Does Image Segmentation Improve Object Categorization?

Andrew Rabinovich[1], Andrea Vedaldi[2] and Serge Belongie[1]

[1]Department of Computer Science and Engineering,
University of California, San Diego
{amrabino,sjb}@cs.ucsd.edu

[2]Department of Computer Science,
University of California, Los Angeles
avedaldi@cs.ucla.edu

Abstract

Image segmentation and object recognition are among the most fundamental problems in computer vision, and the potential interaction between these tasks has been discussed for many years. The usefulness of recognition for segmentation has been demonstrated with various top-down segmentation algorithms, however, the impact of bottom-up image segmentation as pre-processing for object recognition is not well understood. One factor impeding the utility of segmentation for recognition is the unsatisfactory quality of image segmentation algorithms. In this work we take advantage of a recently proposed method for computing multiple stable segmentations and illustrate the application of bottom-up image segmentation as a preprocessing step for object recognition and categorization. We extend a popular bag-of-features recognition model to provide multiple class categorization and localization of objects in images. We compare our categorization results to that of a conventional bag-of-features recognition model on the Caltech and PASCAL image databases.

1. Introduction

The interplay between image segmentation and object recognition has been an active area of research for several decades, both in computer vision and cognitive psychology. The benefits of object recognition have been exploited in top-down image segmentation approaches. Combining object model knowledge and the initial low level segmentation has been shown to improve segmentation accuracy [3]. However, the effects of image segmentation as a pre-processing step for object recognition and categorization are still not clear.

Discovering global structure is at the heart of most approaches to image segmentation. For example, image segmentation methods based on spectral clustering proceed by computing local measurements around each pixel followed by a partitioning step that aims to minimize a global cost function defined over pairwise affinities over these measurements [2, 19, 27]. In this setting, the global structure is represented concretely by a set of partition vectors indicating group membership. Many leading recognition engines, however, are solely based on local feature descriptors [5, 7]. Yet in contrast, the principle of global precedence suggests that global image structure and configurations dominate local feature processing in human pattern perception and recognition [8, 18].

Recently, there have been efforts that leverage manually segmented foreground objects from the cluttered background to improve categorization. In Nilsback et al. [20], for example, flowers are segmented from the background to increase recognition accuracy. By segmenting the objects of interest, the noise introduced by the background around the object is minimized. Yet, methods of unsupervised image segmentation have not been popular as preprocessing for recognition and categorization. One reason for this is the unsatisfactory quality of image segmentation algorithms. It is generally hard to find segmentations that capture all correct object boundaries in images of real world scenes. If the segments were satisfactory, an ideal segmentation based recognition system would resemble the sketch in Figure 1. After perfect segmentation, each segment (representing an object) is labeled by the recognition engine. Segment boundaries are used for localization and the scene category label is inferred from the individual object labels.

Existing recognition algorithms that advocate the use of segmentation appear to work well if strong initial object hypotheses are built into the segmentation engine [11, 30]. For the task of detecting and recognizing objects in still images without object knowledge, however, the recognition capability is still very weak, perhaps due to the segmentation performance. For example, the approach of Martin et al. [15] attempts to integrate all necessary visual cues together to produce one “best” segmentation. The work of Mori et al. [17] acknowledges that an erroneous segment

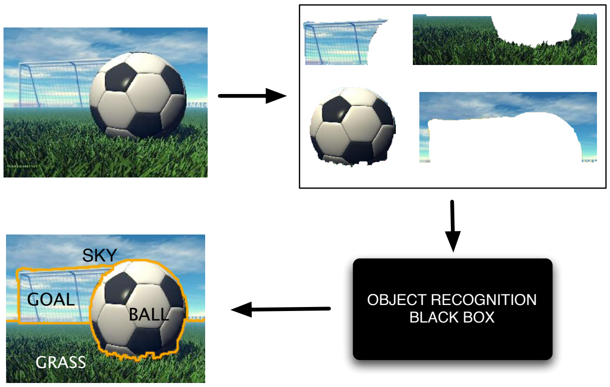


Figure 1. Illustration of a segmentation-based object recognition system. **Top Left:** Original image with four objects: soccer ball, goal, grass and sky. **Top Right:** Ideal image segments. **Bottom Right:** Discriminative object recognition system, e.g. "Bag of Features". **Bottom Left:** Multi-class object recognition with localization.

boundary will degrade recognition accuracy, and thus proposes to oversegment an image into super-pixels to increase the potential quality of a single merged segmentation. Alternatively, works such as Viola and Jones [29] suggest that attempts to calculate a segmentation for an input image are likely to introduce more harm than good, and that a bounding box, at every possible location and scale in the image, must be considered as an object outline for satisfactory object recognition and categorization performance.

