Multi-Scale Improves Boundary Detection in Natural Images

Xiaofeng Ren

Intel Research Seattle
1100 NE 45th Street, 6th Floor, Seattle, WA 98105
xiaofeng.ren@intel.com

Abstract. In this work we empirically study the multi-scale boundary detection problem in natural images. We utilize local boundary cues including contrast, localization and relative contrast, and train a classifier to integrate them across scales. Our approach successfully combines strengths from both large-scale detection (robust but poor localization) and small-scale detection (detail-preserving but sensitive to clutter). We carry out quantitative evaluations on a variety of boundary and object datasets with human-marked groundtruth. We show that multi-scale boundary detection offers large improvements, ranging from 20% to 50%, over single-scale approaches. This is the first time that multi-scale is demonstrated to improve boundary detection on large datasets of natural images.

1 Introduction

Edge detection is a fundamental problem in computer vision that has been intensively studied in the past fifty years. Traditional approaches were built on the analysis of ideal edge models and white sensing noise. A variety of edge detectors were developed, mostly based on image gradients and Gaussian derivative filters, leading to the popular *Canny* edge detector [1] that we still use today.

Edges, like any other image structures, are multi-scale in nature. Early work on multi-scale edge detection used Gaussian smoothing at multiple scales [2]. *Scale-Space* theory gradually emerged [3] and evolved into a field of its own [4, 5]. The Scale-Space theory states that, under a set of mild conditions, the Gaussian function is the unique kernel to generate multi-scale signals. The Scale-Space theory also provides guidelines on the selection and integration of signals across scales [6].

In practice, Gaussian-based edge detectors have considerable difficulty dealing with natural scenes, where idealized edge and noise models do not hold. To address the challenges of natural images, recent approaches have adopted a learning paradigm: large datasets of natural images have been collected and hand-labeled, such as the Berkeley Segmentation Dataset [7], providing both training data and evaluation benchmarks. Boundary detection is formulated as learning to classify salient boundaries against background [8]. State-of-the-art detectors combine local brightness, color and texture contrasts and have been shown to outperform traditional gradient-based approaches [9].

It would be a natural step to combine the strengths of learning-based boundary operators with the insights from classical multi-scale edge detection. Surprisingly, very few efforts have been devoted to this line of research. The analysis in [10] is based on gradient magnitudes only and does not report detection performance. The learning approach in [11] uses a large number of features without any specific discussion on multi-scale. The work of [12] focuses on finding elongated structures in medical images. No benchmarking results have been reported to show that multi-scale detection is better than single-scale approaches. On the other hand, the authors of [9] commented that they found *no* benefit using multi-scale signals.

Beyond local classification, multi-scale has also been extensively explored in mid-level grouping, especially in image segmentation (e.g.[13, 14]). The focus of multi-scale segmentation is typically on the pyramid representation of images, such as for Markov random fields, and on global inference and smoothing. More recently, multi-scale intensity cues have been successfully used (e.g. [15]) in segmenting natural images. However, segmentation algorithms typically do not produce boundary-based benchmarking results, partly because many of them focus on large-scale salient regions/boundaries and tend to ignore details in an image.

Is boundary detection in natural images so difficult that it eludes multi-scale processing? That would be hard to believe. Studies in natural image statistics have strongly suggested that scale-invariance or multi-scale structure is an intrinsic property of natural images (e.g. [16]). In particular, scaling phenomena have been observed in the statistics of boundary contours [17, 18]. This multi-scale nature must have important implications for boundary detection.

In Figure 1, we show several examples of multi-scale boundary contrast signals, by running the (publicly available) *Probability-of-Boundary* operator (Pb) [9] at several disk sizes. Just as we have learned from gradient-based multi-scale processing, signals at different scales exhibit different characteristics: at a large scale, edge detection is reliable, but its localization is poor and it misses small details; at a small scale, details are preserved, but detection suffers greatly from clutters in textured regions. No doubt there is information in multi-scale. The challenge is how to combine the strengths of small and large scales, so as to improve boundary detection performance, not on a single image, but on large collections of natural images.

In this work we empirically study the problem of multi-scale boundary detection in natural images. We explore a number of multi-scale cues, including boundary contrast, localization and relative contrast. We find that multi-scale processing significantly improves boundary detection. A linear classifier combining multi-scale signals outperforms most existing results on the Berkeley Segmentation Benchmark [7]. Extensive experimentation shows that the benefits of multi-scale processing are large and ubiquitous: we improve boundary detection performance by 20% to 50% on four other boundary and object recognition datasets.

Our empirical work is important and of sufficient interest because this is the first time, after over 20 years of active research, that multi-scale is shown to improve boundary detection on large collections of natural images. Our results are useful both in quantifying the benefits of doing multi-scale in local boundary classification, as well as in comparing to alternative approaches to boundary detection (such as mid-level grouping). We obtain these results with a simple algorithm, which can have immediate applications in edge-based object and scene recognition systems.

