Efficient object detection and segmentation for fine-grained recognition

Anelia Angelova
NEC Labs America
Cupertino, CA
anelia@nec-labs.com

Shenghuo Zhu
NEC Labs America
Cupertino, CA
zsh@nec-labs.com

Abstract

We propose a detection and segmentation algorithm for the purposes of fine-grained recognition. The algorithm first detects low-level regions that could potentially belong to the object and then performs a full-object segmentation through propagation. Apart from segmenting the object, we can also 'zoom in' on the object, i.e. center it, normalize it for scale, and thus discount the effects of the background. We then show that combining this with a state-of-the-art classification algorithm leads to significant improvements in performance especially for datasets which are considered particularly hard for recognition, e.g. birds species.

The proposed algorithm is much more efficient than other known methods in similar scenarios [4, 21]. Our method is also simpler and we apply it here to different classes of objects, e.g. birds, flowers, cats and dogs.

We tested the algorithm on a number of benchmark datasets for fine-grained categorization. It outperforms all the known state-of-the-art methods on these datasets, sometimes by as much as 11%. It improves the performance of our baseline algorithm by 3-4%, consistently on all datasets. We also observed more than a 4% improvement in the recognition performance on a challenging large-scale flower dataset, containing 578 species of flowers and 250,000 images.

1. Introduction

This paper addresses the problem of classifying objects that belong to the same basic level category, e.g. species of birds, flowers, etc. This task is often referred to as fine-grained recognition [8, 26] and requires expert, domain-specific knowledge, which very few people generally have. Therefore, developing automated recognition systems for such tasks is of much benefit to non-experts.

The main challenge of fine-grained classification is undoubtedly the very fine differences between species. However, an automatic system will encounter additional challenges. For example, images are taken in natural settings with rich and challenging backgrounds, where the background features may become prominent and serve as distractors to the recognition algorithm. While the background may be useful, e.g. the leaves of the flowers provide informative context, for other super-categories, e.g. birds who are mobile, different classes often share the same background, so segmenting out the background will be beneficial. Segmentation is also helpful to extract the contours of the object of interest, which can provide good features for recognition. Another benefit of a detection and segmentation algorithm is that it can localize the object, which will be beneficial, especially if the object is not in the center of the image or is of size, different from the other objects' sizes.

In this paper we propose an efficient object detection and segmentation algorithm which is effectively used to localize the object and normalize it for scale (Figure 1). Our method segments the possible object of interest *before* trying to recognize it, is much faster than previous counterparts, is applicable to a variety of different super-categories, e.g. birds, flowers, and cats and dogs, and improves the recognition performance for fine-grained classification tasks.

Our approach is based on identifying regions, specific of the categories of interest, at the time of detection. Here the idea is to create feature-based rudimentary detections for the super-class of objects, e.g. birds. These detections are good indicators of the presence of the object and can help point to the possible location of the object. We further apply a Laplacian-based propagation [28] which segments the full object based on low level cues. The key here is that the propagation process is guided by the initially detected regions, but at the same time is capable of preserving object boundaries and thus effectively segments the full object. Furthermore, the obtained segmentation is used to localize the object, normalize it for scale and discount the effects of the background. This is quite beneficial for the final recognition, as shown in our experiments.

The key contributions of this paper are:

- We propose a region-guided detection and segmentation of the object. Apart from providing the object contours, it is beneficial because we can then re-normalize all

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.110

Figure 1. The main idea of this paper is to detect potential object regions and then do a full-object segmentation. This allows the subsequent classification algorithm to 'zoom in' on the object, i.e. re-normalize it to be in the center and take up the whole image. This is particularly beneficial when the object takes a small area of the image, is not in the center, or when the background is shared among different classes (as is the trees background for birds or indoors environment for cats and dogs). The figure shows segmentations produced by our algorithm.

objects to be in the center and take up the whole image, and thus make it comparable to other objects. We combine the feature extracted from the segmented image with a state-of-the-art recognition algorithm and obtain an efficient and reliable recognition pipeline which leads to large improvements in performance.

