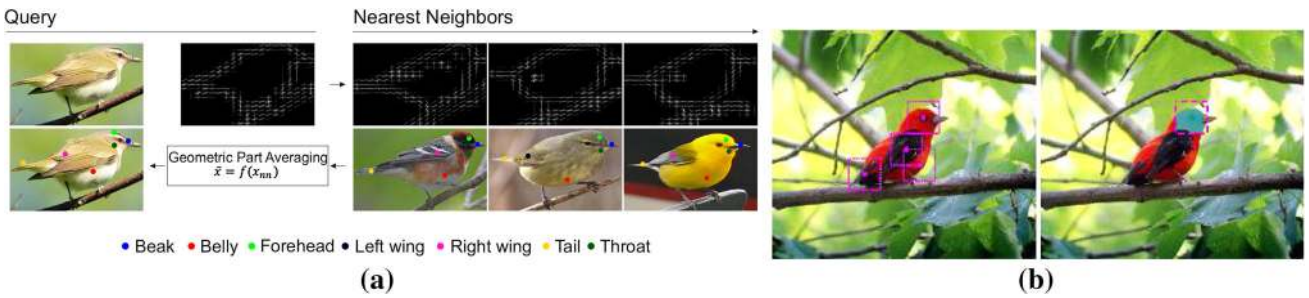

Fig. 4 Supervised alignment. **a** Predicting part locations: in the top left, we have a query image, for which we want to predict part locations. On the right, we have the nearest neighbor training images, their HOG shape representations on which they were retrieved (*top*) and their ground truth part locations (*bottom*). Regressing the locations from the nearest neighbors to the test image we get the predicted parts, shown as the colorful symbols (*bottom left*). Although we rely on exterior shape only, the part locations can be found consistently. **b** Describing parts using all the information within a square patch (shown *left*) gives inferior results compared to using only the information within the square patch that falls inside the object's segmentation mask (shown *right*)

$$\tilde{x} = f(I_q;\ \mathbf{x}_i, V_{\mathbf{x}_i}), i \in I_N \quad (3)$$

The geometric part aggregation function f can vary in sophistication. We can apply simple average aggregation, or we can learn part appearance models in a similar manner to Felzenszwalb et al. (2010); Azizpour and Laptev (2012). With geometric part averaging the predicted part locations are computed as the average of the respective part locations in the nearest neighbor images of the training set. This works well because the nearest neighbour images are well aligned to the query image. Note also that the appearance of the part in the nearest neighbour images is not used in this setting. We have experimentally witnessed that geometric part averaging yields accurate results, accurate enough to recover rough alignments. To ensure maximum compatibility we apply the above procedure for all the training and all the testing images in the dataset, thus acquiring predicted part locations for *all* the objects in the dataset.

Partitioning supervised alignments We know the location of the part centers. Next, we need to define the shape of the parts, given these centers. We consider two strategies, that is square *patches* and square patches refined by segmentation, which we will refer to as *segmented patches*, see Fig. 4b.

Patches The first strategy is related to most part-based models like Felzenszwalb et al. (2010). Given centers α, we sample local descriptors every d pixels from a square region $R_{sq} = \{(x, y)|\alpha_x - T/2 < x < \alpha_x + T/2, \alpha_y - T/2 < y < \alpha_y + T/2\}$. Patches capture both object and background appearance.

Segmented patches The second strategy bears close resemblance to the first one, the difference being that we now take into account also the segmentation mask that gives a spatial support for the objects. For segmented parts we sample only in the common area between the designated part region and the segmentation mask, that is $R_{sg} = R_{sq} \cap S_i$. Segmented parts better capture the object of interest, at the expense of including less context, since descriptors are sampled only within the segmentation mask.

Of course, more strategies can be imagined for partitioning with supervised alignments. Scale invariance could be helpful for example. However, introducing scale invariance for patches comes at the cost of increased complexity and is therefore not considered in the current work.

3.2.2 Unsupervised Alignments

In contrast to the supervised case, in the unsupervised scenario we assume that no ground truth is provided regarding the part locations of the images in the training set. In the absence of such a ground truth, 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 Mikolajczyk et al. (2005). More specifically, given the location x_s of the pixels on the segmentation mask S we fit a 2-D ellipse, whose two axes are computed as

$$a_j = \bar{x}_s + \mathbf{e}_j\sqrt{\lambda_j} \quad (4)$$

where λ_j and $\mathbf{e}_j$ are the j-th eigenvalue and eigenvector of the covariance matrix $C = (x_s - \bar{x_s})^T(x_s - \bar{x_s})$ and $\bar{x}_s$ is the average location of the mask pixels. Ideally, the ellipse should follow the "spine" of the object. We show examples of estimated poses and their local 2-d geometry in Fig. 5a.

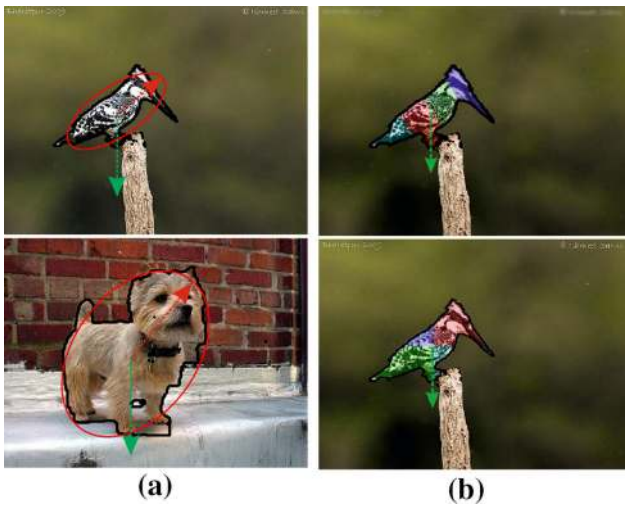

Fig. 5 Unsupervised alignments. Random birds and dogs, after their shape has been recovered, see in (**a**) the *black* contour around the objects. Based on the geometry of the shape we estimate the pose of the object, assuming an elliptical form. Following the gravity vector assumption (Perdóch et al. 2009) of the *green arrows*, we obtain the dominant pose orientation, see *red arrows*. Different strategies for aligning unsupervised partitionings in (**b**). In the *top* image we have gravitational alignments, that adopt an upwards dominant orientation after the gravity vector assumption. In the *bottom* image we have grid alignments, centered according to the center of the elliptical pose

Exploiting the preferred pose and posture can be a disadvantage when an object category appears in a great variety of poses. For objects that appear in a variety of poses, often placed in confusing backgrounds, the segmentation masks are usually not perfect. To minimize such negative influence, we use all the pixels of the foreground segmentation mask for fitting the ellipse.

Partitioning unsupervised alignments For unsupervised alignments one does not have much certainty regarding the object pose. Hence, simple, yet consistent alignment geometries are required to robustly describe similar object locations in previously unseen images.

Gravitational partitionings Given an elliptical pose for the fine-grained object, we need to define a reasonable orientation. Following anatomical observations we first consider

the longer axis to be the principal one. Having chosen the direction of the principal axis, we need to define the starting point. We follow the gravity vector assumption, see (Bay et al. 2008; Perdóch et al. 2009), and adopt the highest end point of the principal axis as its origin to arrive at gravity vector alignments. All partitionings are orthogonal to the principal axis of the fine-grained object. Since this principal axis is often similar to the "spine" of the object, each partitioning captures indirectly a specific anatomical part. For example in the case of four gravitational partitionings on birds, we roughly capture the "head", the "torso", the "belly" and the "tail" of the bird.

Grid partitionings Gravity vector alignments are supposed to follow the principal direction of the object's pose. Often, however, objects are photographed in a wild variety of poses, in which case gravitational alignments might return less consistent results. In this case, and since spatial pyramids have shown excellent result in image-level classification, see (Lazebnik et al. 2006), one can compute grid partitionings centered in the centre of gravity for the estimated elliptical pose. Given an accurate local frame of reference, the grid partitionings capture in their quadrants semantically meaningful regions of the fine-grained object. Furthermore, by vertically mirroring the training images we inject invariance regarding the pose and directionality of the fine-grained object regions. For example, the upper quadrants capture the appearance of the head, while lower quadrants encode the appearance of the belly and the tail, no matter where the object is facing to. Our strategies for aligning unsupervised, gravitational or grid, partitionings are visually summarized in Fig. 5b.

