Quality Evaluation Methods for Crowdsourced Image Segmentation

Doris Jung-Lin Lee, Akash Das Sarma, Aditya Parameswaran

Abstract

Instance-level image segmentation provides rich information crucial for scene understanding in a variety of real-world applications. In this paper, we propose and evaluate several crowdsourced algorithms, including novel worker-aggregation based algorithms and retrieval-based methods based on prior work, for the image segmentation problem. We characterize the different types of worker errors observed, and present a clustering algorithm that is able to capture semantic errors and filter workers with different semantic perspectives. We demonstrate that aggregation-based algorithms attain better performance than existing retrieval-based approaches, while scaling better with increasing numbers of collected 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 and autonomous vehicles, surveillance, image organization and retrieval, and medicine (Irshad and et. al. 2014; Yamaguchi 2012). To this end, there has been a lot of work on employing crowdsourcing to generate training data for computer vision, 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 crowdsourced image processing tasks are known to be noisy due to varying degrees of worker skills, attention, and motivation (Bell *et al.* 2014; Welinder *et al.* 2010b).

In order to deal with these challenges, many have employed heuristics indicative of segmentation quality (Cabezas *et al.* 2015; Sameki *et al.* 2015; Sorokin and Forsyth 2008). While this approach identifies the segmentation drawn by more diligent workers, since it simply picks one *best* bounding box as the solution, it ends up discarding the majority of the worker responses and is limited by what the best worker can do. In this paper, we introduce a novel class of aggregation-based methods, capable of incorporating portions of responses from multiple workers into a combined segmentation and compare its performance with existing retrival-based methods. In addition, we propose a preprocessing technique that resolves different worker perspectives in multiple segmentations.

2 Related Work

Many large-scale efforts in crowdsourced image segmentation contain little to no information on the quality characterization and evaluation of the collected dataset (Torralba *et al.* 2010; Martin *et al.* 2001; Li *et al.* 2009; Gurari *et al.* 2015), which indicate the lack of standardized approaches for quality evaluation. As shown in Figure 1, we break down the existing quality evaluation methods into several categories:

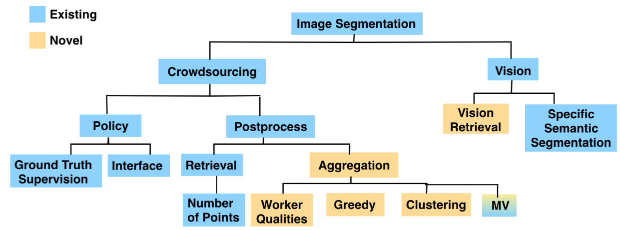

Figure 1: Flowchart summarizing the classes of existing algorithms for image segmentation (blue) and a novel class of algorithms proposed in this paper (yellow). Majority-vote (MV) is colored both blue and yellow, since a common algorithm in crowdsourcing literature, but have not been extensively applied to crowdsourced image segmentation.

Policy-based methods include specialized segmentation interfaces or workflows that ensures that the data collected are of good quality, including periodic verification workflows (Lin *et al.* 2014; Everingham *et al.* 2015), specialized segmentation interfaces (Song *et al.* 2018), and vision supervision of crowdsourced segmentation(Russakovsky *et al.* 2015; Gurari *et al.* 2016).

Retrieval-based methods seek to pick the "best" worker segmentation based on some scoring criteria that evaluates the quality of each segmentation, including the use of vision information (Vittayakorn and Hays 2011; Russakovsky *et al.* 2015), expectation-maximization (EM) approaches for bounding box quality estimation (Welinder *et al.* 2010b), and click-stream behavior(Cabezas *et al.* 2015; Sameki *et al.* 2015; Sorokin and Forsyth 2008).

Aggregation-based methods use multiple worker segmentations to produce a single combined segmentation. Aggregation-based majority vote have been introduced in Sameki et al. (2015) as a way for aggregating expert segmentations to obtain a ground truth segmentation for evaluation purposes.