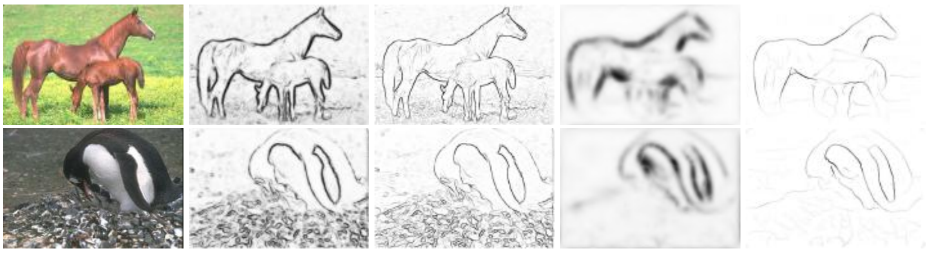


Fig. 1. We run the *Probability-of-Boundary* operator [9] to generate boundary contrast signals at multiple scales. Here we show both the (soft) contrast map and the boundary map (after non-maximum suppression), at a small scale and a large scale. Large-scale signals are reliable but poor in localization, and details are smoothed out. On the other hand, small-scale signals capture detailed structures but suffer from false positives in textured regions and clutter. The challenge is how to combine the strengths of them.

2 Multi-Scale Boundary Detection

A (single-scale) boundary detector finds edges by measuring contrast in a fixed local window. Multi-scale detection varies the scale of the window and combines signals from multiple scales. There are two issues in this process:

Correspondence/Tracking: how do signals across scales correspond to one another? If working with discrete edges (i.e. contrast peaks after non-maximum suppression), one would need to track edge locations across scales.

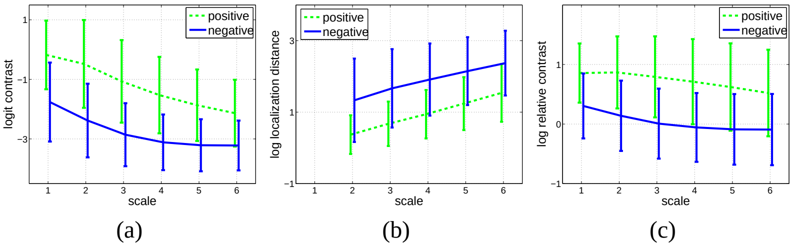
Cue combination: how does one integrate boundary signals from multiple scales?

Traditional approaches to multi-scale edge detection focus on the correspondence (or tracking) problem. Tracking can either be coarse-to-fine, such as in Bergholm's *edge focusing* strategy [19], or be fine-to-coarse, such as in Canny's *feature synthesis* [1]. A large number of multi-scale schemes have been proposed along these lines (see a survey in [20]). Most approaches take a simple view of cue combination: they either accept edges (after thresholding) at all scales, or accept edges that appear at the coarsest scale.

The cue combination problem can be easily accommodated and analyzed in the learning paradigm of boundary detection: cues from multiple scales are inputs to a binary classification problem, and we can maximize performance over the selection of cues and the choice of the classifier. Given the complexities of signals in natural images, the correspondence problem is non-trivial; yet we may quantitatively evaluate and choose between candidate strategies.

We base our analysis on the Berkeley Segmentation Dataset (BSDS) [7] and the *Probability-of-Boundary* (*Pb*) operator [9]. The BSDS collection includes 300 images of various natural scenes, each with multiple human-marked segmentations. The *Pb* operator has been shown to outperform other edge detectors on the BSDS images.

Pb measures local contrast by computing histogram differences between brightness, color and texture distributions in two half-disks. If we vary the radius of the half-disks, we obtain contrast signals at multiple scales. For each scale s, we keep two sets of



(a) (b) (c)

Fig. 2. Empirical distributions of boundary cues across scales. (a) Means and standard deviations of boundary contrast cues (E) at all 6 scales; in average, contrast is higher for positives or true boundaries. (b) Means and standard deviations of localization cues (D), distances to closest edge peaks; true boundaries are better localized. (c) Distributions of relative contrast (R); true boundaries are typically higher in contrast relative to its neighbors. We find all the cues informative. However, they are also noisy as shown by the large standard deviations.

data: $P^{(s)}_{soft}$, soft contrast at each pixel before non-maximum suppression, and $P^{(s)}_{peak}$, sparse/localized edges after non-maximum suppression.

Small-scale signals are typically better for localization. Hence when we set up the classification, we only consider locations that generates a maximal response at the smallest scale (where $P^{(1)}_{peak} > 0$). For each such location, we define a set of cues at multiple scales to classify boundary vs non-boundary. In this work we restrict ourselves to local boundary detection, i.e. making independent decisions at each location.