- We use Laplacian propagation [28] but solve it with fast convergence, which contributes to significant decrease of the run-time for segmentation: 2.5s compared to more than 30s of previous methods [3, 20]. This is of huge importance since segmentation can now be run as part of standard recognition pipelines. Furthermore, the proposed method is simpler and is applicable to variety of datasets, unlike previous works, with the exception of [3], whose methods are designed for specific set of categories e.g. either flowers, or birds, or cats and dogs [8, 17, 20].

We conducted experiments on the well established fine-grained classification datasets: Oxford flower dataset, containing 102 species of flowers [17], the Oxford cats and dogs dataset, containing 37 species of cats and dogs [20], and the Caltech-UCSD-200 birds dataset, containing 200 species of birds [26]. These datasets are very challenging, especially the latter one, as the birds can be encountered in different resolutions with a lot of possible background and object texture variability. The Oxford Cat and Dog dataset additionally contains many pose variations and deformations of the objects of interest. The proposed algorithm outperformed the best known methods in the literature for all three datasets and improves the performance of our baseline algorithm by 3-4%.

Furthermore, our team has collected a large-scale flower dataset which contains 578 different species of flowers and about 250,000 images (Figure 2). This dataset is the largest fine-grained dataset to date. Apart from the significantly larger scale of this dataset, the recognition task is very challenging, because the images exhibit considerable intra-class variabilities, inter-class similarities, scale variations, etc (Figure 2). We tested the proposed algorithm on this 578-class dataset and observed 4.41% improvement in recognition performance compared to the baseline algorithm.

2. Previous work

Fine-grained recognition is a topic of large practical importance and many recent works have addressed such tasks including recognition of flowers [17], birds [2, 8, 26], cats and dogs [19, 20], tree-leaves [15].

Segmentation has played an important role in object recognition with many algorithms available [1, 4, 21]. In another body of works, called co-segmentation [3, 13], better models are trained by exploiting shared appearance features in images containing the same class of objects. These approaches are either too slow or are targeted for segmentation during training.

Recent works have proposed object segmentation for the purposes of better classification. In [5, 19] the authors propose to detect some specific part of the object, e.g. a cat's head, and then segment the object by extrapolating from the textures and colors observed. Another work, again on cat and dog categorization [20], proposes to do segmentation prior to recognition. This work used the famous GrabCut algorithm [21] whose running time is slow for online applications. Our work falls most closely in this category.

Segmentation has also been popular as an initial step for object detection [10] or scene interpretation [11]. Those methods typically work with small coherent regions on the image (called super-pixels) and feed the low-level segmentations to object detection pipelines [10]. Although those methods have provided many insights and useful tools for recognition [10], they have stopped short of providing efficient algorithms for full-object segmentation for either object recognition or detection.

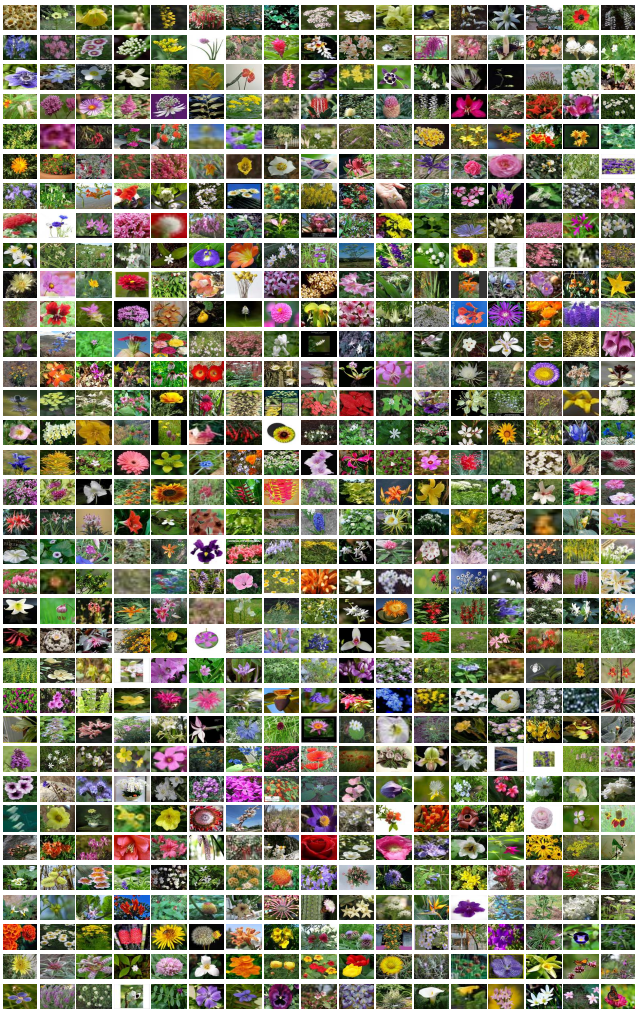

Figure 2. Example images from each class of the large-scale 578 flower dataset.