In theory, extracting unsupervised alignments is less accurate than extracting supervised ones. However, given an accurate spatial support provided by the obtained local frame of reference, and a robust set of rules for defining the pose of the fine-grained objects in different images, we are still able to obtain robust and consistent alignments over the entire database. Another advantage of such unsupervised alignments and their partitionings is that they are consistently found in *all* the images of the whole dataset and not just a small number of them at a time. This contrasts to part detection methods like that of Felzenszwalb et al. (2010); Yang et al. (2012), which require several part templates to ensure high precision. Since such templates are normally activated only for a portion of the training set, the number of available training data for learning the part appearance is effectively reduced.

3.3 Description

3.3.1 Color Fisher Vectors

The proposed alignments, supervised or unsupervised, are designed to be rough. Thus, comparing corresponding regions of objects from different images is bound to be a noisy procedure. Relying on features, such as HOG (Dalal and Triggs 2005), that are designed to return precise representations, but also sensitive to common image transformations are likely to be suboptimal. This is a problem which orderless descriptors, such as Fisher vectors, (Perronnin et al. 2010), do not face, as by design they do not encode any spatial properties of the appearance information. Nonetheless, in a fine-grained categorization setting describing localities is important. To inject such spatial awareness to orderless descriptors, we extract Fisher vectors from the well aligned, and therefore spatially constrained, partitionings. By doing so we maintain a good amount of the spatial extent of the appearance, while avoiding being overly vulnerable to occasions where feature matching is challenging.

Fisher vectors are composed of the derivatives of the likelihood, as measured with a gaussian mixture codebook model, with respect to the model parameters. The Fisher kernel then measures the similarity between two gradient vectors (using the Fisher information matrix). For a gaussian mixture codebook model, with mean terms μ_k and variances σ_k, the Fisher vector representation is $\phi = [\frac{\partial x}{\partial \mu_k}, \frac{\partial x}{\partial \sigma_k}]^T$. The derivatives are computed on local image intensity SIFT Lowe (2004) or color SIFT van de Sande et al. (2010) descriptors.

As we cannot foretell which is the color representation that best reflects the differences between fine-grained species, we opt for extracting three different color SIFT descriptors, namely *RGB-*, *Opponent-* and *C-SIFT*. All these models are built on the *diagonal model*, $I_t = D \cdot I_u$. In the diagonal model I_u is the image taken in the unknown light source, I_t is the transformed image to the color space of interest, and D is a 3×3 diagonal matrix that stands for the color space model. Given a location in an image the SIFT operator is applied on each of the color channels independently. Then, all the SIFT descriptors per channel are concatenated into a single column vector.

RGB color model The RGB color model is the concatenation of the 3 color channels, namely *red*, *green* and *blue*.

Opponent color model Opponent-SIFT uses the opponent color space, namely $[O_1, O_2]^T = [\frac{R-G}{\sqrt{2}}, \frac{R+G-2B}{\sqrt{6}}]^T$. The third channel O_3 stands for the intensity color space. Hence, we do not consider O_3 as intensity SIFT is independently computed. Interestingly, the subtraction operation of O_1 and O_2 cancels any light intensity offset that is added to all channels.

C color model Although the opponent space is shift invariant to light changes, there is still intensity information contained in the channels O_1 and O_2. To make the color space invariant also with respect to light intensity scale changes, the C-color space divides O_1 and O_2 by the intensity channel O_3, thus having the color space $[\frac{O_1}{O_3}, \frac{O_2}{O_3}]^T$. Due to the division by the intensity the scaling factor for the diagonal matrix is canceled out, thus rendering the C-space also scale invariant to the light intensity.

The combination of these color space SIFT descriptors, especially when intensity SIFT is also considered, has been shown (van de Sande et al. 2010) to be fruitful. For a more comprehensive study on the various color descriptors we refer to (van de Sande et al. 2010).

3.3.2 Normalization

Due to the generally small number of words that Fisher codebooks use, unnormalized Fisher vectors are characterized by an over-burstiness of certain visual words. Therefore, for optimal performance, Fisher vectors are *(a)* first, power-normalized so that the large Fisher vector values become less accentuated, then *(b)* ℓ_2-normalized, see (Perronnin et al. 2010). These two subsequent normalizations can be viewed as a single, recursive transformation u, that is $\hat{\phi} = u(\phi)$. Inspired by the findings of Perronnin et al. (2010), we follow a similar normalization procedure that applies ℓ_2 transformations to the feature vector recursively T times.

4 Experiments

4.1 Datasets

Animal categories and their sub-categories provide a challenging testbed for fine-grained categorization, as their taxonomy is usually connected to specific visual appearances. We evaluate our proposed methods on popular fine-grained datasets for recognition of bird species and dog breeds. These datasets capture different aspects of fine-grained categorization, as birds exhibit low inter-class variation of visual patterns and dogs exhibit high intra-class pose variation. For this reason we consider the two datasets complementary.

4.1.1 Birds

The Caltech-UCSD Birds-200-2011 dataset introduced by Wah et al. (2011b), is one of the most extensive ones in the fine-grained literature. The *Birds* dataset is composed of 200 sub-species of birds, several of whom bear tremendous similarities, see Fig. 1a. The bird images in this dataset are distinguished only on a fine-grained level, since several of the sub-species belong to the same family. A characteristic example are the *Forster's Tern* and the *Least Tern* sub-species in the far right of Fig. 1a. As one description reads for the *Forster's Tern* for example, *"the comma-shaped black ear patch in winter plumage is distinctive, but some other plumages are very confusing."*[1] Recognizing, therefore, such nuances is the key for their recognition. For each of the classes in the *Birds* dataset there are 30 training images and approximately 30 testing images. We use the standard training/test split provided by the authors of Wah et al. (2011b). In our experiments we use the ground truth part locations *only* during learning, unless stated otherwise. Furthermore, we use the ground truth segmentations, *only* for evaluation and not for any kind of learning.

[1] http://www.allaboutbirds.org/guide/forsters_tern/id.

4.1.2 Dogs

The Stanford Dogs dataset by Khosla et al. (2011) contains images from 120 different breeds. The dogs are visually easier to distinguish than birds, as only few breeds belong to a common, larger family. See for example how different the *Norwich Terrier* and the *Scotch Terrier* are in the right of Fig. 1b. Dogs, however, are difficult to categorize for other reasons. Since, they are domestic animals, they are photographed in a great variety of poses, scales, viewpoints and often with other objects occluding them. Hence, for the fine-grained categorization of *Dogs*, before anything else, one needs first to recover poses accurately. In the *Dogs* dataset there are in total 12,000 annotated images provided for training and 8,580 images for testing. We use the standard training/test split provided by Khosla et al. (2011).

4.2 Technical Details

Following common practice in the fine-grained literature (Yao et al. 2011, 2012; Yang et al. 2012) we mirror the training images in the datasets to double the size of the training set. We use the bounding boxes to normalize the size of the images, unless stated otherwise. Furthermore, we do not downscale the image like in (Yao et al. 2012; Yang et al. 2012), as we found this has a severe impact on the accuracy. For example downscaling images with the maximum dimension being 250 pixels drops accuracy by 23 % for *Birds*. Last, we note that only for the *Birds* dataset there exist ground truth part locations as well as ground truth segmentations. Therefore, for the experiments where such ground truth information is needed, whether for evaluation or learning, we report results on the *Birds* dataset only.

We extract SIFT descriptors using the VLFeat library (Vedaldi and Fulkerson 2010). We sample densely every 3 pixels and at multiple scales ([16 × 16], [24 × 24], [32 × 32], [40 × 40]). We reduce by PCA the dimensionality of the intensity SIFT descriptors to 64 and of the larger color SIFT descriptors to 80. To arrive at Fisher vectors we use a Gaussian mixture model with 256 components. We use both the derivatives with respect to μ and σ, obtaining Fisher vectors of 32,768 and 40,960 dimensions, when using intensity and color SIFT respectively. For the Fisher vectors we evaluate recursive normalizations for a varying number of recursions, as described in Sect. 3.3. For HOG features we use the VLFeat implementation on a standard spatial grid of 8 pix-

els width per tile and then ℓ_2 normalize them. Unless stated otherwise, we apply the standard normalizations per feature type, that is power and ℓ_2 normalization for Fisher vectors and ℓ_2 normalization for HOG. Finally, as a classifier we use the linear SVM PEGASOS implementation (Shalev-Shwartz et al. 2007) with a fixed parameter $C = 10$.

We use the standard evaluation metric for these datasets, that is the category normalized mean accuracy over all the sub-categories within a dataset. Accuracy is defined as the number of correctly classified pictures for a certain sub-category, divided by the total number of pictures of that sub-category. All results are reported strictly on the test sets.

