Boundary Detection Benchmarking: Beyond F-Measures

Xiaodi Hou
Computation and Neural Systems, Caltech
MC 216-76, Caltech, Pasadena, CA, 91125
xiaodi.hou@gmail.com

Alan Yuille
Department of Statistics, UCLA
8967 Math Sciences Building
yuille@stat.ucla.edu

Christof Koch
Computation and Neural Systems, Caltech
MC 216-76, Caltech, Pasadena, CA, 91125
koch@klab.caltech.edu

Abstract

For an ill-posed problem like boundary detection, human labeled datasets play a critical role. Compared with the active research on finding a better boundary detector to refresh the performance record, there is surprisingly little discussion on the boundary detection benchmark itself.

The goal of this paper is to identify the potential pitfalls of today's most popular boundary benchmark, BSDS 300. In the paper, we first introduce a psychophysical experiment to show that many of the "weak" boundary labels are unreliable and may contaminate the benchmark. Then we analyze the computation of f-measure and point out that the current benchmarking protocol encourages an algorithm to bias towards those problematic "weak" boundary labels. With this evidence, we focus on a new problem of detecting strong boundaries as one alternative. Finally, we assess the performances of 9 major algorithms on different ways of utilizing the dataset, suggesting new directions for improvements.

1. Introduction

Boundaries in an image contain cues that are very important to high level visual tasks such as object recognition and scene understanding. Detecting boundaries has been a fundamental problem since the beginning of computer vision. In the development of boundary detection, datasets [16, 8, 5, 1] - along with their evaluation criteria[1] - have played critical roles. These datasets are responsible for our progress in the problem of boundary detection, not only because they provide an objective quantity to judge the value of each newly proposed algorithm, but also because the images, the labels, and the evaluation standards they set forth have heavily influenced the researchers during the development of a boundary detection algorithm.

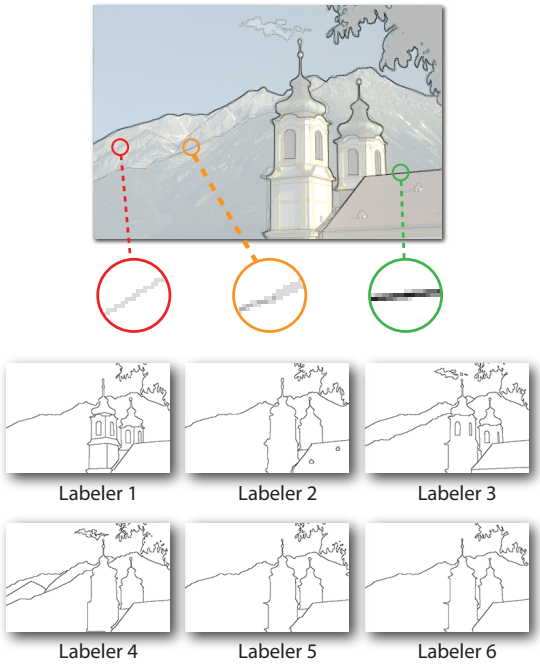

Figure 1. An example image and the corresponding labels from BSDS 300. Top figure shows the original image overlapping with all 6 boundary maps from labelers. There is a clear difference among different labelers. Red circle gives an example boundary segment that is labeled by only one out of 6 labelers (labeler 4). Boundary segment in the orange circle is labeled by two labelers (labeler 3 and 4). The boundary segment in green circle is unanimously labeled by all 6 labelers.

[1] In this paper, we refer to the images and the labels as *datasets*, while the term *benchmark* includes images, labels as well as the corresponding evaluation criteria.

1.1. Boundary detection is ill-defined

What is a boundary? A universally accepted definition of a boundary may not exist. No matter how the definition is made, one can always find counter-examples on which people disagree. In today's most popular benchmark BSDS 300 [16], 28 human labelers contributed a total number of 1667 high quality boundary maps on 300 images of natural scenes (200 training, 100 testing). Within the entire dataset, it is hard to find a image where different people have perfectly matched labels.

In high-level vision tasks such as object recognition or scene classification, human annotation has been traditionally considered reliable. However, the ill-posed nature of boundary detection makes this problem a different scenario. There is surprisingly little discussion about ground-truth data reliability for boundary detection. It is commonly held that human annotations from BSDS 300 are reliable. Previously, [16, 17] have the following observations regarding to the reliability of BSDS 300:

1. Labelers are well trained and correctly instructed. Examined separately, each boundary seems to be aligned to some underlying edge structure in the image. The effect of an adversarial labeler (labelers with totally irrelevant output) is minimal.

