Quality Evaluation Methods for Crowdsourced Image Segmentation

Authors removed for anonymity

Abstract

Instance-level image segmentation provides rich information crucial for scene understanding in a variety of real-world applications, such as robotics and surveillance. 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 also 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 attains better performance than existing retrieval-based approaches, while scaling better with increasing numbers of collected worker segmentations.

1 Introduction

The goal of visual scene understanding is to enable computers to achieve high-level comprehension from images or videos. While object localization and detection identifies *where* an object is in the image, object segmentation provides rich information regarding *what* the object looks like. Precise, instance-level object segmentation is crucial for identifying and/or 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). Unfortunately, at present, vision-based object segmentation algorithms often suffer from oversegmented regions and perform poorly for occluded (hidden) objects (Torralba *et al.* 2010), for cluttered images with many objects (Russakovsky *et al.* 2015), and under undesirable lighting conditions (Bell *et al.* 2015).

To this end, there has been a lot of work on employing crowdsourcing to generate training data for computer vision. In particular, a number of popular computer vision datasets involve fine-grained segmentation derived from crowdsourcing, including Pascal-VOC (Everingham *et al.* 2015), LabelMe (Torralba *et al.* 2010), OpenSurfaces (Bell *et al.* 2015), and MS-COCO (Lin *et al.* 2012). Indeed, fine-grained segmentations are more valuable than more coarse-grained approaches: Kovashka et al. (2016) state that *"detailed annotations enable the development and evaluation of computer vision algorithms that are able to understand the image on a much finer level than what is possible with just simple binary image-level annotations or with rough bounding box-level localization."*

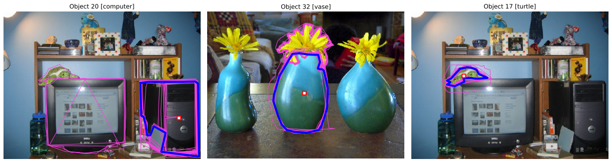

Figure 1: Pink is the segmentation from individual workers. Blue solid line delineates the ground truth. The red boxed pointer indicates the task of interest shown to users. Examples demonstrating common error patterns among crowdsourced image segmentation, including 1) annotations on the wrong semantic object, 2) ambiguity in regional inclusion and exclusion, and 3) imprecision at the object boundary.

Unfortunately, the use of crowdsourcing for fine-grained segmentation is rife with challenges. As shown in Figure 1, workers often (i) make unintentional mistakes while drawing the boundaries, either due to low precision of the image, small area of the object, or lack of drawing skills, (ii) have differing opinions about what constitutes the boundary of an object (e.g., is the stalk of a banana part of the fruit?); (iii) or annotate the wrong object entirely (e.g., drawing a bounding box around a dog when the task requests for one around a car). Doris: make the e.g. lines align with what's in the figure.

In order to deal with these challenges, many have employed heuristics indicative of segmentation quality, including the idea that segmentation boundaries with more points are likely to be more precise (Cabezas *et al.* 2015; Sameki *et al.* 2015; Sorokin and Forsyth 2008).While this approach does enable us to identify the more diligent workers and their annotations, since it is simply picking one "best" bounding box based on these heuristics as the solution, it ends up discarding the majority of the worker responses. Aditya: Comment on whether the focus should be evaluation or new algos (i.e. aggregation based methods) Instead, in this paper, we introduced the notion of an *aggregation-based* approach capable of incorporating portions of responses from different workers into an overall segmentation. By overlaying the segmentations from workers on top of each other, we can decompose the image into non-overlapping tiles, where each tile has some workers who believe this tile belongs to the object, and others who do not. At this point, we can treat each tile as an independent boolean question, deriving an answer from a worker—does this tile belong to the object or not, following which we may be able to apply Expectation-Maximization (EM) (Dawid and Skene 1979) to derive maximum likelihood tiles and worker accuracies, a greedy approach for tile picking based on worker fraction votes, and simple majority vote aggregation. Aditya: this sounds too simple. I propose we describe the entire set of alternative algos we evaluate Doris: maybe talk about perspective clustering, greedy + advanced worker model intuition here? depends on our story The contributions of this paper is as follows:

- Our paper is the first to formulate crowdsourced segmentation problem in terms of aggregating worker bounding boxes at a tile level. Working at this sub-worker-segmentation granularity enables better performance scaling with the number of annotations collected as well as better overall accuracy.
- We provide a survey, comparison, and evaluation of existing approaches for crowdsourced image segmentation, while focusing on the comparison between aggregation-based methods and the existing retrieval-based models.
- We formally characterize the types of worker error in crowdsourced image segmentation and resolve the well-known multiple perspective issue in crowdsourced image segmentation through the aggregation-based approach of spectral clustering, which can be applied as a preprocessing step to any quality evaluation methods.

