A BENCHMARK FOR SEMANTIC IMAGE SEGMENTATION

Hui Li[1], Jianfei Cai[1], Thi Nhat Anh Nguyen[2], Jianmin Zheng[1]
[1]Nanyang Technological University, Singapore, [2]Danang University of Technology, Vietnam

ABSTRACT

Though quite a few image segmentation benchmark datasets have been constructed, there is no suitable benchmark for semantic image segmentation. In this paper, we construct a benchmark for such a purpose, where the ground-truths are generated by leveraging the existing fine granular ground-truths in Berkeley Segmentation Dataset (BSD) as well as using an interactive segmentation tool for new images. We also propose a percept-tree-based region merging strategy for dynamically adapting the ground-truth for evaluating test segmentation. Moreover, we propose a new evaluation metric that is easy to understand and compute, and does not require boundary matching. Experimental results show that, compared with BSD, the generated ground-truth dataset is more suitable for evaluating semantic image segmentation, and the conducted user study demonstrates that the proposed evaluation metric matches user ranking very well.

***Index Terms*—** Benchmark, Evaluation, Semantic Image Segmentation, Dataset

1. INTRODUCTION

Semantic image segmentation refers to the task of segmenting an image into a set of non-overlapped meaningful regions corresponding to objects or parts of the objects which can deliver semantics or high-level structure information. A good semantic image segmentation can benefit many other computer vision tasks and multimedia applications such as object recognition, content-based image indexing, summary and retrieval, and image editing.

Although there is no universal answer on what a good semantic image segmentation should be since the concept of "semantic" is subjective and content-dependent, there are some general criteria. First, a good semantic segmentation should be able to achieve high similarity within segments and low association across the segments. Second, the segmentation boundary should match human perception. Third, semantic segmentation should reflect significant features and small-scale fluctuation should be ignored according to the part salience theory [1]. In other words, semantic image segmentation should decompose an image into a small set of meaningful regions (different from over-segmentation or clustering pixels), each of which is of considerable size. Fig. 1 shows a few examples of semantic image segmentation.

Fig. 1. Examples of semantic image segmentation. Each nature image is followed by a few semantic segmentations at different levels. In general, each image is segmented into a small set of meaningful segments with considerable sizes.

With more and more semantic image segmentation algorithms being developed in the past few years, there is a need to construct a benchmark to evaluate the performance of different algorithms. Although there are already several image segmentation benchmark datasets available, they are not suitable for evaluating semantic image segmentation. This is because the existing datasets are either of fine granularity such as the Berkeley Segmentation Dataset (BSD) [2] (often having 10 - 30 segments) widely used for boundary detection, or having 1-2 objects such as in [3, 4] for object cutout, or for some particular object classes such as in [5, 6], which cannot be directly used for evaluating the considered general semantic image segmentation that typically has less than 10 meaningful segments.

On the other hand, a good benchmark should not only have a representative dataset but also have an effective and efficient evaluation metric. In general, the existing evaluation metrics can be classified into three categories: region-based, boundary-based and hybrid-based. The popular region-based metric in [2] is tolerant to refinement, but is insensitive to boundary complexity and requires similar number of segments in both test segmentation and ground-truth. The widely used boundary-based metric in [7] is sensitive to boundary complexity, but requires a complex math model to match the boundaries in test segmentation and ground-truth.

Hybrid-based metric [8] combines the region and boundary measurements through a weighted summation, which also introduces the additional issue of how to set the combination weight.

In this paper, we develop a benchmark for semantic image segmentation. Particularly, we construct a new dateset that is suitable for testing and evaluating semantic segmentation. Unlike BSD, which costs 30 people 8 months to generate the ground-truths by manual labelling, we construct our ground-truth dataset by making use of the existing fine granular ground-truths in BSD as well as generating ground-truth for new images via an interactive segmentation tool that supports unlimited refinements. We also propose a percept-tree-based region merging strategy that allows to dynamically adapt the stored ground-truth so as to provide the most suitable ground-truth for the input test segmentation. Moreover, we propose a new evaluation metric that is easy to understand and compute, and does not require boundary matching. Experimental results show that, compared with BSD, the generated ground-truth dataset is more suitable for evaluating semantic image segmentation, and the user study demonstrates that the proposed evaluation metric matches user ranking very well.