2. Label variability can be explained by a perceptual organization hierarchy. Even though different labelers may annotate boundaries in different levels of details, they are consistent in a sense that the dense labels "refine" the corresponding sparse labels without contradicting to them. In other words, the same image always elicits the same perceptual organization across different labelers.

Nevertheless, none of these observations are strong enough to legitimatize the BSDS 300 as a benchmark. To be able to evaluate an algorithm faithfully, the benchmark has to be free from both type I (false alarm) and type II (miss) statistical errors. Aforementioned observation #1 rules out type I errors. However, the risk of type II remains unchecked errors in human labels. It is possible that the labelers may miss some equally important boundaries. Once we benchmark an algorithm, the incomplete data may incorrectly penalize an algorithm that detects true boundaries.

As for observation #2, the hierarchical organization of boundaries raises more fundamental questions: Can we give equal weights to the strong boundaries where everyone agrees, and the weak boundaries where only one or two labelers have noticed? When we say "boundary detection", are we trying to solve one single problem with different thresholds? Or different problems at different levels of the perceptual hierarchy?

1.2. The perceptual strength of a boundary

In this paper, the *perceptual strength* of a boundary segment refers to the composite effect of all factors that influence personal decision during boundary annotation. Such factors may include border contrast, object type, or line geometry. One simple way to approximate the perceptual strength of each boundary segment is to take the proportion of labelers who have labeled that specific segment. To get rid of local alignment noise, we match each pair of human boundary maps using the assignment algorithm proposed in [11], with the same parameter set [15] used for algorithm evaluation. For instance, given an image with N labelers, if a boundary pixel from one subject matches with M other labelers, it has a perceptual strength of $\frac{M+1}{N}$. The weakest boundary labels are the ones annotated by only one labeler. These boundaries are referred to as *orphan labels*. In BSDS 300, 29.40% of the boundary labels are orphan labels. In comparison, the second largest population (28.99%) are *consensus labels* that are labeled by everyone.

Clearly, the orphan labels and the consensus labels are not equal. In Sec. 3, we use a psychophysical experiment to assess the statistical difference of weak/strong boundaries. Our experimental results indicate that weak (especially orphan) labels are not capable of evaluating today's algorithms.

Based on this novel discovery, in Sec. 4 we investigate the impact of these weak boundaries on the current evaluation system. A disappointing yet alarming result is that all of the 9 algorithms experience significant performance drops if we test them on strong boundaries only. Furthermore, we pinpoint a mechanism called *precision bubble* in the original BSDS 300 benchmarking algorithm. This mechanism tends to exaggerate the precision of an algorithm, especially when the weak labels are included in the groundtruth.

We raise an important yet largely neglected question: *are we ready to detect strong boundaries?* Our analysis shows that none of the 9 algorithms is capable of discovering strong boundaries significantly better than random selection. The output values of the algorithms are either independent or weakly correlated with the perceptual strength. This result is in sharp contrast to many of today's popular practice of using the output of a boundary detector algorithm as an informative feature in high-level boundary analysis. We conclude our discussion with a comparison of pB v.s. retrained-pB and BSDS 300 v.s. BSDS 500.

2. Related works

Over the last 12 years, a great number of boundary detection algorithms have been proposed. The benchmark's F-measure, according to the measurements proposed in [15], has increased 7 percent, from 64.82% [15] to 71.43% [20].

In this paper, we focus on 9 major boundary detection algorithms (shown in Tab. 1).

All of these algorithms, except cCut, provide very competitive F-measures at the time when they were first introduced. F-measure, also known as F-score, or the harmonic mean of precision and recall, is recommended in [15] as a summary statistics for the precision recall property. Over the past 10 years, it has been accepted as the most important score to judge a boundary detector.

Along with boundary detection, a parallel line of work [25, 27, 12] focuses on the detection of "salient boundaries". These works emphasize on finding salient 1-D structures from the ensemble of line segments discovered by a boundary detector. The stated advantage of these algorithms is to gain extra precision scores at low-recall regions. Therefore, it is interesting to include cCut [12], one of the latest algorithms in this line, and evaluate it under our quantitative framework.

Name	F-measure	Year
pB [15]	0.65	2002
UCM [2]	0.67	2006
Mincover [9]	0.65	2006
BEL [7]	0.66	2006
gPB [4]	0.70	2008
XRen [19]	0.67	2008
NMX [13]	0.71	2011
cCut [12]	0.45	2011
SCG [20]	0.71	2012

Table 1. The list of boundary detection algorithms referred in this paper. Their F-measures increase over time.