A reason for the inadequate performance of image segmentation is the ambiguity of image representation, model parameterization, and the task itself. As described in [23], in general there does not exist a single correct segmentation of an image, but rather a shortlist of meaningful image partitionings. Thus, unlike the above mentioned approaches of using a single segmentation or all possible bounding boxes, the idea of using several segmentations has recently emerged [23, 25, 26]. A handful of segmentations is chosen in hope that a collection of all segments from these few segmentations will result in adequate object boundaries. Russel et al. rely on a collection of random segments to perform object detection, while we use stability as a predictor of "goodness" of a particular set of parameters, cue weightings and model order, as done in [9, 23] to perform object recognition and categorization. Only the most stable segmentations that depict various aspects of the image are chosen to describe object boundaries. In this regard, the segmentations we use go beyond what is available via a simple oversegmentation or superpixel representation in terms of capturing salient image structure.

Partitioning images into segments has been proposed for learning the joint distribution of image regions and words for image region annotation [1]. Recently, a work by Roth and Ommer [25] suggested using multiple segmentations for object recognition. They build a segmentation based recognition system and report competitive results. However, they do not show the performance of their system without segmentation. Thus the effects of segmentation on object categorization remain unclear. Also they do not leverage segmentation for object localization and multi-class object recognition.

In this work we show that preprocessing test images by representing them as a shortlist of segmentations increases the accuracy of object categorization. Having classified each of the segments we infer the following from the shortlist of segmentations: (a) a label for the entire image, (b) object localization via the segment boundary, and (c) multi-class object localization and categorization. We evaluate the benefits of image segmentation, as pre-processing, for object categorization on the Caltech and PASCAL databases.

Finally, we investigate the importance of image segmentation for object categorization, and answer the following questions:

▸ Can segmenting an image improve object recognition?

▸ How does the number of segments affect recognition accuracy?

▸ Does the quality of segmentation affect recognition accuracy?

▸ Is it beneficial to perform localization and multi-class recognition using segmentation?

The organization of this paper is as follows. In Section 2 we review the proposed image segmentation algorithm and how it is used with the bag-of-features (BoF) recognition model. We also address object localization and multi-class object recognition using image segmentation. In Section 3 we describe the experiments and present results of object recognition on the Caltech and PASCAL databases. We conclude with the discussion and future work in Section 4.

2. Segmentation for Recognition

To understand the effects of image segmentation on object recognition and categorization, we consider the stability based image segmentation framework and the BoF object recognition model. Although our results are influenced by these choices, we believe that the conclusions will carry over to other object based recognition models.

2.1. Shortlist of Stable Segmentations

The goal of an unsupervised clustering algorithm is to partition the data based on some criterion that, by definition, does not use labeled examples. Open problems in this area include choosing the appropriate grouping criterion (cue selection and combination) and the number of clusters (model order). Recent advances in stability based clustering algorithms have shown promising results for choosing these parameters. In this work we adopt the framework of [23] to generate a shortlist of stable segmentations.

Next we review the basics of stability based image segmentation. Cues are combined into one similarity measure using a convex combination:

$$W_{ij} = \sum_{f=1}^{F} (p_f \cdot C_{ij}^f), \text{ subject to } \sum_{f=1}^{F} p_f = 1,$$

where W_{ij} is the overall similarity between pixels i and j, C_{ij}^f is the similarity between the i-th and j-th pixels according to some cue f, and F is the number of cues. Since the "correct" cue combination $\vec{p}$ and the number of segments k yielding "optimal" segmentations are unknown *a priori*, we would like to explore all possible parameter settings. However, this is not computationally viable and we adopt an efficient sampling scheme. Nonetheless, we are still left with defining the optimal segmentations, which we do next.

Stability Based Clustering. For each choice of cue combination $\vec{p}$ and number of segments k one obtains different segmentations of the image. Of all possible segmentations arising in this way, one or more can be considered "meaningful." Here we use stability as a heuristic to define and compute the meaningful segmentations.

For a choice of the parameters $\vec{p}$ and k, the image is segmented using Normalized Cuts [14, 27] using the implementation of [4]. The segmentation is considered stable if small perturbations of the image do not yield substantial changes in the segmentation. This condition is evaluated as follows [23]. The image is perturbed and segmented T times and the following score is evaluated:

$$\Phi(k, \vec{p}) = \frac{1}{n - \frac{n}{k}} \left(\sum_{i=1}^{n} \sum_{j=1}^{T} \delta_{ij} - \frac{n}{k} \right).$$

Here n is the number of pixels and δ_{ij} is equal to 1 if the i-th pixel is mapped to a different segment in the j-th perturbed segmentation and zero otherwise. Thus Φ is a properly normalized[1] measure of the probability of a pixel to change label due to a perturbation of the image. Segmentations with high stability score are considered meaningful and are retained. Thus, in general, there may exist several stable segmentations.