4.3 Experiment 1: What Descriptors?

Setup In this first experiment we evaluate whether rigid descriptors, such as HOG Dalal and Triggs (2005), or distribution-based descriptors such as Fisher vectors Perronnin et al. (2010) are more accurate for describing parts in fine-grained categorization. For completeness, we also compare with Bag-of-Words computed on 4,000 words and with a χ^2 kernel. 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 locations for both the training and test sets. We also investigate different parameterizations for Fisher vectors. We experiment on the *Birds* dataset using the provided part annotations. To minimize redundancy due to the overlap, we use the following seven parts only, which together cover the complete silhouette of a bird: *beak, belly, forehead, left wing, right wing, tail* and *throat*. Fisher vectors, HOG and Bag-of-Words are extracted on 100×100 pixel windows. We empirically found this to be a reasonable tradeoff between capturing sufficient content and context, while avoiding the influence of drastic deformations. Fisher vectors, HOG and Bag-of-Words are also extracted from the whole bounding box. In the end we concatenate the Fisher vectors together into a single vector and the HOGs together into a single vector.

We also evaluate the effect of applying recursive normalizations to the final accuracy. To avoid irrelevant factors influencing the results, we conduct this experiment again under an oracle setting and use the ground truth segmentation masks provided for *Birds* to compute a single Fisher vector representation per fine-grained object.

Results We show the results for the different parameterization of the Fisher vectors in Fig. 6. For 128-PCA, we apply the PCA matrix, thus de-correlating only and not reducing the SIFT vectors. We observe that having more gaussian components and more dimensions after PCA has a positive impact on the accuracy. To control the final feature dimensionality, as well as to be compatible with the state-of-the-art, in the following we will make use of 256 gaussian components,

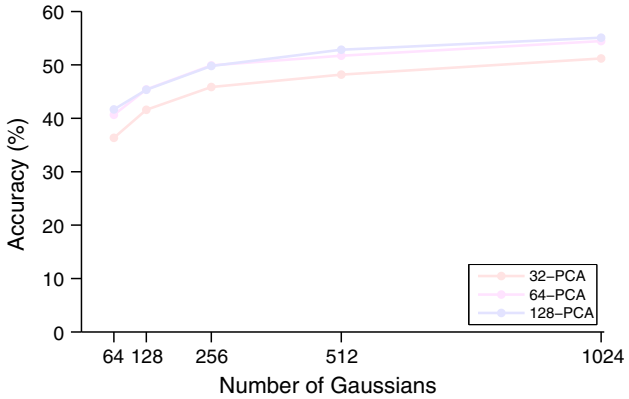

Fig. 6 Influence of Fisher vector parameters. Increasing the number of gaussian components and dimension after PCA improves the final accuracy on the oracle segmentations, reaching up to 55.1 % for 1,024 gaussians and 128 dimensions after PCA

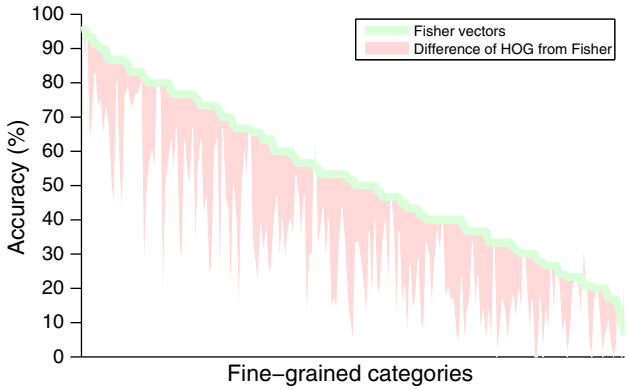

Fig. 7 A fine-grained category-by-category comparison using parts encoded by Fisher vectors or by HOG. We report results on the 200 *Birds* sub-categories measured in terms of accuracy. Fisher vectors perform consistently better on parts than HOG, having an average accuracy of 52.5 % versus 31.8 %. We, furthermore, repeat the same experiment using a Bag-of-Words model computed on 4,000 words and with a χ^2 kernel, obtaining an average accuracy of 25 % (data not shown in the plot)

with the dimensionality reduced to 64 for intensity SIFT and 80 for color SIFT descriptors.

In Fig. 7 we visualize the comparison between Fisher vectors and HOG. Clearly, Fisher vectors are better in describing parts for fine-grained categorization than rigid descriptors like HOG. Where HOG scores an accuracy of 31.8 % on average, the Fisher vectors result in a final average score of 52.5 %. The reason is that HOG descriptors require quite precise part detection, so that the gradients are representative of the appearance. Fisher vectors, however, aggregate the information from a larger area, adding more flexibility to the representation. Two notable exceptions, where HOG outperforms Fisher vector, are shown in Fig. 8. In the majority of cases, however, Fisher vectors are clearly better for describing fine-grained subcategories than HOG, as Fig. 7 reveals, outperforming for 184 out of the 200 bird categories. With

Fig. 8 Occasionally, encoding parts by HOG is better than Fisher vectors. The *Rhinoceros Auklet* birds in the first row have a very characteristic white horn on their beaks and two elongated white feather brows next to their eyes and their beaks. The shape-sensitive HOG better captures the appearance of those birds. Similarly, the *Brandt Cormorant* species also has a very distinctive sigmoid shape, also better described with HOG. In the majority of cases, however, Fisher vectors are significantly more accurate, see Fig. 7

Bag-of-Words we obtain a lower accuracy of 25 %. Since Fisher vectors can be viewed as an extension of the Bag-of-Words model by considering the codebook derivatives, we conclude that these additional statistics of Fisher vectors are essential for fine-grained categorization. The above results and conclusion will be added in the first experiment. From now on we report results using Fisher vectors for describing the appearance of parts and alignments.

Regarding the recursive normalization on Fisher vectors, we observe that optimal results are obtained after two recursions, that is $T = 2$, improving recognition over the standard power normalization by an absolute 2–3 %. This conclusion was also confirmed in subsequent non-oracle experiments, improving recognition even up to 4 % for color based features.

4.4 Experiment 2: What Type of Regions?

Setup In this experiment we evaluate various partitionings for the description of fine-grained objects. The majority of the approaches in fine-grained categorization are evaluated considering the bounding boxes available both during training and testing. For completeness, we also first evaluate the case when bounding boxes are always available.

For the supervised alignments we follow the same setup as in the previous experiment, using the same seven parts plus a Fisher vector extracted from the whole bounding box. We predict the location of these parts in unseen images using the top-20 nearest neighbors. When the majority of the nearest neighbors does not have a certain part, it is marked as absent for the unseen image and the corresponding part of the Fisher vector is set to the zero vector. Also, we repeat the same experiment using only the predicted location of the *beak*.

For the unsupervised alignments no ground truth part annotation is required, so we evaluate on both *Birds* and *Dogs*. After extracting the principal axis of the object of interest, we split the segmentation mask into aligned partitionings. For the object-level Fisher vector we use only the pixels within the segmentation mask and not the whole bounding box. We also examine what is the effect of a varying number of parts on the final accuracy.

Finally, we provide comparisons with state-of-the-art methods reported on the same datasets. For this purpose, we first evaluate the significance of color in fine-grained categorization. Apart from grayscale SIFT features, we additionally extract SIFT features from the *RGB*, *Opponent* and *C*-spaces (van de Sande et al. 2010).

Results. We show the results of this experiment for *Birds* in Table 1. When considering supervised patch alignments, we obtain 50.2 % accuracy, a large improvement over the 39.8 % from the 2×2 spatial pyramid. Comparing the individual accuracy differences, the supervised alignments perform consistently better than spatial pyramids for 141 of the 200 classes (data not shown). The reason is that birds are well aligned, so the Fisher vectors computed on the respective parts capture the same nuances that differentiate sub-classes more consistently.

We measure the accuracy of the estimated part locations with respect to the ground truth locations. To cancel out the different bounding box geometries we normalize the part locations. After normalization the average location error is 12 %.

