Aggregating Crowdsourced Image Segmentations

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

Abstract

Instance-level image segmentation provides rich information crucial for scene understanding in a variety of real-world applications. In this paper, we evaluate multiple crowdsourced algorithms for the image segmentation problem, including novel worker-aggregation-based methods and retrieval-based methods from prior work. We characterize the different types of worker errors observed in crowdsourced segmentation, and present a clustering algorithm as a preprocessing step that is able to capture and eliminate errors arising due to workers having different semantic perspectives. We demonstrate that aggregation-based algorithms attain higher accuracies than existing retrieval-based approaches, while scaling better with increasing numbers of worker segmentations.

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 (Natonek 1998), image organization and retrieval (Yamaguchi 2012), and medicine (Irshad and et. al. 2014). To this end, there has been a lot of work on employing crowdsourcing to generate training data for segmentation, including Pascal-VOC (Everingham *et al.* 2015), LabelMe (Torralba *et al.* 2010), OpenSurfaces (Bell *et al.* 2015), and MS-COCO (Lin *et al.* 2012). Unfortunately, raw data collected from the crowd is known to be noisy due to varying degrees of worker skills, attention, and motivation (Bell *et al.* 2014; Welinder *et al.* 2010).

To deal with these challenges, many have employed heuristics indicative of crowdsourced segmentation quality to pick the best worker-provided segmentation (Sorokin and Forsyth 2008; Vittayakorn and Hays 2011). 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 such errors in Section 3 and propose a clustering-based preprocessing technique that resolves them in Section 5.

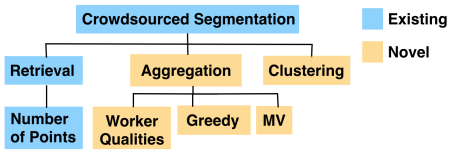

Figure 1: Taxonomy of quality evaluation algorithms for crowdsourced segmentation, including existing methods (blue) and our novel algorithms (yellow).

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 (Vittayakorn and Hays 2011; Russakovsky *et al.* 2015), and clickstream behavior (Cabezas *et al.* 2015; Sameki *et al.* 2015; Sorokin and Forsyth 2008).
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.

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 Error:** 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 that eliminates semantic ambiguities and errors. We present our experimental evaluation in Section 6.

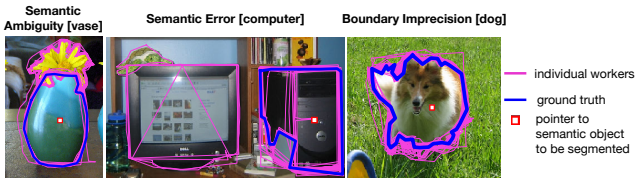

Figure 2: Examples of common worker errors.

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 (left), we display three worker segmentations for a toy example with 6 resulting tiles. Any subset of these tiles can contribute towards the final segmentation.

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 the worker has 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 workers.

Aggregation: Expectation-Maximization (EM)
Unlike MV, which assumes that all workers perform uniformly, EM approaches infer the likelihood that a tile is part of the ground truth segmentation, while simultaneously estimating hidden worker qualities. In Section 6 we evaluate an EM variant which assumes that each worker has a (different) fixed probability for a correct vote. Details of this, and more fine-grained variants can be found in our technical report (Lee *et al.* 2018).

Aggregation: Greedy Tile Picking (greedy)
The greedy algorithm picks tiles in descending order of the tiles' ratios of (estimated) overlap area with the ground truth to (estimated) non-overlap area with ground truth, for as long as the (estimated) Jaccard similarity of the resulting segmentation continues to increase. Intuitively, tiles that have a high overlap area and low non-overlap area contribute to high recall with limited loss of precision. Since tile overlap and non-overlap areas, and Jaccard similarity of segmentations with ground truth are unknown, we use different heuristics to estimate these values. We discuss details of this algorithm and its theoretical guarantees in our technical report.

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 (Vittayakorn and Hays 2011; Sorokin and Forsyth 2008).

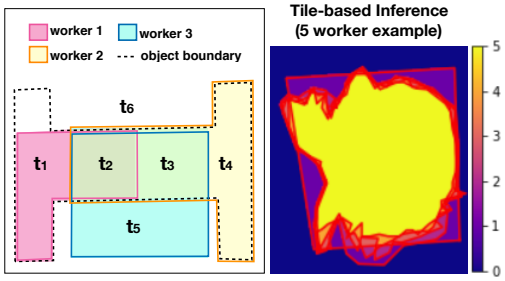

Figure 3: Left: Toy example demonstrating tiles created by three workers' segmentations around an object delineated by the black dotted line. Right: Segmentation boundaries drawn by five workers shown in red. Overlaid segmentation creates a mask where the color indicates the number of workers who voted for the tile region.

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 a 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.

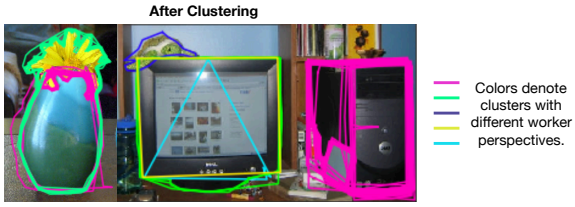

Figure 4: Example image showing clustering performed on the same object from Figure 2 left and middle.