2.1 Multi-scale boundary cues

There are two perspectives of "scale" in boundary detection: (1) intuitively, edges have intrinsic scales; and (2) we use measurements at multiple scales to capture the intrinsic scales and to improve detection performance.

In the examples in Figure 1, the back of the horse is a typical large-scale edge, and the textured region under the penguin contain typical small-scale edges. A large-scale edge is much more likely to be an object boundary, and small-scale edges are mostly false positives. There are several cues that can help us make this distinction:

1. **contrast**: a large-scale edge has consistent contrast measurements at multiple scales. A small-scale edge would have high contrast at a small observation scale but lower contrasts at larger scales.
2. **localization**: if we look at the peak response (after non-maximum suppression), a large-scale edge tends to have a consistent location and does not shift much. Peak locations of small-scale edges become unreliable in large-scale measurements or disappear altogether.
3. **relative contrast**: also known as *contrast normalization*, the strength of an edge relative to its surroundings is a cue for boundary saliency. A weak contrast boundary may be salient, at a large observation scale, if other locations around it are significantly lower in contrast. Texture edges, though maybe high contrast, is not salient because many other edges nearby have comparable contrasts.

For **contrast**, we use the soft Pb contrast computed using multiple disk sizes, converted (back) to a linear scale:

$$E^{(s)} = \log \left[P^{(s)}_{soft} / (1 - P^{(s)}_{soft}) \right]$$

We also define a localization cue: we threshold [1] the peak signal $P^{(s)}_{peak}$ into a binary edge map, and compute its distance transform. The result is a distance $d^{(s)}$ from each pixel to the closest peak location, at each scale s. We define **localization** as:

$$D^{(s)} = \log \left(d^{(s)} + 1 \right)$$

Note that in defining contrast $E^{(s)}$, we have avoided the correspondence problem by using signals before non-maximum suppression. There are, of course, issues associated with this simple strategy. The soft contrasts at large scales are fairly blurry. That is, contrast generated by a single boundary may extend spatially and boost false positives in surrounding areas. The localization cue compensates this lack of correspondence: for an off-boundary point, even though the contrast may be high at a large scale, it is far away from peak locations and hence will be suppressed.

Finally, for contrast normalization, we compute average contrasts $P^{(s)}_{avg,L}$ and $P^{(s)}_{avg,R}$ in the “left” and “right” half disks around each point, and define **relative contrast** as:

$$R^{(s)} = \log \left[P^{(s)}_{soft} / \min \left(P^{(s)}_{avg,L}, P^{(s)}_{avg,R} \right) \right]$$

where “left” and ”right” disks[2] using the maximum *Pb* orientation.

2.2 Cue validation

Having defined multi-scale boundary cues, the first thing to ask is whether these features are informative for boundary detection. To answer this question empirically, we use the 200 training images in BSDS and compare distributions around on-boundary and off-boundary locations. In Figure 2 we visualize the empirical means and standard deviations of the distributions. Two observations can be made:

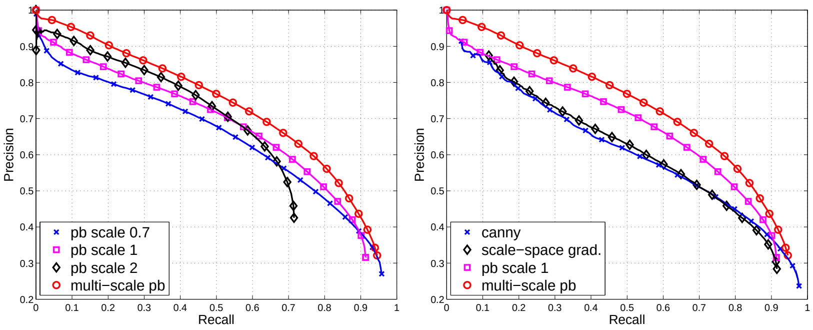
- The cues we have defined are indeed informative for boundary detection, as one would have expected from intuition. All the cues, at all scales, exhibit different distributions for positive and negative examples.
- The signals are nevertheless noisy. Individually these boundary cues are weak, as the standard deviations are large comparing to the differences between the means.

2.3 Cue combination

In the learning paradigm of boundary detection, finding boundaries is formulated as a classification between boundary and non-boundary pixels. The boundary classification

[1] We choose thresholds such that 95% of the groundtruth edges are kept.

[2] We set the scale to be 2.5 times the disk radius in *Pb*.



(a) (b)

Fig. 3. Precision-recall curves for performance evaluation. (a) Comparing to the standard (scale 1) *Pb* operator, large-scale *Pb* has a higher precision (at the low recall range), and small-scale *Pb* has a higher (asymptotic) recall. Our multi-scale approach combines the strengths of both, dominating any fixed scale *Pb*. (b) Comparing to gradient-based approaches: scale-space edge detection of Lindeberg [6] offers minor improvements over Canny [1]. Pb performs much better than gradient-based methods. Our multi-scale classifier offers a large and consistent improvement over Pb and achieves an F-measure of 0.68, outperforming existing approaches.

problem has the characteristics of having a large amount of training data, relatively low dimension, and poor separability (significant overlap between classes). Consistent with the observations in [9], we have found in our experiments that logistic regression, linearly combining cues across scales, performs as good as a fair number of other standard techniques. We choose logistic regression for its simplicity.

