Tool Diversity as a Means of Improving Aggregate Crowd Performance on Image Segmentation Tasks

Jean Y. Song, Raymond Fok, Fan Yang,
Kyle Wang, Alan Lundgard, Walter S. Lasecki
CROMA Lab | MISC Group
Electrical Engineering and Computer Science, University of Michigan – Ann Arbor
{jyskwon,rayfok,yangtony,wangkyle,arlu,wlasecki}@umich.edu

Abstract

Crowdsourcing is a common means of collecting training data, such as image segmentations, for many computer vision applications. However, designing accurate crowd-powered image segmentation systems is challenging because defining the boundaries of an object in an image requires considerable fine motor skills and hand-eye coordination that leads to some level of errors from every participant. Typically, answers from multiple workers are used to generate a more accurate combined result, but biases in how people make mistakes result in shared errors that remain even after aggregation. In this paper, we introduce an approach that leverages *multiple segmentation tools* for the same task to avoid systematic biases introduced by the tools themselves. We illustrate the efficacy of the approach through FourEyes, a hybrid intelligence system that leverages a set of four image segmentation tools. We present a series of studies that evaluate the feasibility of our multi-tool approach, and show that it is able to significantly improve aggregate accuracy in semantic image segmentation.

Introduction

Image segmentation demarcates objects in a visual scene from the background (Figure 1), allowing computer vision (CV) systems to learn to recognize these specific objects. These CV systems can in turn enable autonomous cars to identify pedestrians, surveillance drones to recognize potential threats, and help people with disabilities more easily interact with their surroundings (Zhong et al. 2015).

Perceiving object boundaries in visual scenes comes naturally to people, but remains a challenging open problem for CV systems due to their inability to understand scene semantics. Crowd-powered object segmentation tools can bridge this gap by using human understanding of scenes to produce large manually-demarcated training data sets for automated systems (Gurari, Sameki, and Betke 2016; Lin et al. 2014; Bell et al. 2013). However, designing highly accurate crowdsourcing systems that scale efficiently (with respect to cost / human time) for segmentation tasks is challenging because the manual task of tracing the boundaries requires considerable hand-eye coordination and fine motor skills that result in many errors if performed quickly. Many web-based image segmentation tools (Bearman et al. 2016; Bell et al. 2013; Carlier et al. 2014; Russell et al. 2008; Gouravajhala et al. 2017) have been designed to help workers reduce the effort needed to complete a task and to increase the accuracy of their output. However, different tool designs may induce different biases in worker performance, which could lead to systematic errors when only a single tool is used.

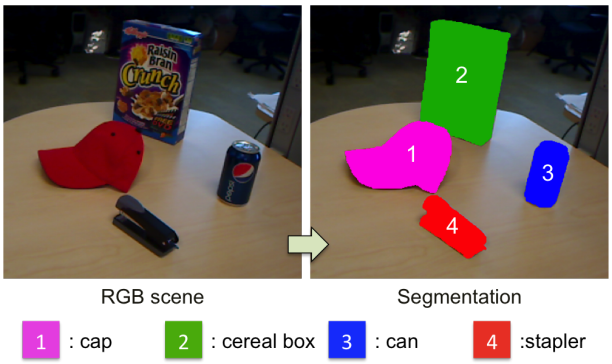

Figure 1: Example of the target image (left), the ground truth object segmentations (right), and the color codes mapped to object annotations (bottom).

In this paper, we present the idea of *tool diversity* as a means of improving aggregate crowd performance. Unlike the standard aggregation methods in crowdsourcing, which tries to design and use the best single tool available with many workers to reach high accuracy (Lasecki et al. 2011), we show that using multiple fairly reasonable tools can diversify the patterns in worker responses, and help systems achieve higher *combined* accuracy (Figure 2). This insight is motivated by ensemble learning methods in machine learning that use multiple learning algorithms to obtain better prediction than obtained from any of the component algorithms alone. To illustrate the efficacy of this approach, we introduce a multi-tool crowd-powered image segmentation system (FourEyes) to demonstrate the proposed idea. We show that heterogeneous tool aggregation provides more accurate segmentations than any individual base tool, even with a simple voting strategy.