Other related work, although not doing segmentation per se, has proposed to first localize the potential object region and utilize this information during recognition [14, 22, 23].

3. Object detection and segmentation

This section describes how to detect and segment the object, or objects, in an image. As a first step, a set of rudimentary region-based detection of parts of the object are done (Section 3.1). Then, using those regions as initialization, the Laplacian propagation method, presented in Section 3.2, is applied. Finally, the segmented image (which contains the detected and segmented object, possibly cropped and resized) and input image are processed through the feature extraction and classification pipeline (Section 4) and the final classification is obtained.

3.1. Detecting object-specific regions

We start our method with an initial search for regions possibly belonging to an object from the super-class. For simplicity we use the super-pixel segmentation method by Felzenszwalb and Huttenlocher [9] to over-segment the image into small coherent regions. Each super-pixel region is described by the following set of feature descriptors [3]: average color (R,G,B) of all the pixels within the region, global pooling of all HOG features [6] in the region, after encoding them by the LLC method [25], shape mask of the region obtained by normalizing the region's area bounding box to 6x6 pixels, and size and boundary features as in [3]. We use here the encoded HOG features, instead of SIFT features of [3], because we believe they have better generalization capabilities and because in our classification method (Section 4) these features are already precomputed in the image and can be reused.

Using the features above, we train a classification model to decide if a region belongs to a super-class (e.g. all flowers) or the background. Using ground truth segmentation of training images, we consider super-pixel regions with large overlap with the foreground and background ground truth areas, as positive and negative examples, respectively. Then a linear SVM classifier [7] is trained. When no ground truth is available, we start from an approximate segmentation and iteratively improve the segmentation by applying the trained model. That is, each model is used to segment the training images anew; the newly segmented images are used as 'ground truth' for building an improved model, and so on. This procedure is standard in other segmentation works [3]. The birds and cats and dogs dataset have ground truth segmentation provided, so we built a single model. For the Oxford 102 flowers dataset we used the segmentation images in [17] as seed an improved it iteratively. The training of the model is done offline.

An advantage of this model is that it is general and can be trained on different types of datasets and is not specialized to characteristics of one super-class. As shown later in our experiments, we have the same algorithms for both training of the model and detection for flowers, birds, cats and dogs.

3.2. Full-object segmentation

3.2.1 Setup

Let I_j denote the j-th pixel in an image and f_j denotes its feature representation. The goal of the segmentation task is to find the label X_j for each pixel I_j, where $X_j = 1$ when the pixel belongs to the object and $X_j = 0$, otherwise. For the optimization we relax the requirement on X_j and allow them to be real-valued. We form the affinity matrix W, using the feature representations f_i of each pixel.

$$W_{ij} = exp\left(-\frac{|f_i - f_j|^2}{2\sigma^2}\right)$$

The terms W_{ij} are nonzero for only neighbouring pixels, e.g. in our case we use the 8-connected component neighborhood for each pixel. Additionally, we set $W_{ii} = 0$. Here we set f_i to be the (R,G,B) color values of the pixel, mostly motivated by speed of computation, but other choices are possible too.

The goal is to minimize the cost function $C(X)$ with respect to all pixel labels X:

$$C(X) = \sum_{i,j=1}^{N} W_{ij} \left| \frac{X_i}{\sqrt{D_{ii}}} - \frac{X_j}{\sqrt{D_{jj}}} \right|^2 + \sum_{i=1}^{N} \lambda \left| X_i - Y_i \right|^2$$

where $D_{ii} = \sum_{j=1}^{N} Wij$ and Y are the desired labels for some (or all) the pixels. Those label constraints can be very useful to impose prior knowledge of what is an object and background (we use the SVM margins produced by the model from Section 3.1). This is a standard Laplacian label propagation formulation [28], and the equation above is often written in an equivalent and more convenient form:

$$C(X) = X^T(I - S)X + \lambda |X - Y|^2 \qquad (1)$$

where S is set to $S = D^{-1/2} W D^{-1/2}$.

