

Image-Segmentation Evaluation From the Perspective of Salient Object Extraction

Feng Ge, Song Wang, and Tiecheng Liu
Department of Computer Science and Engineering
University of South Carolina, Columbia, SC 29208
{gef, songwang, tiecheng}@cse.sc.edu

Abstract

Image segmentation and its performance evaluation are very difficult but important problems in computer vision. A major challenge in segmentation evaluation comes from the fundamental conflict between generality and objectivity: For general-purpose segmentation, the ground truth and segmentation accuracy may not be well defined, while embedding the evaluation in a specific application, the evaluation results may not be extended to other applications. We present in this paper a new benchmark for evaluating image segmentation. Specifically, we formulate image segmentation as identifying the single most perceptually salient structure from an image. We collect a large variety of test images that conforms to this specific formulation, construct unambiguous ground truth for each image, and define a reliable way to measure the segmentation accuracy. We then present two special strategies to further address two important issues: (a) the most salient structures in some real images may not be unique or unambiguously defined, and (b) many available image-segmentation methods are not developed to directly extract a single salient structure. Finally, we apply this benchmark to evaluate and compare the performance of several state-of-the-art image-segmentation methods, including the normalized-cut method, the level-set method, the efficient graph-based method, the mean-shift method, and the ratio-contour method.

1. Introduction

As a central step in computer vision, image segmentation has been extensively investigated in the past decades with the development of a large number of image-segmentation methods [1, 2, 3, 10, 11, 15, 16, 17, 19, 20]. However, general-purpose image segmentation is still an unsolved problem. We still lack reliable ways in performance evaluation for quantitatively positioning the state of the art of image segmentation. In early days, segmentation performance

is usually evaluated by subjectively judging on several sample images. Such subjective evaluation on a small data set lacks both generality and objectivity [7, 14, 18, 22, 23, 24]. To address this problem, it has been widely agreed that a benchmark, which includes a large set of test images and some objective performance measures, is necessary for image segmentation evaluation.

Unfortunately, benchmark-based segmentation evaluation [6, 13] usually suffers from a well-known dilemma between objectivity and generality. On the one hand, the test images are expected to have a large variety so that the evaluation results reflect the real performance on various applications. However, for such a general-purpose segmentation, the ground truth may not be well defined and therefore, it is difficult to define an objective segmentation-performance measure. On the other hand, by embedding the segmentation evaluation in a certain class of images and/or in a specific application, the evaluation results may not be useful for more general applications, although the well-defined ground truth and segmentation-performance measures are available.

One important prior work on the segmentation benchmark is Berkeley benchmark presented by Martin et al. [13]. This benchmark contains more than 1000 various natural images, each of which is manually processed by a group of people to get the ground-truth segmentations. Such manual segmentations reflect the general human perception and therefore, different people may partition the images into different number of segments. Although this benchmark achieves good generality, it may have some problems on the evaluation objectivity. Given non-unique ground truths, this benchmark uses a global consistency error (GCE) and a local consistency error (LCE) to measure the segmentation accuracy. These two measures tolerate unreasonable refinement of the ground truth, i.e., if the segmentation is a refined version of the ground truth, or vice versa, the segmentation error is always zero. Therefore, trivial segmentations, where each segment only contains one pixel or the whole image is a single segment, always produce “perfect”

100% segmentation accuracy in this benchmark. In the recent work by Estrada and Jepson [6] and Martin [12], the precision-recall figures are used to better reflect the trade-offs between the segmentation performance and the number of regions.

The goal of this paper is to develop a new image-segmentation benchmark by seeking a balance between the objectivity and the generality in evaluating image segmentation. Particularly, *we formulate image segmentation as extracting the single most salient structure in the image*. In this formulation, the ground truth is a segment with a closed boundary, thus image segmentation is reduced to an image-bipartitioning or a boundary-detection problem. Similar to Berkeley benchmark, the saliency of the identified structure may come from a combination of various cues, such as intensity, texture, shape, size, or familiarity, and we do not put any explicit bias on any one of them in constructing the ground truth. By treating the salient structure as the foreground *figure*, and the remaining portion as the *background*, such a formulation is usually referred to as *figure-ground* segmentation in prior literatures. For convenience, we will continue to use this terminology in this paper. However, it must be emphasized that the “background” in our test images has a more general meaning than a trivial segment of homogenous intensity, random noise, or uniform texture as assumed in most prior literatures. Actually, the background segment may contain many other salient structures that are not as salient as the one shown by the foreground segment.