2.2. Bag of Features

In this work we utilize the BoF object recognition framework [7, 22] due to its popularity and simplicity. This method consists of four steps: (i) images are decomposed into a collection of "features" (image patches); (ii) features are mapped to a finite vocabulary of "visual words" based on their appearance; (iii) a statistic, or *signature*, of such visual words is computed; (iv) the signatures are fed into a classifier for labeling. All four steps can be implemented in a variety of ways. Here we adopt the implementation and default parameter settings provided by [28].

2.3. Integrating Bag of Features and Segmentation

We integrate segmentation with BoF as follows. Each segment is regarded as a stand-alone image by masking and zero padding the original image. Then the signature of the segment is computed as in regular BoF, but any features that fall entirely outside its boundary are discarded. Eventually, the image is represented by the ensemble of the signatures of its segments.

This simple idea has a number of effects: (i) by clustering features into segments we incorporate coarse spatial information; (ii) masking often enhances the contrast of segment boundaries, making features along the boundaries more shape-informative; (iii) computing signatures on homogeneous segments improves their signal-to-noise ratio.

Next we discuss how segments and their signatures are used to classify segments and whole images and to localize objects in them.

Labeling Segments. Let I be a test image and S_q its q-th segment, with i being the image index and c the category index, such that I_{ic} is the i-th training image of the c-th category. Let $\phi(I)$ (or $\phi(S)$) be the signature of image I (or segment S) and $\Omega(I)$ (or $\Omega(S)$) the number of features extracted in image I (or segment S).

Segments are classified based on a simple nearest neighbor rule. Define the un-normalized distance of the test segment S_q to class c as:

$$d(S_q, c) = \min_i d(S_q, I_{ic}) = \min_i \|\phi(S_q) - \phi(I_{ic})\|_1$$

So $d(S_q, c)$ is the minimum l_1 distance of the test segment S_q to all the training images I_{ic} of category c. We assign the segment S_q to its closest category $c_1(S_q)$:

$$c_1(S_q) = \operatorname*{argmin}_c d(S_q, c).$$

In order to combine segment labels into a unique image label it is useful to rank segments by classification reliability. To this end we introduce the following confidence measure.

Labeling Confidence. Define the *second best labeling* of segment S_q as the quantity:

$$c_2(S_q) = \operatorname*{argmin}_{c \neq c_1(S_q)} d(S_q, c).$$

[1] In particular, $\Phi \in [0, 1]$ and it is not biased towards a particular value of k.

In order to characterize the ambiguity of the labeling $c_1(S_q)$ we compare the distance of S_q to $c_1(S_q)$ and $c_2(S_q)$, defining:

$$p(c_1(S_q)|S_q) = (1-\gamma)+\gamma/C, \quad \text{where } \gamma = \frac{d\left(S_q, c_1(S_q)\right)}{d\left(S_q, c_2(S_q)\right)}$$

and C is the number of categories. This is the belief that S_q has class $c_1(S_q)$; for other labels, $c \neq c_1(S_q)$:

$$p(c|S_q) = \frac{1 - p\left(c_1(S_q)|S_q\right)}{C-1}.$$

So $p(c|S_q)$ is a probability distribution over labels and it is uniform when $d\left(S_q, c_1(S_q)\right) \approx d\left(S_q, c_2(S_q)\right)$ and peaked at $c_1(S_q)$ when $d\left(S_q, c_1(S_q)\right) \ll d\left(S_q, c_2(S_q)\right)$.

Labeling Whole Images. Let $\{S_1, ..., S_K\}$ be all the segments of a test image I. We let the segments vote for the image label as follows. Each segment S_q votes for class c proportionally to its confidence $p(c|S_q)$ and has an amount of votes $w(S_q)$ to use. The label of the image I is then given by:

$$c(I) = \underset{c}{\operatorname{argmax}} \sum_{q=1}^{K} p(c|S_q)w(S_q).$$

The weights $w(S_q)$ encode both the importance and the reliability of the segment S_q, irrespective of the class label. As both of these factors are roughly proportional to the number of features of the segment, we define $w(S_q) = \Omega(S_q)/\Omega(S_{\max})$ where $S_{\max}$ is the largest segment (in terms of number of features).

Localization. In many approaches to object localization, the bounding box that yields highest recognition accuracy is used to describe objects' location [16, 29]. Here we use the segment boundaries instead.

Given the labels of each segment, $c_1(S_q)$, and the overall image label, $c(I)$, we look for segments whose labels match the image label, i.e., $c(I) = c_1(S_q)$. Among these, we check for overlapping segments and we return the first k unique segment boundaries. Note that this method is not limited to BoF and could be used to add localization capabilities to other recognition methods.

