Quality Evaluation Methods for Crowdsourced Image Segmentation

Doris Jung-Lin Lee
University of Illinois,
Urbana-Champaign
jlee782@illinois.edu

Akash Das Sarma
Facebook, Inc.
akashds@fb.com

Aditya Parameswaran
University of Illinois,
Urbana-Champaign
adityagp@illinois.edu

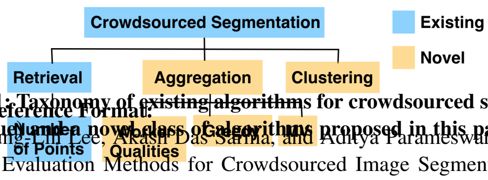

Figure 1: Taxonomy of existing algorithms for crowdsourced segmentation (blue) and a novel class of algorithms proposed in this paper (yellow).

ACM Reference Format:
Doris Jung-Lin Lee, Akash Das Sarma, and Aditya Parameswaran. 1997. Quality Evaluation Methods for Crowdsourced Image Segmentation. In *Proceedings of ACM Woodstock conference (WOODSTOCK'97)*, Jennifer B. Sartor, Theo D'Hondt, and Wolfgang De Meuter (Eds.). ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Precise, instance-level object segmentation is crucial for identifying and tracking objects in a variety of real-world emergent applications of autonomy, including robotics [10], image organization and retrieval [18], and medicine [7]. To this end, there has been a lot of work on employing crowdsourcing to generate training data for segmentation, including Pascal-VOC [4], LabelMe [15], OpenSurfaces [2], and MS-COCO [8]. Unfortunately, raw data collected from the crowd is known to be noisy due to varying degrees of worker skills, attention, and motivation [1, 17].

To deal with these challenges, many have employed heuristics indicative of crowdsourced segmentation quality to pick the best worker-provided segmentation [14, 16]. However, this approach ends up discarding the majority of the worker segmentations and is limited by what the best worker can do. In this paper, we make two contributions: First, we introduce a novel class of aggregation-based methods that incorporates portions of segmentations from multiple workers into a combined one described in Section 4. To our surprise, despite its intuitive simplicity, we have not seen this class of algorithms described or evaluated in prior work. We evaluate this class of algorithms against existing methods in Section 6. Second, our analysis of common worker errors in crowdsourced segmentation shows that workers often segment the wrong objects or erroneously include or exclude large semantically-ambiguous portions of an object in the resulting segmentation. We discuss these semantic errors and ambiguities in Section 3 and propose a clustering-based preprocessing technique that resolves such errors in Section 5.

2 RELATED WORK

As shown in Figure 1, quality evaluation methods for crowdsourced segmentation can be classified into two categories:

Retrieval-based methods pick the "best" worker segmentation based on some scoring criteria that evaluates the quality of each segmentation, including vision information [11, 16], and click-stream behavior [3, 12, 14].

Aggregation-based methods combine multiple worker segmentations to produce a final segmentation that is not restricted to any single worker segmentation. An aggregation-based majority vote approach was employed in Sameki et al. (2015) to create an expert-established gold standard for characterizing their dataset and algorithmic accuracies, rather than for segmentation quality evaluation as described here.

Orthogonal methods to improve segmentation quality include periodic verification [4, 9], specialized interfaces [13], and vision-based supervision [6, 11]. These methods could be used for quality improvement on top of any of the algorithms in this paper.

3 ERROR ANALYSIS

On collecting and analyzing a number of crowdsourced segmentations (described in Section 6), we found that common worker segmentation errors can be classified into three types: (1) **Semantic Ambiguity:** workers have differing opinions on whether particular regions belong to an object (Figure 2 left: annotations around 'flower and vase' when 'vase' is requested); (2) **Semantic Mistake:** workers annotate the wrong object entirely (Figure 2 right: annotations around 'turtle' and 'monitor' when 'computer' is requested.); and (3) **Boundary Imperfection:** workers make unintentional mistakes while drawing the boundaries, either due to low image resolution, small area of the object, or lack of drawing skills (Figure 3 left: imprecision around the 'dog' object).