3 Preliminaries

3.1 Data & Goals

We collected crowdsourced segmentations from Amazon Mechanical Turk where each HIT consisted of one segmentation task for a specific pre-labeled object in the image. There

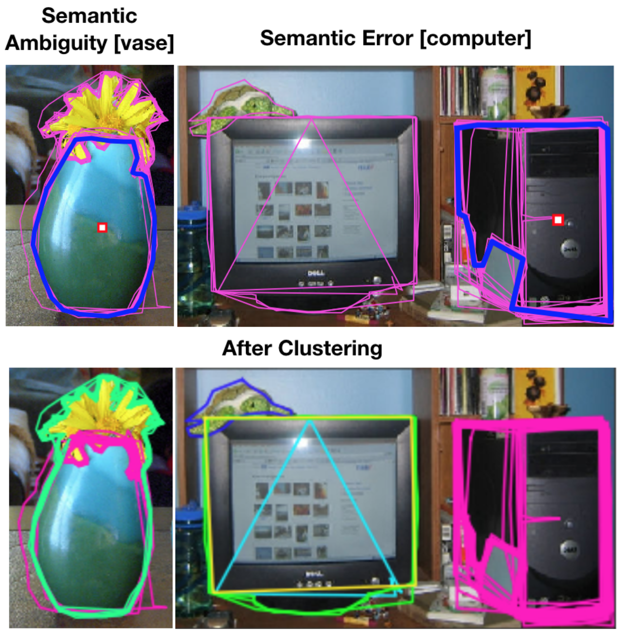

Figure 2: Top: Pink is the segmentation from individual workers. Blue solid line delineates the ground truth. The red boxed pointer is the interface icon indicating the semantic object to be segmented. Bottom: Boundary colors highlight different worker perspectives resulting from clustering.

were a total of 46 objects in 9 images from the MSCOCO dataset (Lin *et al.* 2014). For each object, we collected segmentation masks from a total of 40 workers. Each task contains a semantic keyword and a pointer indicating the object to be segmented.

3.2 Evaluation Metrics

Evaluation metrics used in our experiment measures how well the final segmentation (S) produced by these algorithms compare against ground truth (GT). The most common evaluation metric used in literature are area-based methods which take into account the intersection, $IA = area(S \cap GT)$, or union, $UA = area(S \cup GT)$, between the user and the ground truth segmentations. Specifically, we use Precision (P) $= \frac{IA(S)}{area(S)}$, Recall (R) $= \frac{IA(S)}{area(GT)}$, and Jaccard (J) $= \frac{UA(S)}{IA(S)}$ metrics to evaluate our algorithms.

3.3 Error Analysis

Common worker errors can be classified into three types: (1) **Semantic Ambiguity:** differing opinions on whether particular regions belong to part of an object (Figure 2 Left); or (2) **Semantic Mistakes:** annotate the wrong object entirely (Figure 2 right); or (3) **Boundary Imprecision:** 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). In the following section, we will discuss a preprocessing method that we have developed to resolve semantic ambiguity and errors observed in prior work (Sorokin and Forsyth 2008; Lin *et al.* 2014; Gurari *et al.* 2018). Since quality evaluation in past literature have been largely focused on minimizing boundary precision issues, we will then describe novel aggregation-based algorithms that we have developed for this purpose and compare them with existing retrieval-based methods.

4 Perspective Resolution in Crowdsourced Image Segmentation

4.1 Worker Clustering

Our clustering-based approach is based on the intuition that workers with similar perspectives will have segmentations that are closer to each other. We capture the similarity between a pair of workers by computing the Jaccard score between their segmentations and perform spectral clustering to separate workers into clusters. Figure 2 bottom illustrates how spectral clustering is capable of dividing the worker responses into clusters with meaningful semantic associations, reflecting the crowd's diversity of perspectives in completing same task.