3.2.2 Optimization

The optimization problem in Equation 1 can be solved iteratively as in [28]. Alternatively, it can be solved as a linear system of equations, which is the approach we chose. After differentiation of Equation 1 we obtain an optimal solution for X, which we solve as a system of linear equations:

$$(I - (1 - \alpha)S)X = \alpha Y, \quad \alpha = \frac{\lambda}{1 + \lambda}$$

In our implementation we use the Conjugate Gradient method, with preconditioning, and achieve very fast convergence.

Since the diffusion properties of the foreground and background of different images (and datasets) may vary, we consider separate segmentations for the detected foreground only-areas and background-only areas, respectively. This is done since the segmentation with respect to one of them could be good but not with respect to the other and combining the results of foreground and background segmentations produces more coherent segmentation and takes advantage of their complementary functions. Denoting $Y_{fg} = Y$ when $Y > 0$ and 0 otherwise, and $Y_{bg} = -Y$ when $Y < 0$ and 0, otherwise, we solve:

$$X_{fg} = \alpha(I - (1-\alpha)S)^{-1} Y_{fg},\ X_{bg} = \alpha(I - (1-\alpha)S)^{-1} Y_{bg}$$

In practice those two segmentations are done simultaneously, by applying the following normalization on the labels

Figure 3. The region detection and segmentation algorithm. Top: Input image and the initial regions which are classified with high score to belong to either a flower or the background. Bottom: Label propagation on this image and the final segmentation result.

Y_{fg}, Y_{bg} prior to the optimization as follows:

$$X_{segm} = \alpha(I - (1 - \alpha)S)^{-1} \left(\frac{Y_{fg}}{||Y_{fg}||_2} - \frac{Y_{bg}}{||Y_{bg}||_2} \right) \qquad (2)$$

This makes the algorithm faster since it avoids separate optimizations. At the same time, it gives equivalent results to the individual foreground and background segmentations which are more stable. To obtain the final segmentation X_{segm} is thresholded at 0.

Figure 3 visualizes the results of the label propagation algorithm and the final segmentation. The top right image shows the score of each super-pixel region (here we use the classification margin). The bottom right image shows the solution of the Laplacian propagation, given the initial regions (i.e. the solution to Equation 2). Note that not all of the object regions have high scores initially. This is also true for the background regions. After the Laplacian propagation, a stronger separation between foreground and background is obtained. Figure 4 shows example segmented images. Note that not all segmentations are successful, especially for birds. However, as seen later in the experiments, even partial segmentations are helpful and our method offers improvement in performance.

4. Fine-grained recognition with segmentation

This section describes how we use the segmented image in the final fine-grained recognition task. For simplicity, we first describe the baseline algorithm.

Baseline. We apply a feature extraction and classification pipeline which is an implementation of the algorithm in Lin et al. [16]. In our feature extraction pipeline we first extract HOG [6] features at 4 different scales, then those features are encoded in 8K dimensional global feature dictionary using the LLC method [25]. After that, a global max