Quality evaluation methods in prior work have largely focused on minimizing boundary imperfection issues. So, we first describe our novel aggregation-based algorithms designed to reduce boundary imperfections in Section 4. Next, in Section 5, we discuss a preprocessing method eliminates semantic ambiguities and errors, also observed in prior work [5, 9, 14]. We present our experimental evaluation in Section 6.

4 FIXING BOUNDARY IMPERFECTIONS

At the heart of our aggregation techniques is the *tile data representation*. A tile is the smallest non-overlapping discrete unit created by overlaying all of the workers' segmentations on top of each other. The tile representation allows us to aggregate segmentations from multiple workers, rather than being restricted to a single worker's segmentation—allowing us to fix one worker's errors with help from another. In Figure 3 (right), we display three worker segmentations for a toy example with 6 resulting tiles. Any subset of these tiles can contribute towards the final segmentation.

WOODSTOCK'97, July 1997, El Paso, Texas USA

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

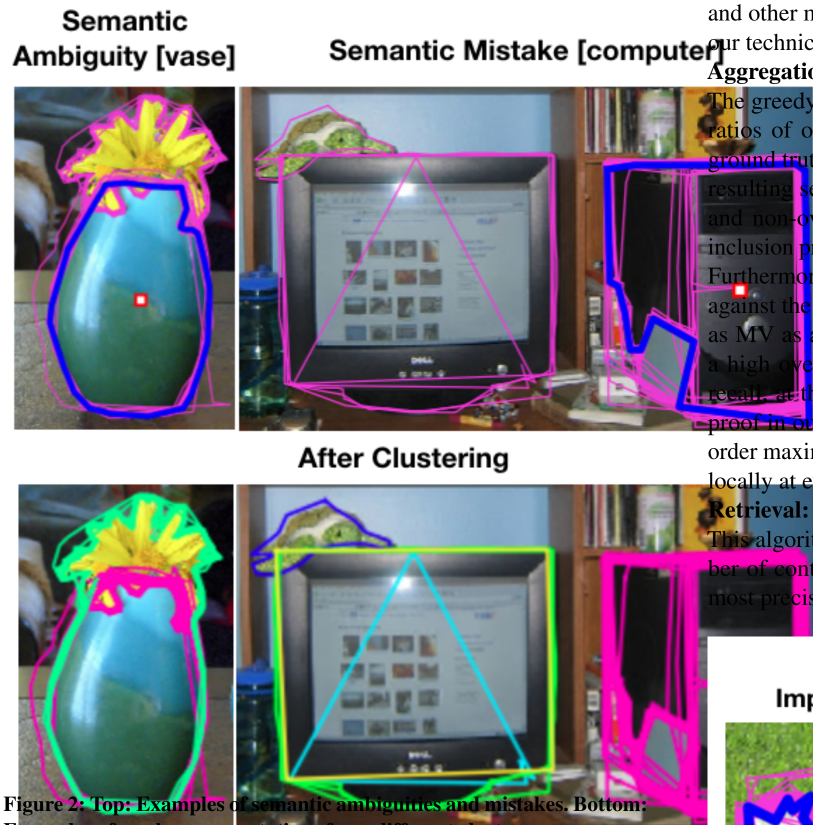

Figure 2: Top: Examples of semantic ambiguities and mistakes. Bottom: Examples of worker segmentations from different clusters.

This simple but powerful idea of tiles also allows us to reformulate our problem from one of "generating a segmentation" to a setting that is much more familiar to crowdsourcing researchers. Since tiles are the lowest granularity units created by overlaying all workers' segmentations on top of each other, each tile is either completely contained within or outside a given worker segmentation. Specifically, we can regard a worker segmentation as multiple boolean responses where they have voted 'yes' or 'no' to every tile independently. Intuitively, a worker votes 'yes' for every tile that is contained in their segmentation, and 'no' for every tile that is not. As shown in Figure 3 (right), tile t_2 is voted 'yes' by worker 1, 2, and 3; tile t_3 is voted 'yes' by worker 2 and 3. The goal of our aggregation algorithms is to pick an appropriate set of tiles that effectively trades off precision versus recall.