The following considerations make this benchmark more appropriate for image-segmentation evaluation. First, the figure-ground segmentation is a much better defined problem than the general-purpose segmentation. As shown in Fig. 1, for the figure-ground segmentation, human perception can produce an unambiguous ground truth on many natural images. By collecting only such images into the benchmark, we can construct a unique unambiguous ground-truth segmentation for each image. Second, in the figure-ground segmentation, an image is always partitioned into two segments. This property, together with the uniquely-defined ground truth, facilitates the definition of segmentation-performance measures that are more robust and objective than the GCE and LCE measures adopted in Berkeley benchmark. Third, the figure-ground segmentation can be regarded as a special case of the general-purpose segmentation. Therefore, given the fact that there are still no good solutions to figure-ground segmentation, we believe it may be too early to evaluate the general-purpose segmentation where the unique ground truth is not available. Finally, the performance on the figure-ground segmentation, to some extent, reflects the performance on the general-purpose segmentation, because, if one method can perform well in segmenting the most salient structure from an image, then it can be applied iteratively to the same image to extract multi-

ple structures to accomplish the general-purpose image segmentation.

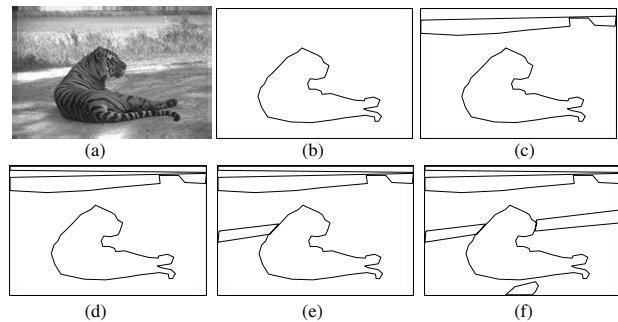


Figure 1. The figure-ground segmentation is usually better defined than the general-purpose segmentation: (a) a sample image. (b) The unambiguous ground truth in the figure-ground segmentation. (c-f) Four different ground-truth segmentations produced by different people in the general-purpose segmentation.

We need to address two important problems in applying this benchmark to evaluate various image-segmentation methods. First, most available image-segmentation methods are not specifically designed to produce a figure-ground segmentation. For example, many of them partition the input image into a set of disjoint segments, from which the foreground and background are usually not specified. Second, many real images contain multiple salient structures and in some images, and the most salient structure may not be unambiguous defined. Although we intentionally collect only the images with unambiguous most salient structures to construct the proposed benchmark, we expect that the images with multiple choices of salient structures can also be included and evaluated. In this paper, we introduce two special strategies in Section 3 to address these two important problems.

2. Benchmark Construction

In this section, we introduce the two basic components of the proposed segmentation benchmark: test-image database and the segmentation-performance measure.

2.1. Test-Image Database

As the first stage of the benchmark construction, we collected 1023 real natural images from internet, digital photos, and some well-known image databases. We carefully examined each image before including it into the database. A particular requirement is that each image contains a single most salient foreground structure that is unambiguous in human visual perception. This way, the ground-truth segmentation can be easily constructed by manually extracting the closed boundary of this salient structure. To make this benchmark suitable for evaluating a large variety of image-segmentation methods, color information is removed, and

all the images are unified to 256-bit gray-scale images in PGM format, with a size in the range of 80×80 to 200×200 .

We hired two computer-science undergraduate students to build this test-image database. They use the following strategy to decide whether to include an image into the database. First, both of them look at the considered image and select the most salient structure independently. Second, if both of them select the same structure without any reservation, this image will be included into the database. Otherwise, if they choose different salient structures or any one of them has reservations in determining the most salient structure, this image will not be included. After one image is decided to be included into the database, they work together to construct a single ground-truth segmentation by extracting the closed boundary of the identified salient structure.