2.1. Relevant theories on dataset analysis

In contrast to the perennial efforts in breaking benchmark performance records, theoretical analysis on benchmark reliability is brought to people's attention only in recent years. These studies can be roughly categorized into either human annotation analysis, or benchmark design analysis. The first problem of human annotation comes with the recent trends of obtaining annotation data via crowdsourcing [22]. Many seminal models [18, 26] have been proposed to analyze the crowdsourced annotation process in general. Specifically, [24] has proposed strategies to estimate the quality of crowdsourced boundary annotation. On the other hand, [23] has raised a series of interesting questions to the design philosophy of today's object recognition benchmarks. Their alarming results suggest the potential pitfalls of some widely adopted benchmarks.

3. A psychophysical experiment

While collecting the human annotation, BSDS 300 [16] gave the following instructions to each of the labelers:

> *Divide each image into pieces, where each piece represents a distinguished thing in the image. It is important that all of the pieces have approximately equal importance. The number of things in each image is up to you. Something between 2 and 20 should be reasonable for any of our images.*

The instruction is intentionally made vague in order to minimize potential labeling bias towards any specific sub-type of boundaries. However, the absence of precise instruction also leads to a considerable labeling variation. As we have discussed in Sec. 1, 31.39% of the boundary labels are *orphan labels*. On one hand, we know that these boundaries are labeled by well-educated Berkeley students chosen from a graduate level computer vision class. On the other hand, we also aware that the annotation of these orphan labels is due to a pure random assignment of labelers. How well can we trust these relatively weak labels?

In this section, we introduce a two-way forced choice paradigm to test the reliability of a boundary dataset. In each trial, a subject[2] is asked to compare the relative perceptual strength of two local boundary segments with the following instruction:

> *Boundaries divide each image into pieces, where each piece represents a distinguished thing in the image. Choose the relatively stronger boundary segment from the two candidates.*

One of the two boundary segments is chosen from the human label dataset, and the other is a boundary segment produced by an algorithm. The advantage of this two-alternative experiment is that it cancels out most of the cognitive fluctuations, such as spatial attention bias, subject fatigue, and decision thresholds that are different among subjects. Moreover, compared to the tedious labeling process, this paradigm is much simpler and cheaper to implemented via crowdsourcing. In our experiment, the average response time for each trial is 5 seconds. One caveat is that the comparison experiment requires the algorithm generated candidate segment to have a similar appearance to the human labels. Among the 9 benched algorithms, BEL is the only algorithm that does not produce thinned edges, and therefore is skipped for the experiment.

3.1. Easy and hard experiments for boundary comparison

Using different boundary sampling strategies, we can design two experiments: hard and easy. In the hard experiment, each algorithm is first thresholded at its optimal F-measure, and then matched to the original human labels to

[2] We refer to *labelers* as the people who originally labeled the BSDS300 dataset, while *subjects* refers to people we recruited to perform our two-way forced choice experiment.

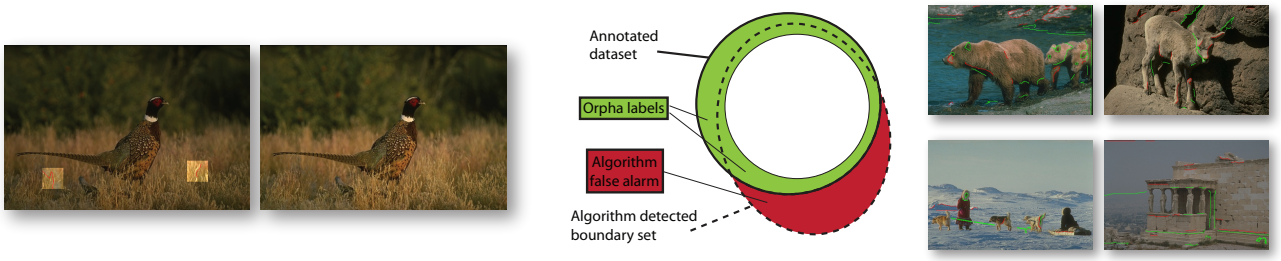

Figure 2. An illustration of the two-way, forced choice experiment (hard mode). **A) The experiment interface**: In each trial, a subject is presented with two images. On the left image, two boundary segments (high contrast squares with red lines) are superimposed onto the original photo. The subject is asked to click on one of two boundary segments that she/he feels stronger. At the same time, the original image is also presented in a separate window. **B) The Venn diagram of sets of boundary segments**: The thick circle encompasses the full human labeled boundary set of the dataset. The subset of orphan labels is shown in the green area. The algorithm detected boundary set is the dotted ellipsoid. The subset of algorithm false alarms is highlit in red. In each trial, we randomly select one boundary segment from the green area, and the other one from the red area. **C) Orphan labels v.s. algorithm false alarms**: Some example images with both human orphan labels (shown in green lines) and false alarms of PB algorithm (shown in red lines). In many examples, the relative strength between algorithm false alarm and human orphan labels is very hard to tell.

find false alarms – boundary segments that are considered weaker than human labels. And then, for each testing image, we randomly draw one instance of algorithm false alarm, and compare it against another randomly selected human orphan label. Fig. 2 gives a detailed illustration of this process. This experiment is called "hard experiment" because the relative order between human labeled orphan label and algorithm detected false alarms is not easy to determine (as one can see in Fig. 2.C).

