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

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). 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

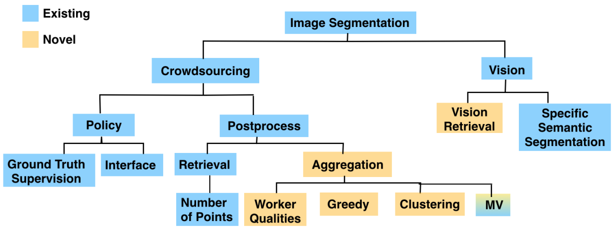

Figure 1: Flowchart summarizing the classes of existing algorithms for image segmentation (blue) and a novel class of algorithms proposed in this paper (yellow). Majority-vote (MV) is colored both blue and yellow, since a common algorithm in crowdsourcing literature, but have not been extensively applied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

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 1, 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, 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).

Retreival-based methods Retreival-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 Aggregation-based methods makes use of multiple worker segmentations to produce a single combined segmentation. Our paper formulate a novel "tiles" approach for aggregation methods that operates on the discrete non-overlapping units composed of all worker segmentations overlaid on top of each other. 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, rather than being used for aggregating worker segmentations.

3 Preliminaries

3.1 Data & Goals

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. As shown in Figure **??**, 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.

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, $IA = area(S \cup GT)$, 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.

4 Precision-savvy algorithms: Aggregation v.s. Retrieval Comparison

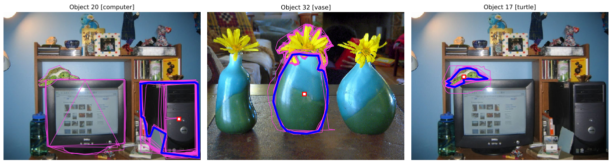

Figure 2: Pink is the segmentation from individual workers. Blue solid line delineates the ground truth. The red boxed pointer indicates the task of interest shown to users. Examples demonstrating common error patterns among crowdsourced image segmentation, including 1) annotations on the wrong semantic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

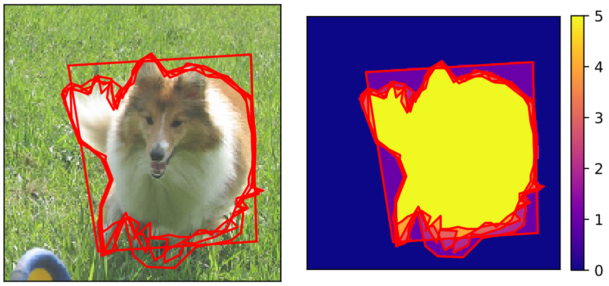

Figure 3: Left: Red boundaries shows the segmentation boundaries drawn by five workers overlaid on the image. Right: Segmentation boundaries still shown in red. The overlaid segmentation creates a masks where the color indicates the number of workers who voted for the tile region.

As shown in Figure 2, common worker errors can be classified into three types: (1) **Boundary Imprecision:** unintentional mistakes while drawing the boundaries, either due to low precision of the image, small area of the object, or lack of drawing skills , (2) **Semantic Ambiguity:** have differing opinions about whether particular regions belong to part of an object; or (3) **Semantic Mistakes:** annotate the wrong object entirely.

Out of the 46 objects in our dataset, 9 objects suffer from semantic ambiguity, 18 objects from semantic mistakes, and almost all objects suffer from some form of boundary imprecision to varying degrees. Since the main focus for quality evaluation in past literature have been focused on finding worker segmentation with minimal boundary precision issues, we will first describe novel aggregation-based algorithms that we have developed and compare them with existing retrieval-based methods for addressing type three errors. In the following section, we will discuss a preprocessing method that we have developed to resolve the semantic ambiguity and mistakes, which have been observed in prior work (Sorokin and Forsyth 2008; Lin *et al.* 2014; Gurari *et al.* 2018).

4.1 Retrieval-based 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).

4.2 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 3, 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 **??**.

4.3 Majority Vote Aggregation (MV)

The aggregation-based majority vote algorithm examines — 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.

4.4 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. 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.

4.5 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.

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

Figure 4 shows the comparisons between the best performing algorithm amongst aggregation-based (greedy, EM) and retrieval-based (num points) algorithms. The solid line in

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.

Figure 4 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 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.

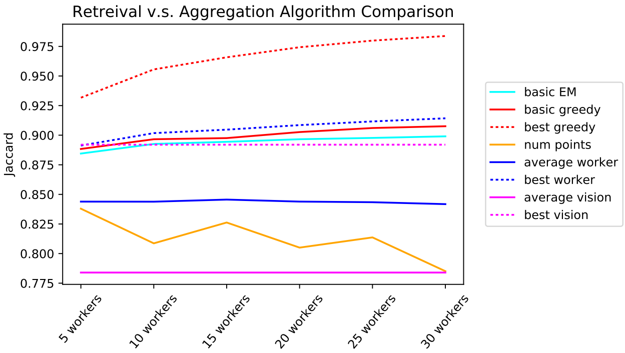

Figure 4: Jaccard performance comparison between best-performing algorithms from retrieval and aggregation-based methods, with clustering as a preprocessing step where possible. Color denotes the type of algorithm used.

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.

As shown in the dotted and solid line pairs in Figure 4, 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).

5 Perspective Resolution in Crowdsourced Image Segmentation

5.1 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, while workers with different perspectives from each other will have segmentations that differ from each other. We capture the similarity between a pair of workers by computing the Jaccard score between their segmentations. Then, 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.

Figure 5 illustrates 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.

Clustering results can also be used as a preprocessing step to any of the quality evaluation algorithms that we discussed earlier. On average, clustering results in —–% increase in — across the –NUMBER— aforementioned algorithms. In particular, we see a greater improvement with clustering preprocessing for algorithms that are not very robust in resolving semantic errors or ambiguity, such as for the `num pts` retrieval algorithm, than compared to the aggregation-based methods.

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 5, 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

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

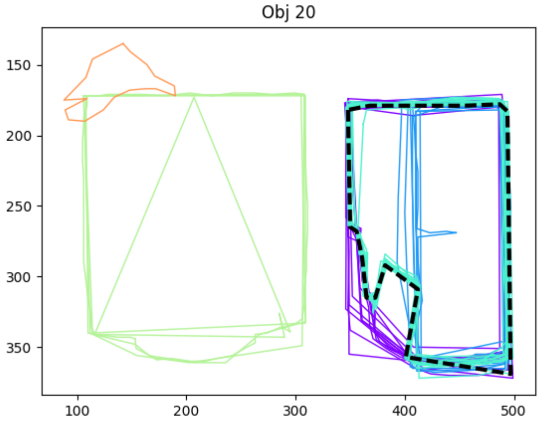

Figure 5: Example image showing clustering performed on the same object from Figure 2.

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.

6 Conclusion & Future Work

Given the success of aggregation-based models, immediate future work investigates why and how majority vote —, we have preliminary studies that shows that worker qualities are good indicators of actual —-. why 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 behavior.

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 perspectives, 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

[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.

[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.