3 Experiments and Empirical Analysis

We carry out empirical analysis of our multi-scale boundary detection scheme using the Berkeley Segmentation Dataset, 200 training and 100 test images of resolution 480-by-320. We run the *Pb* operator at 6 scales, half-octave apart ($\sqrt{2}$), starting at one half-octave lower than the default *Pb* scale. We obtain positive training examples by running a minimum distance matching between groundtruth boundary pixels and edge pixels in the smallest scale (i.e. $P_{peak}^{(1)}$), with a distance threshold of 3 pixels. For negative examples, we use edges in $P_{peak}^{(1)}$ that are 20 pixels away from groundtruth boundaries.

Boundary detection performance is evaluated in the precision-recall framework of [9]. We use both the F-measure (harmonic mean of precision and recall) and average precision (area under a P/R curve) [21]. F-measure selects the best trade-off point of precision vs recall. Average precision summarizes performance using entire curves and is better at revealing differences in sub-ranges of recall.

3.1 Multi-scale improves boundary detection

In Figure 3(a), we show the precision-recall curve for our multi-scale detection, along with the performance of *Pb* at three scales (0.7, 1 and 2). The performance of *Pb* at multiple scales is exactly as expected: at a small scale, the precision is lower, but more details are recovered and the asymptotic recall is higher. At a large scale, the precision is higher in the beginning, showing salient boundaries being more reliably detected; however, the curve saturates at a lower recall. Our multi-scale approach combines the strengths of both small- and large-scale, producing a P/R curve that dominates that of *Pb* at any fixed scale. Our approach achieves the asymptotic recall rate of the smallest scale while maintaining high precision in the entire range. In particular, the performance in the mid-range (0.5-0.8 recall) is much higher than any of the fixed scale *Pb*.

In our evaluation we also include two gradient-based approaches: the Canny edge detector [1] which uses single-scale gradients, and the scale-space edge detector of Lindeberg [6] [3]. The scale (sigma) in Canny is set to be 0.0025. For scale-space edges we use 4 scales, initial search scale 1.5, and search grid spacing of 8,16 and 32 pixels.

In Figure 3(b), we show the precision-recall curves for all four approaches: Canny, scale-space edges, Pb, and our approach "Multi-Pb". We observe that in our setting, scale-space edge detection finds only minor improvements over Canny and drops below Canny at the end. In comparison, our multi-scale approach offers a large and consistent improvement over single-scale Pb [4]. The F-measure of our results is 0.68, higher than most existing results on the Berkeley benchmark [22, 23, 11, 24, 25, 12]. On the gray-scale version of the BSDS benchmark, our F-measure is 0.66.

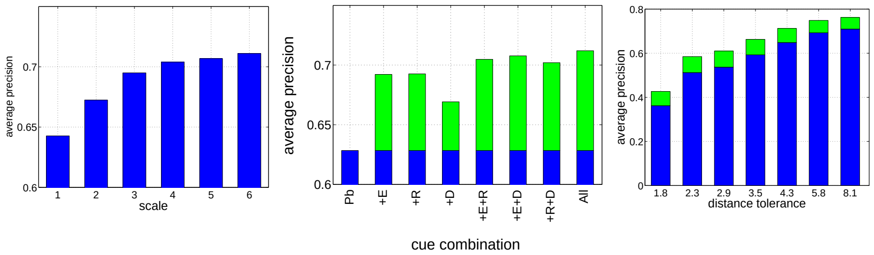
It is interesting to compare our results to that in [9], where the authors did not find much advantage of using multi-scale. The two approaches are similar in spirit; the main differences are: (1) they worked with smaller images (half resolution); (2) they used only three scales, half octave apart; (3) they did non-maximum suppression after combining the signals, where we use small-scale Pb only; and (4) we have used additional features such as relative contrast or localization. Table 1 shows an empirical analysis of the differences. We find that our success is a combination of these factors, each contributing to the final improvement. In particular, our non-maximum suppression scheme is better, as it preserves details and prevents them from being smoothed out by large-scale signals (which typically have higher weights).

Avg. Precision	Pb (single)	Pb 3 (soft)	Pb 3	Pb 6	Multi-Pb
full-res (480x320)	0.648	0.647	0.683	0.687	0.712
half-res (240x160)	0.643	0.641	0.666	0.677	0.683

Table 1. We compare the average precision of standard Pb and Multi-Pb (this work) with several variants: Pb combined at 3 scales, with non-maximum suppression applied after combination; Pb combined at 3 scales (with our non-maximum suppression scheme); and Pb combined at 6 scales. We show results on both the full-sized BSDS images (with 0.75% distance threshold) and half-sized ones (with 1% threshold). These results help explain why our conclusion of multi-scale here is positive while that in [9] was negative.