Similarly, we also design an easy experiment. First, we remove all the human labels that are not unanimously labeled by everyone. This leaves us with a very small but strong subset of labels (perceptual strength equals 1). Then, with this new dataset, we re-benchmark all 8 algorithms, determining their optimal F-measures and thresholds (higher than their original thresholds), and find each algorithm's false alarms under its new optimal threshold. Finally, the competition is made between strong human labels and confident output of algorithm false alarms.

For each algorithm on either easy/hard experiment, we produce one trial per image for all 100 test images. 5 subjects participated in the experiment, and a total number of 8000 responses are collected. The final ordering for each trial is determined by majority voting of all 5 subjects. To interpret the result, we introduce a term called *dataset risk*. This value measures the probability that an algorithm false alarm wins over a human label. Ideally, a perfectly constructed dataset should have zero risk, because it does not miss any strong boundary segments, and algorithm false alarms are always weaker than any instance from the perfect boundary dataset. However, our experiment results in Fig. 3 show that the BSDS 300 – especially those orphan labels, are far away from being perfect.

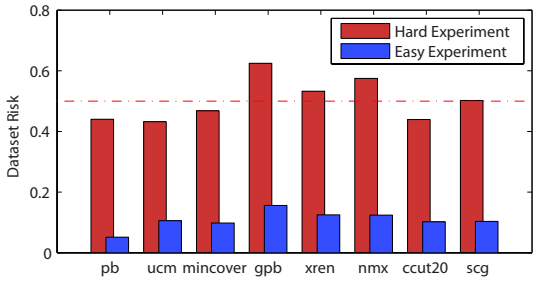

Figure 3. Results of our hard and easy experiments. The average risks over all algorithms are 0.5017 and 0.1082 for hard and easy experiments, respectively. Dotted red line indicates the 50% chance performance. The average result of the hard experiment is even greater than chance level.

3.2. Interpreting the risk of a dataset

From Fig. 3, we observe high risks in the hard experiment for all algorithms that we have tested. The first conclusion one can draw from this observation is rather depressing – the orphan labels are extremely unreliable since they falsely classify good algorithm detections into false alarms (or falsely include weak algorithm detections into hits, depending on the thresholds). Yet, we can also interpret the results of hard experiments in a more optimistic way: the computer vision algorithms have performed so good that

their results look as good as some of the human's. In other words, these algorithms have passed a restricted Turing test if the dataset risk is equal to or greater than 0.5.

No matter whether to choose the pessimistic or the optimistic perspective, it is clear that the orphan labels are not appropriate to serve as a benchmark – or even parts of a benchmark. Instead, we should put more focus on the consensus boundaries because the risk is much lower.

It is worth mentioning that our results on the easy experiment does not necessarily imply that the consensus boundaries is a perfect dataset. However, as long as the missed boundaries of consensus labels cannot be accurately detected by an algorithm, this data remains to be valid for a benchmark. In other words, given the performance of today's top algorithms, detecting strong boundaries is a meaningful Turing test that is not yet solved.

4. F-measures and the precision bonus

Given the fact that the orphan labels are unreliable, what role do those labels play in the benchmarking process? How much can they affect the result of F-measure? In this section, we show that the orphan labels can create a "precision bonus" during the calculation of the F-measure..

In the original benchmarking protocol of BSDS 300, the false negative is defined by comparing *each* human boundary map with the thresholded algorithm map, and count the unmatched human labels. In comparison, the false positive is defined by comparing the algorithm map with *all* human maps, and then count the algorithm labels that are not matched by *any* human. In other words, the cost of each algorithm missing pixel is proportional to the human labelers who have detected that boundary, whereas the cost of each false alarm pixel is just one. This protocol exaggerates the importance of the orphan labels in the dataset, and encourages algorithms to play "safely" by enumerating an excessive number of boundary candidates. Strategically, detecting strong boundaries has become a much more risky endeavor under the current framework of F-measure.