2 Related Work

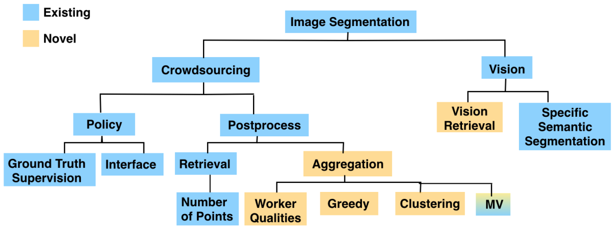

Figure 2: Flowchart summarizing the classes of existing algorithms for image segmentation (blue) and a novel class of algorithms proposed in this paper (yellow). Majority-vote (MV) is colored both blue and yellow, since a common algorithm in crowdsourcing literature, but have not been extensively applied to crowdsourced image segmentation.

Many large-scale efforts in 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 indicates the lack of standardized approaches for quality evaluation in crowdsourced image segmentation. As shown in Figure 2, we break down the existing quality evaluation methods into several categories:

Policy-based methods Policy-based quality evaluation methods are specialized segmentation interfaces or workflows that ensures that the data collected are of good quality. Several large-scale crowdsourced segmentation efforts have employed verification techniques within the data collection workflow, such as supervising workers periodically by evaluating worker responses against known ground-truth segmentation during data collection (Lin *et al.* 2014; Everingham *et al.* 2015). Common scoring functions for characterize the quality of the worker's segmentation against the ground truth includes segmentation accuracy(Everingham *et al.* 2015) or Jaccard index (Sameki *et al.* 2015; Gurari *et al.* 2016). Specialized interfaces for object segmentation have also been developed to ensure high-quality data collection. For example, Song et al. (2018) makes use of worker's responses from four different image segmentation interfaces to derive an aggregated bounding box. Other segmentation workflows also employ vision information to supervise the crowdsourced response(Russakovsky *et al.* 2015; Gurari *et al.* 2016).

Since these policy-based methods are interface-dependent, require expensive expert-drawn ground-truth annotations or vision information, the results are not easily reproducible. In addition, the segmentations collected by the simple click-and-draw interface in many of the large scale segmentation efforts can not be improved with this technique as a post-processing method. Due to the lack of reproducibility, our paper do not compare against these policy-based methods in extensive details.

Retreival-based methods Retreival-based methods seeks to pick the "best" worker segmentation based on some scoring criteria. (Vittayakorn and Hays 2011) proposed several heuristic scoring functions that included vision information such as edge detection or color mixture to assess the quality of crowdsourced segmentations. (Vittayakorn and Hays 2011) provides a comparison of retrieval-based algorithms for image segmentation. Doris: Need to describe Vittyakorn et al in more detail Other heuristic approaches that don't require ground truth segmentations include characterizing the user types and studying their click-stream behavior to determine work quality(Cabezas *et al.* 2015; Sameki *et al.* 2015; Sorokin and Forsyth 2008), and comparing the worker responses to features extracted from computer vision algorithms(Vittayakorn and Hays 2011; Russakovsky *et al.* 2015). As we demonstrate in this paper, aggregating worker responses not only eliminates the need for costly expert segmentations, but also leads to better results than individual workers alone.

Doris: this paragraph is probably not necessary, if so merge a couple sentence into retreival sectionThe other methods we evaluated draw on prior work in worker quality evaluation adapted to the segmentation context. Starting from the seminal work of Dawid and Skene (Dawid and Skene 1979), applying Expectation-Maximization (EM) to derive both worker accuracies and true answersAditya: not a complete sentence.. Welinder and Perona (Welinder and Perona 2010a) extend this Aditya: what? to separately model worker quality and the biases applied to binary, multivalued and continuous annotations. Welinder and Perona (Welinder *et al.*

2010b) develop a multidimensional array of worker accuracies and task difficulties by considering object-presence labeling as a noise generation processAditya: do you mean a noisy generative process?. The objective truth label is captured by a multidimensional quantity of task-specific measurements and deformed by worker and image related noise, the noisy vector obtained after this process is projected onto the vector of user expertise (which summarizes how well the user perceives each of these measurements), and finally the score is binarized into an inferred label Aditya: fairly confusing and really long sentence.

Aggregation-based methods Pixel-based majority vote Aditya: since you may not have defined what pixel based majority vote is until this point, i would explicitly describe it was introduced in (Sameki *et al.* 2015) as a way for aggregating expert-bounding boxes to obtain the ground truth, rather than for aggregating worker segmentations. Many have extended this line of work beyond binary classifications by developing EM-like approaches that works on multiple-choice (Karger *et al.* 2013) as well as free-form responses (Lin *et al.* 2012), but these have not been directly applied to the task of object segmentation.