Table 1 Experiment 2: What type of partitioning for Birds? 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 that score 52.5 %. When considering the segmentation masks for the description of the supervised alignments as in the right picture of Fig. 4b, the accuracy improves even further

Method	Accuracy (%)
Supervised segmented patch alignments	57.6
Unsupervised gravitational alignments	51.6
Supervised patch alignments	50.2
Unsupervised grid alignments	49.2
Fisher vector from segmentation masks	42.6
2×2 spatial pyramid	39.8
Supervised alignment on beak	37.8
Fisher vector from bounding box	32.1

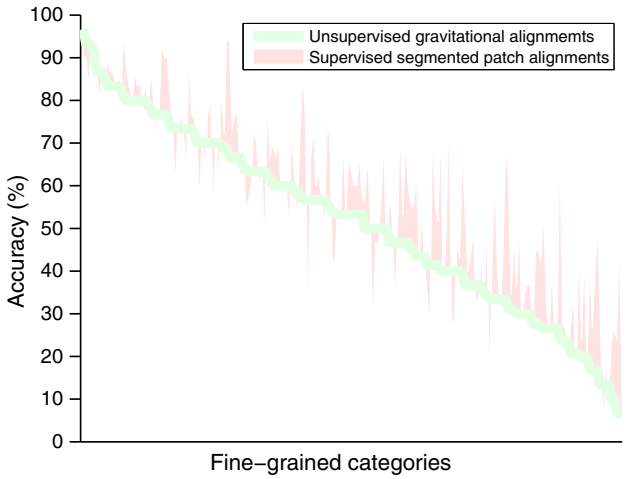

Fig. 9 Comparison between supervised segmented patch alignments and unsupervised gravitational alignments for *Birds*. We observe that the supervised ones are especially beneficial for those classes where unsupervised alignments exhibit lower accuracy (*right* part of the figure)

Interestingly, when considering the supervised segmented patch alignments using the GrabCut based segmentations the recognition accuracy improves further, reaching 57.6 % and outperforming all other methods. This translates to a 7 % gain as compared to supervised patch alignments. We can therefore deduce that segmentation masks are helpful not only for describing whole objects, as they are normally used (Chai et al. 2012, 2013), *but also for the description of individual parts or regions* of the fine-grained object of interest. With an exception of the work from Arbelaez et al. (2012), who use poselet-inspired region detectors, we are not aware of any works that researched the potential of segmented parts for recognition.

We focus now on the case when no ground truth of the part locations is provided, neither for training nor for testing. For unsupervised gravitational alignments we reach an accuracy of 51.6 %. Having fewer partitionings leads to a lower accuracy (48.4 % for two partitionings), whereas too many alignments bring little extra benefit (51.7 % for seven partitionings). Extracting four partitionings therefore suffices and we will use this number throughout the rest of the experiments where we extract unsupervised alignments, unless stated otherwise. Comparing the supervised and unsupervised alignments when using their optimal settings, we show the differences in Fig. 9. We observe that the supervised ones improve the accuracy especially for the classes where unsupervised alignments exhibit lower accuracy visible in the right part of the figure.

For the *Dogs* dataset we present the results in Table 2. The unsupervised grid alignments outperform the unsupervised gravitational alignments. The reason is that dogs are seen in a considerably larger variety of poses, scales and occlusions. In fact, as it is often the case that only the dog face is visible, any method that attempts to discover semantically meaningful parts becomes weaker, as also observed from Chai et al. (2013). Hence, for super-categories like *Dogs*, where the subcategories are found in varying and peculiar poses, precise pose normalization should precede the extraction of fine-grained details.

Table 2 Experiment 2: What type of partitioning for Dogs? The unsupervised grid alignments outperform the unsupervised gravitational alignments. As also noted by Chai et al. (2013), the reason is that dogs are seen in a considerably larger variety of poses, scales and occlusions

Method	Accuracy (%)
Unsupervised grid alignments	45.2
Unsupervised gravitational alignments	42.9
2×2 spatial pyramid	42.8
Fisher vector from segmentation mask	40.1
Fisher vector from bounding box	36.2

We conclude that extracting localized alignments or parts matters in a fine-grained categorization setting. Furthermore, given their high accuracy, as well as their independence from ground truth part annotations, unsupervised alignments are appealing compared to supervised ones.

Adding color First, we evaluate the importance of color descriptors in fine-grained categorization tasks. In this experiment, we use the ground truth bounding boxes, as this is also done by the methods we are comparing against. The results after the addition of color are available in Fig. 10a for *Birds* and in Fig. 10b for *Dogs*. We observe that color consistently improves accuracy. From individual color channels only Opponent-SIFT performs well, increasing accuracy from 51.6 to 62.7 % for *Birds* and from 45.2 to 51.5 % for *Dogs*. When fusing the model predictions of the Fisher vectors from all color channels by averaging, we reach an

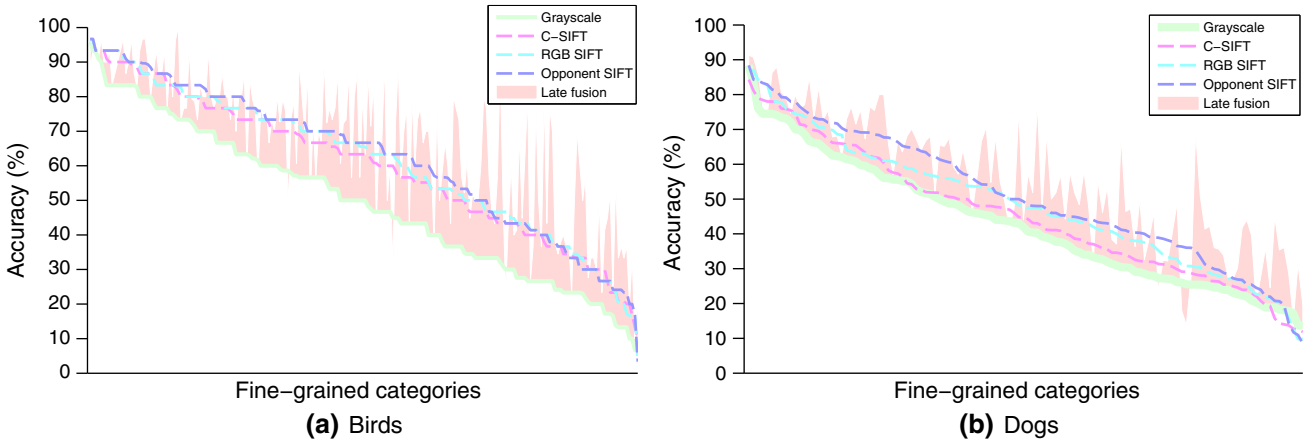

Fig. 10 Experiment 2: Adding color given bounding boxes. **a** For *Birds* when considering the color information the accuracy becomes higher than the 51.6 % obtained with grayscale only. More specifically, we obtain 60.0 % by using C-SIFT, 61.5 % by using RGB-SIFT and 62.7 % by using Opponent-SIFT. When fusing the Fisher vectors computed on different color spaces with late fusion, the accuracies improve further to 67.0 %. **b** For *Dogs* we make similar observations: 45.3 % with C-SIFT, 48.3 % with RGB-SIFT, 50.1 % with Opponent-SIFT and 55.1 % with average late fusion, as compared to 42.9 % when only grayscale SIFT is used. Color is beneficial for fine-grained categorization

Table 3 Experiment 2: Comparison with state-of-the-art for Birds given bounding boxes. Unsupervised alignments outperform the state-of-the-art. Note here that the deep learning method of Donahue et al. (2013) makes use of extra labeled data

Method	Accuracy (%)
Unsupervised gravitational alignments	**67.0**
Donahue et al. (2013)+ Zhang et al. (2013)	65.0
Chai et al. (2013)	59.4
Donahue et al. (2013)	58.8
Berg and Belhumeur (2013)	56.9
Zhang et al. (2013)	50.1
Jia et al. (2013)	38.9

Highest scores marked with bold

Table 4 Experiment 2: Comparison with state-of-the-art for Dogs given bounding boxes. Unsupervised alignments outperform the state-of-the-art

Method	Accuracy (%)
Unsupervised grid alignments	**57.0**
Chai et al. (2013)	45.6
Yang et al. (2012)	38.0
Bo et al. (2010)	36.0
Khosla et al. (2011)	22.0

Highest scores marked with bold

accuracy of 67.0 % for birds and 57.0 % for dogs. Hence, using multiple color channels brings a clear advantage over only grayscale information, as known for general object and scene detection (van de Sande et al. 2010). In fact the experimental results reveal that a right use of color has an even stronger impact on the categorization of fine details, at least when animal species are considered.

State-of-the-art comparison given bounding boxes Next, we compare state-of-the-art methods on fine-grained categorization, which also assume that the bounding box around the object is available at runtime. The results are available in Tables 3 and 4 for *Birds* and *Dogs* respectively. We observe that for birds unsupervised gravitational alignments arrive at good recognition rates of 67.0 % compared to the very recent state-of-the-art. The closest competitor, the deep learning approach of Donahue et al. (2013) combined with pose normalization from Zhang et al. (2013), reaches an accuracy of 65 %. DeCAF makes use of large deep learning networks composed of 7 layers that require elaborate pre-training on many labeled images from 1,000 classes from ImageNet. Similar results are observed for *Dogs*, where unsupervised grid alignments score 57.0 % average accuracy. The closest competitor is the recent work of Chai et al. (2013), reporting an accuracy of 45.6 %. We conclude that unsupervised alignments achieve state-of-the-art recognition rates for fine-grained categorization.