2. RELATED WORK

2.1. Review of image segmentation datasets

Several image segmentation datasets have been constructed for testing and evaluating different algorithms [9, 5, 3, 6, 4, 2]. In particular, the dataset of image segmentation with a bounding box prior contains 50 images with only one salient object in each image and it targets for foreground extraction. PASCAL VOC 2009 dataset [5] is for recognizing objects from a number of visual object classes in realistic scenes. It has 20 object classes and totally 14743 images. With object segmentation ground-truths included, PASCAL VOC 2009 dataset can also be used to test and evaluate object-level segmentations. CMU-Cornell iCoseg dataset [3] is a recent co-segmentation dataset with 38 groups and totally 643 images. Each group contains a common foreground and each image has a ground-truth that separates foreground and background. Weizmann horses dataset [6] contains 328 images of horses. To avoid potential ambiguities when evaluating different segmentation algorithms, the segmentation evaluation dataset [4] selects 200 gray level image with only 1 or 2 objects in each of the images. BSD [2] is a widely used image segmentation benchmark which includes 500 images with 5 to 10 manually labelled ground-truths for each image. BSD is targeted for boundary detection and general image segmentation.

2.2. Review of evaluation metrics

Here we briefly review the two widely used segmentation evaluation metrics. The first one is a region-based metric called consistency error proposed in [2] which evaluates the consistency of segment interiors. One property of this metric is that it is tolerant to refinement. For example, suppose we have two segmentations S_1 and S_2 for an image. Let $R(S, p)$ denote the region in segmentation S that contains pixel p. If $R(S_1, p_i)$ is a proper subset of $R(S_2, p_i)$, then p_i lies in a region of refinement and there is no local error; otherwise, the local error is non-zero. Two different ways named Global Consistency Error (GCE) and Local Consistency Error (LCE) are introduced to combine individual local errors into an overall error measurement for the entire image. Both GCE and LCE are symmetric and tolerant to refinement. The problem with this metric is that it is meaningless when the number of segments in the test segmentation is quite different from that of the ground-truth. For example, the segmentation with each pixel being a segment is a refinement of any segmentation, which results in an overall zero error.

The second metric is a boundary based metric called boundary matching proposed in [7]. Boundary matching evaluates a segmentation algorithm by matching the boundaries of the test segmentation to the boundaries of ground-truths, and then sums the matching quality. It first converts the boundary matching problem into a minimum cost bipartite assignment problem. After solving the problem, by using a particular threshold as localization error, boundary pixels are classified into hits, misses, and false positives. With these numbers, precision and recall for a single segmentation can be computed as a summary of the matching quality. The advantages of this metric is that it is sensitive to boundary complexity and does not need to match interior pixels.

3. GROUND-TRUTH GENERATION

Before constructing the benchmark dataset, we need to address two issues: what type of ground-truths is needed and how to generate the ground-truths. For the first issue, as aforementioned, the semantic image segmentation is expected to segment an image into a small number of meaningful regions with considerable size. With such a definition, different people are still likely to draw different semantic segmentation results for the same image since the semantic interpretations are different (see Fig. 1 for example). One way to deal with this is to store multiple ground-truths for one image, just like that in the Berkeley segmentation dataset, which however misses the connections among different ground-truths and is also storage-inefficient.