We can better evaluate the impact of such *precision bonus* by re-benchmarking the algorithms on different levels. First we threshold the human labels by different perceptual strengths, from $0, 0.2, 0.4 \ldots$ to 1. And then use each of these subset of the human labels as the ground-truth to benchmark all 9 algorithms. At each perceptual strength, an algorithm find its optimal threshold that produces the maximal F-measure. Fig. 4 plots the precision and recall values at the optimal algorithm thresholds for all 9 algorithms.

Despite its strong influence on the benchmark scores, the precision bubble by itself should not be considered as a "mistake" in the design. What makes today's benchmarking practice questionable is the joint cause of the following facts: 1) weak boundaries in BSDS 300 are not reliable enough to evaluate today's algorithms; and 2) precision bonus gives

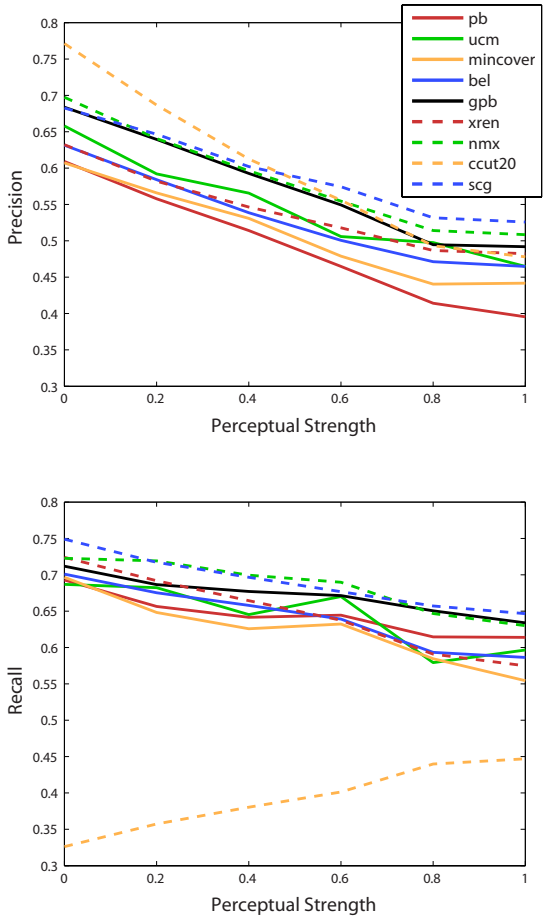

Figure 4. The optimal precision and recall values for all algorithms, benchmarked under different label strengthes. By increasing the perceptual strength, we transform the problem from "boundary detection" to "strong boundary detection". The precision values for algorithms dropped 28.7% in average. In contrast, the recall values, which are not affected by the precision bonus, only dropped 9% in average.

extra credits to algorithms working on the low perceptual strength boundaries – which according to fact 1, is not a good practice.

5. Detecting strong boundaries

The simplest way to avoid the problem of weak labels is to benchmark the algorithms using consensus labels only, as shown in Fig. 5. However, the performances of the tested algorithms have dropped so significantly that it stimulates us to ask another question: *are we detecting strong boundaries better than random?*

To compute the baseline performance of a null hypothesis, we design a control experiment called *partial labels*. In this experiment, we crop out a part of each human boundary map to make the total number of pixels in the remaining