4.5 Experiment 3: Automatic Fine-Grained Categorization

Having the bounding box location is a useful piece of information, as it separates, albeit roughly, the object of interest from the majority of the background. However, in most realistic scenarios bounding boxes are not available. In this experiment we examine the effectiveness of fully automatic fine-grained categorization, a process that entails automatic detection, segmentation and categorization of the fine-grained objects. To this end we first evaluate the importance of accu-

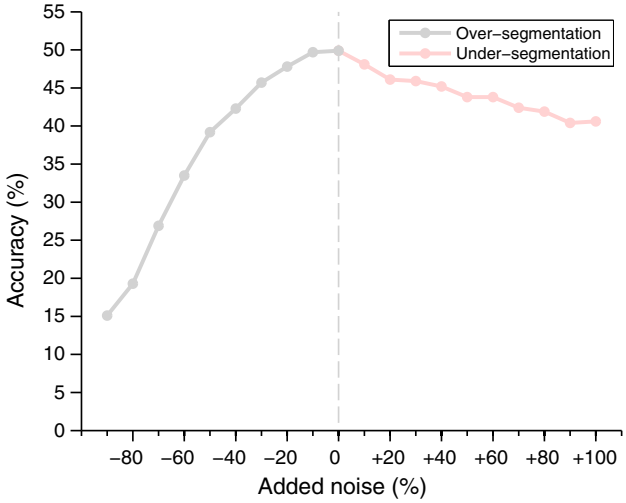

Fig. 11 Experiment 3A: The effect of segmentation accuracy in fine-grained categorization oracle segmentations on birds. Noisy segmentation masks always hurt accuracy. However, missing superpixels of the ideal object segmentation (over-segmentation) is noticeably more harmful than including excessive background (under-segmentation)

rate segmentation in an oracle setting, by simulating added noise on ground truth segmentation masks. Then, we evaluate automatically detecting, segmenting and categorizing fine-grained objects.

4.5.1 Experiment 3A: Segmentation Accuracy

Setup In this experiment we evaluate the significance of accurate segmentations in a theoretical fine-grained categorization setting, where we assume that perfect segmentations for all fine-grained objects are available. We perform this experiment on the *Birds* dataset, as it is the only one for which ground truth segmentation masks are available. To make sure that conclusions reflect only the importance of segmentation accuracy, we extract a single Fisher vector from within the segmentation mask area, *without considering any kind of partitionings*. We start from the perfect ground truth segmentations, then generate artificially foreground or background noise. This way we simulate scenarios of over-segmentation, where part of the object is overlooked, and under-segmentation, where part of the background is considered as object. To generate the artificial noise we first decompose an image into a large number of superpixels using (Felzenszwalb and Huttenlocher 2004). Then, for under-segmentation we include extra superpixels neighboring the perfect segmentation mask, while for over-segmentation we exclude superpixels from the foreground mask. The superpixels are chosen such that the desired level of artificial noise is reached.

Results We plot the results of this experiment in Fig. 11. Over-segmentation appears to be quite harmful, see the left part of Fig. 11. Losing a little bit of foreground, up to −20 % has little impact on accuracy. However, when more foreground information is missing, the accuracy drops rapidly. When focusing on the right part of Fig. 11, where background noise is added to simulate under-segmentation, we observe a noticeable but not dramatic decrease in accuracy. Indeed, adding 100 % background noise, that is an area equal to the size of the bird, decreases the accuracy from 49.9 to 40.6 %. If we expect the segmentation to be imperfect, either because of the low imaging quality or the challenging viewing conditions, a bias in favor of adding background than omitting foreground should be preferred.

4.5.2 Experiment 3B: Fine-Grained Categorization Without Human Intervention

Setup In this experiment we make no assumptions regarding the location of the object and want to compute a probability map, that encodes how likely is an object to be present at a particular image region. The first candidate is objectness (Alexe et al. 2012), which was designed particularly for this purpose. We use the objectness parameters suggested in the latest release software, version 2.0, by the authors. For the objectmaps we use three state-of-the-art object proposal algorithms. Firstly, we use the deformable part model (Felzenszwalb et al. 2010). We lower the DPM detection threshold to −1.0, decided after visual inspection, to increase the number of detections returned. Secondly, we use selective search (Uijlings et al. 2013) to generate object proposals. Last, we use the recently proposed prime proposals (Manén et al. 2013). For fairness of comparison we use the pre-configured object proposal and detector models as proposed by the respective authors. For the supervised deformable part model we use the bird and dog detectors as trained on PASCAL VOC. As objectness, DPM, selective search and prime objectmaps serve the same purpose, for clarity we will refer to all of them as objectmaps during the evaluation. We include comparisons with state-of-the-art methods that also do not require a location for the fine-grained object at runtime.

Results We present the results for the *Birds* dataset in the first two rows of Table 5. The highest accuracy is obtained using the selective search and the prime objectmaps with unsupervised gravitational and grid alignments respectively. Their accuracy in the range of 40.5–44.1 % (for compactness these numbers are not included in the tables) is a competitive result, when compared to the 51.6 % accuracy obtained from the same method when the bounding box locations are given, and the 57.6 % when supervised segmented patch alignments are employed, see Table 1. As in the previous experiment, we also consider the addition of three color spaces for the selective search objectmaps, see Fig. 6. The results are consistent with the conclusions of the previous experiment. Extract-

Table 5 Experiment 3B: Fine-grained categorization without human intervention. For birds unsupervised bounding box proposals (Uijlings et al. 2013) suffice for computing an accurate location for the object of interest. For dogs, however, where often multiple objects appear in the image, supervised bounding box proposals, (Felzenszwalb et al. 2010), are more accurate

	Alignments	Objectmaps			
		Objectness (%)	DPM (%)	Selective search (%)	Prime proposals (%)
Birds	Unsupervised gravitational	32.7	36.6	**40.6**	39.8
	Unsupervised grid	31.7	33.4	38.6	**40.8**
Dogs	Unsupervised gravitational	29.4	**36.8**	30.4	30.0
	Unsupervised grid	31.4	**36.8**	34.0	32.6

Highest scores marked with bold

ing Fisher vectors from the Opponent-, RGB and C-SIFT spaces increases accuracy to 51.6, 49.0 and 48.9 % respectively. Applying late fusion using all color spaces as well as grayscale SIFT, we arrive at a final accuracy of 53.6 %. For comparison, the automatic system from Zhang et al. (2012), that requires several part annotations during training, reports an accuracy of 28.2 %. Note here that the selective search and prime objectmaps *are fully unsupervised, requiring no human provided boxes, not even for training images*, keeping the amount of human intervention to the minimum of providing *only image-level* annotations for the training set. The reason for their good performance in recovering bird locations is that birds often appear in isolation, with few other objects in the image. As a result, the selective search and prime bounding boxes usually concentrate around the most prominent object, which is a bird in most cases.

For the dogs dataset the results are shown in the last two rows of Table 5. For dogs, that often appear in a cluttered environment with many other objects, deformable part objectmaps work best, be it for gravitational or grid alignments, reaching an accuracy of 36.8 % for both cases. After the addition of color on deformable part objectmaps, we obtain similar improvements as before, arriving at 47.2 and 49.0 % for gravitational and grid alignments respectively.

We conclude that fully automatic fine-grained categorization is within reach. Using objectmaps as spatial priors allows unsupervised alignments to have a competitive accuracy, while requiring *no* user interaction regarding the parts nor the location of the fine-grained objects.

4.6 Qualitative Analysis

Best recognized fine-grained objects. In Fig. 12 we plot pictures from the *Birds* and *Dogs* categories for which unsupervised alignments reach the highest accuracy. The results for *Birds* are obtained with unsupervised gravitational alignments, whereas for *Dogs* with unsupervised grid alignments.

The fifteen birds with the highest recognition accuracy are characterized by an extensive color palette on their plumage. For example the *European Goldfinch* is easy to distinguish based on the intricate color patterns of red patches on their heads, followed by a black and white ring around their necks, their white belly, brown back and black and yellow wings. It appears that having several colors in different combinations and on different bird locations explains why these specific birds are easier to recognize than other species.

For *Dogs* we derive similar conclusions. First, as expected the different dog species have different colors, yet their chromatic palette is significantly more limited than for birds. Nevertheless, from the experimental results, see Fig. 10, we know that color is also an asset. We conjecture that this is because for dogs *the color gradients are more important than the color itself*. The reason is that the color gradients locally reveal a particular type of texture, usually characteristic of the dog's type of fur. For example the long, thin, "rasta"-like hair colored with different gradients of gray identify a *Komondor*, whereas the different gradients of brown and yellow identify the shiny fur of a *Sussex spaniel*. Hence, for *Dogs* extracting gradient based SIFT descriptors from different color spaces appears to be a good design choice as well, although the color variety is not as exotic as in the case of *Birds*.