In this research, we use percept tree to represent the object hierarchy in an image. In a percept tree, each node represents an object and a child node is a sub-segment of its parent node. Nodes at the same level of the percept tree should have equal importance. Fig. 2 gives an example of the percept tree, where the image is first segmented into the foreground and the background in level 1, the foreground is then further segmented into left man and right man in level 2, and at the

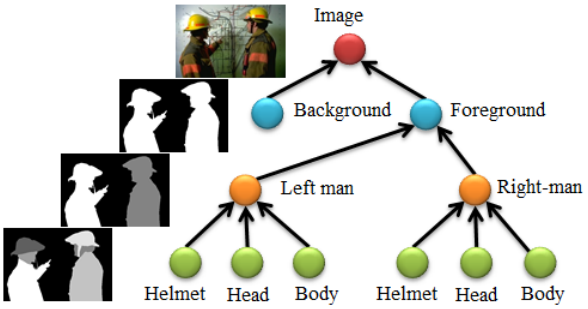
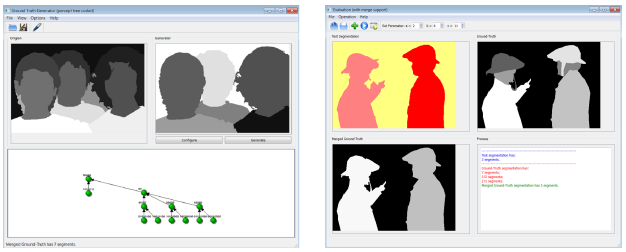


Fig. 2. An example of percept tree.

last level, each person is further segmented into three objects: helmet, head and body. The number of levels is small since our semantic image segmentation requires small number of segments with considerable sizes. Based on the percept tree, we propose to use the segmentation corresponding to the leaf-level of the percept tree as the ground-truth for our semantic image segmentation benchmark. Moreover, we embed the percept tree into the group-truth map by using different color codes for different nodes at the same level and using parent's color codes as a suffix for its children's color codes. In this way, one ground-truth map can generate multiple ground-truths corresponding to different levels of the percept tree through region merging, which will be discussed in Section 4.1.

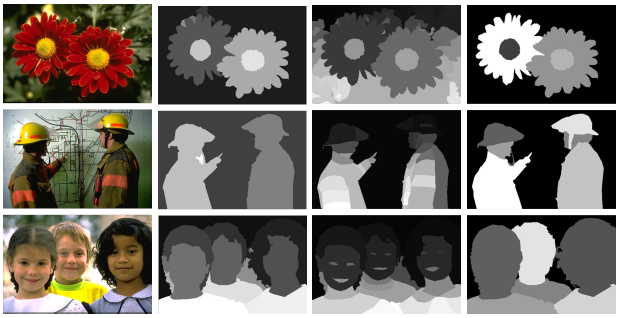
For the second issue on how to generate ground-truths, a common approach is through manual labelling to reach pixel-level segmentation accuracy, which is very tedious and time-consuming. Here, we consider two methods to generate the ground-truths. One method is to leverage the existing interactive segmentation tools such as [10], which can achieve high segmentation accuracy with a small amount of user effort. The other method is to adapt the ground-truths available in the existing segmentation datasets such as BSD that have been produced by laborious manual labelling. These two methods are complement to each other, where the former is to generate ground-truths for new images and the latter is for the images in the existing datasets. The combination of these two methods makes the construction of our semantic image segmentation benchmark faster and easily extendable.

In particular, we make use of the fine granularity ground-truths available in BSD. A C++ software is developed to merge the BSD ground-truths into those suitable for evaluating semantic segmentations as well as embedding the percept tree information into each ground-truth. Fig. 3(a) shows a snapshot of the software, where the upper left window, upper right window and the lower window are the original ground-truth from BSD with 16 segments, adapted ground-truth for our benchmark with 7 segments and the percept tree, respectively. Fig. 4 illustrates a few generated ground-truths by adapting those in BSD.



(a) Ground-truth generator (b) Region merging

Fig. 3. (a) Software snapshot for generating the semantic ground truths from BSD. (b) Dynamic region merging for segmentation evaluation.