However, while EM algorithms assign probabilities regarding *how good a worker's segmentation is*, for the task of object segmentation, we are ultimately more interested the end goal of *what is the best segmentation that we can get from these data*. To our knowledge, our work is the first to compare various aggregation-based approaches for segmentation quality evaluation extensively against existing approaches, and study how clustering could be used for the issues of multiple perspectives described in (Sorokin and Forsyth 2008; Lin *et al.* 2014; Gurari *et al.* 2018), for crowdsourced image segmentationAditya: what is described? the issue of perspectives? or clustering?.

Vision-based approaches Doris: Describe state-of-the-art precise instance-level segmentation (Facebook Detectron, CRFs). Argue that crowd response is still important

3 Goal and Data

3.1 Goal

For any specified object in an image, there exists a *tight*Doris: why is specifying "tight" segmentation? Not sure purpose of this paragraph segmentation which exactly outlines the object; we call this the *ground-truth* segmentation for this object. Workers are asked to provide segmentations for objects; they often do not provide the ground-truth segmentation, and their segmentation is often noisy. Thus our goal is the following: given a raw image and multiple noisy worker segmentations for a specific object, estimate the ground truth segmentation for that object.

Show one image + ground truth here

3.2 Data Collection

We collected crowdsourced segmentation data from Amazon Mechanical Turk where each HIT consisted of one segmentation task for a specific pre-labeled object in the image. There 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 worker was paid 5 cents per annotation. After eliminating segmentation masks that contains self-intersecting polygon contours, our final dataset contains 1784 bounding boxes made by 198 unique workers. As shown in Fig.3, each task contains a semantic keyword and a pointer indicating the object to be segmented. These tasks represent a diverse set of task difficulty (different levels of clutteredness, occlusion, lighting) and levels of task ambiguity.

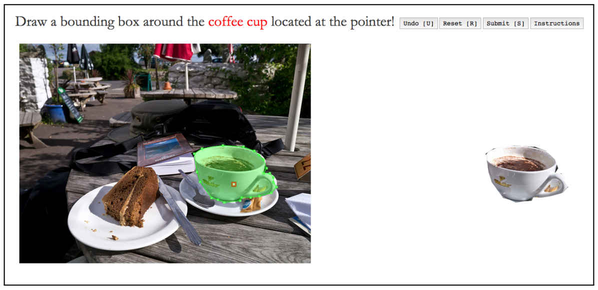
Draw a bounding box around the coffee cup located at the pointer!

Figure 3: An example interface for the segmentation webapp can be seen here.

3.3 Worker Error patterns

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). The average precision, recall and Jaccard similarity of the worker segmentation against the ground truth across all of the objects was 92%, 94%, 86% respectively. While workers have equal rates of overbounding or underbounding behavior, workers tend to overbound by a larger amount (7212 pixels on average) than underbound (1306 pixels on average).

Visual examination of worker bounding boxes reveals several common error patterns evident across different objects. As shown in the example in Figure 1, common worker errors can be classified into three types:

1. **Semantic error:** Workers annotate the wrong semantic object.
2. **Regional semantic ambiguity:**Workers annotate the correct semantic object, but included a portion connected to that object that should not have been included as part of the annotation.
3. **Boundary imprecision:** Workers annotate the correct semantic object, but segmentation boundaries are imprecise.

Type 1 and 2 errors have also been observed in prior work (Sorokin and Forsyth 2008; Lin *et al.* 2014; Gurari *et al.* 2018), which noted that disagreement in worker responses can come from questions that are ambiguous or difficult to answer, such as segmenting a individual person from a crowd. Since there are multiple workers annotating each object, each object can suffer from multiple types of error: we found that out of the 46 objects in our dataset, 9 objects suffer from type one error and 18 objects from type two error. Almost all objects suffer from some form of type three error of varying degree of imprecision around the object boundary. The main evaluation methods highlighted in this paper focuses on resolving the type three imprecise, “sloppy” bounding box errors. However, since type one and two errors are also fairly severe in contributing to the recall lost, we did not want to simply eliminate objects that suffer from these issues. We will discuss a preprocessing procedure used to address these errors in Section 4.

Aditya: after having read all this, i wonder if we want to simply move the dataset description to when the evaluation happens; here just say that we look at some example worker segmentations for some tasks that we issued and try to identify general principles that can motivate the design of the algorithms, described next.Doris: No, I agree that this would help highlight these observation as one of our contribution, this section needs to appear early on, to motivate why we have developed certain algos for resolving perspectives (e.g. clustering)

4 Methods

Image annotation problems, and in particular, the segmentation problem can be approached in many ways. In this section we classify and discuss several methods that we use to perform segmentation in images. At a broad level, segmentation can be performed using by computer-vision based methods or by using crowdsourcing. We discuss each of these approaches and outline multiple segmentation algorithms below. We evaluate all of these algorithms and report our findings on their performances in Section 6.

Give names to all methods that can be used in experiments section

Figure 2 depicts the classification of approaches as well as the specific algorithms that we will discuss below.