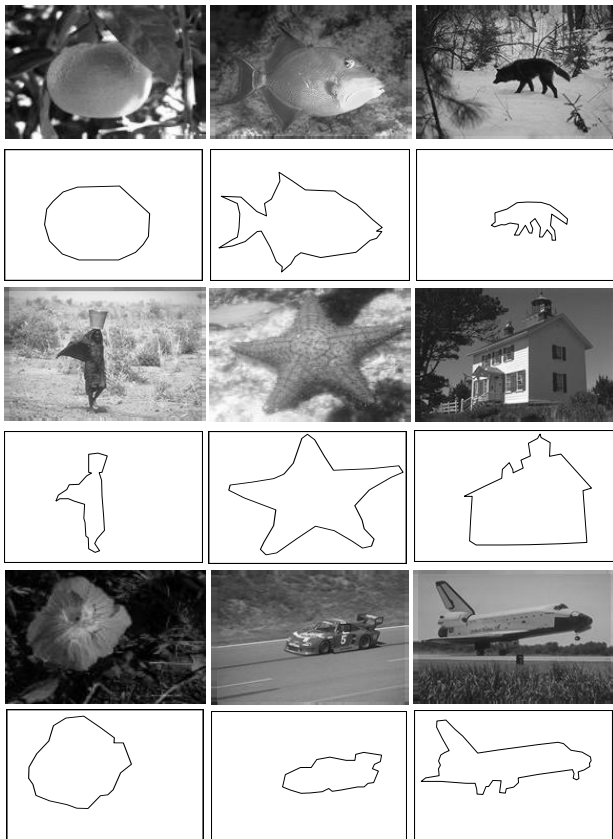


Figure 2. Nine sample images in our image database and the ground truth produced manually.

Figure 2 demonstrates several sample images and their ground-truth segmentations in the current image database. Note that we intentionally collect images with various foreground structures, such as human, animal, vehicle, building, etc., and various backgrounds. Also note that, in the collected images, the most salient structure may not be the only structure in the image, and the background may contain some structures that are not as perceptually salient as

the foreground one. Certainly, the decision made by these two students may not always be psychophysically consistent with other people, i.e., some collected images, when presented to other viewers, may still result in a different foreground structure. In Section 3, we will develop a special strategy to handle this problem. With this special strategy, an image with multiple salient structures, from which the most salient one may not be unambiguously defined, can still be evaluated and the only requirement is that one of the salient structures is labelled as the ground truth. We believe the ground truths constructed by these two students will satisfy this lower requirement.

2.2. Performance Measure

Under the figure-ground assumption, we expect an image-segmentation method to produce only two segments: one for the foreground and the other for the background. While the ground-truth segmentation is also a figure-ground bipartition of this image, the segmentation performance can be measured by the coincidence between the segmentation result and the ground truth. The basic performance measure we implement for this benchmark is based on the region coincidence. Let the region A be the ground-truth foreground structure and the segment R be the detected foreground structure, we define the region-based segmentation accuracy as

$$P(R; A) = \frac{|A \cap R|}{|A \cup R|} = \frac{|A \cap R|}{|A| + |R| - |A \cap R|}, \quad (1)$$

where $|\cdot|$ is the operation of computing the segment area. Different from the region-coincidence-based GCE and LCE measures used in Berkeley benchmark, this measure has no bias to the segmentations that produces overly large or small number of segments. The numerator, $|A \cap R|$, measures how much the ground-truth structure is detected. The denominator, $|A \cup R|$, is a normalization factor which normalizes the accuracy measure to the range of $[0, 1]$. With this normalization factor, the accuracy measure penalizes the error of detecting irrelevant regions as the foreground segments (false positives). It is easy to see that this region-based measure is insensitive to small variations in the ground-truth construction and incorporates the accuracy and recall measurement into one unified function.

3. Segmentation-Evaluation Strategies

Based on the above benchmark, we evaluate the following six image-segmentation methods: (1) **Normalized-cut method (NC)** [19] implemented by Shi and Malik [5]; (2) **Efficient graph-based method (EG)** [16] implemented by Felzenszwalb and Huttenlocher [9]; (3) **Mean-shift method (MS)** [3] implemented by Comaniciu and Meer [4]; (4) **Level-set method (LS)** [17] implemented by

Fan [8]; (5) **Ratio-contour method (RC)** [20] implemented by Wang et al. [21]; (6) A trivial uniform-partition method (UP) without considering any image features.

We choose the non-trivial methods of NC, EG, MS, LS, and RC based on three considerations: (a) They well represent different categories of image-segmentation methods, (b) all of them are relatively new methods and/or implementations that well represent the current state of the art of the general-purpose image segmentation, and (c) the softwares of these five methods are publicly available. The trivial method of UP is also included because we want to investigate to what extent the state-of-the-art image-segmentation methods, such as NC, EG, MS, LS, and RC, outperform the trivial segmentation method.

To evaluate segmentation using this benchmark, the most desirable form of segmentation output is certainly a figure-ground-style segmentation, i.e., the image is partitioned into two segments with one as the foreground and the other as the background. In this case, we can directly measure the segmentation accuracy using Eq. (1). However, different image-segmentation methods generate segmentation results in different forms. Among the above six selected methods, LS and RC produce figure-ground segmentations and therefore, we may directly apply Eq. (1) to evaluate their performance on each image. Note that the foreground segment produced by LS may not be a single connected one, because LS does not preserve the foreground topology. NC, EG, MS, and UP partition an image into a set of disjoint segments without labelling the foreground and background. Consequently, we need to develop new strategies so that they can be fairly and convincingly evaluated in the benchmark.