(a) Images (b) BSD (c) BSD (d) Ours

Fig. 4. Generating the semantic ground-truths from the BSD ground-truths. (a) original images; (b) corresponding BSD ground-truth with the least number of segments; (c) corresponding BSD ground-truth with the most number of segments; (d) adapted ground-truths for evaluating semantic image segmentation.

4. PROPOSED EVALUATION METHODOLOGY

Based on the generated ground-truths, in this section we describe our proposed evaluation methodology, which takes two segmentations, i.e. test segmentation and ground-truth, as the input and produces a score indicating how good the test segmentation result is. Our evaluation methodology consists of two steps: region merging and metric computation.

4.1. Region merging strategy

The first step in our proposed evaluation is to dynamically merge regions in the ground-truth so as to provide the most suitable ground-truth for the input test segmentation. Recall that for each image we only store one ground-truth map/image at the leaf-level of the percept tree. However, the input test segmentation results by different semantic segmentation algorithms could be quite diverse in terms of the number of segments. It is hard to have a fair evaluation if the numbers of segments in the test segmentation and the ground-truth are quite different. Thus, we propose to merge

regions according to the percept tree so as to generate the most suitable ground truth that has the closest number of segments compared to the test segmentation.

This region merging step follows the rule that nodes at the same level of the percept tree are either all merged or merged none since all the segments at the same level have approximately equal importance. Fig. 3(b) is a snapshot taken during the evaluating process which shows the merged ground-truth (bottom-left) automatically generated from the ground-truth (up-right) for evaluating test segmentation (up-left).

4.2. Evaluation Metric

Our proposed evaluation metric belongs to the type of boundary-based metrics, which compare boundaries between the test segmentation and the ground-truth. Unlike the state-of-the-art boundary metric proposed in [7], which seeks the optimal boundary matching, here we propose an intuitive and simple boundary metric that is easy to understand and compute.

Our basic idea is to check each boundary pair in a segmentation against that in the other segmentation, reward the matching cases and punish the mismatching cases. Mathematically, we define the evaluation function as

$$E = \sum_{i} \sum_{j \in N_k(i)} W_{ij}(p_i \otimes p_j) \tag{1}$$

with:

$$p_i \otimes p_j = \begin{cases} 1, & R(S_1, p_i) \neq R(S_1, p_j) \\ 0, & otherwise \end{cases}$$

$$W_{ij} = \begin{cases} W_{ij}^{+}, & R(S_2, p_i) \neq R(S_2, p_j) \\ W_{ij}^{-}, & otherwise. \end{cases}$$

$p_i \otimes p_j$ can be considered as an XOR operation which returns 1 when pixel i and its k-ring neighbor pixel j do not belong to the same segment in the segmentation S_1, i.e. $R(S_1, p_i) \neq R(S_1, p_j)$, and returns 0 otherwise. Eq. (1) enforces that a reward weight W_{ij}^{+} (e.g. 1) is given to a boundary pair, i.e. $p_i \otimes p_j = 1$, if the corresponding pair in the other segmentation S_2 also belongs to different segments, i.e. $R(S_2, p_i) \neq R(S_2, p_j)$; otherwise, a punish weight W_{ij}^{-} (e.g. -1) is given. A higher E value means a better matching between the two segmentations.

With S_1 being the test segmentation and S_2 being the ground-truth, (1) becomes computing precision P, which measures how accurate the boundaries of the test segmentation are. On the other hand, with S_1 being the ground-truth and S_2 being the test segmentation, (1) becomes computing recall R, which measures how many ground-truth boundaries are correctly labelled. Following [7], we use F_α-measure to combine the precision and the recall into one score:

$$F_\alpha = \frac{\hat{P} \cdot \hat{R}}{(1-\alpha)\hat{R} + \alpha\hat{P}} \tag{2}$$

(a) Test segmentation (b) Ground-truth segmentation

Fig. 5. Segmentation evaluation for a pixel, where the black rectangle represents the checking window (1-ring neighborhood here), solid lines indicate reward and dashed lines indicate punish.

where $\hat{P}$ and $\hat{R}$ are normalized P and R values, and α is the tradeoff factor. Since P and R are defined in the same way and of the same importance, we set $\alpha = 0.5$ and use it for all the experiments.