To recognize and localize objects of classes other than the image class, all segments S_q are ranked with respect to their label confidence $p(c_1(S_q)|S_q)$ and the first k segment boundaries are returned irrespective of the whole image label.

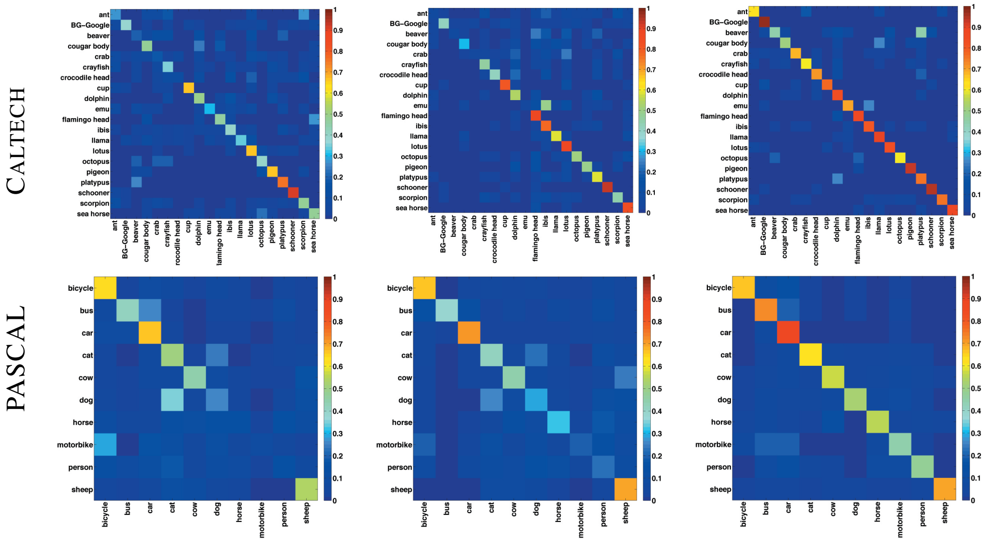
3. Experimental Results

For our experiments we use images from the standard datasets Caltech-101 and PASCAL. For the Caltech-101 database we picked the twenty most difficult categories, as reported by [31]. For both databases, we used 30 images per category for training. The implementation details of [28] for the BoF model are the following. 5000 random patches at multiple scales (from 12 pixels to the image size) are extracted from each image such that larger patches are sampled less frequently (as these would be redundant). The feature appearance is represented by SIFT descriptors [13] and the visual words are obtaining by quantizing the feature space using hierarchical K-means with $K = 10$ at three levels [21]. The image signature is an histogram of such hierarchical visual words, L_1 normalized and TFxIDF reweighed [21]. In an unoptimized MATLAB/C implementation, the computation of SIFT and the relevant signatures, takes on average 1 second for each segment in the image on on a Pentium 3.2 GHz. Finally, the signatures are fed to a k-nearest-neighbor classification algorithm. Implemented in MATLAB, training the classifier and constructing the vocabulary takes under 1 hour for 20 categories with 30 training images in each category. Classification of test images, however, is done in just a few seconds.

To understand the importance of image segmentation quality for object categorization accuracy we consider the following two segmentation methods. The first is the stability based image segmentation using normalized cuts [27]. Images are segmented using brightness and texture cues. We consider a varying number of segments per segmentation, $k = 2, \ldots, 10$, which together results in 54 segments $(2 + 3 + 4 + \cdots + 10)$. Implemented in MATLAB, each segmentation takes between 10-20 seconds per image with $T = 100$ restarts, on a Pentium 3.2 GHz , depending on the image size. Typical images in the Caltech database are at least 600 × 400 pixels. We'll refer to this method as "Stable Segmentations" (Sseg). The second segmentation method is a simple grid-like image partitioning method, similar to that of [10]. In real time, an image is broken into $k = 4, 9, 16, 25$ equal sub-images, which together results in 54 segments $(4 + 9 + 16 + 25)$. We refer to this method as "Block Segmentations" (Bseg).

3.1. Average Recognition Accuracy

We compare the categorization results of the BoF with and without segmentation pre-processing to quantify the effects of image segmentation on the accuracy of object categorization. Figure 2 shows the confusion matrices of 20 most difficult categories from the Caltech-101 and PASCAL databases simply using the BoF model. Confusion matrices of average recognition with no pre-processing, with "Block Segmentations", and with "Stable Segmentations" are shown in columns (a), (b), and (c) respectively. The results of average recognition accuracy are summarized in Table 1. The reported results are based on 54 segments per image. In the case of "Stable Segmentations" segments are taken from 9 segmentations, and for "Block Segmentations"



(a) (b) (c)

Figure 2. Confusion matrices of object categorization accuracy using the BoF model. **Top row:** 20 hardest categories of Caltech101. **Bottom row:** PASCAL dataset. (a) BoF model with no preprocessing. (b) BoF model with test images represented by "Block Segmentations". (c) BoF recognition model with test images represented by "Stable Segmentations".

from 4.