3.1. Strategy 1 for Estimating an Upper-Bound Performance

Many image-segmentation methods can produce a small set of candidate foreground segments R_1, R_2, \dots, R_k , which may or may not overlap with each other. For example, we can repeat RC on the input image to obtain multiple candidates of the most salient structure. Since we do not know which one of them should be used to evaluate against ground truth, we define the segmentation accuracy by

$$P_1(R_1, R_2, \dots, R_k; A) = \max \left\{ \frac{|A \cap R_i|}{|A \cup R_i|}, i = 1, 2, \dots, k \right\} \quad (2)$$

where A is the ground-truth foreground segment. We can see that Eq. (2) is essentially an upper-bound performance because it can be achieved only when an ideal postprocessing step helps pick the best candidate from R_1, R_2, \dots, R_k . In real application, more cues or knowledge of the desirable structure may or may not be available to implement

this postprocessing. Note that the identified candidate foreground segments R_1, R_2, \dots, R_k may overlap with each other.

This strategy is particularly useful in addressing an important problem mentioned in Section 1: Many real images contain multiple salient structures in which the most salient one may not be unambiguously defined from the human perception. Using this strategy, we can still include such images into the database and simply label one salient structure to construct the ground truth. In evaluating an image-segmentation method, we can iteratively apply it to produce a set of candidates of the foreground structure. Specifically, we can repeat the segmentation process more times than the number of salient structures in the images, then we evaluate the performance of each candidate using this strategy and find the best-performed candidate. The basic assumption underlying this evaluation strategy is that a good segmentation method should be able to detect a specified salient structure in an image after a finite number of iterations even if this image contains multiple salient structures.

3.2. Strategy 2 for Estimating an Upper-Bound Performance

Many available image-segmentation methods, such as NC, EG, MS, and UP, partition an image into a set of disjoint segments $\{R_1, R_2, \dots, R_n\}$, where $R_i \cap R_j = \emptyset$ for $i \neq j$, and $\cup_{i=1}^n R_i = \Omega$, the whole image. In this case, the ground-truth foreground segment usually corresponds to a subset of these disjoint segments. To evaluate these methods in our benchmark, we find a best possible combination of a subset of the segments $R_i, i = 1, 2, \dots, n$ to construct a figure-ground segmentation. In this paper, we simply count a segment R_i into the foreground R if it has more than 50 percent overlap with the ground-truth foreground A in terms of the area, i.e.,

$$R = \bigcup_{i: \rho(R_i, A) > 0.5} R_i.$$

where

$$\rho(R_i, A) = \max \left\{ \frac{|R_i \cap A|}{|R_i|}, \frac{|R_i \cap A|}{|A|} \right\}.$$

From the resulting foreground R , we define the performance measure $P_2(R_1, R_2, \dots, R_n; A)$ as $P(R; A)$ given in Eq. (1). An example of performance evaluation using this strategy is illustrated in Fig. 3.

We can see that, just like Strategy 1, Strategy 2 also provides an upper-bound of the segmentation performance by assuming a postprocessing step of identifying and merging some segments to form the foreground. Note that these two upper-bound performances may not be achieved or even approached in real applications, where the ground truth is not

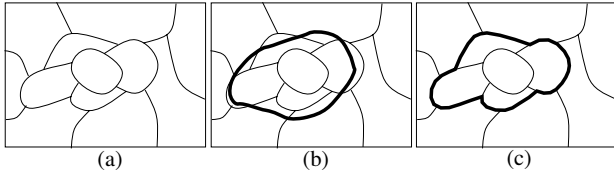


Figure 3. An illustration of Strategy 2: (a) an image-segmentation result; (b) the boundary of the ground-truth segmentation (the thick curve) overlapped on the segmentation result; (c) the figure-ground segmentation (the thick curve) derived using Strategy 2.

a priori known. However, the postprocessing steps required by these two strategies are of different complexity. Picking the best candidate, required in Strategy 1, is essentially a verification problem with linear number of choices, while region merging required in Strategy 2 has an exponential number of choices. Therefore, we believe the upper-bound performance obtained using Strategy 1 is more likely to be achieved than the one obtained using Strategy 2. For example, the upper-bound performance calculated using Strategy 2 is useful only when the total number of segments, n , is small. Considering the extreme case that each pixel is partitioned as a segment, the upper-bound performance obtained using Strategy 2 is a meaningless value of 100 percent, which is similar to the GCE and LCE measures developed in Berkeley benchmark. But the difference is that GCE and LCE also result in meaningless high accuracy when too fewer segments are produced, such as the case where the whole image is partitioned as a single segment. In this paper, we always set the segmentation parameters to produce a reasonably small number of segments when applying these two strategies.