Fig. 5 illustrates the segmentation evaluation of individual pixels. Particularly, at each location, we evaluate those pixels within a checking window that belong to a segment different from that of the center pixel. Since the computation is only needed for boundary pairs, the computation cost is low.

4.3. Weight Function

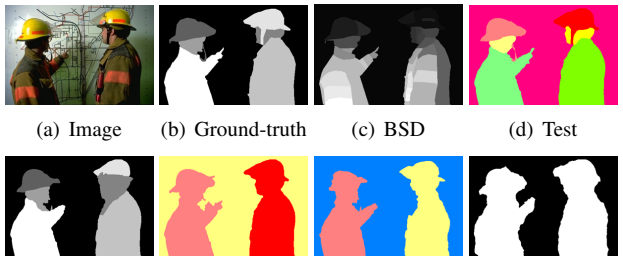
In this subsection, we discuss how to design the weight function W_{ij}. One intuitive way is to set the reward/punish weights to be crisp values (e.g. 1 and -1). However, such a setting does not distinguish different test segmentations with different distances away from the ground-truth boundaries. For example, for two test segmentations where one is close to the ground-truth boundary and the other is far away, the intuitive setting might lead to similar evaluation scores, which is not reasonable. In addition, considering that the boundaries in digital images are imperfect due to digitization and a boundary with 1 or 2 pixels away from the ground-truth might not be perceived as a difference, emphasizing perfect boundary alignment by crisp values is not necessary.

Inspired by the Fuzzy-evaluation function proposed in [8], we define our reward and punish functions as:

$$W_{ij}^{+} = \begin{cases} 1, & 0 \leq d(i,j) < a \\ 1 - 0.5(\frac{d-a}{b-a})^2, & a \leq d(i,j) < b \\ 0.5(\frac{c-d}{c-b})^2, & b \leq d(i,j) < c \\ 0, & c \leq d(i,j) \end{cases} \tag{3}$$

$$W_{ij}^{-} = \begin{cases} 0, & 0 \leq d(i,j) < a \\ -0.5(\frac{d-a}{b-a})^2, & a \leq d(i,j) < b \\ 0.5(\frac{c-d}{c-b})^2 - 1, & b \leq d(i,j) < c \\ -1, & c \leq d(i,j) < c+a \\ 0, & c+a \leq d(i,j) \end{cases} \tag{4}$$

where a, b, c are parameters for localization error, controlling the function slope, and dense evaluation complexity, respec-



(a) Image (b) Ground-truth (c) BSD (d) Test

(e) Test (f) Test (g) Test (h) Test

Fig. 6. Examples of a few test segmentations. (c)-(h): different test segmentations.

Table 1. Evaluations of the test segmentations in Fig. 6.

Tests	Precision	Recall	F-measure
(c)	0.43	0.77	0.55
(d)	0.80	0.78	0.79
(e)	0.59	0.54	0.57
(f)	0.79	0.78	0.79
(g)	0.50	0.48	0.49
(h)	0.38	0.35	0.36

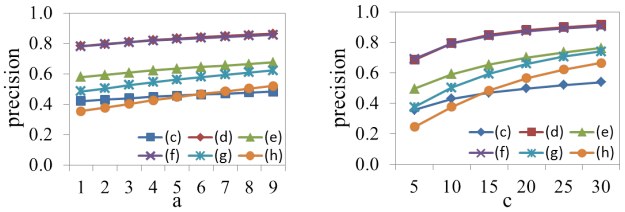
tively, and $d(i,j)$ is the Euclidean distance between p_i and p_j.