	No Seg.	Bseg	Sseg
Caltech	44.9%	50.6%	75.5%
PASCAL	38.5%	43.5%	61.8%

Table 1. Average categorization accuracy for both the Caltech and PASCAL datasets.

Caltech	Bseg	Sseg
ant	0.24	0.47
BG Google	0.25	0.84
Beaver	0.23	0.65
Cougar body	0.27	0.69
Crab	0.27	0.51
Crayfish	0.24	0.53
Crocodile Head	0.37	0.72
Cup	0.31	0.77
Dolphin	0.31	0.78
Emu	0.19	0.64
Flamingo Head	0.17	0.62
Ibis	0.27	0.58
Llama	0.28	0.73
Lotus	0.40	0.65
Octopus	0.11	0.66
Pigeon	0.13	0.78
Platypus	0.19	0.71
Schooner	0.34	0.72
Scorpion	0.12	0.56
Sea Horse	0.16	0.62

Pascal	Bseg	Sseg
Bicycle	0.24	0.51
Bus	0.34	0.70
Car	0.34	0.66
Cat	0.26	0.62
Cow	0.30	0.64
Dog	0.23	0.60
Horse	0.26	0.52
Motorbike	0.14	0.43
Person	0.22	0.59
Sheep	0.19	0.67

Table 2. Average object localization accuracy for Caltech and PASCAL datasets.

3.2. Localization and Multiclass Categorization

The quality of object localization, whether for single or multi-class recognition, can be evaluated in a number of ways. Some compare object centroid location, while others attempt to maximize the overlap between predicted bounding box around the object and the ground truth one [16]. However, objects are generally not rectangular and should be localized by their boundary contour, which we do here. To quantify the accuracy of object localization, we adopt a method from the PASCAL Challenge [6] and consider the overlap, ρ, between ground truth localization, GT, and the retrieved localization, R, is $\rho = \frac{GT \cap R}{GT \cup R}$. Note that ρ is misleading in cases where the objects' contour area is smaller than that of its bounding box (Fig. 5). In Table 2 we report the average localization accuracy for each category in both the Caltech and PASCAL datasets. For each image, the segment R which is more likely to have a given label is compared to the ground truth bounding box GT. We have also explored the relationship between number of segmentations per image and object localization accuracy, but we cannot report the results due to space constraints. Generally, categories of objects with complex boundaries are localized more accurately as the number of segments increase, while blob-like objects do not benefit as much from an increase in the number of segments. Fig. 4 and 5 show example of objects localized by our method.

3.3. Quality of Image Segmentation

Due to the principle of global precedence and the importance of the shape cue, it is expected that the object categorization accuracy based on "Stable Segmentation" should outperform that of the trivial "Block Segmentations". Indeed, the results in Table 1 indicate that the improvement with "Stable Segmentations" is significant.

The localization based on "Stable Segmentations" is also superior to that of "Block Segmentations". The "Stable Segmentations", shown in Fig. 4 and 5, are capable of identifying objects' boundaries relatively accurately. Us-