4. Evaluation Results

4.1. Performance Comparison

We first compare the average performance of different image-segmentation methods. To make a fair comparison, we measure the average performance when 1023 testing images are segmented into the same number of segments. Table 1 shows the average upper-bound performance \bar{P}_2 (using Strategy 2) of NC, EG, MS, and UP in terms of the number of produced segments. For EG and MS, the number of produced segments are controlled by the parameter S , the minimum allowed segment area. Therefore, we continuously vary S to achieve segmentation with different number of segments. The average upper-bound performance \bar{P}_1 (using Strategy 1) of RC in terms of the number of iterations is shown in Table 2. For LS, the implementation used in our evaluation has many parameters and it is very difficult to achieve multiple candidate foreground structures or a specified number of disjoint segments. Therefore, the default parameters of the LS implementation [8] are used in evalu-

ation, and the average performance of LS is 0.33. For NC, EG, and MS, their implementations contain other parameters besides the number of regions. In our evaluation, we fix the number of regions but vary the other parameters to get the best performance on each image.

| # segments | NC | EG | MS | UP |
|------------|------|------|------|------|
| 2 | 0.39 | 0.28 | 0.32 | 0.27 |
| 5 | 0.58 | 0.52 | 0.46 | 0.30 |
| 10 | 0.70 | 0.65 | 0.57 | 0.42 |
| 20 | 0.78 | 0.76 | 0.66 | 0.54 |
| 40 | 0.82 | 0.83 | 0.71 | 0.64 |
| 80 | 0.85 | 0.87 | 0.73 | 0.72 |
| 160 | 0.88 | 0.89 | 0.76 | 0.79 |
| 320 | 0.89 | 0.89 | 0.77 | 0.84 |

Table 1. Comparison of the average upper-bound performance \bar{P}_2 of NC, EG, MS, and UP.

| | | | | | | |
|-------------|------|------|------|------|------|------|
| Iterations | 1 | 2 | 3 | 4 | 5 | 6 |
| Performance | 0.4 | 0.50 | 0.53 | 0.55 | 0.57 | 0.58 |
| Iterations | 7 | 8 | 10 | 12 | 16 | 20 |
| Performance | 0.59 | 0.59 | 0.60 | 0.61 | 0.62 | 0.62 |

Table 2. The average upper-bound performance \bar{P}_1 of RC in terms of the number of iterations.

From Tables 1 and 2, we can see that the average upper-bound performance of the three non-trivial methods — NC, EG, and MS — are better than that of the trivial UP, when a relatively small number of segments are produced. When an image is segmented into more than 80 segments, the average upper-bound performance of the NC, EG, and MS methods are close to that of UP. In addition, with the increase of image segments, the upper-bound performance becomes much more difficult to reach through region merging or best-candidate selection. From this perspective, the upper-bound performance derived from over-segmentation (≥ 160 segments) is largely meaningless.

Table 1 also suggests the choices of the segmentation parameters. It shows that, for the collected 1023 images, NC, EG, and MS all get close to their respective limits of the upper-bound performance when images are segmented into around 80 segments, i.e., segmenting images into more than 80 segments cannot noticeably increase the upper-bound performance any more even using the proposed two strategies. With less than 80 segments, we can also see both NC, EG, and MS have an upper-bound performance better than UP. Particularly, around 40 segments are expected to be the target for NC, EG, and MS. This table shows that the appropriate range of the number of resulting segments is 10 – 80.

From Table 1, we can surely draw the conclusion that even this simplified figure-ground segmentation is still far

from being solved with the state-of-the-art segmentation methods. Note that the performances in Table 1 are still some kind of upper bounds that are usually difficult to reach in real applications. Also be reminded that these 1023 images are carefully examined beforehand so that human visual system is able to unambiguously extract the single ground-truth foreground structure. We believe that, only after knowing how to solve this simplified yet well-defined figure-ground segmentation, can we make real progress on general-purpose image segmentation, where the ground truth may not be well defined. Additionally, an effective figure-ground segmentation method itself can facilitate many important computer-vision applications, such as content-based image retrieval.