5. EXPERIMENTAL RESULTS

In this section, we evaluate the quality of the generated benchmark ground-truths and the effectiveness of the proposed quantitative metric. We choose 100 images from the BSD, which contain unambiguous objects in human vision perception. The original ground-truths in BSD are adapted according to the percept tree to generate the wanted ground-truths for evaluating semantic image segmentation, as described in Section 3. For the proposed evaluation metric, unless it is specified, the default values for parameters a, b, c are set to $a = 2$, $b = (c + a)/2$ and $c = 10$.

5.1. An example

Fig. 6 gives an example of a natural image with our ground-truth, one original BSD ground-truth, and several test segmentations, which are generated by based on the interactive segmentation method in [10] with different user strokes. Table 1 lists the corresponding quantitative evaluation results of precision P, recall R and F-measure F_α with respect to our generated ground-truth. In particular, Fig. 6(c) is a case that one original BSD ground-truth is taken as a test segmentation, which is considered as an over-segmentation. Thus, (c) has a higher recall value than its precision. Fig. 6(d) and (e) are the test results with seven segments, where (d) has better visual quality than (e). Correspondingly, the scores of (d) in Table 1



(a) with $c = 10$ (b) with $a = 2$

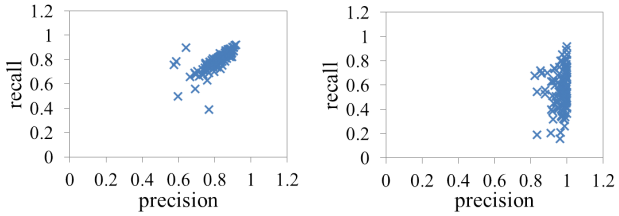
Fig. 7. Evaluation results under different parameter values for the test segmentations in Fig. 6.

are higher than those of (e). Fig. 6(f) and (g) are the test segmentations with three segments, and (h) is the one with two segments. All these results indicate that our quantitative evaluation metric matches the relative visual quality well. In addition, comparing (d) and (f), our metric gives similar scores although their numbers of segments are quite different. This is because of the region merging operation described in Section 4.1. Giving similar scores to (d) and (f) also matches the human perception since both segmentations are good in their respective semantic levels.

We also use Fig. 6 as an example to study the impact of the parameter variation on the evaluation results. Fig.7 shows the precision vs. a and precision vs. c curves. It can be seen that the precision scores increase with the increase of parameter a. This is understandable since relatively larger a means more local errors can be tolerated. From this point of view, our benchmark is able to control the degree of concerns on the segmentation accuracy. Similarly, the precision scores increase as parameter c increases since relatively larger c corresponds to slower reward decreasing and slower punish increasing in the weight functions. Note that larger c also results in higher computational complexity since the checking window becomes larger. Thus, c can be used to control the complexity and the differentiation ability of our proposed evaluation metric.

5.2. Ground-truth comparison

Here, we compare our generated benchmark with the Berkeley Benchmark for semantic image segmentation. Specifically, for each of the 100 test images, we generate one good-quality semantic image segmentation result based on the interactive segmentation tool in [10] with some local refinements. Then, we quantitatively evaluate these 100 test segmentations with reference to our generated ground-truths as well as the original BSD ground-truths. Fig. 8 shows the results of recall versus precision under different benchmarks. It can be seen that Berkeley benchmark gives high precision for most of the test segmentations but with generally low recall values. This is because many BSD ground-truths are of fine granularity and are generated for boundary detection purpose. On the contrary, our benchmark provides fairly



(a) Using our ground-truth (b) Using BSD ground-truth

Fig. 8. Evaluation results by using different benchmarks.