4.1 Vision-based methods

There has been a lot of prior work in segmenting objects based on color boundaries. These approaches, however, are typically non-exact, and far from robust. Furthermore, while they segment the entire image into several disjoint pieces, they do not serve to identify objects. Another class of prior works aim to find rectangular bounding boxes for objects of a specified type, for instance, cars in traffic surveillance images. Our goal, however, is find tight segmentations around specified objects. Doris: the comparison that we made in related works was v.s. specific object segmentation rather than rectangular BB object detection. Maybe we should not talk about rectangular BB to avoid confusion? Object segmentation using purely automated techniques would require training computer vision models on specific object types.

We implement a semi-supervised algorithm that can produce segmentations for arbitrary objects in the absence of large volumes of tailor-made training data. While this algorithm works largely on raw image data, it requires some external help in the form of one “reference” segmentation. Intuitively, a rough segmentation can be thought of as a pointer for the algorithm to the relevant regions of the image. The algorithm then uses the color profile of the image to segment out the similarly colored regions of the image that overlap with the reference segmentation. Specifically, we begin by splitting the input image into multiple regions, or *tiles* that have the same color using the work of (Felzenszwalb and Huttenlocher 2004)—the desired number of output tiles can be modified using a tuning parameter k, to produce finer or coarser tiles.Doris: would it be worthwhile to discuss tiles separately at the beginning of section 4 for both aggregation based methods and vision. Or use separate terminology in different the two cases. Figure 4 shows one example of the vision color tiling for different chosen granularities. Next, we overlay the given rough segmentation on top of the color tiles.

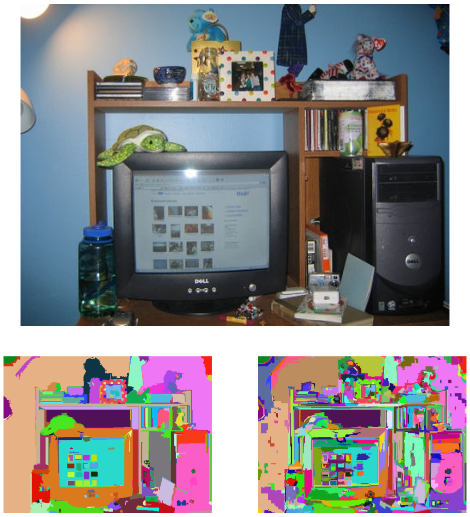

Figure 4: Left: Raw image. Center: Vision tiles with $k = 500$. Right: Vision tiles with $k = 100$.

img for ref segmentation overlaid on vision tiles (incorporate with Fig.4)

Now, the algorithm focuses on *choosing the right set of tiles based on the given reference segmentation*. Intuitively the algorithm picks color tiles that have significant overlap with the given reference segmentation, i.e., returns the union of all tiles for which greater than a certain area threshold of the tile is intersecting with the reference segmentation. We experiment with different granularities for the vision preprocessing as well as scan a variety of tile filtering area thresholds.

We also implemented a second algorithm that looks at this data from the other side, by trying to *fix or improve the boundaries of the given reference segmentation using the color tiles information*, discussed in more detail in our technical report.

4.2 Crowdsourcing methods

Crowdsourcing based approaches use human workers to provide segmentations for objects in images. Since human workers are not perfect and often make mistakes, crowdsourcing approaches typically elicit multiple worker segmentations for each object to reduce output variance resulting from the errors of an individual worker. As we saw in Section 3.3, different worker segmentations for the same object can differ from each other due to differences in perspective as well as errors in tracing the outline of the object. Crowdsourcing algorithms need to take these multiple differing worker segmentations as input and output a single, accurate segmentation.

First, we discuss a preprocessing step that helps identify and eliminate the semantic errors (described in Section 3.3) resulting from multiple perspectives.

Worker Clustering

Intuitively, workers that have similar perspectives, will have segmentations that are closer to each other, while workers that have different perspectives from each other will have segmentations that differ from each other. We capture the "similarity" of a pair of workers by computing the Jaccard coefficient between their segmentations. We perform *spectral clustering* to separate workers, using their pairwise similarities, into clusters. We find that the resulting clusters accurately separate and group workers based on their perspectives or the type of semantic errors they make. We also find that the largest cluster is typically free of any semantic errors. Therefore, our preprocessing step consists of clustering workers based on their mutual pairwise Jaccard similarity scores, and filtering away the workers that do not belong to the largest cluster.

Next, we discuss two classes of crowdsourcing algorithms that fundamentally differ in the way they handle multiple worker segmentations to generate a single output segmentation.

Worker retrieval methods

This class of algorithms tries to identify good and bad workers, and then chooses the best worker segmentation as the output segmentation. In this paper, we look at two different ways of ranking workers and choosing the best worker. First, we use the *number of control points*, i.e. number of vertices in a worker's segmentation polygon to rank workers. This is a ranking scheme that (Vittayakorn and Hays 2011) showed performs well in practice. Intuitively, workers that have used a larger number of points are likely to have been more precise, and provided a more complex and accurate segmentation. Other heuristic ranking scheme is described in more detail in our technical report (Anonymous 2018). In Section 5 we model worker qualities in terms of their probabilities of annotating pixels of the image correctly, and estimate the worker qualities by utilizing an Expectation-Maximization algorithm. We also use the estimated worker qualities (for different quality models) to predict segmentation accuracy and rank workers, and choose the best worker segmentation based on this ranking as another alternative algorithm.