To compare the relative performance of different image segmentation methods, we also count the number of images on which one method outperforms the others. For example, if NC achieves the best performance on an image I_j among all the methods, we consider NC the winner on I_j . We then count the number of winning images of each segmentation method and show the result in Table 3. Note that Strategy 2 is used for NC, EG, UP and MS and Strategy 1 is used for RC. In this table, RC is compared to the other five by setting different RC iterations. As discussed before, we only include LS for comparison when two segments are produced.

| K | NC | EG | MS | UP | RC | LS |
|--------|-----|-----|-----|----|-----|-----|
| 2(1) | 197 | 86 | 144 | 42 | 394 | 160 |
| 5(2) | 285 | 232 | 194 | 56 | 256 | NA |
| 10(4) | 361 | 258 | 218 | 38 | 148 | NA |
| 20(8) | 336 | 351 | 242 | 15 | 79 | NA |
| 40(16) | 264 | 446 | 265 | 6 | 42 | NA |

Table 3. The number of winning times of each method. In the table, K is the number of produced segments for NC, EG, MS, and UP (or the number of iterations for RC and LS).

From Tables 1 and 3, we can also see that the trivial UP has the worse average upper-bound performance. When two segments are produced, RC wins more times than the other methods. When targeting for more than 20 segments, EG wins more times than the other methods. Since the average upper-bound performance of EG is very close to that of NC, it indicates that EG wins only marginally on most images. Basically, for the NC, EG, and MS methods, there is no strong evidence (based on Table 1 and Table 3) showing that one specific method is apparently superior to the others; in fact, their average (upper-bound) performances are very close.

Several other reasons prohibit us from ranking the five non-trivial segmentation methods: (a) Most performances listed here are estimations of upper bounds; whether we can

reach or approach the upper bounds largely depends on specific applications. (b) Many methods are not especially developed for figure-ground segmentation; their performance may still be significantly improved if they are tuned to the figure-ground segmentation. (c) We use two different strategies in comparing the upper-bound performance of different methods. These two strategy have different complexity. Strategy 1 used in RC is basically a *verification* process with linear-complexity search space. For the Strategy 2, a *region merging* postprocessing is required to find a figure-ground segmentation; clearly, the search space is of exponential size in terms of the number of segments, and therefore, the upper-bound performance obtained by Strategy 2 is usually more difficult to reach than the one obtained by Strategy 1.

4.2. Combination of Image Segmentation Methods

Besides evaluating and comparing the performance of individual image-segmentation method, it is also important to know whether and how these methods are statistically related. If these six methods can complement each other, then it would be worthwhile for researchers to further investigate ways to boost the performance by combining them. To better understand the correlation of these methods, we introduce a virtual method, which automatically picks the best method (out of the selected six ones) for each individual image. We name this virtual method as the combined method and its performance as the *combined performance*. This combined performance indicates the best performance we can get by “ideally” combining these six methods.

In combining the above six methods, we specify the the number of resulting segments for NC, EG, MS, and UP and the number of iterations for RC, as discussed above. For LS, we only consider the case where two segments are produced. In Figure 4, we illustrate the segmentation performance using the *cumulative-performance* histogram curve $p(x) : [0, 1] \rightarrow [0, 1]$ (or *performance curve* in short), which describes the (upper-bound) performance distribution on all 1023 images. A specific point $(x, p(x))$ on the curve indicates that $100 \cdot x$ percent of the images are segmented with an accuracy lower than $p(x)$. Equivalently, this also means that $100 \cdot (1 - x)$ percent of the images are segmented with an accuracy higher than $p(x)$. Clearly, the higher a performance curve in the Cartesian coordinate system, the better the performance of the corresponding segmentation method and of the parameters setting. Note that, the performance (or the accuracy) shown in this figure is the upper-bound performance estimated using the proposed two strategies when an image is segmented more than two segments.

Figure 4 (a-d) shows the upper-bound performance of the combined method when the resulting segments for NC, EG, MS, and UP are set to 2, 5, 10, and 20, respectively and the number of iterations for RC is set to 1, 2, 4, 8, corre-

spondingly. We can see that the performance of this “ideal” combined method is substantially better than that of each individual method when two segments are produced, but only slightly better when 20 segments are produced. The results in Figure 4 indicate that the combination of different image-segmentation methods cannot significantly boost the segmentation performance when images are over-segmented.

To summarize, we have the following observations and conclusions on the performance-evaluation experiments.