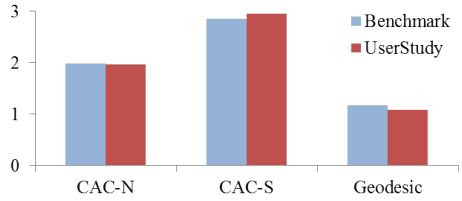


Fig. 9. Comparisons of the average ranking scores between the user study and our benchmark evaluation.

balanced precision and recall values, concentrated at the high end. This comparison demonstrates that our benchmark is more suitable for evaluating semantic segmentations.

5.3. User study

In order to show our quantitative evaluation results match human vision perception, a user study is conducted among 14 subjects. Among these subjects, 7 of them have no/little experience on image processing, 4 of them have experience on image processing but not on image segmentation, and 3 of them have research experience on image segmentation.

Multi-label segmentation tools based on Geodesic [11], Convex Active Contour (CAC) [10] are used to generate 3 groups of test segmentations: Geodesic, CAC with normal mode(CAC-N), CAC with smooth mode(CAC-S). The difference between the normal mode and the smooth mode is the setting of the smooth parameter which controls the smoothness of the image segmentation.

A user study software is implemented which randomly picks 1 image along with its 3 test segmentations. Subjects are asked to give scores (1 to 3, the higher the better) to each test segmentation. The scores are saved into a log file for analyzing. Scores are averaged among 14 log files for each algorithm over all the images. The average ranking scores of our benchmark are also generated based on our objective evaluation metric. Fig. 9 shows the average ranking scores of the three groups of the test segmentations under both user study and our benchmark evaluation. It demonstrates that our benchmark evaluation matches the subjective user ranking very well.

6. CONCLUSION

The contributions of this paper are twofold. First, we have built a benchmark dataset for semantic image segmentation. This is done by utilizing existing ground-truths and an interactive segmentation tool, which makes the construction faster and easily extendable. Our generated ground-truth is also embedded with the concept of percept tree, which makes the ground-truths adaptable via simple region merging. Second, we have proposed a new metric that is simple and intuitive to evaluate semantic image segmentation, and matches human perception well. We will release our benchmark to the public and we believe such a benchmark will greatly benefit the community to develop better semantic image segmentation.

Acknowledgements:

This research is supported by MoE AcRF Tier-1 Grant RG 30/11, Singapore.

7. REFERENCES

[1] D. D. Hoffman and M. Singh, "Salience of visual parts," *Cognition*, vol. 63, no. 1, pp. 29 – 78, 1997.

[2] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *ICCV*, July 2001, vol. 2, pp. 416 – 423.

[3] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen., "icoseg: Interactive co-segmentation with intelligent scribble guidance," in *CVPR*, 2010, pp. 3169 – 3176.

[4] S. Alpert, M. Galun, R. Basri, and A. Brandt, "Image segmentation by probabilistic bottom-up aggregation and cue integration," in *CVPR*, June 2007, pp. 1 – 8.

[5] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results," http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html.

[6] E. Borenstein and S. Ullman, "Class-specific, top-down segmentation," in *ECCV*, 2002, pp. 109 – 122.

[7] D. Martin, C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color,and texture cues," *IEEE Trans. PAMI*, vol. 26, no. 5, pp. 530 – 549, May 2004.

[8] X. Jin and C. H. Davis, "A genetic image segmentation algorithm with a fuzzy-based evaluation function," May 2003, vol. 2, pp. 938 – 943, ICFS.

[9] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr, "Interactive image segmentation using an adaptive gmmrf model," in *ECCV*, 2004, pp. 428 – 441.

[10] T. Nguyen, J. Cai, J. Zhang, and J. Zheng, "Robust interactive image segmentation using convex active contours," in *IEEE Trans. on Image Processing*, 2012, vol. 21, pp. 3734 – 3743.

[11] X. Bai and G. Sapiro, "A geodesic framework for fast interactive image and video segmentation and matting," in *ICCV*, 2007, pp. 1 – 8.