Now that we have modeled segmentation as a collection of worker votes for tiles, we can now develop familiar variants of standard quality evaluation algorithms for this setting.

Aggregation: Majority Vote Aggregation (MV)

This simple algorithm includes a tile in the output segmentation if and only if the tile has 'yes' votes from at least 50% of all worker segmentations.

Aggregation: Expectation-Maximization (EM)

Unlike MV, which assumes that all workers perform uniformly, EM approaches use worker quality models to infer the likelihood that a tile is part of the ground truth segmentation. While simultaneously estimating worker qualities and tile likelihoods as hidden variables, our basic worker quality model that we evaluate in Section 6 assumes a fixed probability for a correct vote. Details of the formal derivation and other more fine-grained worker quality models can be found in our technical report.

Aggregation: Greedy Tile Picking (greedy)

The greedy algorithm picks tiles in descending order based on the ratios of overlap area to non-overlap area (both with respect to ground truth), for as long as the estimated Jaccard similarity of the resulting segmentation continues to increase. Since the tile overlap and non-overlap against ground truth are unknown, we use tile-inclusion probabilities from EM to estimate these areas as a heuristic. Furthermore, since we cannot compute the actual Jaccard similarity against the unknown ground truth, we use a heuristic baseline such as MV as a proxy for the ground truth. Intuitively, tiles that have a high overlap area and low non-overlap area contribute to high recall, at the cost of relatively little precision error. We include a proof in our technical report showing that picking tiles in such an order maximizes the Jaccard similarity of the resulting segmentation locally at every step.

Retrieval: Number of Control Points (num pts)

This algorithm picks the worker segmentation with the largest number of control points around the segmentation boundary (i.e., the most precise drawing) as the output segmentation [14, 16].

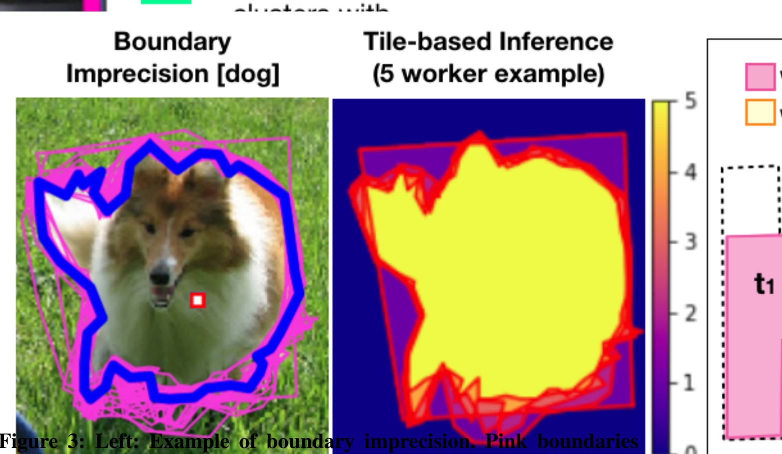

Figure 3: Left: Example of boundary imprecision. Pink boundaries show worker segmentations and blue delineates the ground truth. Middle: Segmentation boundaries drawn by five workers shown in red. Overlaid segmentation creates a mask where the color indicates the number of workers who voted for the tile region. Right: Toy example demonstrating tiles created by three workers' segmentations around a dumbell object delineated by the black dotted line.

5 PERSPECTIVE RESOLUTION

As discussed in Section 3, disagreements often arise in segmentation due to differing worker perspectives on large tile regions. We developed a clustering-based preprocessing approach to resolve this issue. Based on the intuition that workers with similar perspectives will have segmentations that are close to each other, we compute the Jaccard similarity between each pair of segmentations and perform spectral clustering to separate the segmentations into clusters. Figure 2 (bottom) illustrates how spectral clustering divides the worker segmentations into clusters with meaningful semantic associations, reflecting the diversity of perspectives for the same task. Clustering results can be used as a preprocessing step for any quality evaluation algorithm by keeping only the segmentations that belong to the largest cluster, which is typically free of semantic errors.