The key contributions of this work are: 1) a novel crowdsourcing approach that combines input across different tool types to improve aggregate quality; 2) FourEyes, a crowd-

powered image segmentation system that implements our approach, combining the output of four different tool types to improve on the accuracy of a group of workers using a single segmentation tool; and 3) experimental results validating our system's effectiveness, and suggesting the benefits of our multi-tool approach.

Approach

Prior work has used task decomposition—the process of breaking down larger tasks into more manageable, focused pieces of work called subtasks—to make tasks more approachable for non-expert crowd workers. Once task decomposition has been used to break down a larger unit of work as much as possible within a corresponding workflow, most crowdsourcing systems then use multiple workers in parallel to improve accuracy further by aggregating their answers. Our proposed approach fills in the gap where traditional task decomposition leaves off. When a task (or subtask) can no longer be broken down, we propose using multiple different tools across different workers to complete the same [sub]task, instead of having all parallel workers complete the same task with the same interface or tool.

While we demonstrate this new crowdsourcing paradigm using an image segmentation task, it can benefit any task where different approaches to solving the same problem can be devised. Specifically, tasks that have the following properties would be especially amenable to our approach:

1. The task response correctness is cumulative with worker input. In other words, quality improves (converges to correct) as more worker inputs are collected. Problems where majority voting works would belong to this class.

2. The task is tractable enough to yield close-to-correct responses from workers, but responses can be expected to have a high chance of imperfection. That is, tasks for which humans are good at providing decent heuristic responses would benefit most from our approach.

3. The task has an objectively correct answer, but also tolerates imperfections from workers' responses. Handwriting recognition or re-assembling a shredded document can be example tasks. On the other hand, tasks like creative writing do not have a single correct answer, and thus are not appropriate for our approach.

4. The expected human error can be distributed differently between tools. This way, the diverse tool set can complement a broad range of error types. If this were not the case (i.e., if the errors were all biased in the same direction), then multiple tools would not be more effective than one.

Many common crowdsourcing problems (e.g., language annotation) have these properties, suggesting that a range of domains beyond the one explored in this paper may also be able to benefit from our approach. In the following sections, we introduce FourEyes to demonstrate that our crowdsourcing paradigm is beneficial to image segmentation tasks as one example of the potential of this approach.

System Design

FourEyes consists of four crowd-powered object segmentation tools (Figure 2), each with different levels of input required from workers to complete the task (different levels of autonomy). The first tool, **Basic Trace**, allows workers to draw boundaries of objects by holding the mouse button, which is a method commonly found in manual image segmentation tools (Gurari, Sameki, and Betke 2016). The second and third tools, **Drag-and-Drop** and **Pin-Placing**, are motivated by image registration techniques and use less manual interaction as compared to Basic Trace. For the template-based tools, we construct an icon list by downloading images of a particular object from an established image search engine like Google or Bing. These icon images are then filtered for transparency and size, and the first ten are used to construct each icon list. Workers are asked to select the icon that most accurately matches that object in the scene based on the shape, proportion of dimensions, and perspective. Drag-and-Drop allows workers to drag the icon image to place it in desired location, and rotate/scale to best align it with the object in scene. Pin-Placing allows worker to click four locations on their selected icon, and pair them with four corresponding points on the object in the scene. Then an automatic transformation algorithm will run to transform icon image to align corresponding points. The fourth tool, **Floodfill**, requires the least manual interaction. Workers are first asked to click on the object they want to segment, which triggers a flood fill algorithm to highlight all neighboring pixels sharing a RGB value similar to the RGB value of the pixel that was clicked. Workers can then adjust a slider to modify the algorithm's color tolerance parameter.

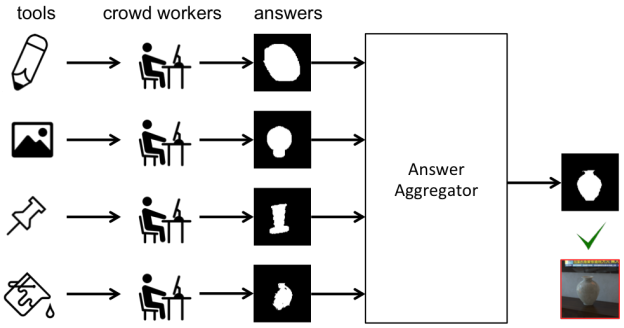

Figure 2: This work introduces a tool diversity approach that leverages multiple different tools for the same task to improve aggregate crowd performance by reducing systematic error biases that might otherwise result from using any constituent tool type alone.