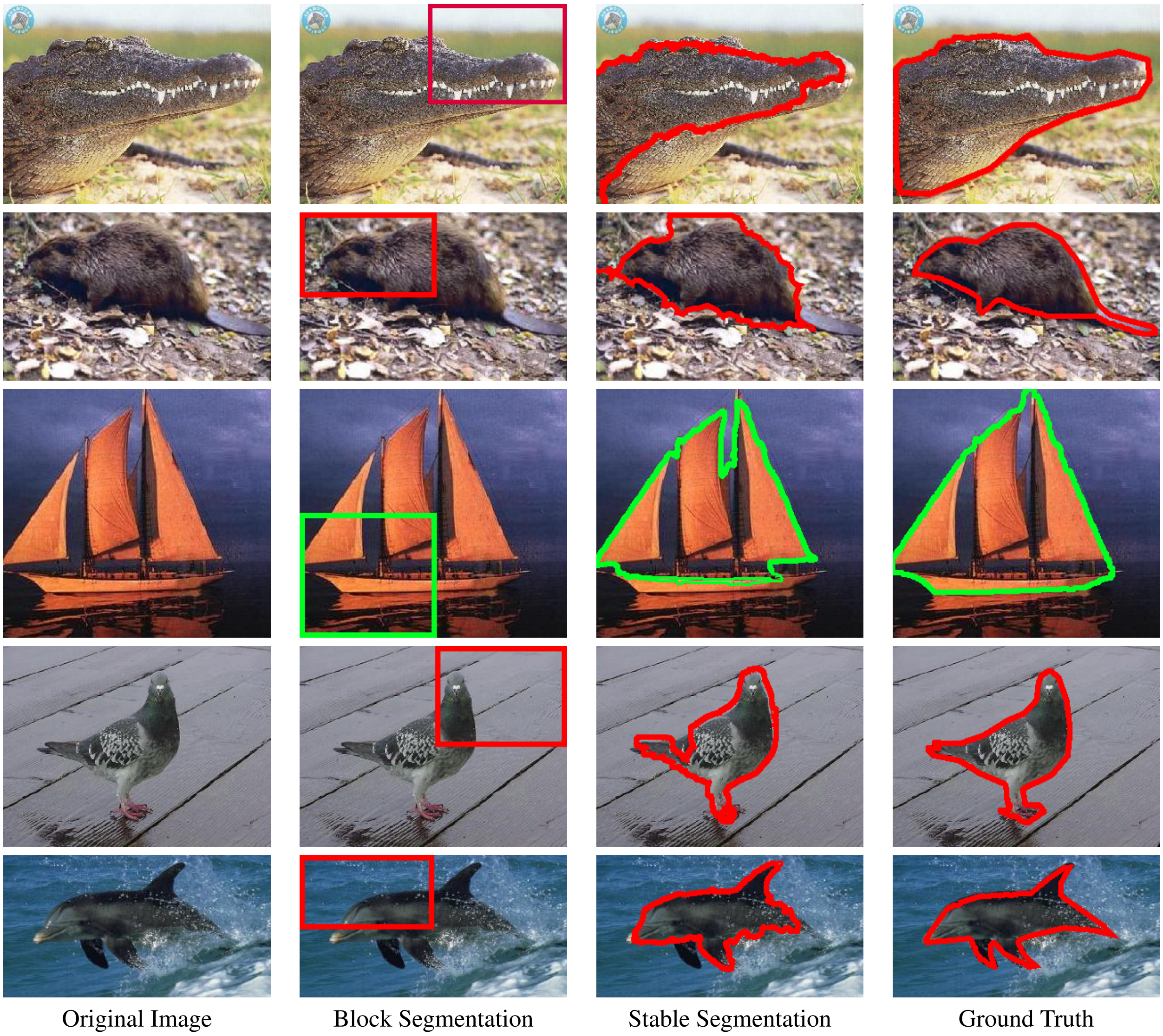


Figure 4. Object localization using "Stable Segmentations" as pre processing for the BoF categorization model. Examples from the Caltech101 dataset. *Please view in color*.

ing "Block Segmentations", however, localization results are poor: the centroids of segments often do not match the objects' center and segments' boundaries truncate the objects. Average per category localization results are reported in Table 2.

Regardless of the particular segmentation algorithm, the size of the shortlist or the number of segments used to represent a test image can play an important role in determining object recognition accuracy. On one hand, as the number of possible segmentations increases, the chance of having a segment perfectly represent the object increases as well. On the other hand, an increase in the number of segments also increases the noise, namely, segments with incorrect category assignment. Figure 3 illustrates the effect of increasing the number of segments to represent the test images. The recognition accuracy of all categories significantly increases with the number of segments. However, around the 35 segment mark, the effect of the more accurate segment boundaries is cancelled out by the noise from meaningless segments. Thus, for most categories, the recognition accuracy saturates past 35 segments per image (note that 35 segments are distributed among 7 different segmentations).

4. Discussion

Although a link between image segmentation and object recognition has been discussed for many years, the effects of low-level global image segmentation on recognition and