In addition, clustering offers the additional benefit of preserving worker's semantic intentions. For example, while the green cluster in Figure 2 (bottom right) would be considered *bad* segmentations for the particular task ('computer'), this cluster can provide more data for another segmentation task corresponding to 'monitor'. A potential future work direction would be to crowdsource the semantic labels for the computed clusters to enable the reuse of segmentations across multiple objects to lower costs.

6 EXPERIMENTAL EVALUATION

Dataset Description

We collected crowdsourced segmentations from Amazon Mechanical Turk; each HIT consisted of one segmentation task for a specific pre-labeled object in an image. There were a total of 46 objects in 9 images from the MSCOCO dataset [9] segmented by 40 different workers each, resulting in a total of 1840 segmentations. Each task contained a keyword for the object and a pointer indicating the object to be segmented. Two of the authors generated the ground truth segmentations by carefully segmenting the objects using the same task and interface.

Evaluation Metrics

Evaluation metrics used in our experiments measure how well the final segmentation (S) produced by these algorithms compare against ground truth (GT). The most common evaluation metrics used in the literature are area-based methods that take into account the intersection area, $IA = area(S \cap GT)$, or union area, $UA = area(S \cup GT)$ between the worker and ground truth segmentations, including Precision (P) = $\frac{IA(S)}{area(S)}$, Recall (R) = $\frac{IA(S)}{area(GT)}$, and Jaccard (J) = $\frac{UA(S)}{IA(S)}$.

Experiment 1: Aggregation-based methods perform significantly better than retrieval-based methods

In Figure 4, we vary the number of worker segmentations along the x-axis and plot the average Jaccard score on the y-axis across different worker samples of a given size across different algorithms. Figure 4 (left) shows that the performance of aggregation-based algorithms (greedy, EM) exceeds the best-achievable through existing retrieval-based method (Retrieval). Then, in Figure 4 (right), we estimate the upper-bound performance of each algorithm by assuming that the 'full information' based on ground truth was given to the algorithm. For greedy, the algorithm is aware of all the actual tile overlap and non-overlap areas against ground truth, and does not need to approximate these values. For EM, we consider the performance of the algorithm if the true worker quality parameter values (under our worker quality model) are known. For retrieval, the full information version directly picks the worker with the highest Jaccard similarity with respect to the ground truth segmentation. By making use of ground truth information (Figure 4 right), the best aggregation-based algorithm can achieve a close-to-perfect average Jaccard score of 0.98 as an upper bound, far exceeding the results achievable by any single 'best' worker (J=0.91). This result demonstrates that aggregation-based methods are able to achieve better performance by performing inference at the tile granularity, which is guaranteed to be finer grained than any individual worker segmentation.

The performance of aggregation-based methods scale well as more worker segmentations are added.

Intuitively, larger numbers of worker segmentations result in finer granularity tiles for the aggregation-based methods. The first row in Table 1 lists the average percentage change in Jaccard between 5-workers and 30-workers samples, demonstrating a monotonically increasing relationship between number of worker segmentations used and the performance. However, retrieval-based methods do not benefit from more segmentations.

Experiment 2: Clustering as preprocessing improves algorithmic performance.

The average percentage change between the no clustering and clustering results is shown in Table 1. Clustering generally results in an accuracy increase. Since the 'full information' variants are already free of semantic ambiguity and errors, clustering does not assist with further improvement.

	Retrieval-based		Aggregation-based			
Algorithm	num pts	worker*	MV	EM	greedy	greedy*
Worker Scaling	6.30	2.38	1.63	1.64	2.16	5.59
Clustering Effect	5.92	-0.02	2.05	1.38	5.55	-0.06

Table 1: Percentage change due to worker scaling and clustering. Algorithms with * makes use of ground truth information.