[3] We thank Mark Dow for his code at http://lcni.uoregon.edu/ mark/SS_Edges/SS_Edges.html.

[4] Note that the BSDS evaluation has changed considerably from that in [9]. In particular, the F-measure for Pb is 0.65 under the current implementation.



(a) (b) (c)

Fig. 4. Cue combination evaluated with average precision: (a) improvements in performance as we gradually add the number of scales; (b) improvements over default scale *Pb* by adding subsets of features, including contrast (E), localization (D), and relative contrast (R); (c) improvements over *Pb* with different choices of distance tolerance in evaluation.

3.2 Cue evaluation

We quantitatively measure individual contributions of the cues by running multi-scale detection with subsets of the features. All these experiments use logistic regression as the classifier. We use average precision for evaluation.

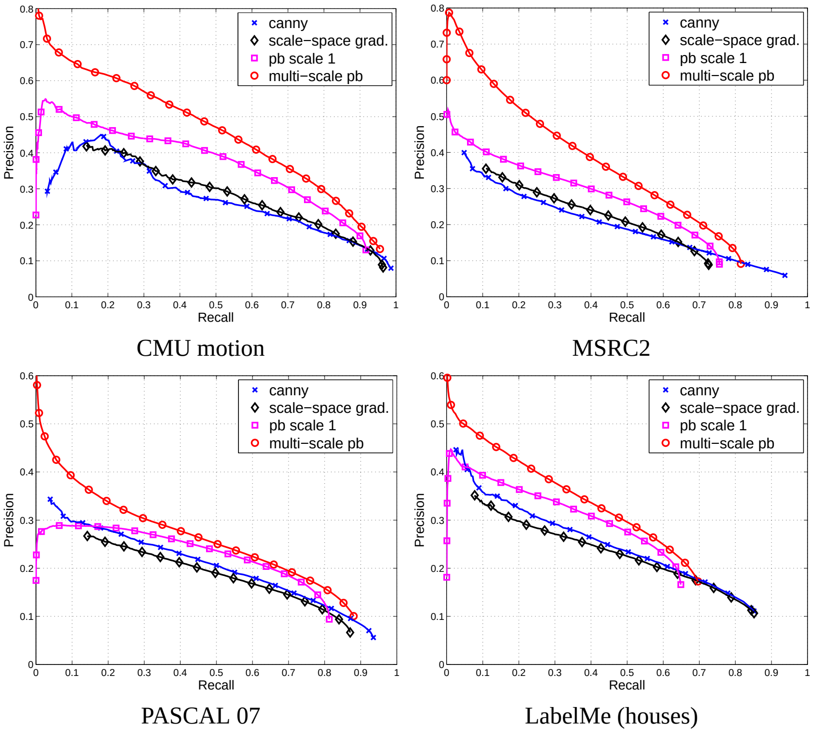
Figure 4(a) shows the average precision of detection when we vary the number of scales used. We start with a single scale (the smallest) and gradually expand to 6 scales. We observe that the improvement is much larger in the first few scales, showing diminishing returns. Nevertheless, the large scales still make contributions, possibly indicating the existence of large-scale boundaries in the dataset that are best captured at a large scale of observation.

In Figure 4(b), we evaluate the contributions of the three sets of cues: contrast (E), localization (D), and relative contrast (R). They are combined with the default (second smallest) scale *Pb* to show the improvements over *Pb*. Individually, contrast (E) and relative contrast (R) work better. However, there seems to be a fair amount of redundancy between contrast (E) and relative contrast (R). Localization (D) by itself does not work well; however, it improves performance when combined with contrast (E).

In the precision-recall evaluation, detections are matched to groundtruth boundaries with a distance threshold. The default in the BSDS benchmark is 0.75% of image diagonal, about 5 pixels. We vary this tolerance and show the results in Figure 4(c). It appears that the (absolute) improvement in average precision is about the same for all the choices of tolerance. Relative to *Pb*, the improvement is larger at small distance tolerances, indicating that the multi-scale approach is better at localizing boundaries.

3.3 Additional experiments

The empirical evaluations on the Berkeley Segmentation Dataset is very encouraging: we show large and consistent improvements using multi-scale cues. To further verify the benefits of multi-scale detection, we test our approach on four other large datasets with groundtruth segmentation: 30 images from the CMU motion boundary dataset [26], 519



CMU motion

MSRC2

PASCAL 07

LabelMe (houses)

Fig. 5. Precision-recall evaluation on four other datasets with groundtruth segmentation, covering a variety of objects and scenes. We use parameters trained from BSDS. Our multi-scale approach significantly improves over Pb in all four cases, in the entire range of recall.

images from the MSR Cambridge object dataset [27, 28] (MSRC2), 422 images from the segmentation competition in the PASCAL challenge 2007 [29], and 218 images from a subset (Boston houses 2005) of the LabelMe database [30].