Aggregation-based methods

Rather than simply identifying and picking a single worker's segmentation, aggregation-based methods seek to combine multiple workers' segmentations into a single merged segmentation. At the heart of all our aggregation techniques is the following data representation: we logically overlay all workers' segmentations on top of each other within the framework of the overall image. As illustrated in 5, the overlaid worker segmentations can be thought of as a Venn diagram that represents a partitioning of the entire image into multiple worker *tiles* formed by the intersections of different worker segmentations. We then choose and merge a subset of the tiles to give the final output segmentationDoris: vague. 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—this allows for the potential of choosing the best partial segmentations for an object and joining them, or fixing one worker's errors by taking the help of another worker's segmentation. The problem of choosing a good set of tiles is, however, non-trivial. Since aggregation based methods are the least studied methods by previous work, we discuss them in further detail in Section 5.

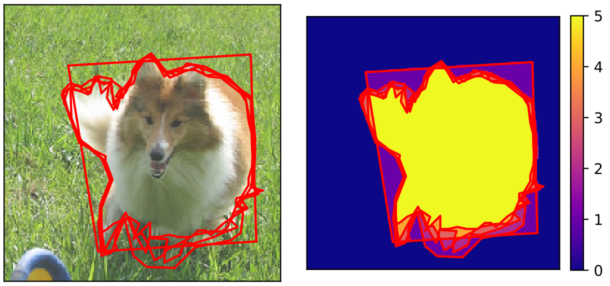

Figure 5: Left: Red boundaries shows the segmentation boundaries drawn by five workers overlaid on the image. Right: Segmentation boundaries still shown in red. The overlaid segmentation creates a masks where the color indicates the number of workers who voted for the tile region.

5 Aggregation Methods: A Closer Look

The key insight of aggregation-based methods is that they perform inference at a more fine-grained "tile" level, rather than at the bounding box level as is in the case of retrieval-based methods. In this section, we discuss three different algorithms for choosing a "good" set of worker tiles in aggregation-based methods.

5.1 Majority Vote Aggregation (MV)

A common strategy for aggregating multiple worker responses in crowdsourced algorithms is *Majority Vote*(MV)Doris: Akash: add citation for MV from general crowdsourcing domain. In the context of choosing worker tiles to produce a segmentation, the natural Majority Vote algorithm looks at each tile, and includes the tile in the output segmentation if and only if the tile is covered by at least 50% of all worker segmentations.

5.2 Expectation-Maximization

While Majority Vote is a very useful algorithm in practice, it does not distinguish between workers in any way. In reality, however, not all workers are equal. Now, we try to model worker quality, and use worker quality information to infer the likelihood that a tile is part of the ground truth segmentation. Since both, the worker qualities, as well as the likelihoods of tiles being part of the ground truth are hidden quantities, we employ an Expectation-Maximization based approach to simultaneously esimtate both of these sets of quantities.

techreport

We intuitively describe three worker models that we experiment with below. In our technical report, we formalize the notion of the probability that a set of tiles forms the ground truth, and solve the corresponding maximum likelihood problem, for each of these worker models.

Worker quality models.
We can think of workers as agents that look at each pixel in an image and label it as part of the segmentation, or not. Their actual segmentation is the union of all the pixels that they labeled as being part of their segmentations. Each pixel in the image is also either included in the ground truth segmentation or not included in the ground truth segmentation. We can now model worker segmentation as a set of boolean pixel-level (include or don't include) tasks, each having a ground truth boolean value. Based on this idea, we explore three worker quality models:

- *Basic model:* Each worker is captured by a single parameter Bernoulli model, $< q >$, which represents the probability that a worker will label an arbitrary pixel correctly.
- *Ground truth inclusion model (GT):* Two parameter Bernoulli model $< qp, qn >$, capturing false positive and false negative rates of a worker. This helps to separate between workers that tend to overbound and workers that tend to underbound segmentations.
- *Ground truth inclusion, large small area model (GTLSA):* Four parameter model $< qp_l, qn_l, qp_s, qn_s >$, that distinguishes between false positive and false negative rates for large and small tiles. In addition to capturing overbounding and underbounding tendencies, this model captures the fact that workers tend to make more mistakes on small tiles, and penalizes mistakes on large tiles more heavily.