7 CONCLUSION AND FUTURE WORK

We identified three different types of errors for crowdsourced image segmentation, developed a clustering-based method to capture the semantic diversity caused by differing worker perspectives, and introduced novel aggregation-based methods that produce more accurate segmentations than existing retrieval-based methods.

Our preliminary studies show that our worker quality models are good indicators of the actual accuracy of worker segmentations. We also observe that the greedy algorithm is capable of achieving close-to-perfect segmentation accuracy with ground truth information. Given the success of aggregation-based methods, including the simple majority vote algorithm, we plan to use our worker quality insights to improve our EM and greedy algorithms. We are also working on using computer vision signals to further improve our algorithms.

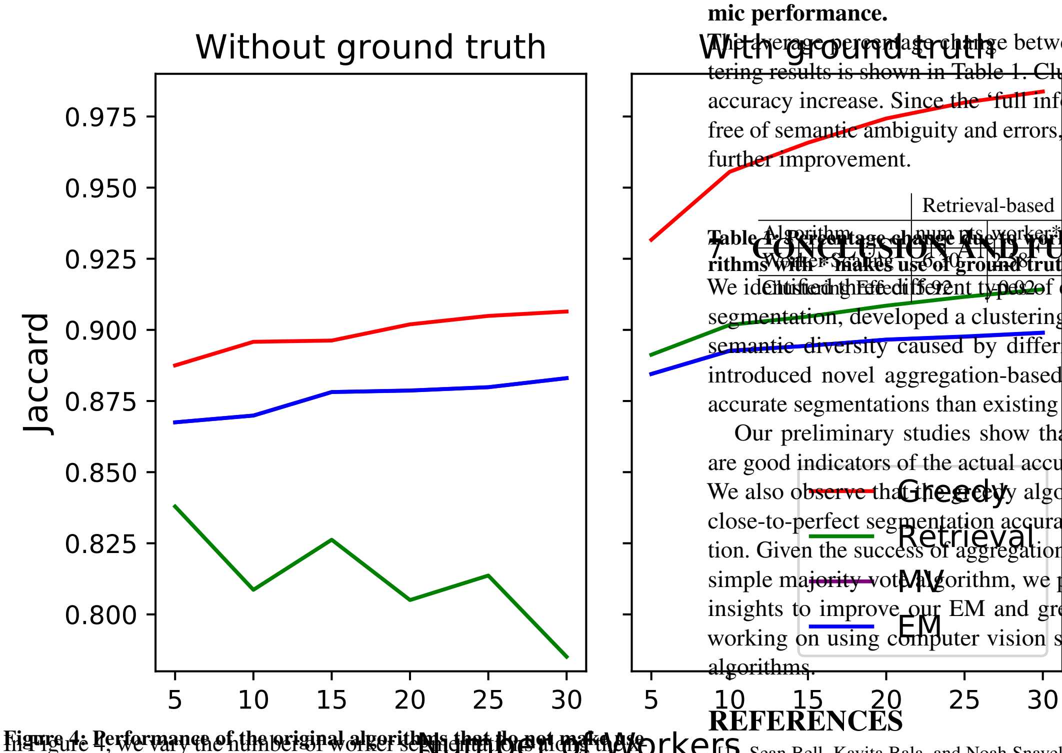

Figure 4: Performance of the original algorithms that do not make use of ground truth information (Left) and ones that do (Right). MV and EM results are so close that they overlay on each other.

REFERENCES

[1] Sean Bell, Kavita Bala, and Noah Snavely. 2014. Intrinsic Images in the Wild. *ACM Trans. on Graphics (SIGGRAPH)* 33, 4 (2014).

[2] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. 2015. Material Recognition in the Wild with the Materials in Context Database. *Computer Vision and Pattern Recognition (CVPR)* (2015).