These datasets add a lot of varieties to our empirical evaluation. They include both large vs. small objects, low- vs. high-resolution images, single-object photos vs. complex outdoor and indoor scenes, and detailed boundary labeling vs coarse polygonal contours. To show the robustness of our approach, in all these experiments we use the same parameters trained from BSDS.

Figure 5 shows precision-recall evaluations on the four datasets. The CMU motion dataset and the MSRC2 dataset contain large objects; we show P/R curves using distance tolerance 0.75% in benchmarking. The PASCAL and LabelMe datasets have high resolution and tend to have small objects in scenes; we show P/R curves using distance tolerance 0.6%. Table 2 lists average precisions for the five datasets at three thresholds.

These experiments show that our multi-scale approach is robust and offers large improvements over single-scale approaches. The amount of improvement differs, from about 20% (in average precision) in PASCAL 07 to about 45% in MSRC2. We see more improvements in CMU motion and MSRC2, probably because they both tend to have

Dist. Threshold	Th=0.6%				Th=0.75%			
Methods	Canny	S Grad.	Pb	Multi-Pb	Canny	S Grad.	Pb	Multi-Pb
BSDS (test)	0.554	0.527	0.593	0.663	**0.605**	**0.589**	**0.648**	**0.712**
CMU Motion	0.245	0.252	0.315	0.413	**0.271**	**0.287**	**0.350**	**0.448**
MSRC2	0.170	0.156	0.197	0.283	**0.193**	**0.182**	**0.228**	**0.325**
PASCAL 07	**0.197**	**0.173**	**0.196**	**0.242**	0.226	0.204	0.233	0.277
LabelMe (houses)	**0.223**	**0.206**	**0.211**	**0.251**	0.254	0.235	0.245	0.283

Table 2. Average precision evaluation on all five datasets, comparing four approaches at two distance thresholds in benchmarking. Our approach improves single-scale Pb by about 10% on BSDS, 20% on PASCAL and LabelMe, about 30% on CMU motion and 45% on MSRC2. Highlighted numbers correspond to the curves in Figure 5. Qualitatively there is little difference when we vary the distance threshold to other values.

Avg. Precision	BSDS	CMU Motion	MSRC2	PASCAL07	LabelMe
Multi-Pb	0.712	0.448	0.325	0.277	0.283
Multi-Pb (Pyramid)	0.707	0.443	0.323	0.275	0.295

Table 3. Multi-scale processing can be made efficient by working with sub-sampled images at large scales. Average precision hardly decreases when we use an image pyramid.

large objects. Our results on the CMU motion boundary dataset are comparable to what have been reported in [26], remarkable because we use no motion at all. The precisions on PASCAL and LabelMe images are lower for all approaches, probably because only a subset of objects and boundaries in these images are marked. Nevertheless, we still improve performance there proportionally.

We observe that improvements over Pb are most prominent in the low-recall/high-precision range. Similar phenomena have been found in many related studies on natural image boundary detection [22, 24, 25]. These approaches typically focus on the low-recall range and show little improvements near high-recall. In comparison, our approach offers consistent improvements for all recall.

3.4 Pyramid-based multi-scale processing

Our approach is based on the Pb operator, which computes histogram differences between two half disks. The computational cost increases linearly with disk area, and at large scales it becomes prohibitive for large images. A standard solution is to use a pyramid and work with sub-sampled images at large scales. We have tested this pyramid approach: at a large scale $s > 1$, we resize images by a factor of $1/\sqrt{s}$, hence keeping the cost constant across scales. An empirical comparison is shown in Table 3. As expected, we find there is little loss in performance.

4 Discussions

We have studied multi-scale boundary detection in the context of natural images. Conceptually our approach is straightforward: we compute contrast and localization cues for a number of scales, and use logistic regression to linearly combine them. Our multi-scale approach combines the strengths of both large-scale (high precision) and small-scale (high recall and good localization). Our approach outperforms most reported results on the Berkeley Segmentation Benchmark. Significant improvements (20% to 45%) have been demonstrated on four other boundary and object recognition datasets.

Our work has answered two important empirical questions on boundary detection:

Does multi-scale processing improve boundary detection in natural images? Intuition says yes, because boundaries are multi-scale in nature; but we need more than intuition to move the field forward. Previous studies did not find any empirical evidence in benchmarking [9]. Our work gives an affirmative answer to this question. Moreover, we show that performance continuously improves as we add more scales. This implies that, because there are a wide range of scales in natural image structures, having a large range of observation scales would be useful.