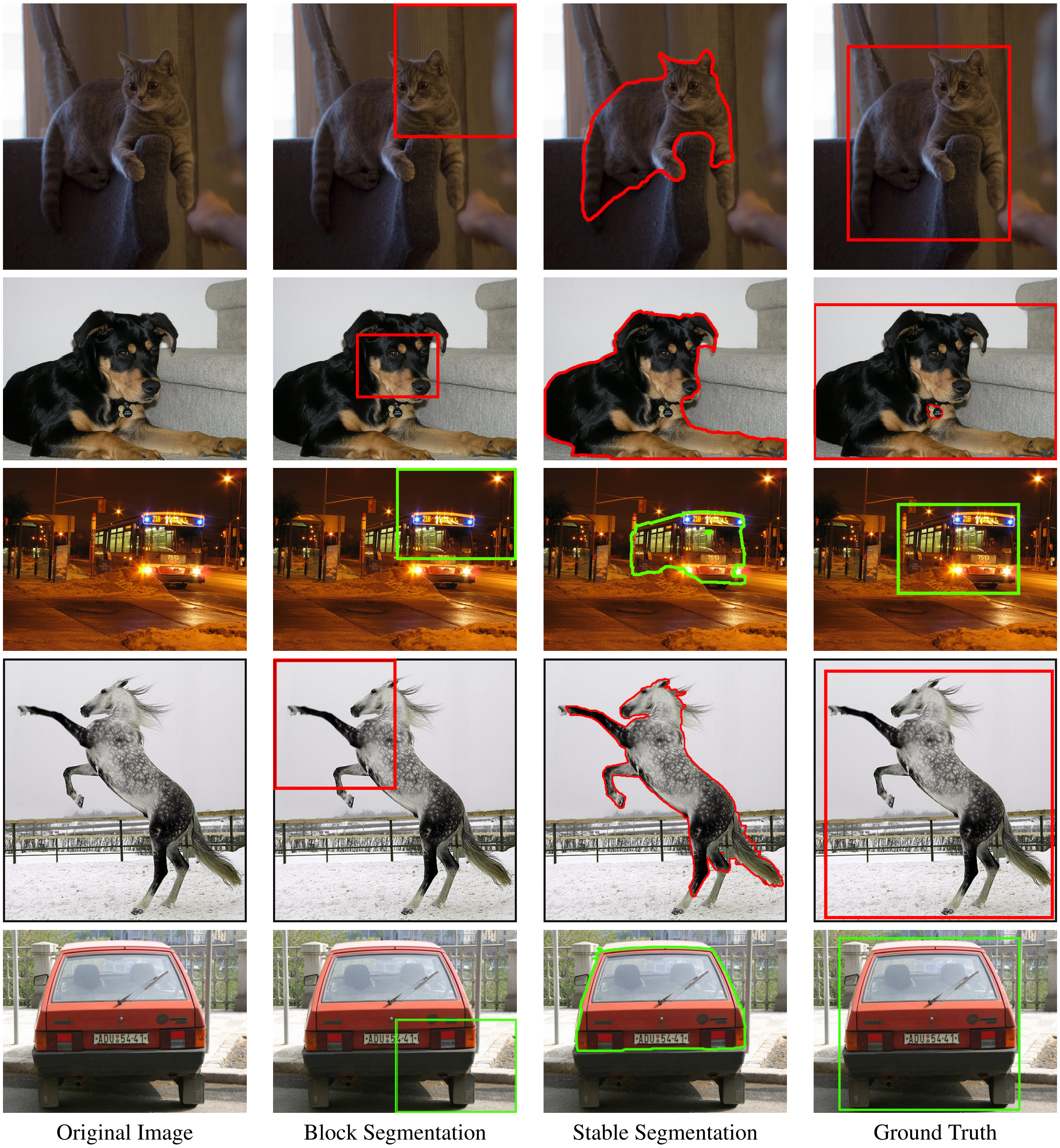


Figure 5. Object localization using "Stable Segmentations" as pre processing for the BoF categorization model. Examples from the PASCAL dataset. *Please view in color.*

categorization have not been convincingly shown. In this work we demonstrated that image segmentation can in fact improve object recognition and categorization and it also adds object localization and multi-class categorization capabilities to an off-the-shelf categorization system.

Often segmentation has not been used in recognition because of the difficulty of obtaining segments corresponding to the objects of interest. In this work we solve this problem by relying on a shortlist of potentially meaningful segmentations (identified by a stability criterion) which significantly increase the chance of extracting suitable segments. Incorporating this segmentation method with a simple BoF model was shown to bring the recognition accuracy to a level comparable with the state-of-the-art (Table 1, [31]).

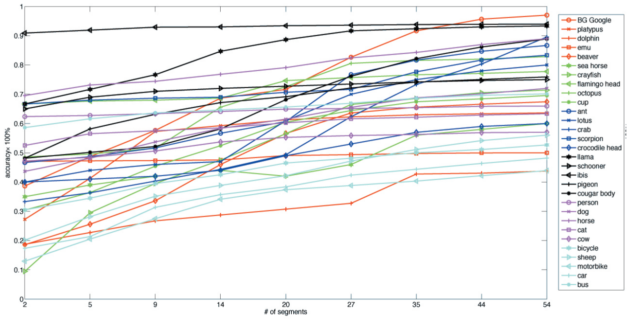


Figure 3. Object recognition vs. length of segmentation shortlist. Only Stable Segmentations results are shown. Note the general trend of accuracy improvement as the number of segments increases. The accuracy improvement saturates at around 35 segments.

To summarize, the effects of image segmentation on object categorization are the following:

▶ Segmenting an image does improve object categorization accuracy.