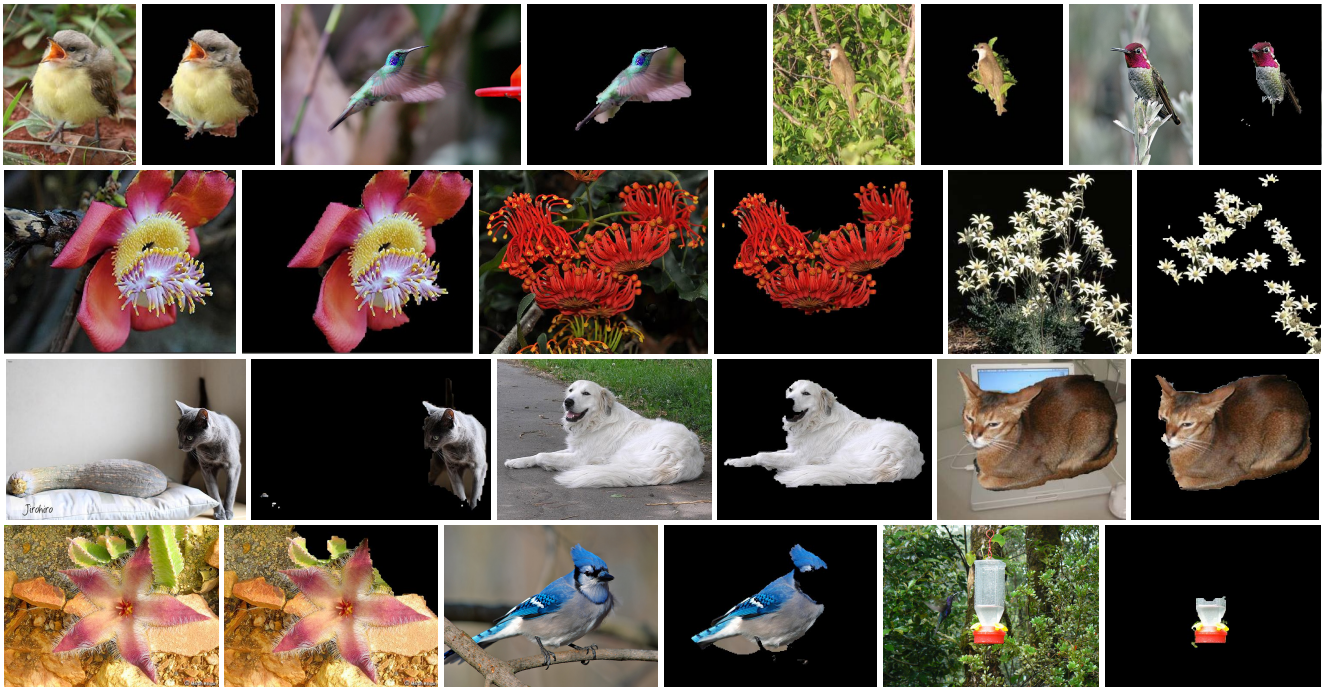

Figure 4. Example segmented images from the datasets tested in this paper. Although not necessarily perfect, these segmentations are sufficient to remove most of the background. Examples of failed segmentations are shown in the bottom row: only a portion of the background is removed, this is typical of flowers since they take larger areas of the image; parts of the object are missing, e.g. for birds' heads or tails, especially if they are of different colors; finally, some segmentations may completely fail (rightmost pair of images).

pooling of the encoded features in the image is done, as well as, max poolings in a 3x3 grid of the image. Our classification pipeline uses the 1-vs-all strategy of linear SVM classification and we used the Liblinear SVM implementation [7]. For the very large 578-flowers dataset, we used a Stochastic Gradient Descent algorithm, since Liblinear cannot load the whole data into memory.

The segmented image is processed through the same feature extraction pipeline as the original image, that is, new features are extracted for the segmented image. We then combine the two sets of extracted features, by concatenating their feature representations. One thing to note here is that, because of our decision to apply HOG type features and pooling to the segmented image, the segmentation helps with both providing shape of the contour of the object to be recognized, as well as, ignoring features in the background that can be distractors. On the other hand, by keeping both sets of features from the original and the segmented image, we can avoid losing precision due to mis-segmentation, and can also include the background for cases for which it may provide useful context (e.g. the leaves and stems of some flowers may be useful for recognition).

In our experiments we found that it is sufficient to keep a global pooling of the segmented image, in addition to the full set of poolings for the original image. We note here that we re-extract features from the segmented image (with their new HOG, encodings etc) and since much 'cleaner' local features are extracted at the boundary, they provide very useful signal, although pooled globally. We believe this is crucial for the improvements we achieved. We cropped the image to the segmented region (+20 pixels margin), so as to account for scale variability. The latter is very beneficial since these datasets have variabilities in scale and one of the purposes of our segmentation is to be able to localize the object and normalize for its scale. No cropping is done for the two flower datasets, since the flowers are assumed to take most of the image area (even for small 'cluster' flowers). We did not do cropping for the experiment which uses ground truth bounding box information (for birds).