Is there room for improvement in local boundary detection? The comprehensive and meticulous experiments in [9], along with the psychological evidences in [31], suggest that there is a limited amount of information in local image neighborhoods, and the *Pb* boundary detector is already close to the limit. This has led many researchers to pursue other paths, such as exploring mid-level grouping [22, 32, 23, 33, 24, 25], complex example-based models [11], or scene knowledge [34]. Our approach stays within the framework of local boundary detection, making decisions independently at each pixel, and we show significant improvements over *Pb*. Our results also compare favorably to those of the more sophisticated algorithms above.

In retrospective, there are three reasons why multi-scale edge detection has not become popular: (1) having a high cost; (2) possibly losing details; and most importantly, (3) lack of empirical support. We have “proved”, with extensive experimentation on a variety of datasets, that multi-scale processing improves boundary detection, boosting precision at salient boundaries while preserving details. Using an image pyramid keeps computational cost within a constant factor. It is our hope that multi-scale will soon become a standard component of boundary detection approaches.

References

1. Canny, J.: A computational approach to edge detection. IEEE Trans. PAMI **8** (1986) 679–698
2. Witkin, A.: Scale-space filtering. In: Int’l. J. Conf. on Artificial Intell. Volume 2. (1983) 1019–1022
3. Koenderink, J.: The structure of images. Biological Cybernetics **50** (1984) 363–370
4. Lindeberg, T.: Scale-Space Theory in Computer Vision. Kluwer Academic Publishers (1994)
5. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. IEEE Trans. PAMI **12** (1990) 629–639
6. Lindeberg, T.: Edge detection and ridge detection with automatic scale selection. Int’l. J. Comp. Vision **30** (1998) 117–156
7. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: ICCV. (2001) II:416–423
8. Konishi, S., Yuille, A., Coughlan, J., Zhu, S.: Fundamental bounds on edge detection: an information theoretic evaluation of different edge cues. In: CVPR. (1999) 573–579

9. Martin, D., Fowlkes, C., Malik., J.: Learning to detect natural image boundaries using local brightness, color and texture cues. IEEE Trans. PAMI **26(5)** (2004) 530–549
10. Konishi, S., Yuille, A., Coughlan, J.: A statistical approach to multi-scale edge detection. In: ECCV Workshop on Generative-Model-Based Vision. (2002)
11. Dollar, P., Tu, Z., Belongie, S.: Supervised learning of edges and object boundaries. In: CVPR. Volume 2. (2006) 1964–71
12. Galun, M., Basri, R., Brandt, A.: Multiscale edge detection and fiber enhacement using differences of oriented means. In: ICCV. (2007)
13. Bouman, C., Shapiro, M.: A multiscale random field model for bayesian image segmentation. IEEE Trans. Im. Proc. **3**(2) (1994) 162–177
14. Koepfler, G., Lopez, C., Morel, J.: A multiscale algorithm for image segmentation by variational method. SIAM J. Numer. Anal. **31**(1) (1994) 282–299
15. Sharon, E., Brandt, A., Basri, R.: Segmentation and boundary detection using multiscale intensity measurements. In: CVPR. (2001)
16. Ruderman, D.L., Bialek, W.: Statistics of natural images: Scaling in the woods. Physics Review Letters **73**(6) (1994) 814–817
17. Elder, J., Goldberg, R.: Ecological statistics of gestalt laws for the perceptual organization of contours. Journal of Vision **2**(4) (2002) 324–353
18. Ren, X., Malik, J.: A probabilistic multi-scale model for contour completion based on image statistics. In: ECCV. Volume 1. (2002) 312–327
19. Bergholm, F.: Edge focusing. IEEE Trans. PAMI **9** (1987) 726–741
20. Basu, M.: Gradient-based edge detection methods - a survey. IEEE. Trans. System, Man and Cybernatics **32**(3) (2002) 252–60
21. Rijsbergen, C.V.: Information Retrieval, 2nd ed. Univ. of Glasgow (1979)
22. Ren, X., Fowlkes, C., Malik, J.: Scale-invariant contour completion using conditional random fields. In: ICCV. Volume 2. (2005) 1214–1221
23. Arbelaez, P.: Boundary extraction in natural images using ultrametric contour maps. In: Workshop on Perceptual Organization in Computer Vision (POCV). (2006)
24. Felzenszwalb, P., McAllester, D.: A min-cover approach for finding salient curves. In: Workshop on Perceptual Organization in Computer Vision (POCV). (2006)
25. Zhu, Q., Song, G., Shi, J.: Untangling cycles for contour grouping. In: ICCV. (2007)
26. Stein, A., Hoiem, D., Hebert, M.: Learning to find object boundaries using motion cues. In: ICCV. (2007)
27. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: ECCV. (2006)
28. Malisiewicz, T., Efros, A.: Improving spatial support for objects via multiple segmentations. In: BMVC. (2007)
29. Everingham, M., Van Gool, L., Williams, C., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC2007). http://www.pascal-network.org/challenges/VOC/voc2007/workshop/
30. Russell, B., Torralba, A., Murphy, K., Freeman, W.: LabelMe: a database and web-based tool for image annotation. Technical Report AI Memo AIM-2005-025, MIT (2005)
31. Martin, D., Fowlkes, C., Walker, L., Malik, J.: Local boundary detection in natural images: Matching human and machine performance. In: ECVP. (2003)
32. Yu, S.: Segmentation induced by scale invariance. In: CVPR. (2005) I:445–451
33. Estrada, F., Elder, J.: Multi-scale contour extraction based on natural image statistics. In: Workshop on Perceptual Organization in Computer Vision (POCV). (2006)
34. Hoiem, D., Efros, A., Hebert, M.: Recovering occlusion boundaries from a single image. In: ICCV. (2007)