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. Workers were compensated $0.05 per task. There were a total of 46 objects in 9 images from the MSCOCO dataset (Lin *et al.* 2014) 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 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). We use the Jaccard score Jaccard (J) $= \frac{UA(S)}{IA(S)}$, which accounts for the intersection area, $IA = area(S \cap GT)$ and union area, $UA = area(S \cup GT)$ between the worker and ground truth segmentations.

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

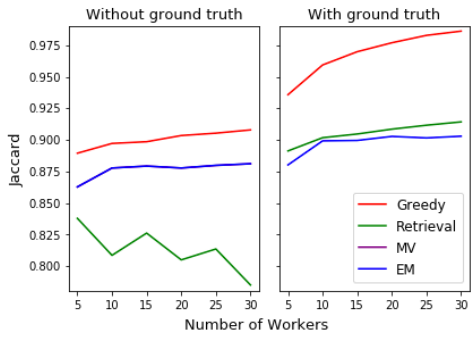

Figure 5: Performance of the original algorithms that do not make use of ground truth information (Left) and ones that do (Right). Here, the EM result overlaps with MV as they exhibit similar performance. Other diverging variants of EM is described in our technical report.

In Figure 5, 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 5 (left) shows that the performance of aggregation-based algorithms (greedy, EM) exceeds the best achievable through existing retrieval-based methods (Retrieval). Then, in Figure 5 (right), we estimate the upper-bound performance of each algorithm by assuming that 'full information' based on ground truth is given to the algorithm. For greedy, the algorithm is aware of all the actual tile overlap and non-overlap areas against ground truth. For EM, 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. By making use of ground truth information (Figure 5 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 shows the average percentage change in performance between 5-workers and 30-workers samples. We observe that aggregation based methods typically improve in performance with an increase in number of workers, while this is not generally true for retrieval-based methods.

Experiment 2: Clustering as preprocessing improves algorithmic performance.

The second row in Table 1 shows the average percentage Jaccard change when clustering preprocessing is used. While clustering generally results in an accuracy increase, since the 'full information' variants are already free of semantic errors, we do not see further improvement for these variants.

	Retrieval-based		Aggregation-based			
Algorithm	num pts	worker*	MV	EM	greedy	greedy*
Worker Scaling	-6.30	2.58	2.12	1.78	2.07	5.38
Clustering Effect	5.92	-0.02	2.05	0.03	5.73	0.283

Table 1: Jaccard percentage change due to worker scaling and clustering. Algorithms with * use 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.

References

[Bell *et al.* 2014] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Trans. on Graphics (SIGGRAPH)*, 33(4), 2014.

[Bell *et al.* 2015] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[Cabezas *et al.* 2015] Ferran Cabezas, Axel Carlier, Vincent Charvillat, Amaia Salvador, and Xavier Giro-I-Nieto. Quality control in crowdsourced object segmentation. *Proceedings of International Conference on Image Processing, ICIP*, 2015-Decem:4243–4247, 2015.

[Everingham *et al.* 2015] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, January 2015.

[Irshad and et. al. 2014] H Irshad and Montaser-Kouhsari et. al. Crowdsourcing Image Annotation for Nucleus Detection and Segmentation in Computational Pathology: Evaluating Experts, Automated Methods, and the Crowd. *Biocomputing 2015*, pages 294–305, 2014.

[Lee *et al.* 2018] Doris Jung-Lin Lee, Akash Das Sarma, and Aditya Parameswaran. Aggregating crowdsourced image segmentations. Technical report, Stanford InfoLab (ilpubs.stanford.edu:8090/1161/), 2018.

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

[Lin *et al.* 2014] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. *European Conference on Computer Vision (ECCV)*, 8693 LNCS(PART 5):740–755, 2014.

[Natonek 1998] E. Natonek. Fast range image segmentation for servicing robots. In *Proceedings. 1998 IEEE International Conference on Robotics and Automation (Cat. No.98CH36146)*, volume 1, pages 406–411 vol.1, May 1998.

[Russakovsky *et al.* 2015] Olga Russakovsky, Li-Jia Li, and Li Fei-Fei. Best of Both Worlds: Human-Machine Collaboration for Object Annotation. pages 2121–2131, 2015.

[Sameki *et al.* 2015] Mehrnoosh Sameki, Danna Gurari, and Margrit Betke. Characterizing Image Segmentation Behavior of the Crowd. pages 1–4, 2015.

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

[Torralba *et al.* 2010] Antonio Torralba, Bryan C. Russell, and Jenny Yuen. LabelMe: Online image annotation and applications. *Proceedings of the IEEE*, 98(8):1467–1484, 2010.

[Vittayakorn and Hays 2011] Sirion Vittayakorn and James Hays. Quality Assessment for Crowdsourced Object Annotations. *Procedings of the British Machine Vision Conference*, pages 109.1–109.11, 2011.

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

[Yamaguchi 2012] Kota Yamaguchi. Parsing clothing in fashion photographs. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, pages 3570–3577, Washington, DC, USA, 2012. IEEE Computer Society.