What are the limits of visual features? Here we examine the other extreme, namely the categories which were difficult to recognize. In Fig. 13 we show images of the two most confused pairs of bird categories, when only grayscale information is used: *Forster's Tern* versus *Least Tern* and *Pelagic Cormorant* versus *Red faced Cormorant*. We observe that all the confused pairs belong to the same family of species. Indeed, their main differences are some colored details, *e.g.*, the color of the beak. This is illustrated by a one-vs-one comparison of the birds in Fig. 13 and the color versions of them in Fig. 1.

Now, we turn our attention to the case when also color is considered. In Fig. 14 we show images of two highly confused categories, when Opponent SIFT color features are considered: *Great Grey Shrike* versus *Loggerhead Shrike* and *Caspian Tern* versus *Elegant Tern*. These categories look very similar. It is likely that these birds are taxonomized based on some physiological, rather than purely visual, characteristics. Indeed, when looking up the taxonomical motivation for the *Loggerhead Shrike* and the *Great Grey Shrike*, we found that

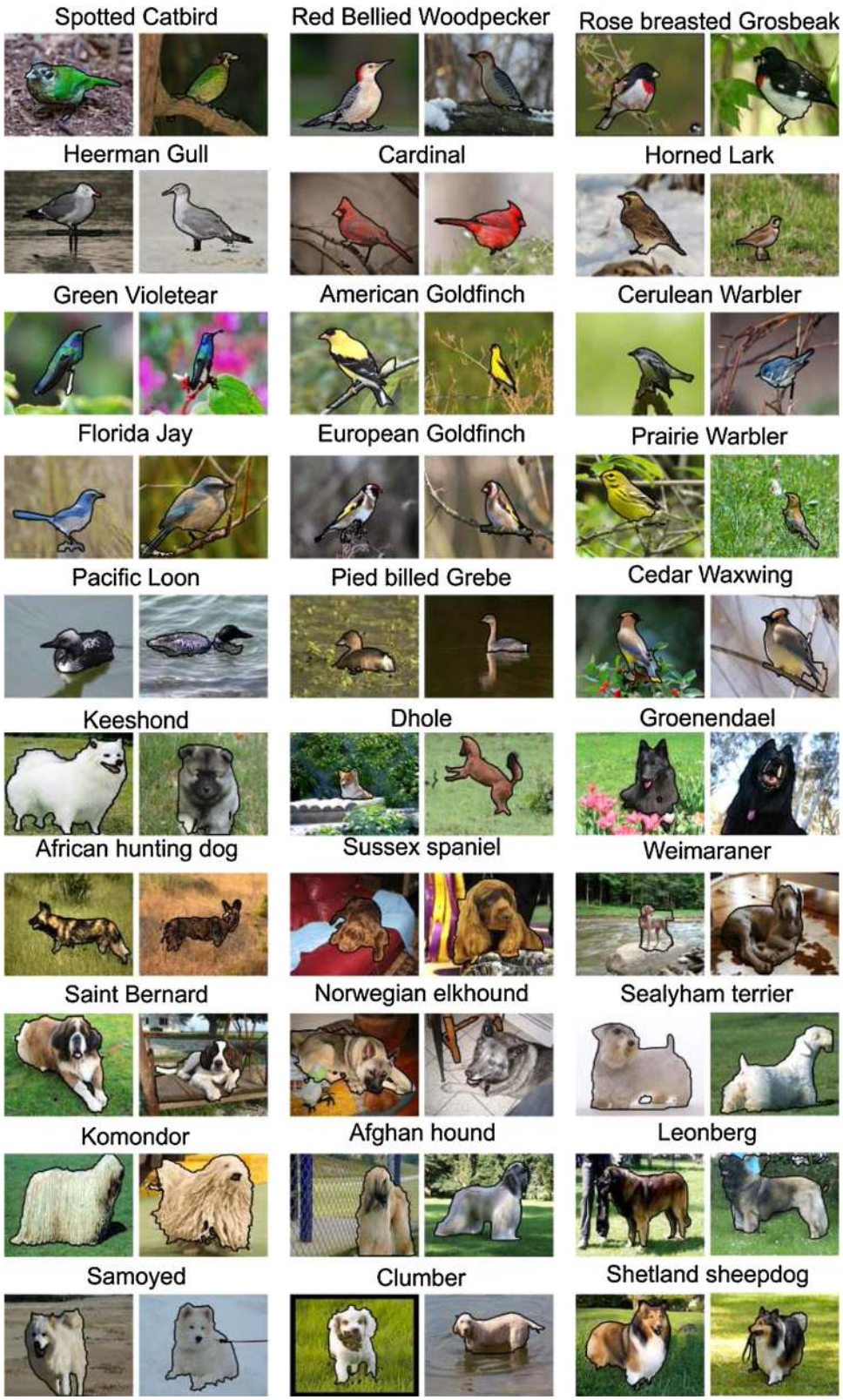

Fig. 12 Experiment 4: Some of the best classified categories for unsupervised alignments for Birds and Dogs. For completeness we draw the detected boundaries after segmentation, see black contours. We observe that birds and dogs in these sub-categories have consistent appearance. It is noteworthy, especially for *Birds*, that most sub-species have very distinctive color patterns, which are well described by the color Fisher vectors we extract. To draw conclusions regarding the limitations of visual features, we present failure cases in Fig. 13 and 14.

Fig. 13 Experiment 4: Two of the most confused pairs of bird categories, when only grayscale information is used. On the left we have the *Forster's Tern* and *Least Tern* species, while on the right we have the *Pelagic Cormorant* and the *Red faced Cormorant*. The visual similarities between classes are remarkable, especially when no color is considered. Color is often necessary for telling such sub-categories apart

Fig. 14 Experiment 4: Two of the most confused pairs of bird categories after adding color with Opponent SIFT. The first pair of confused birds contains the *Great Grey Shrike* and *Loggerhead Shrike* species, whereas the second one the *Caspian Tern* and the *Elegant Tern* species. These birds species seem very similar to each other, even after the addition of color. It is likely that they are taxonomized based also on non-visual criteria, such as anatomical or geographical ones. Indeed, the main two differences between the *Great Grey Shrike* and *Loggerhead Shrike* are **a** the proportion between their head and their beak and **b** their habitat, with *Great Grey Shrike* living in the north and *Loggerhead Shrike* in the south

their main two differences are anatomical and geographical. First, for the *Loggerhead Shrike* the proportion between the head and the beak is *usually* larger[2]. Second, the two species are parapatric[3]. The *Great Grey Shrike* appears in Northern Eurasia and America, whereas the *Loggerhead Shrike* lives in the southern Mediterranean zone. This type of anatomical or geographical information is unlikely to be recovered from single pictures, where the birds appear in all sorts of angles, viewpoints and scales and the context is limited. We conclude that when this is the level of recognition required, expert knowledge, metadata, or perhaps analysis of the environment, as Darwin (1859) would argue, might be necessary for guiding the machine further. For example, to recognize the *Great Grey Shrike* from the *Loggerhead Shrike* we could examine whether the surroundings correspond to a subarctic or a temperate habitat respectively[4], either in an automatic fashion or via questions posed to the user (Branson et al. 2010).

What makes a *Bobolink* a *Bobolink*? Here we exploit the properties of the linear SVM classifier, more specifically the additivity of the classification scores per feature dimension (Maji et al. 2008; Gavves et al. 2012). Given a sub-category c and its classification model w^c, we retrieve the dimensions d with the largest, positive weight values $d = \arg_{d'} \max w^c_{d'}$, since they contribute the most to the final classification score. We then identify those pixels that have the strongest Fisher response for the dimensions d of the sub-category classifier w^c_d. Due to monotonicity, the power and ℓ_2 normalization do not influence the outcome of this qualitative evaluation. We visualize in Fig. 15 results for the top 20 dimensions ($|d| = 20$) for the 20 pixels with the strongest Fisher response using unsupervised alignments and Opponent SIFT.

Given the rough nature of the alignments we make several observations from the visualizations. First, it appears that the distinctive details appear consistently on similar locations on the fine-grained objects. For the *Boat tailed Grackle* the wide, round tail is the most distinctive detail. For the *Red face Cormorand*, it is the red patch on the bird's head. An interesting case is the *Hooded Marganser*. What is considered very distinctive for this bird are the bright yellow eyes and secondarily the black and white stripes on its breast. As most birds have dark eyes, a brightly colored eye makes the difference. On the contrary, the large back of the head is not considered very discriminative and would probably be better captured by HOG. Overall, it appears that Fisher operates as a spatial hashing function, that builds a correspondence between spatial details and certain feature dimensions. As a

[2] http://en.wikipedia.org/wiki/Great_Grey_Shrike.