In terms of computation, our algorithm performs much better compared to competitors [3, 19]. The segmentation takes about 2.5 seconds, and our baseline algorithm runs within 1-2 seconds. We note that the segmentation procedure is much faster than previously known segmentation methods, which take at least 30 seconds [4, 21]. Furthermore, our segmentation run-time allows it to be run as a part of standard recognition pipelines at test time, which had not been possible before, and is a significant advantage.

5. Experiments

In this section we show experimental results of our proposed algorithm on a number of fine-grained recognition benchmarks: Oxford 102 flowers [17], Caltech-UCSD 200 birds [2, 26], and the recent Oxford Cats and Dogs [20] datasets. In each case we report the performance of our baseline classification algorithm, the best known benchmark results achieved on this dataset, and our proposed algorithm in the same settings. We compare to our baseline algorithm, because it measures how much the proposed segmentation has contributed to the improvement in classification performance. In addition, we measure our performance on the large-scale 578-category flower dataset.

5.1. Oxford 102 flower species dataset

Oxford 102 flowers dataset is a well-known dataset for fine-grained recognition proposed by Nilsback and Zisserman [17]. The dataset contains 102 species of flowers and a total of 8189 images, each category containing between 40 and 200 images. It has well established protocols for training and testing, which we adopt in this paper too.

A lot of methods have been tested on this dataset [3, 12, 17, 18], including some segmentation-based [3, 17]. Some of the segmentation methods are designed to be very specific to the appearance of flowers [17] (with the assumption that a single flower is in the center of the image and takes most of the image), while others [3] are more general and can also be applied to other types of datasets. Our approach, too, does not make assumptions about the set of categories for classification or the initial location or size of the objects in the image.

The performance of our approach on this dataset (see Table 1) is 80.66%, which outperforms all previous known methods in the literature (some by as much as 4 to 8%) [3, 12, 17, 18]. One important thing to note is that the improvement of our algorithm over our baseline is about 4%, and the only difference between the two is the addition of the proposed segmentation algorithm and the features extracted from the segmented image.

5.2. Caltech-UCSD 200 birds species dataset

Caltech-UCSD-200 Birds dataset [26] is a very challenging dataset containing 200 species of birds. Apart from very fine-differences between different species of birds, what makes the recognition hard in this dataset is the variety of poses, large variability in scales, and also very rich backgrounds in which the birds often blend in. The best classification performance achieved on this data is 16.2% classification rate by [3]. Even when using ground truth bounding boxes, provided as annotations with the dataset [26], the reported results have been around 19% [26, 27] and most recently 24.3% [3], but the latter result additionally uses crude ground truth segmentation of each bird.

Method	Accuracy (in %)
Our baseline (no segmentation)	76.7
Nilsback and Zisserman [17]	72.8
Ito and Cubota [12]	74.8
Nilsback and Zisserman [18]	76.3
Chai, Bicos method [3]	79.4
Chai, BicosMT method [3]	80.0
Ours	**80.66**
Ours: improvement over our baseline	**+**3.94

Table 1. Classification performance on Oxford 102 flower dataset.

The proposed algorithm improves the performance both with and without using ground truth bounding boxes (see Tables 2 and 3). Our algorithm achieves 30.17% classification performance compared to 19.2 [27] in the same setting, which in an improvement of 11% over the best known baselines in this scenario. Another interesting observation is that our algorithm achieves a performance of 27.60% when applying segmentation alone (i.e. without combining it with the baseline algorithm). This is by itself an impressive improvement over the other known algorithms for this dataset (even when not taking advantage of our baseline performance). When considering the benefits of the segmentation, we noticed that examples which have more cluttered backgrounds are helped most by the segmentation.

Most importantly, our algorithm shows improvement over all known prior approaches, when no ground truth bounding boxes are used. In this case we observed 17.5% classification rate compared to previous 15.7% and 16.2%, Our baseline algorithm here achieves only 14.4% which in on par with the performance of SPM-type methods in this scenario. Another thing to notice here is that the improvement over our baseline, when no bounding boxes information is known, is larger than the improvement with bounding boxes. This improvement is consistent across the other datasets tested in this paper, which do not have bounding box information. We attribute this to the fact that the bounding boxes have perfect object localization and scaling, and to large extent have background elimination capabilities. This underlines the importance of our proposed automatic detection and segmentation of the object, which then allows to 'zoom in' on the object, especially for large-scale datasets for which providing bounding boxes or other ground truth information will be infeasible.