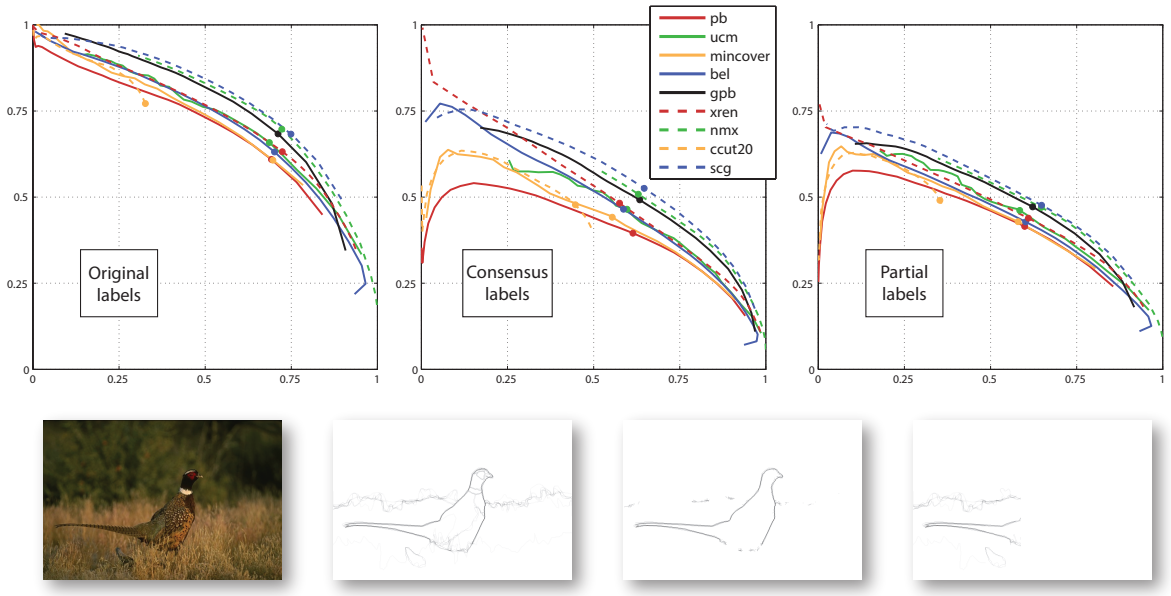

Figure 5. Benchmarking all 9 algorithms using different ground truths. The top figure shows the precision recall curves, with solid dots indicating the maximal F-measure location. The bottom figure gives an example image and the ground-truth labels: original labels, consensus labels, and partial labels. The partial label (bottom right figure) of this image is clearly an unrealistic ground-truth because the majority of the bird boundary is discarded.

map equals to that of a strong boundary map (see Fig. 5). Because such cropping operation is completely independent of the image content, it can be considered as a random subsampling from an algorithm perspective.

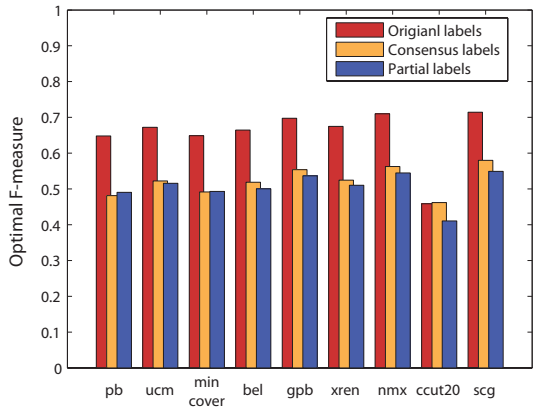

Figure 6. Algorithm performances (optimal F-measures) evaluated under different ground-truths.

With the PR curves shown in Fig. 5, the optimal F-measures of all three experiments are compared in Fig. 6. Except for cCut, all other algorithms have suffered severe performance decreases when shifting from detecting all labels to detecting consensus labels only. Such performance drop is so devastating that the F-measures are no better (even worse for pB algorithm) than the control experiment with randomly contaminated ground-truth.

In this experiment, the salient boundary algorithm cCut has the most significant performance drop on partial labels. However, the overall performance of cCut is not comparable with the state-of-the-arts detectors (such as gPB, NMX, or SCG), even if we benchmark them on the consensus labels.

The comparative results of consensus and partial labels contradict our intuitions that algorithm detection strength is correlated with the perceptual strength of a boundary. It also questions the practices in computer vision that use boundary detector output as a feature for high-level visual tasks. For instance, intervene contour [14, 6] is a well-established method that computes the affinity of two points in the image by integrating the boundary strengths along the path that connects those two points. Many other works such as [21, 10, 3] also included pB (or gPB) boundary intensity in their feature design. To understand the relationship between algorithm output and the perceptual strength of a boundary, we further plot the perceptual strength distribution with respect to algorithm detector output for all 9 algorithms. In Fig. 7, we can see that the correlation between algorithm output and perceptual strength of the boundary is rather weak.

5.1. Retrain on strong boundaries

Another useful test to evaluate our current progress on strong boundary is to retrain an algorithm. Because of its great popularity, we focus on pB algorithm for the retraining