5.3 Greedy Tile Picking

Doris: the terminology "overlap" can be a bit confusing with the abbrev that we chose, since overlap area would be OA (rather than outside area). Maybe introduce it as intersection area or introduce terms "inside" and "outside" to correspond with the abbrev OA,IA. Next, we present a greedy tile picking algorithm that grows the output set of tiles by adding in one tile at a time. Suppose tile t, overlaps with the ground truth segmentation with intersection area of $IA(t)$, and has area $OA(t)$ not overlapping with the ground truth. The greedy algorithm sorts tiles in decreasing order of their $\frac{IA(t)}{OA(t)}$ ratio and iteratively adds the next tile to the growing set of output tiles, until the Jaccard value of the current set of tiles will decrease with the next added tile. Doris: explain intuition of why I/O is used. The key idea behind this algorithm is the following statement

techreport

(proof available in our technical report): It can be shown that given a set of tiles, T, the tile t that maximizes Jaccard($T \cup t$) score of the union of the set of tiles against the ground truth, is the tile with maximum value of $\frac{IA(t)}{OA(t)}$. The primary challenge with this approach is that we do not know the actual $IA(t)$, $OA(t)$ values for any tile. We implement a heuristic version of this algorithm, where we estimate the intersection area of any tile, $IA(t)$, by using the fraction of workers that have voted for a tile, and greedily maximize for estimated Jaccard value at every step.

techreport

In our technical report, we also discuss variants of this algorithm where we use different techniques to estimate the intersection areas of tiles, resulting in corresponding variants of the greedy algorithm.

6 Results and Discussion

6.1 Experimental Setup

A sub-sampled dataset was created from the full dataset to determine the efficacy of these algorithms on varying number of worker responses. Every object was randomly sampled worker with replacement. For small worker samples, we average our results over larger number of batches than for large worker samples (which have lower variance, since the sample size is close to the original data size). The number of workers plotted indicates the number of distinct worker segmentation used in the algorithm, note that in the clustered cases, the actual number of segmentation in the cluster may be less than what is cited in the sample.

6.2 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, $IA = area(S \cup GT)$, or union, $UA = area(S \cap GT)$, 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.

6.3 What is the difference in performance between retrieval and aggregation-based methods?

Figure 6 shows the comparisons between the best performing algorithm amongst aggregation-based (greedy, EM) and retrieval-based (num points) algorithms. The solid line in Figure 6 shows algorithms that does not make use of ground truth information as part of the inference, while the dotted line shows the corresponding algorithm that makes use of

Retrieval-based			Aggregation-based			
num pts	avrg worker	best worker	MV	EM	greedy	best greedy
-6.30	-0.25	2.58	1.63	1.64	2.16	5.59

Table 1: Percentage change in Jaccard between 5 workers samples and 30 workers sample averages.

ground truth information. Amongst the algorithms that do not make use of ground truth information, the performance of the greedy and EM algorithms exceeds the best achievable through existing retrieval-based method via the `num points` scoring heuristic and the vision-based algorithms.

By examining the dotted ground-truth algorithms, we learn the best achievable aggregation-based algorithm performs far better than the best worker segmentation. This result demonstrates since aggregation-based methods performs inference at a finer *tile* granularity, it is able to achieve better performance than compared to retrieval-methods. Doris: We should split w/ GT and no GT information into two plots side by side with same color scheme. Also consider merging greedy and EM into just one that says aggregation based method, and renaming num points as retrevial based methods.

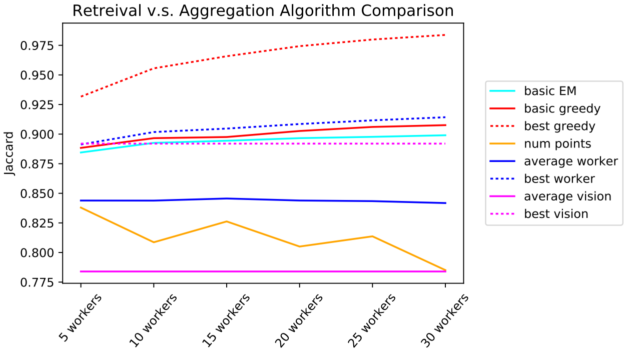

Figure 6: Jaccard performance comparison between best-performing algorithms from retrieval and aggregation-based methods, with clustering as a preprocessing step where possible. Color denotes the type of algorithm used.

Table 1 shows that the three retrieval-based methods on the left do not improve the resulting Jaccard significantly when more annotations are used, whereas the right four aggregation-based methods improves significantly from the 5 worker to 30 worker sample. Intuitively, the worker scaling of retrieval-based methods is not guaranteed [1]. On the other hand, since larger worker samples results in finer granularity tiles for the aggregation-based methods, there is an monotonically increasing relationship between number of worker segmentation used in the sample and performance due to the finer tiles set created by multiple segmentations.

Takeaway: Since aggregation-based methods operate at a finer tile-based granularity than whole-segmentation retrieval methods, the performance of aggregation-based approaches is better and also scales well as more annotations are collected.