5.3. Oxford Cats and Dogs dataset

Oxford Cats and Dogs [20] is a new dataset for fine-grained classification which contains 6033 images of 37 breeds of cats and dogs. Parkhi et al, who collected the dataset, showed impressive performance on this dataset [20]. They apply segmentation at test time, as is done here, but their algorithm is based on Grabcut [21],

Method	Accuracy (in %)
Our baseline (no segmentation)	14.4
Chai, Bicos segmentation [3]	15.7
Chai, BicosMT segmentation [3]	16.2
Ours	**17.5**
Ours, improvement over our baseline	**+**3.1

Table 2. Classification performance on Caltech-UCSD 200 birds dataset with automatic segmentation. Please refer to Table 3 for comparison to other baselines which additionally use ground truth bounding box information.

Method (with ground truth boxes)	Accuracy (in %)
Our baseline (no segmentation)	29.06
Branson et al [2]	19.00
Yao et al. [27]	19.20
Chai et al. [3]	23.30
Ours, segmentation only, see text	27.60
Ours	**30.17**
Ours, improvement over our baseline	**+**1.11

Table 3. Classification performance on Caltech-UCSD 200 birds dataset, when **ground truth** bounding boxes are used (the result in [3] uses crude ground truth segmentation masks in addition to bounding boxes).

which is slow. Also, the methods proposed in [20] are specific to recognizing cat and dog breeds and utilize head and body layout information.

We compared our performance on this dataset with the prespecified protocol proposed in the paper (Table 4). For this dataset too, we see that our general method outperforms the best category-specific one from [20] and is far better than their more general approach (denoted as 'image info only') or a bag of words-based method. Note that [20] also reported classification when using cat and dog head annotations or ground truth segmentation during testing, whereas here our experiments do not use such information.

5.4. Large-scale 578 flower species dataset

This dataset consists of 578 species of flowers and contains about 250,000 images and is the largest and most challenging such dataset we are aware of. The goal of developing this data is to build a recognition application which can recognize and/or provide top K suggestions (e.g., for K=5, 10, etc.) for an input flower image, and be available for general use.

We tested our baseline algorithm vs the proposed segmentation-based algorithm on this data, see Table 5. The improvement provided by our segmentation method is 4.41% for the top 1 returned result. Figure 5 shows the recognition performance for top K, where $K = 1, \ldots, 10$. As seen, we obtain improvements across all of them, with

Method	Accuracy (in %)
Our baseline (no segmentation)	50.79
VLFeat [24] + Bag of words [20]	38.45
Parkhi et al. (image info only) [20]	39.64
Parkhi et al. [20]	54.05
Ours	**54.30**
Ours, improvement over our baseline	**+**3.51

Table 4. Classification performance on Oxford Cats and Dogs dataset.

Method	Accuracy (in %)
Our baseline (no segmentation)	52.35
Ours	**56.76**
Ours, improvement over our baseline	**+**4.41

Table 5. Classification performance on the large-scale 578 flowers dataset for the top returned result.

top 1 having an improvement of about 4.41%, top 5 of about 2.7% and top 10 of about 2%.

Note that this large-scale data has no segmentation ground truth or bounding box information (since it contains 250,000+ images and obtaining those would be prohibitive or at least very expensive). Thus, here the advantage that an automatic segmentation algorithm can give in terms of improving the final classification performance is really important. Another interesting fact is that here we have used the same initial region detection model that was trained on the Oxford 102 flowers dataset, which contains fewer species of flowers (102 instead of 578). This was motivated again by the lack of good ground truth for such a large volume of data. Naturally, the performance of the segmentation algorithm can be further improved after adapting the segmentation model to this specific dataset.

Discussion. As seen by the improvements over the baseline, our segmentation algorithm gives advantage in recognition performance. This is true even if the segmentation may be imperfect for some examples. This shows that segmenting out the object of interest during testing is of crucial importance for an automatic algorithms and that it is worthwhile exploring even better segmentation algorithms.