Clustering offers an additional benefit of preserving worker's semantic intentions in the case where there are multiple instances of different errors. For example, the mistakened clusters in Figure 2 bottom right contained semantic concepts "monitor" and "turtle". While these are considered *bad* annotations for this particular task, this cluster of annotation can provide more data for another semantic segmentation task "monitor". A potential future work includes adding additional crowdsourcing tasks for semantic labeling of clusters (which is cheaper and more accurate than segmentation) to enable reuse of annotations across multiple objects and lower the cost of data collection.

5 Precision-focused algorithms: Aggregation v.s. Retrieval Comparison

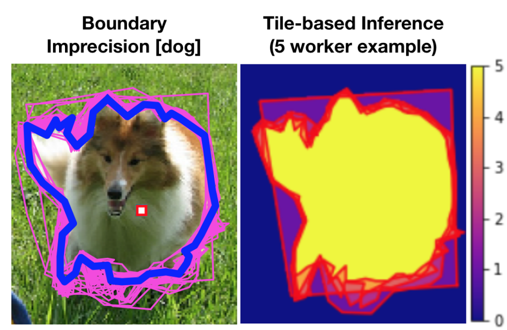

Figure 3: Left: Pink boundaries shows worker segmentations and blue delineates the ground truth. Right: Segmentation boundaries drawn by five workers shown in red. Overlaid segmentation creates a masks where the color indicates the number of workers who voted for the tile region.

At the heart of our aggregation techniques is the "tile" data representation, where we logically overlay all workers' segmentations on top of each other, as illustrated in Figure 3 right, to create non-overlapping discrete tile units. The intuition here is that by splitting the image into tiles, we get finer granularity information than by looking at complete segmentations. This also allows us to aggregate data from multiple workers rather than having to choose a single worker bounding box—enabling the opportunity to choose partial segmentations by fixing one worker's errors via the help from another worker's segmentation. Now, we will describe several algorithms for picking a good set of tiles.

Aggregation: Majority Vote Aggregation (MV)
Include tiles in the output segmentation if and only if the tile is covered by at least 50% of all worker segmentations.

Aggregation: Expectation-Maximization (EM)
Unlike MV, which assumes that all workers performs uniformly, EM approaches use worker quality models to infer the likelihood that a tile is part of the ground truth segmentation. The EM algorithm simultaneously estimate both worker qualities and tile likelihoods as hidden variables. Details of the formal derivation and three worker quality models that we have developed can be found in our technical report.

Aggregation: Greedy Tile Picking (greedy)
Using tile probabilities from EM to estimate intersection area between ground truth and tile, then greedily pick tiles with the largest intersection area ratio until Jaccard score begins to decrease (effectively picking the largest and most probable tiles that should be included first). The Jaccard score is computed between the merged output from the selected set of tiles and MV segmentation.

Retrieval: Number of Control Points (num pts)
Pick the worker segmentation with the largest number of control points around the segmentation boundary (i.e. most precise drawing) as the output segmentation.

5.1 Retrieval v.s. Aggregation-based Comparison

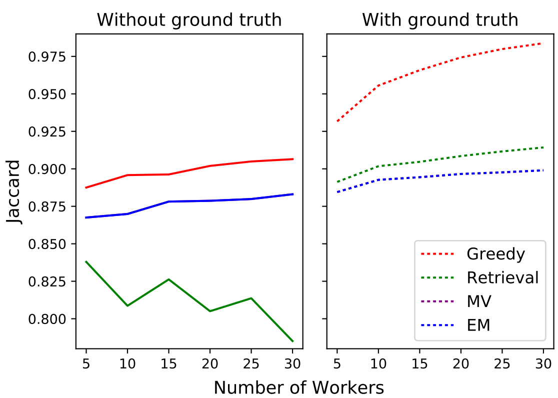

Figure 4: Jaccard performance comparison between best-performing algorithms from retrieval and aggregation-based methods with clustering as a preprocessing step where possible. We compare between the original algorithms that do not make use of ground truth information (Left) and ones that do (Right). Note that MV and EM results are so close that they overlay on each other.