Experimental Settings

To understand the effect of tool diversity on improving aggregate crowd performance, we recruited 288 crowd workers from Amazon Mechanical Turk to use FourEyes. Workers were given one of four tools to perform the task of image segmentation. We recruited six unique workers for each tool-scene pair (four tools and 12 scenes), resulting in a total of 1224 object segmentations. Each scene contains three to seven objects, totaling 51 objects. The scenes were gath-

ered from publicly-available data sets [1,2], and represent typical indoor scenarios with commonplace objects. The scenes ranged from a living room to a tabletop, and contained everyday objects (e.g., plant, laptop, soda can, cereal box, flashlight, etc). Each worker was shown one scene and a series of objects to segment depending on the number of objects in the scene. For each task, the order of objects in each list was randomized to avoid bias. Each worker was given one scene with one tool to perform a segmentation task.

Before crowd workers can begin the task, they are shown a short instructional video demonstrating the goal of the task, and how to use the tool they will be provided with. Workers are also shown examples of desirable and undesirable segmentations. If the worker decides to proceed, they are directed to FourEyes and their subsequent interactions with the system are logged. Task instructions are also accessible at any time if necessary. Each worker was paid between \$0.35 and \$0.60 per task, proportional to the number of objects they had to segment or to the level of difficulty of a given tool (a pay rate of $\sim$\$10/hr). The level of difficulty of each tool was determined by looking at their average latency time from a dozen preliminary experiments.

Results and Discussion

To measure success on the image segmentation task, we primarily care about the accuracy of the resulting segmentation. To measure accuracy, we use precision, recall, and F_1 score (the harmonic mean of precision and recall). To calculate these measures, we manually generated a ground truth segmentation for each object in each scene (as in Figure 1). Precision and recall of worker responses were both measured using per-pixel comparisons between worker answers and the ground truth. F_1 is computed from the same measures (e.g., true positive rate) as precision and recall.

Performance of Individual Tools

There was a statistically significant difference in accuracy measures across the different tools (all $p < 0.01$). Floodfill's precision was significantly better than the other three tools. On the other hand, its recall was significantly worse than the other three tools. The tool with the highest F_1 score was Basic Trace, performing significantly better than the other three. We observed that with Basic Trace, Drag-and-Drop, and Pin-Placing, workers tended to select objects by putting large margins around the objects, resulting in high recall but low precision. On the other hand, Floodfill gave high precision but low recall because the selection area tended to be smaller than the actual object boundaries due to boundaries that were shaded or colored differently.

Single-Tool vs Two-Tool Aggregation

We explored the aggregation result of two different team sizes (four workers and six workers) and all possible agreement thresholds. We implement a pixel-level uniform voting algorithm, with each answer weighted equally. For four workers, we tested agreement thresholds of 25%, 50%, 75%, and 100%. For six workers, we tested agreement thresholds of 16.7%, 33.3%, 50%, 66.7%, 83.3%, and 100%. Notably, the two extreme thresholds (lowest and highest) always give poor F_1 score (under 0.7) regardless the team size or tool pair. Since the extreme cases were so inaccurate, the rest of our experiments used only moderate agreement thresholds.

Team Size 4				Team Size 6			
	Prec.	Recall	F_1		Prec.	Recall	F_1
T_1	0.679	**0.989**	**0.759**	T_1	0.606	**0.990**	**0.728**
T_2	0.630	0.943	0.725	T_2	0.591	0.940	0.679
T_3	0.633	0.840	0.639	T_3	0.608	0.848	0.593
T_4	**0.856**	0.654	0.691	T_4	**0.830**	0.664	0.679