6. Conclusions and future work

We propose an algorithm which combines region-based detection of the object of interest and full-object segmentation through propagation. The segmentation is applied at test time and is shown to be very useful for improving the classification performance on four challenging datasets.

We tested our approach on the most contemporary and challenging datasets for fine-grained recognition improved the performances on all of them. We further tested with

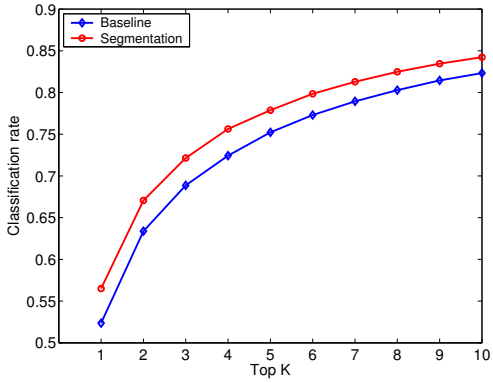

Figure 5. Classification performance on the large-scale 578 flowers dataset for top $K = 1, \ldots, 10$ retrieved results.

578-category flower dataset which is the largest collection of flower species we are aware of. The improvements in performance over the baseline are about 3-4%, which is consistent across all the experiments. Our algorithm is much faster than previously used segmentation algorithms in similar scenarios, e.g. [4, 21]. It is also applicable to a variety of types of categories, as shown in this paper on birds, flowers, and cats and dogs.

Our future work will consider improvements to the feature model, e.g. represent it as a mixture of submodels, each one responsible for a subset of classes that are very similar to each other but different as a group from the rest.

Acknowledgements

We thank the team at the Department of Plant and Microbial Biology at UC Berkeley, lead by Prof. Chelsea Specht. The collection of the large-scale 578-class flower dataset would not have been possible without their critical expertise in flower identification. We also thank Dr. Yuanqing Lin and Olga Russakovsky for their help and for valuable discussions.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. Classcut for unsupervised class segmentation. *ECCV*, 2010.

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

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

[4] G. Csurka and F. Perronnin. An efficient approach to semantic segmentation. *IJCV*, 2011.

[5] Q. Dai and D. Hoiem. Learning to localize detected objects. *CVPR*, 2012.

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

[7] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 2008.

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

[9] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 2004.

[10] C. Gu, J. Lim, P. Arbelaez, and J. Malik. Recognition using regions. *CVPR*, 2011.

[11] D. Hoiem, A. Efros, and M. Hebert. Closing the loop on scene interpretation. *CVPR*, 2008.

[12] S. Ito and S. Kubota. Object classfication using heterogeneous co-occurrence features. *ECCV*, 2010.

[13] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. *CVPR*, 2010.

[14] F. Khan, J. van de Weijer, and M. Vanrell. Top-down color attention for object recognition. *ICCV*, 2009.

[15] N. Kumar, P. Belhumeur, A. Biswas, D. Jacobs, J. Kress, I. Lopez, and J. Soares. Leafsnap: A computer vision system for automatic plant species identification. *ECCV*, 2012.

[16] Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, and T. Huang. Large-scale image classification: fast feature extraction and svm training. *CVPR*, 2011.

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

[18] M.-E. Nilsback and A. Zisserman. An automatic visual flora - segmentation and classification of flower images. *DPhil Thesis, University of Oxford, UK*, 2009.

[19] O. Parkhi, A. Vedaldi, C. V. Jawahar, and A. Zisserman. The truth about cats and dogs. *ICCV*, 2011.

[20] O. Parkhi, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Cats and dogs. *CVPR*, 2012.

[21] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graphics*, 2004.

[22] O. Russakovsky, Y. Lin, K. Yu, and L. Fei-Fei. Object-centric spatial pooling for image classification. *ECCV*, 2012.

[23] J. Sanchez, F. Perronnin, and T. de Campos. Modeling the spatial layout of images beyond spatial pyramids. *Pattern Recognition Letters*, 2012.

[24] A. Vedaldi and B. Fulkerson. Vlfeat library.

[25] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. *CVPR*, 2010.

[26] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-ucsd birds 200. *Technical Report CNS-TR-2010-001, California Institute of Technology*, 2010.

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

[28] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Scholkopf. Learning with local and global consistency. *NIPS*, 2004.