1. The figure-ground segmentation is still far from a solved problem with the state-of-the-art segmentation methods. When two segments are produced in each image, the performances of the five tested non-trivial methods are low and are very close.
2. When more than two segments are produced, the average (upper-bound) performances of the NC, EG, and MS are very similar. The performance differences among them are marginal and there is no obvious winner. All of them produce much better performance than the trivial UP when the number of produced segments in an image is no more than 80.
3. The experimental results provide useful information on selecting appropriate parameters for each method. For RC, more iterations can significantly improve the upper-bound performance, and 10 iterations is an appropriate setting. For NC, EG, and MS, the target number of segments should be in the range of 10 – 80, with 40 being an expected number.
4. When fewer segments are produced, a combination of different image segmentation methods may improve the performance. With the increase of the number of image segments, the performance gain resulted from combining these methods decreases. Particularly, when images are segmented into 20 or more regions, the performance of the combined method is only marginally better than that of the EG and NC.

5. Conclusions

In this paper, we presented a new benchmark for evaluating image segmentation. The major contribution is the construction of a benchmark for image segmentation evaluation with an assumption of figure-ground segmentation, i.e., identifying the single most perceptually salient structure from an image. We presented two special strategies to handle two important problems in using this benchmark: (a) the most salient structures in an image may not be unique or unambiguous; (b) many available image-segmentation methods do not directly produce figure-ground-style segmentations. In this benchmark, the image-segmentation problem

is better defined while the test-image collection and ground-truth construction can still be easily and unambiguously achieved. Currently, we have collected 1023 natural images for this benchmark. We applied this benchmark to evaluate the performance of five state-of-the-art image-segmentation methods. The results clearly show that this figure-ground segmentation problem is still far from well solved. While this figure-ground segmentation can be treated as a simplified yet better-defined case of the general-purpose segmentation, we expect that this benchmark sets a more realistic goal to image-segmentation researchers.

Acknowledgement This work was funded, in part, by NSF-EIA-0312861.

References

- [1] E. Borenstein and S. Ullman. Learning to segment. In *European Conference on Computer Vision*, pages 315–328, 2004.
- [2] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Color- and texture-based image segmentation using EM and its application to image querying and classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1026–1038, 2002.
- [3] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.
- [4] D. Comaniciu and P. Meer. Mean-shift image segmentation software. <http://www.caip.rutgers.edu/riul/research/code/EDISON/index.html>.
- [5] T. Cour, S. Yu, and J. Shi. Normalized cut image segmentation software. <http://www.cis.upenn.edu/~jshi/software/>.
- [6] F. J. Estrada and A. D. Jepson. Quantitative evaluation of a novel image segmentation algorithm. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 20–26, 2005.
- [7] M. Everingham, H. Muller, and B. Thomas. Evaluating image segmentation algorithms using the pareto front. In *European Conference on Computer Vision*, pages 34–48, 2002.
- [8] D. Fan. Level-set image segmentation software. <http://www.cs.wisc.edu/~fan/LevelSet/>.
- [9] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based segmentation software. <http://people.cs.uchicago.edu/~pff/segment/>.
- [10] Y. Gdalyahu, D. Weinshall, and M. Werman. Stochastic image segmentation by typical cuts. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 596–601, 1999.
- [11] R. M. Haralick and L. G. Shapiro. Survey: Image segmentation techniques. *Computer Vision Graphics Image Process*, 29:100–132, 1985.
- [12] D. Martin. An empirical approach to grouping and segmentation. Ph.D. Thesis, Berkeley University, 2003.
- [13] 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 *International Conference on Computer Vision*, volume 2, pages 416–425, 2001.