[3] https://en.wikipedia.org/wiki/Parapatric_speciation.

[4] http://www.allaboutbirds.org/guide/loggerhead_shrike/id.

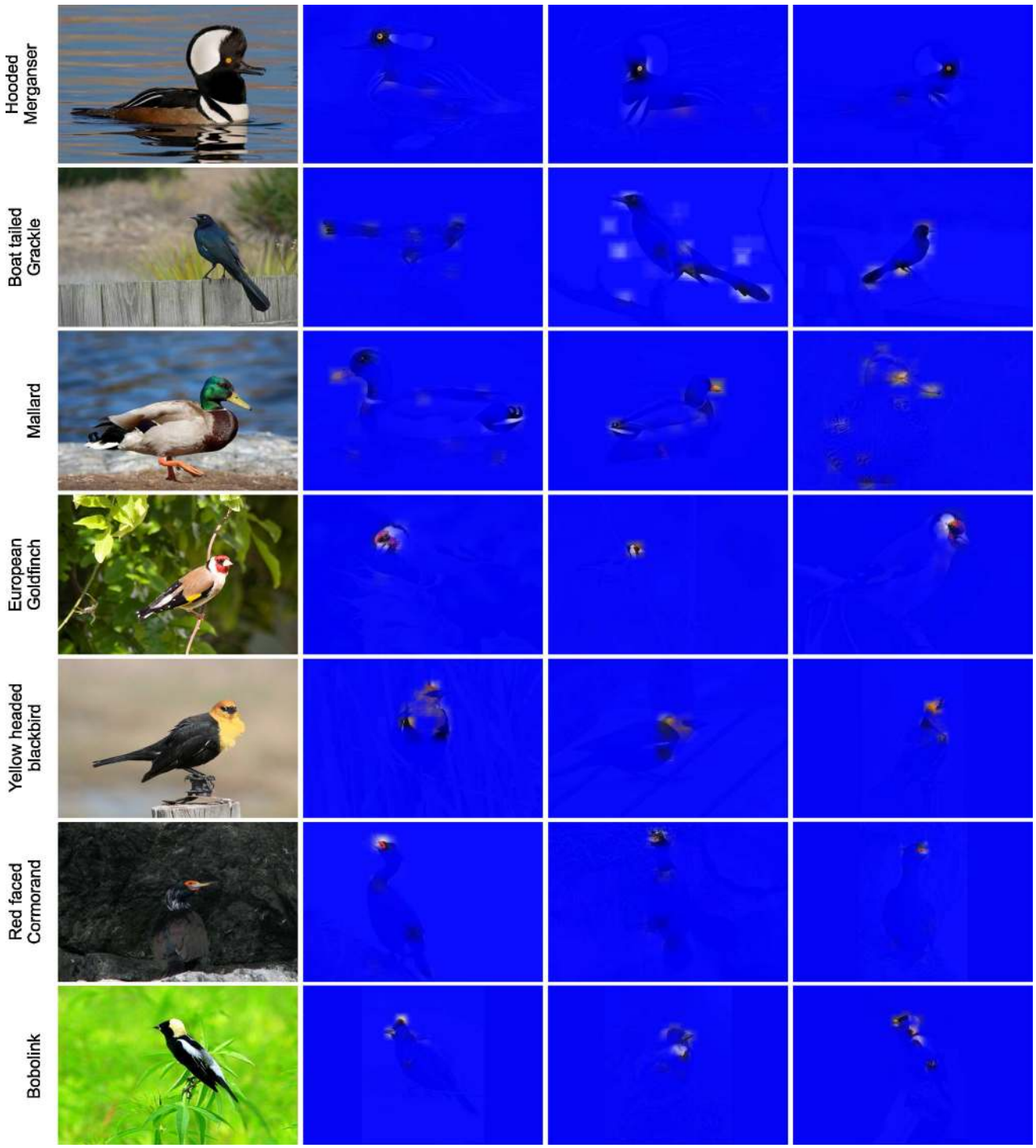

Fig. 15 What makes a Bobolink a Bobolink? Visualizing why birds are recognized as certain sub-species. We first compute the d classifier dimensions with the largest positive weights. Then, we detect the SIFT descriptors for which we observe the maximum response for these d dimensions. Finally, to generate the saliency maps we average the rectangular patches on which the SIFT descriptors were computed. The qualitative results show that the distinctive details appear consistently on similar locations on the fine-grained objects. Furthermore, we generally observe that *the most prominent appearance detail lies usually on the head*

result, although a more precise object or part localization is always welcome, employing features, such as Fisher vectors, may largely have the similar effect.

Furthermore, we generally observe that *the most prominent information lies usually on the head*. Placing special importance on detecting the head is therefore justified and may bring significant accuracy benefits, as has also been shown by Liu et al. (2012); Parkhi et al. (2012); Chai et al. (2013). Finally, we answer that a Bobolink is made by *angular beaks and very sharp, black and yellow edges around the head and the neck of a bird*.

5 Conclusion

We aim in this paper for fine-grained categorization without human interaction. Different from prior work, we show that localizing distinctive details by roughly aligning the object of interest allows for successful categorization of fine-grained categories. In cases when an object pose can be confidently extracted, it is beneficial to focus first on recovering the pose and then detecting the interesting part locations: the anatomical constraints imposed by a detected pose make sure that the parts do not drift away. We also postulate that since fine-grained parts differ usually more in their appearance than in their shape, parts are better described by classification-based encodings than shape-based descriptors.

Furthermore, we explore alternative uses of segmentation for fine-grained categorization. We quantify the link between segmentation accuracy and classification accuracy. We find that when imperfect segmentations are to be expected, it is better to include extra background than to omit part of the foreground (Fig. 11). When one cannot expect bounding boxes, we propose a methodology to recover the spatial support of a fine-grained object, even in the absence of a user-provided bounding box. Further, we show that refining parts by segmentation improves fine-grained categorization further (Table 1).

We perform experiments on the challenging CUB-2011 dataset composed of 200 bird species and on the Stanford Dogs datasets composed of 120 dog breeds. Under a controlled, oracle setting the experimental results indicate that for rough alignments, distribution based features, such as Fisher vectors, are a better choice than rigid features, like HOG (Fig. 7).

We proceed with performing fine-grained categorization on unseen images, obtaining high recognition rates (Table 1, 2). What is more, the experiments reveal the importance of color SIFT in the recognition of fine-grained subspecies (Tables 3, 4). Averaging the outputs of all color based classifiers leads to 67 % mean accuracy in classifying *Birds* and 57 % in classifying *Dogs*, a new state-of-the-art even when compared with deep learning approaches that make use of extra data. We attribute the high recognition rates of supervised and unsupervised alignments encoded with color Fisher vectors to two factors: first, the rough, but consistent grouping of spatially neighboring fine-details and second, the potential of the Fisher vectors in describing such fine-details, even when the latter are not precisely localized.

Table 6 Experiment 3B: Comparison for Birds with state-of-the-art, without human intervention. Late fusion of unsupervised gravitational alignments increases accuracy significantly. Here selective search objectmaps are used

Method	Accuracy (%)
Late fusion	**53.6**
Opponent-SIFT	51.6
RGB-SIFT	49.0
C-SIFT	48.9
Grayscale	40.6
Zhang et al. (2012)	28.2

Highest scores marked with bold

In the absence of bounding boxes the proposed objectmaps, built on off-the-shelf object hypothesis algorithms, provide a good enough spatial support for the fine-grained object of interest. With unsupervised alignments that expect no user input such as bounding boxes, not even during training, we obtain an accuracy of 53.6 %, where the previous best was 28.2 % reported by Zhang et al. (2012) (Table 6).

Finally, our qualitative analysis reveals that Fisher operates as a spatial hashing function, that builds a correspondence between spatial details and certain feature dimensions (Fig. 15). Therefore, even though a more precise object or part localization is always welcome, employing features, such as Fisher vectors, may largely have a similar impact. We, furthermore, observe that computer vision alone cannot solve all categorizations, as the subtle species differences might be anatomical, epochal, or geographical (Fig. 14). In such situations, use of expert knowledge, active learning or metadata would be necessary. For the majority of cases, however, local alignments allow for accurate, and inexpensive, categorization of fine-grained categories.