6.4 How well does the inferred worker qualities predict individual worker performance?

Correlation of worker qualities against performance To further investigate how the EM models are performing, we looked at whether the model-inferred worker qualities is indicative of the actual quality of a segmentation. We performed linear fitting independently for each sample-objects and computed the R^2 statistics to determine whether worker qualities can accurately predict precision, recall, and Jaccard scores. Visual inspection of the basic worker quality model fitting showed that for objects that suffered from type two errors (semantic ambiguity), the single-parameter worker quality was unable to capture the overbounding behavior, which lead to a low precision and Jaccard. The results are listed in Table 2 to highlight how our advanced worker qualities were able to better capture these scenarios. The clustering preprocessing was not performed for the values in Table 2 to demonstrate the sole effect of the EM algorithm. Nevertheless, our clustered results also show a similar trend, with an average of R^2=0.88 and 0.89 for the GT and GTLSA models across all objects respectively. We also find that in general the linear fit improves as the number of data points increases, which indicates consistency in the fitted model.

N	basic	GT	GTLSA	isobasic	isoGT	isoGTLSA
5	0.601	0.907	0.901	0.576	0.907	0.904
10	0.632	0.895	0.899	0.633	0.895	0.898
15	0.622	0.897	0.898	0.622	0.897	0.897
20	0.636	0.894	0.899	0.637	0.894	0.898
25	0.66	0.901	0.905	0.661	0.901	0.904
30	0.673	0.907	0.914	0.676	0.907	0.913

Table 2: Linear correlation of worker qualities against ground truth performance for different quality models across different number of workers (N). The lower worker samples exhibit lower R^2 due to the variance from smaller number of datapoints for each independent fit.

Best worker quality retrieval One application of worker qualities is that it could be used as an annotation scoring function for retrieving the best quality worker segmentation. We explore this approach by training a linear regression model for every sample-object and use the worker qualities to predict the precision, recall, and Jaccard of individual worker annotations against ground truth. Then, we query the model with the inferred worker quality and retrieve the worker with the best predicted Jaccard.

The reason why a linear regression model was chosen rather than simply sorting the worker qualities and picking the best is that sorting based on multiple worker qualities (precision, recall, Jaccard) effectively applies equal weighting to all quality attributes, whereas our advanced models are specifically designed to capture cases of false-positives and false-negatives that can yield drastically different recall

[1] except in the case of picking the best worker, the more samples means higher probability that there would be a better segmentation

and precision values. We have tested that the linear regression model performs better on this task that simple sorting is capable of learning the weights that helps it make better predictions. As shown in Table 3, the performance of worker-quality based retrieval is comparable the performance other aggregation-based methods. We find that amongst the different worker quality models, advanced worker quality models perform the best, agreeing with our intuition regarding correlation results observed in Table 2.

algo/N	5	10	15	20	25	30
num points	0.838	0.809	0.826	0.805	0.814	0.785
best worker	0.891	0.902	0.905	0.909	0.912	0.914
MV	0.885	0.893	0.894	0.897	0.898	0.899
EM[basic]	0.884	0.893	0.894	0.897	0.898	0.899
EM[GT]	0.885	0.893	0.894	0.897	0.898	0.899
EM[GTLSA]	0.871	0.892	0.891	0.896	0.897	0.899
greedy	0.888	0.896	0.896	0.902	0.905	0.906
wqr[basic]	0.878	0.877	0.877	0.877	0.878	0.878
wqr[GT]	0.884	0.885	0.885	0.885	0.887	0.887
wqr[GTLSA]	0.874	0.881	0.883	0.885	0.886	0.887

Table 3: Summary of average performance across workers with clustering applied as preprocessing in all algorithms across different number of workers (N). wqr is the abbreviation for best worker quality retrieval methods.

6.5 How do different families of aggregation-based algorithms relate and compare?

Given the success of aggregation-based models, we wanted to further study how different algorithms perform compared to one another.

> *Takeaway: As shown in Table 3, majority vote, while simple, performs nearly as well as the advanced EM and greedy based approaches.*

This is because both EM and greedy have learned worker qualities that converged to MV behaviorAkash: need help with this explanation.

As shown in the dotted and solid line pairs in Figure 6, when using ground truth to estimate intersection areas, we can achieve an average Jaccard of 0.983 as an upper bound with the 30 workers sample, which indicates that with better probabilistic estimation of intersection area, aggregation-based methods can achieve close to perfect segmentation outputs, exceeding the results than achievable by any single 'best' worker (J=0.91 for 30 workers). Algorithms that gives users the option for collecting highly-accurate segmentation can have several useful applications in the biomedical domain (Gurari *et al.* 2015).

6.6 How well does clustering resolve multiple perspectives of crowdworkers and improve quality evaluation algorithms?

Figure 7 demonstrates how spectral clustering is capable of dividing the worker responses based on pairwise mutual Jaccard into clusters with meaningful semantic associations, reflecting the crowd's diversity of perspectives in completing same task.