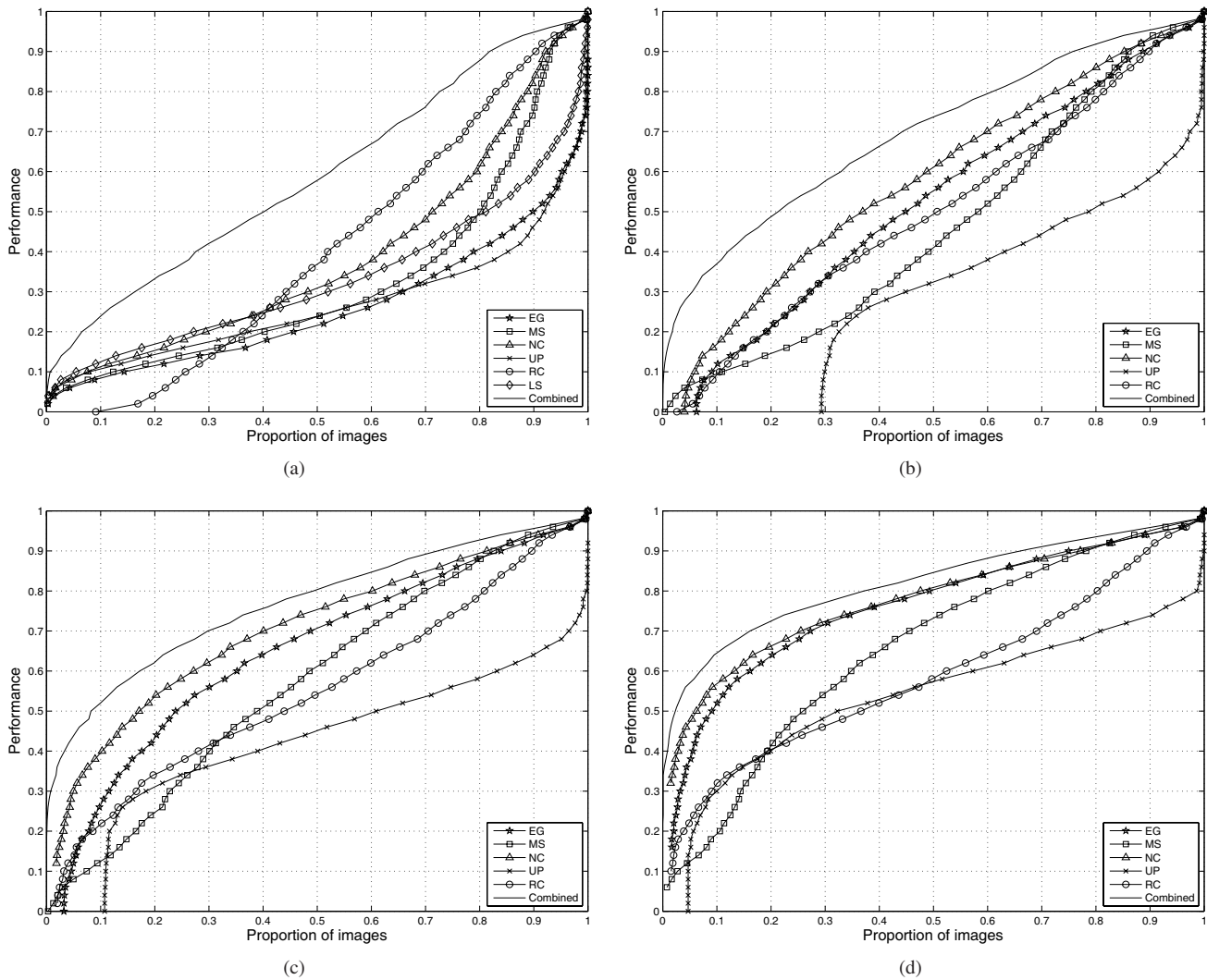


Figure 4. The performance curves of each segmentation method and of the combined method when the average number of produced segments is (a) 2; (b) 5; (c) 10; (d) 20.

[14] B. McCane. On the evaluation of image segmentation algorithms. In *Digital Image Computing: Techniques and Applications*, pages 455–461, 1997.

[15] H. Nguyen, M. Worring, and R. Boomgaard. Watersnakes: Energy-driven watershed segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(3):330–342, 2003.

[16] D. H. P. Felzenszwalb. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[17] J. Sethian. *Level Set Methods and Fast Marching Methods*. Cambridge, U.K.: Cambridge Univ. Press., 1999.

[18] C. W. Shaffrey, I. H. Jermyn, and N. G. Kingsbury. Psychovisual evaluation of image segmentation algorithms. In *Advanced Concepts for Intelligent Vision Systems*, pages 1–7, 2002.

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

[20] S. Wang, T. Kubota, J. Siskind, and J. Wang. Salient closed boundary extraction with ratio contour. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4):546–561, 2005.

[21] S. Wang, T. Kubota, and J. M. Siskind. Ratio-contour software. <http://www.cse.sc.edu/~songwang/>.

[22] L. Yang, F. Albrechtsen, T. Lonnestad, and P. Grottum. Psychovisual evaluation of image segmentation algorithms. In *Lecture Notes in Computer Science*, pages 759–765, 1995.

[23] H. Zhang, J. E. Fritts, and S. A. Goldman. An entropy-based objective segmentation evaluation method for image segmentation. In *SPIE Storage and Retrieval Methods and Applications for Multimedia*, pages 38–49, 2004.

[24] Y. Zhang. A survey of evaluation methods for image segmentation. *Pattern Recognition*, 29(8):1335–1346, 1996.