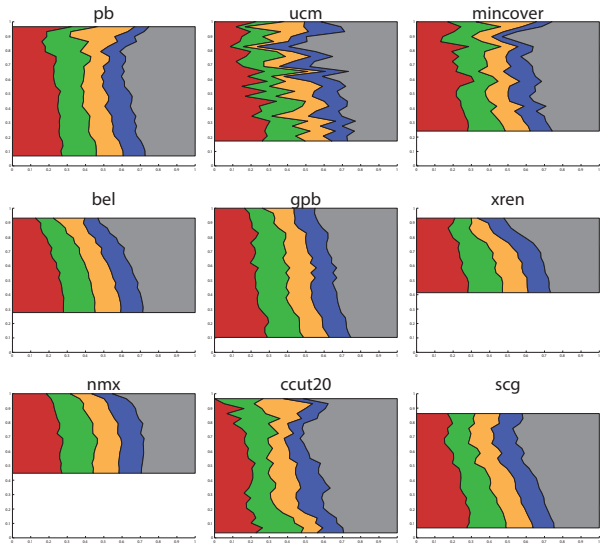

Figure 7. Boundary perceptual strength distribution. This experiment is done with the original (full) labels. In each sub-figure, the X-axis is the percentage of matched human label strength (always summing to 1), the Y-axis is the algorithm output value. If we extract one row with $y = k$ in a sub-figure, the color strips represent the distribution of the human labels that are matched to all algorithm pixels where detection output is equal to k. Red area represents human labels with perceptual strength in $[0, 0.2)$, whereas green represents perceptual strengths in $[0.2, 0.4)\ldots$, and finally the gray area shows the population of consensus labels. Ideally, the gray area should have a upper triangular shape (XREN is the closest) – that is, algorithm output being correlated with human perceptual strength.

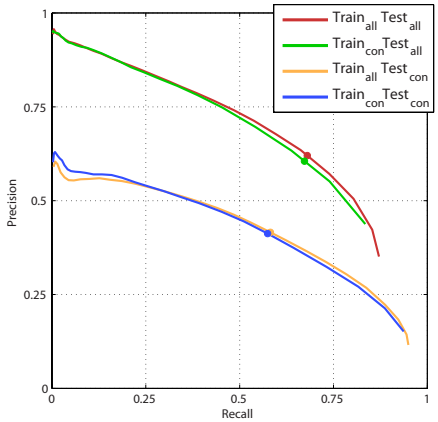

Figure 8. Retrain pB algorithm using consensus labels, and compare the results on original (all) and consensus (con) boundaries respectively.

experiment. Using the publicly available MATLAB codes from the authors' website, we re-generate the training samples with consensus boundaries and learn a new set of parameters. This retrained-pB is then compared against the original pB in the original as well as the consensus label test sets. The retrained-pB does not gain superior F-measure even if we use consensus labels as the ground-truth.

5.2. BSDS 300 and BSDS 500

Recently, BSDS 300 has been enriched to BSDS 500 with 200 additional testing images. According to [4], the protocol used to collect new human labels remains the same as in BSDS 300. According to our analysis, the population of orphan and consensus labels of these 200 new images are 30.58% and 30.15%, respectively. Not only the statistics of BSDS 500 looks very similar to the original BSDS 300, the performance of algorithms on this new dataset is also very close. Since BSDS 500 is fairly new, not many algorithms have provided their results on this new dataset. We choose two most representative algorithms SCG and gPB for our analysis. The optimal F-measure of these algorithms under all boundaries, or consensus boundaries are reported in Fig. 9.

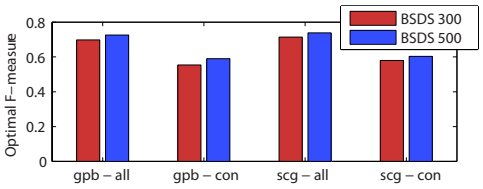

Figure 9. Comparison of SCG and gPB algorithms on BSDS 300 and BSDS 500 datasets. The comparison is also made by either using original (all) boundaries or consensus (con) boundaries only. The difference between BSDS 300 and BSDS 500 is small (mean difference is 0.028) and consistent (STD over all 4 different settings is 0.0058).

6. Discussion

In this paper, we have raised doubts on the current way of benchmarking an algorithm on the most popular dataset of boundary detection (Further results are provided in the supplemental material). With a psychophysical experiment, we show that the weak, especially the orphan labels are not suitable for benchmarking algorithms. However, if we shift from the original problem of boundary detection, to the new problem of strong boundary detection, we are on one hand blessed with a more reliable dataset; but on the other hand, disappointed by the experimental results that none of the current algorithms has shown evidence of good performance.

Our results in Fig. 7 do not conclude that the current algorithms' output value is a useless feature for high-level tasks. The validity of using boundary detector output to reveal high-level semantic information may not have a one-line answer. It depends critically on the specific scenarios as well as the design of the high-level vision algorithms. At

present, researchers from different topics have not yet converged to one common framework.

Acknowledgments

The first author would like to thank Zhuowen Tu, Yin Li and Liwei Wang for their thoughtful discussions. The research was supported by the ONR via an award made through Johns Hopkins University, by the G. Harold & Leila Y. Mathers Charitable Foundation, by Army Research Lab with 62250-CS and the Office of Naval Research N00014-12-10883.