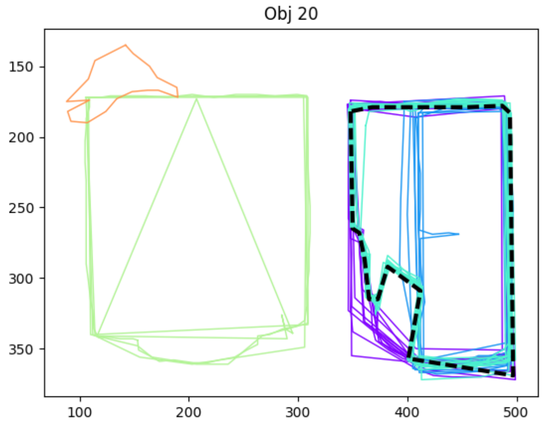

Figure 7: Example image showing clustering performed on the same object from Figure 1.

Compared to using a metric-based heuristic to detect and eliminate these errors, clustering has additional benefit of preserving worker's semantic intentions in the case where there are multiple instances of different errors. For example, in Figure 7, the mistakened clusters included 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 direction of future work includes adding a additional crowdsourcing task for semantic labelling of clusters (which is cheaper and more accurate than segmentation) to enable reuse of annotations across objects and lower the cost of data collection.

In addition to perspective resolution, clustering results can also be used as a preprocessing step to any of the quality evaluation algorithms that we have discussed. The clustering preprocessing can significantly improve algorithms that are not very robust to segmentations that contain semantic errors or regional semantic ambiguity issues, such as the heuristic-based number of points approach. When examining the gap of increase with and without clustering in Figure 8, we find that aggregation-based methods performs better than retrieval-methods exhibits a smaller gap between the performances. This effect is due to aggregation-based method's higher performance in the no cluster case, indicating that it is able to capture some of the semantic ambiguity and errors in the dataset.

7 Conclusion

In this paper, we perform an extensive study of several image segmentation algorithms spanning semi-supervised vision approaches, crowdsourced retrieval approaches, and novel crowdsourced aggregation approaches. We identified three different types of errors that workers typically make on segmentation tasks, some caused by differing perspec-

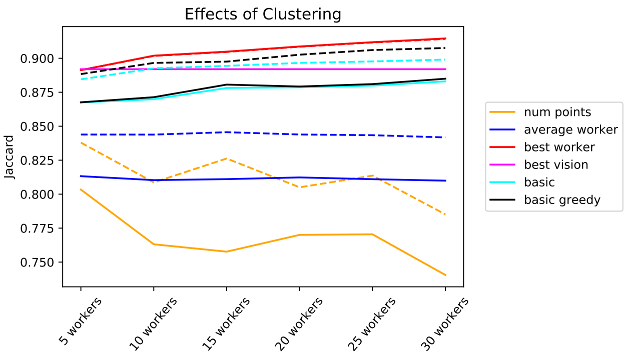

Figure 8: Performance comparisons between averaging over experiments with clustering as a preprocessing step(dotted) and the unclustered cases(solid) for different algorithms.

tives, and developed a clustering-based method to filter out workers that are making semantic errors. We demonstrate the strength of our worker clustering algorithm as well as the aggregation-based segmentation algorithms through extensive experiments in 1) its ability to improve as more worker segmentations are collected and 2) yield better performance than retrieval-based methods. We also found that while majority vote is a fairly simple algorithm, it performs nearly as well as the advanced EM and greedy inference approaches. Our code is open source and available for researchers to benchmark and compare techniques. Our work represents a first step in understanding and comparing the different types of algorithms available for image segmentation tasks. It opens a number of exciting directions for exploration, for instance: (a) Studying the effect of task difficulty, or worker qualities across different, objects, and (b) Designing better hybrid algorithms that combine the different types of algorithms described in this paper.

References

[Adriana Kovashka *et al.* 2016] Adriana Kovashka, Olga Russakovsky, and Li Fei-Fei. Crowdsourcing in Computer Vision. *Foundations and Trends® in Computer Graphics and Vision*, 10(2):103–175, 2016.

[Anonymous 2018] Anonymous. Tile-based quality evaluation methods for crowdsourcing image segmentation: Extended technical report. 2018.

[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.

[Dawid and Skene 1979] A. P. Dawid and A. M. Skene. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. 28(1):20–28, 1979.

[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.

[Felzenszwalb and Huttenlocher 2004] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181, 2004.

[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.

[Karger *et al.* 2013] David R Karger, Sewoong Oh, and Devavrat Shah. Efficient crowdsourcing for multi-class labeling. *ACM SIGMETRICS Performance Evaluation Review*, 41(1):81, 2013.

[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 and Perona 2010a] Peter Welinder and Pietro Perona. Online crowdsourcing: Rating annotators and obtaining cost-effective labels. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010*, pages 25–32, 2010a.

[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.