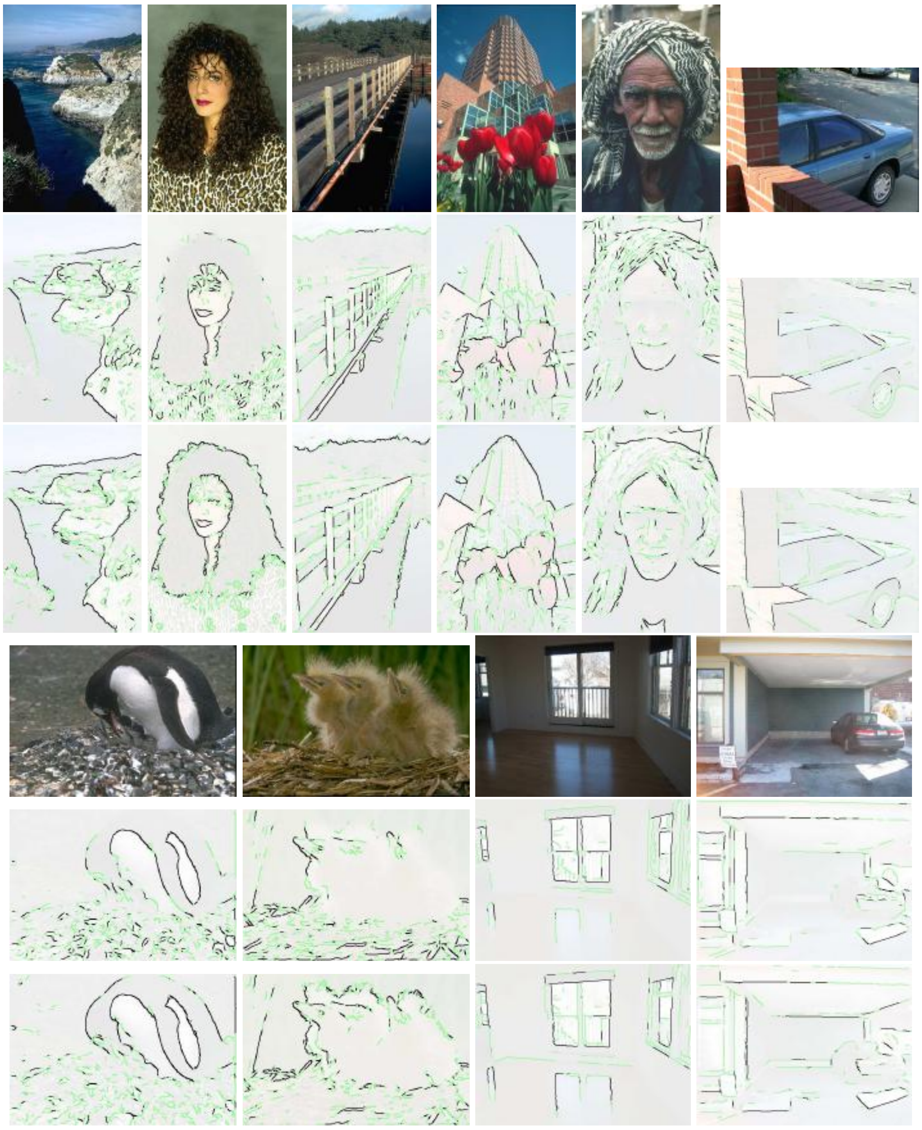


Fig. 6. Examples shown in three rows: input images, *Pb*, and multi-Pb (this work). Detected boundaries are discretized into two types: strong edge pixels (ranked 1-2000 in each image) in black, and weak edge pixels (ranked 2001-6000 in each image) in green. In such a way we visualize two points on the precision-recall curves in Figure 3 and 5. Large-scale boundaries are enhanced (i.e. from green to black, or from none to green), while small-scale boundaries, such as those found in texture, are suppressed (from black to green, from green to none).