[3] Ferran Cabezas, Axel Carlier, Vincent Charvillat, Amaia Salvador, and Xavier Giro-I-Nieto. 2015. Quality control in crowdsourced object segmentation. *Proceedings of International Conference on Image Processing, ICIP* 2015-Decem (2015), 4243–4247. https://doi.org/10.1109/ICIP.2015.7351606 arXiv:arXiv:1505.00145v1

[4] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2015. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision* 111, 1 (Jan. 2015), 98–136.

[5] Danna Gurari, Kun He, Bo Xiong, Jianming Zhang, Mehrnoosh Sameki, Suyog Dutt Jain, Stan Sclaroff, Margrit Betke, and Kristen Grauman. 2018. Predicting Foreground Object Ambiguity and Efficiently Crowdsourcing the Segmentation(s). *International Journal of Computer Vision (IJCV)* (2018). https://doi.org/10.1007/s11263-018-1065-7 arXiv:1705.00366

[6] Danna Gurari, Mehrnoosh Sameki, Zheng Wu, and Margrit Betke. 2016. Mixing Crowd and Algorithm Efforts to Segment Objects in Biomedical Images. *Medical Image Computing and Computer Assisted Intervention Interactive Medical Image Computation Workshop* (2016), 1–8.

[7] H Irshad and Montaser-Kouhsari et. al. 2014. Crowdsourcing Image Annotation for Nucleus Detection and Segmentation in Computational Pathology: Evaluating Experts, Automated Methods, and the Crowd. *Biocomputing 2015* (2014), 294–305. https://doi.org/10.1142/9789814644730_0029

[8] Christopher H Lin, Mausam, and Daniel S Weld. 2012. Crowdsourcing control : Moving beyond multiple choice. *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)* (2012), 491–500. arXiv:arXiv preprint arXiv:1210.4870.

[9] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. *European Conference on Computer Vision (ECCV)* 8693 LNCS, PART 5 (2014), 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[10] E. Natonek. 1998. Fast range image segmentation for servicing robots. In *Proceedings. 1998 IEEE International Conference on Robotics and Automation (Cat. No.98CH36146)*, Vol. 1. 406–411 vol.1. https://doi.org/10.1109/ROBOT.1998.676445

[11] Olga Russakovsky, Li-Jia Li, and Li Fei-Fei. 2015. Best of Both Worlds: Human-Machine Collaboration for Object Annotation. (2015), 2121–2131.

[12] Mehrnoosh Sameki, Danna Gurari, and Margrit Betke. 2015. Characterizing Image Segmentation Behavior of the Crowd. (2015), 1–4.

[13] Jean Y. Song, Raymond Fok, Alan Lundgard, Fang Yang, Juho Kim, and Walter S. Lasecki. 2018. Two Tools are Better Than One : Tool Diversity as a Means of Improving Aggregate Crowd Performance. *Proceedings of the International Conference on Intelligent User Interfaces* (2018).

[14] Alexander Sorokin and David Forsyth. 2008. Utility data annotaton with Amazon Mechanical Turk. *Proceedings of the 1st IEEE Workshop on Internet Vision at CVPR 08* c (2008), 1–8. https://doi.org/10.1109/CVPRW.2008.4562953

[15] Antonio Torralba, Bryan C. Russell, and Jenny Yuen. 2010. LabelMe: Online image annotation and applications. *Proc. IEEE* 98, 8 (2010), 1467–1484. https://doi.org/10.1109/JPROC.2010.2050290

[16] Sirion Vittayakorn and James Hays. 2011. Quality Assessment for Crowdsourced Object Annotations. *Procedings of the British Machine Vision Conference* (2011), 109.1–109.11. https://doi.org/10.5244/C.25.109

[17] Peter Welinder, Steve Branson, Serge Belongie, and Pietro Perona. 2010. The Multidimensional Wisdom of Crowds. *NIPS (Conference on Neural Information Processing Systems)* 6 (2010), 1–9. https://doi.org/10.1.1.231.1538

[18] Kota Yamaguchi. 2012. Parsing Clothing in Fashion Photographs. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (CVPR '12)*. IEEE Computer Society, Washington, DC, USA, 3570–3577. http://dl.acm.org/citation.cfm?id=2354409.2355126