▶ Increasing the number of segments increases categorization accuracy.

▶ Increasing the quality of the segmentation improves object categorization accuracy.

▶ It is possible to achieve good localization and multi-class recognition performance using image segmentation.

We found that the quality of image segmentation does affect the average categorization accuracy for the BoF model. However, even the most trivial spatial grouping of interest points (i.e., Bseg) in the BoF model increases the categorization accuracy (but not as much as for Sseg). Localization is greatly affected by the segmentation quality as well. The number of segmentations/segments and their quality also critically impacts the categorization accuracy.

In ongoing work we are considering databases with many more categories. We are also exploring alternative recognition models that may take advantage of multiple-segment representations more explicitly, e.g. [12]. Finally, the proposed approach of segmenting test images and recognizing individual segments, provides an intuitive framework for semantic context based object categorization, as explored in [24].

Acknowledgements

This work was funded in part by NSF Career Grant #0448615, the Alfred P. Sloan Research Fellowship, NSF IGERT Grant DGE-0333451, AFOSR FA9550-06-1-0138 and ONR 67F-1080868.

References

[1] K. Barnard, P. Duygulu, R. Guru, P. Gabbur, and D. Forsyth. The effects of segmentation and feature choice in a translation model of object recognition. *CVPR*, 2003.

[2] F. Benezit, T. Cour, and J. Shi. Spectral segmentation with multi-scale graph decomposition. In *CVPR*, 2005.

[3] E. Borenstein and S. Ullman. Class-specific, top-down segmentation. *European Conference on Computer Vision*, 2, 2002.

[4] T. Cour, F. Benezit, and J. Shi. http://www.seas.upenn.edu/timothee/software/ncut_multiscale/ncut_multiscale.html.

[5] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. *Workshop on Statistical Learning in Computer Vision, ECCV*, 2004.

[6] M. Everingham et al. The 2005 pascal visual object classes challenge. In *In Proc. of PASCAL Challenge Workshop, LNAI,*, 2006.

[7] R. Fergus, P. Perona, and A. Zisserman. Object Class Recognition by Unsupervised Scale-Invariant Learning. *CVPR*, 2003.

[8] H. Hughes, G. Nozawa, and F. Kittler. Global precedence, spatial frequence channels, and the statistics of naturals scenes. *J. of Cog. Neuroscience*, 8(3):197–230, May 1996.

[9] T. Lange, V. Roth, M. Braun, and J. Buhmann. Stability-based validation of clustering solutions. In *NIPS*, 2002.

[10] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition*, 2006.

[11] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. *Workshop on Statistical Learning in Computer Vision, ECCV*, 2004.

[12] A. Levin and Y. Weiss. Learning to Combine Bottom-Up and Top-Down Segmentation. *ECCV*, 2006.

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

[14] J. Malik, S. Belongie, J. Shi, and T. Leung. Textons, contours and regions: Cue integration in image segmentation. In *Proc. 7th Int'l. Conf. Computer Vision*, pages 918–925, 1999.

[15] D. Martin, C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *PAMI*, 26(5):530–549, May 2004.

[16] K. Mikolajczyk, B. Leibe, and B. Schiele. Multiple object class detection with a generative model. *CVPR*, 2006.

[17] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering human body configurations: combining segmentation and recognition. In *CVPR*, 2004.

[18] D. Navon. Forest before trees: The precedence of global features in visual perception. *Congitive Psychology*, 1977.

[19] A. Y. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, 2002.

[20] M. Nilsback, A. Zisserman, M. Vision, and M. Kumar. A visual vocabulary for flower classification. *CVPR*, 2006.

[21] D. Nister and H. Stewenius. Scalable Recognition with a Vocabulary Tree. In *CVPR*, 2006.

[22] E. Nowak, F. Jurie, and B. Triggs. Sampling Strategies for Bag-of-Features Image Classification. *LNCS*, 2006.

[23] A. Rabinovich, T. Lange, J. Buhmann, and S. Belongie. Model order selection and cue combination for image segmentation. In *CVPR*, 2006.

[24] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *International Conference on Computer Vision*, 2007.

[25] V. Roth and B. Ommer. Exploiting low-level image segmentation for object recognition. In *DAGM*, 2006.

[26] B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. *CVPR*, 2006.

[27] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):888–905, August 2000.

[28] A. Vedaldi. http://vision.ucla.edu/vedaldi/code/bag/bag.html.

[29] P. Viola and M. Jones. Robust real time object detection. In *Second International Workshop on and Computational Theories of Vision*, 2001.

[30] S. Yu and J. Shi. Object-specific figure-ground segregation. *CVPR*, 2003.

[31] H. Zhang, A. C. Berg, M. Maire, and J. Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *CVPR*, 2006.