References

[1] S. Alpert, M. Galun, R. Basri, and A. Brandt. Image segmentation by probabilistic bottom-up aggregation and cue integration. In *CVPR 2007. IEEE Conference on*, pages 1–8. IEEE, 2007. 1

[2] P. Arbelaez. Boundary extraction in natural images using ultrametric contour maps. In *CVPR Workshop, 2006. IEEE Conference on*, pages 182–182. IEEE, 2006. 3

[3] P. Arbeláez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik. Semantic segmentation using regions and parts. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3378–3385. IEEE, 2012. 6

[4] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(5):898–916, 2011. 3, 7

[5] E. Borenstein and S. Ullman. Class-specific, top-down segmentation. *Computer Vision–ECCV 2002*, pages 639–641, 2002. 1

[6] T. Cour, F. Benezit, and J. Shi. Spectral segmentation with multiscale graph decomposition. In *CVPR, 2005. IEEE Conference on*, volume 2, pages 1124–1131. IEEE, 2005. 6

[7] P. Dollar, Z. Tu, and S. Belongie. Supervised learning of edges and object boundaries. In *CVPR, 2006 IEEE Conference on*, volume 2, pages 1964–1971. IEEE, 2006. 3

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 1

[9] P. Felzenszwalb and D. McAllester. A min-cover approach for finding salient curves. In *CVPR Workshop, 2006. IEEE Conference on*, pages 185–185. IEEE, 2006. 3

[10] V. Ferrari, T. Tuytelaars, and L. Van Gool. Object detection by contour segment networks. *Computer Vision–ECCV 2006*, pages 14–28, 2006. 6

[11] A. Goldberg and R. Kennedy. An efficient cost scaling algorithm for the assignment problem. *Mathematical Programming*, 71(2):153–177, 1995. 2

[12] R. Kennedy, J. Gallier, and J. Shi. Contour cut: identifying salient contours in images by solving a hermitian eigenvalue problem. In *CVPR, 2011. IEEE Conference on*, pages 2065–2072. IEEE, 2011. 3

[13] I. Kokkinos. Boundary detection using f-measure-, filter- and feature-(f 3) boost. *Computer Vision–ECCV 2010*, pages 650–663, 2010. 3

[14] T. Leung and J. Malik. Contour continuity in region based image segmentation. *Computer Vision–ECCV 1998*, pages 544–559, 1998. 6

[15] D. Martin, C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(5):530–549, 2004. 2, 3

[16] 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 *Computer Vision–ICCV 2001. IEEE International Conference on*, volume 2, pages 416–423. IEEE, 2001. 1, 2, 3

[17] D. Martin, J. Malik, and D. Patterson. *An Empirical Approach to Grouping and Segmentaqtion*. Computer Science Division, University of California, 2003. 2

[18] V. Raykar, S. Yu, L. Zhao, A. Jerebko, C. Florin, G. Valadez, L. Bogoni, and L. Moy. Supervised learning from multiple experts: Whom to trust when everyone lies a bit. In *ICML, 2009. ACM Conference on*, pages 889–896. ACM, 2009. 3

[19] X. Ren. Multi-scale improves boundary detection in natural images. *Computer Vision–ECCV 2008*, pages 533–545, 2008. 3

[20] X. Ren and L. Bo. Discriminatively trained sparse code gradients for contour detection. *Advances in Neural Information Processing Systems*, 25, 2012. 2, 3

[21] X. Ren, C. Fowlkes, and J. Malik. Figure/ground assignment in natural images. *Computer Vision–ECCV 2006*, pages 614–627, 2006. 6

[22] A. Sorokin and D. Forsyth. Utility data annotation with amazon mechanical turk. In *CVPR Workshops, 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 3

[23] A. Torralba and A. Efros. Unbiased look at dataset bias. In *CVPR, 2011. IEEE Conference on*, pages 1521–1528. IEEE, 2011. 3

[24] S. Vittayakorn and J. Hays. Quality assessment for crowdsourced object annotations. In *Proceedings of the British machine vision conference*, pages 109–1, 2011. 3

[25] S. Wang, T. Kubota, and J. Siskind. Salient boundary detection using ratio contour. *Advances in Neural Information Processing Systems*, 16, 2003. 3

[26] P. Welinder, S. Branson, S. Belongie, and P. Perona. The multidimensional wisdom of crowds. In *In Proc. of NIPS*, pages 2424–2432, 2010. 3

[27] Q. Zhu, G. Song, and J. Shi. Untangling cycles for contour grouping. In *Computer Vision–ICCV 2007. IEEE Conference on*, pages 1–8. IEEE, 2007. 3