Aggregation-based methods performs significantly better than retrieval-based methods
Figure 4 left shows that amongst the algorithms that do not make use of ground truth information, the performance of aggregation-based algorithms (greedy, EM) exceeds the best achievable through the existing retrieval-based method (num pts). 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 than any individual worker segmentation.

Performance of aggregation-based methods scales well as more workers segmentation are added.
Intuitively, larger worker samples results in finer granularity tiles for the aggregation-based methods, resulting in an monotonically increasing relationship between number of worker segmentation used in the sample and performance evident in Table 1. However, worker scaling for retrieval-based methods are not guaranteed.

	Retrieval-based		Aggregation-based			
Algorithm	num pts	worker*	MV	EM	greedy	greedy*
Worker Scaling	-6.30	2.58	1.63	1.64	2.16	5.59
Clustering Effect	5.92	-0.02	2.05	1.38	5.55	-0.06

Table 1: The first row lists the average percentage change in Jaccard between 5 workers samples and 30 workers sample. The second row lists the average percentage change between the no clustering and clustering results. Algorithms with * makes use of ground truth information.

Clustering as preprocessing improves algorithmic performance.
Clustering results can also be used as a preprocessing step to any of the quality evaluation algorithms by keeping only the segmentations that belong to the largest cluster, which is typically free of any semantic errors. As shown in Table 1, on average, clustering generally results in an increase the resulting algorithmic performance. Since the ground-truth supervised variants are already free of semantic ambiguity and errors, there is minimal improvement resulting from clustering.

6 Conclusion & Future Work

In this paper, we identified three different types of errors for crowdsourced image segmentation, and developed a clustering-based method capture the semantic diversity caused by differing worker perspectives. Moreover, we introduced novel aggregation-based methods that performed better when compared to existing retrieval-based methods.

Given the success of aggregation-based methods, our future work investigates why and how our EM and greedy approaches differ from majority vote. Our preliminary studies show that worker qualities are good indicators of actual Jaccard of worker segmentations with ground truth and that the greedy algorithm is capable of achieving close-to-perfect segmentation accuracy with ground truth information. We have also explored using vision information to improve these algorithms. Bridging the gap between our current approach to the maximum potential of aggregation-based methods can result in more accurate and perspective-aware crowdsourced segmentation outputs in the future.

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.

[Gurari *et al.* 2015] Danna Gurari, Diane Theriault, Mehrnoosh Sameki, Brett Isenberg, Tuan A. Pham, Alberto Purwada, Patricia Solski, Matthew Walker, Chentian Zhang, Joyce Y. Wong, and Margrit Betke. How to collect segmentations for biomedical images? A benchmark evaluating the performance of experts, crowdsourced non-experts, and algorithms. *Proceedings of 2015 IEEE Winter Conference on Applications of Computer Vision (WACV) 2015*, pages 1169–1176, 2015.

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

[Gurari *et al.* 2018] Danna Gurari, Kun He, Bo Xiong, Jianming Zhang, Mehrnoosh Sameki, Suyog Dutt Jain, Stan Sclaroff, Margrit Betke, and Kristen Grauman. Predicting Foreground Object Ambiguity and Efficiently Crowdsourcing the Segmentation(s). *International Journal of Computer Vision (IJCV)*, 2018.

[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.

[Li *et al.* 2009] Li Jia Li, Richard Socher, and Li Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*, pages 2036–2043, 2009.

[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.

[Martin *et al.* 2001] 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 *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423, July 2001.

[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.

[Song *et al.* 2018] Jean Y. Song, Raymond Fok, Alan Lundgard, Fang Yang, Juho Kim, and Walter S. Lasecki. Two Tools are Better Than One : Tool Diversity as a Means of Improving Aggregate Crowd Performance. *IUI'18: Proceedings of the International Conference on Intelligent User Interfaces*, 2018.

[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 2011*, pages 109.1–109.11, 2011.

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

[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.