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

References

Alexe, B., Deselaers, T., & Ferrari, V. (2012). Measuring the objectness of image windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(11), 2189–2202.

Arbelaez, P. Hariharan, B. Gu, C. Gupta, S. Bourdev, L. & Malik, J. (2012). Semantic segmentation using regions and parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012 (pp. 3378-3385). IEEE.

Azizpour, H. & Laptev, I. (2012). Object detection using strongly-supervised deformable part models. In *Proceedings of the European conference on Computer Vision*, (pp. 836–849).

Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-up robust features (surf). *Vision and Image Understanding: Computer*, *110*(3), 346–359.

Berg, T. & Belhumeur, P. N. (2013). POOF: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 955–962). IEEE.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115.

Bo, L. Ren, X. & Fox, D. (2010) Kernel descriptors for visual recognition. In *Proceedings of the Neural Information Processing Systems*.

Bourdev, L. & Malik, J. (2009). Poselets: Body part detectors trained using 3D human pose annotations. In *IEEE International Conference on Computer Vision*, (pp. 1365–1372). IEEE.

Branson, S. Wah, C. Schroff, F. Babenko, B. Welinder, P. Perona, P. & Belongie, S. (2010). Visual recognition with humans in the loop. In *Proceedings of the European Conference on Computer Vision*.

Branson, S. Perona, P. & Belongie, S. (2011). Strong supervision from weak annotation: Interactive training of deformable part models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Branson, S. Van Horn, G. Wah, C. Perona, P. & Belongie, S. (2014). The ignorant led by the blind: A hybrid human—machine vision system for fine-grained categorization. *International Journal of Computer Vision*, 1–27.

Carreira, J. (2012). CPMC: Automatic object segmentation using constrained parametric min-cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(7), 1312–1328.

Chai, Y. Lempitsky, V. & Zisserman, A. (2011). BiCoS: A bi-level co-segmentation method for image classification. In *IEEE International Conference on Computer Vision*, (pp. 2579–2586). IEEE.

Chai, Y. Rahtu, E. Lempitsky, V. Van Gool, L. & Zisserman, A. (2012). TriCoS: A tri-level class-discriminative co-segmentation method for image classification. In *Proceedings of the European Conference on Computer Vision*.

Chai, Y. Lempitsky, V. & Zisserman, A. (2013). Symbiotic segmentation and part localization for fine-grained categorization. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE.

Cinbis, R. G. Verbeek, J. & Schmid, C. (2013). *Segmentation driven object detection with fisher vectors*. In *IEEE International Conference on Computer Vision*.

Dalal, N. & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference Computer Vision and Pattern Recognition*.

Darwin, C. (1859) On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.

Deng, J. Dong, W. Socher, R. Li, L.-J. Li, K. & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Donahue, J. Jia, Y. Vinyals, O. Hoffman, J. Zhang, N. Tzeng, E. & Darrell, T. (2013). DeCAF: A deep convolutional activation feature for generic visual recognition. Technical report. arXiv:1310.1531.

Duan, K. Parikh, D. Crandall, D. & Grauman, K. (2012). Discovering localized attributes for fine-grained recognition. In *Proceedings of the IEEE Conference on Vision and Pattern Recognition*.

Everingham, M. Van Gool, L. Williams, C. K. I. Winn, J. & Zisserman, A. (2007). The PASCAL Visual Object Classes Challenge (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007 2007

Farrell, R. Oza, O. Zhang, N. Morariu, V. I. Darrell, T. & Davis, L. S. (2011). Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. In *Proceedings of the IEEE International Conference on Computer Vision*.

Felzenszwalb, P. F., & Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *International Journal of Computer Vision*, *59*(2), 167–181.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(9), 1627–1645.

Gavves, E. Snoek, C. G. M. & Smeulders, A. W. M. (2012). Convex reduction of high-dimensional kernels for visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Gavves, E. Fernando, B. Snoek, C. G. M. Smeulders, A. W. M. & Tuytelaars, T. (2013). Fine-grained categorization by alignments. In *Proceedings of the IEEE International Conference on Computer Vision*.

Gosselin, P. H. Murray, N. Jégou, H. & Perronnin, F. (2013). Inria+Xerox@FGcomp: Boosting the fisher vector for fine-grained classification. Research Report RR-8431, INRIA.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*, 1254–1259.

Jia, Y. Vinyals, O. & Darrell, T. (2013). Pooling-Invariant Image Feature Learning. Technical report. arXiv:1302.5056.

Khosla, A. Jayadevaprakash, N. Yao, B. & Fei-Fei, L. (2011). Novel dataset for fine-grained image categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Lazebnik, S. Schmid, C. & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Conference Computer Vision and Pattern Recognition*.

Liu, J. Kanazawa, A. Jacobs, D. & Belhumeur, P. (2012). Dog breed classification using part localization. In *Proceedings of the European Conference on Computer Vision*.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Maji, S. Berg, A. C. & Malik, J. (2008). Classification using intersection kernel support vector machines is efficient. In *Proceedings of the IEEE Conference on Vision and Pattern Recognition*.

Maji, S. Kannala, J. Rahtu, E. Blaschko, M. & Vedaldi, A. (2013). Fine-grained visual classification of aircraft. Technical report.

Manén, S. Guillaumin, M. & Van Gool, L. (2013). Prime object proposals with randomized prim's algorithm. In *Proceedings of the IEEE International Conference on Computer Vision*.

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., et al. (2005). A comparison of affine region detectors. *International Journal of Computer Vision*, *65*, 43–72.

Nilsback, M.E. & Zisserman, A. (2008). Automated flower classification over a large number of classes. In *ICVGIP*.

Parikh, D. & Grauman, K. (2011). Interactive discovery of task-specific nameable attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Parkhi, O. M. Vedaldi, A. Zisserman, A. & Jawahar, C. V. (2012). Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Perdóch, M. Chum, O. & Matas, J. (2009). Efficient representation of local geometry for large scale object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Perronnin, F. Sanchez, J. & Mensink, T. (2010). Improving the fisher kernel for large-scale image classification. In *Procedings of the European Conference Computer Vision*.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.

Rother, C. Kolmogorov, V. & Blake, A. (2004). Interactive foreground extraction using iterated graph cuts. In *ACM Transactions on Graphics: Grabcut*. ACM

Sanchez, J. Perronnin, F. & Akata, Z. (2011). Fisher vectors for fine-grained visual categorization. In *Proceedings of the IEEE Conference Computer Vision and Pattern Recognition*.

Shalev-Shwartz, S. Singer, Y. & Srebro, N. (2007) Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the International Conference on Machine Learning*.

Swain, M. J., & Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision*, *7*, 11–32.

Uijlings, J. R. R., van de Sande, K. E. A., Gevers, T., & Smeulders, A. W. M. (2013). Selective search for object recognition. *International Journal of Computer Vision*, *104*, 154–171.

van de Sande, K. E. A. Gevers, T. & Snoek, C. G. M. (2010). Evaluating color descriptors for object and scene recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Vedaldi, A. & Fulkerson, B. (2010). VLFeat: An open and portable library of computer vision algorithms. In *Proceedings of the International Conference on Multimedia*. ACM

Vedaldi, A. Gulshan, V. Varma, M. & Zisserman, A. (2009). Multiple kernels for object detection. In *Proceedings of the International Conference on Vision*.

Wah, C. Branson, S. Perona, P. & Belongie, S. (2011a). Multiclass recognition and part localization with humans in the loop. In *Proceedings of the IEEE International Conference on Computer Vision*.

Wah, C. Branson, S. Welinder, P. Perona, P. & Belongie, S. (2011b). The Caltech-UCSD Birds-200-2011 Dataset. Technical report.

Xie, L. Tian, Q. Yan, B. & Zhang, S. (2013). Hierarcical part matching for fine-grained visual categorization. In *Proceedings of the IEEE Conference on Computer Vision*.

Yang, S. Bo, L. Wang, J. & Shapiro, L. (2012). Unsupervised template learning for fine-grained object recognition. In *Proceedings of the Neural Information Processing Systems*.

Yao, B. Khosla, A. & Fei-Fei, L. (2011). Combining randomization and discrimination for fine-grained image categorization. In *Proceedings of the IEEE Conference on Vision and Pattern Recognition*.

Yao, B. Bradski, G. & Fei-Fei, L. (2012). A codebook-free and annotation-free approach for fine-grained image categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zhang, N. Farrell, R. & Darrell, T. (2012). Pose pooling kernels for sub-category recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zhang, N. Farrell, R. Iandola, F. & Darrell, T. (2013). Deformable part descriptors for fine-grained recognition and attribute prediction. In *Proceedings of the IEEE Conference on Computer Vision*.