(a) Homogeneous tool aggregation

Team Size 4				Team Size 6			
	Prec.	Recall	F_1		Prec.	Recall	F_1
T_{12}	0.689	**0.888**	0.750	T_{12}	0.696	**0.929**	**0.774**
T_{13}	0.662	0.882	0.725	T_{13}	0.683	0.862	0.730
T_{14}	**0.818**	0.853	**0.806**	T_{14}	**0.831**	0.792	0.771
T_{23}	0.621	0.838	0.687	T_{23}	0.648	0.837	0.697
T_{24}	0.791	0.794	0.755	T_{24}	0.780	0.796	0.739
T_{34}	0.800	0.728	0.722	T_{34}	0.809	0.697	0.690
				T_{123}	0.660	0.896	0.722
				T_{124}	0.795	0.845	0.774
				T_{134}	0.778	0.814	0.749
				T_{234}	0.766	0.780	0.722

(b) Heterogeneous tool aggregation

Table 1: Average accuracy across different levels of agreement thresholds. The performance pattern was consistent in different team sizes.

Homogeneous Aggregation As a baseline, we explore segmentation accuracy of homogeneous aggregation (same-tool aggregation). The statistical result of the baseline is shown in Table 1(a). For a compressed summary, each team size is averaged across different agreement thresholds. The abbreviations T_1, T_2, T_3, and T_4 represent Basic Trace, Drag-and-Drop, Pin-Placing, and Floodfill, respectively. The performance of tools was consistent in different team sizes. For both team sizes, combining answers from T_4 gave the highest average precision, and combining answers from T_1 gave the highest average recall and F_1 score.

Heterogeneous Aggregation We then combined workers' answers from *multiple* segmentation tools for the same task. We tested all possible two- and three-tool pairs. Table 1(b) shows the results of these combinations. The term T_{ij} represents combination of T_i and T_j, where $i, j = 1, 2, 3, 4$. Note that the three measures for each team size is averaged across different agreement thresholds.

The results show that heterogeneous aggregation improves F_1 score in both team sizes compared to homogeneous aggregation. The maximum F_1 score for homogeneous aggregation was achieved by Basic Trace, and the values were 0.759 and 0.728 for team size four and six, respec-

[1] https://rgbd-dataset.cs.washington.edu/dataset.html/

[2] https://www.doc.ic.ac.uk/ ahanda/VaFRIC/iclnuim.html/

Team Size	Voting Threshold	Best Homo	Best Hetero	p-value
4	**50%**	T_4 0.742	$\boldsymbol{T_{14}}$ **0.837**	**0.00143 (p < 0.005)**
	75%	T_1 (0.776)	T_{14} (0.776)	0.989
6	33.3%	T_4 0.763	T_{14} 0.802	0.182
	50%	T_1 0.759	$\boldsymbol{T_{14}}$ **0.824**	**0.00168 (p < 0.005)**
	66.7%	T_1 0.825	T_{124} 0.835	0.665
	83.3%	T_1 0.797	T_{12} 0.783	0.729

Table 2: The best performing homogeneous tools and heterogeneous tool pairs and their F_1 scores. We ran an ANOVA test to check the statistical significance.

tively. The maximum F_1 score for heterogeneous aggregation was achieved by Basic Trace $\times$ Floodfill for team size four (0.806) and by Basic Trace $\times$ Drag-and-Drop for team size six (0.774). For both team sizes, heterogeneous aggregation performed better.

To compare the statistical significance, we ran an ANOVA test on F_1 scores for each agreement threshold. Table 2 shows the best performing homogeneous and heterogeneous tools for each threshold. From heterogeneous tool aggregation, we get a 9% improvement ($p < 0.005$) when agreement threshold is 50%, and no significant decrease in performance in any case. Notably, 50% agreement was not only the case where the heterogeneous pair performed significantly better than the homogeneous pair, but also the case that returned the highest average accuracy across all conditions.

From our experiments, we observed that the two highest aggregate-performance tools pairs were combinations of a high-precision (but low-recall) tool and a high-recall (but low-precision) tool. Precision and recall often have an inverse relationship, where one can be increased at the cost of reducing the other. In crowdsourcing literature, researchers have investigated different payment schemes to observe a tradeoff between precision and recall on an object annotation task (Mao et al. 2013). Our work suggests that different tools can be built to target either high precision or high recall so that the harmonic means of both can be maximized by aggregating results from two methods. More generally, our experiment indicates that the tool diversity strategy on crowdsourcing tasks can improve aggregate crowd performance by compensating for various types of inherent individual systematic error biases when combined together. Our study demonstrates that tool diversity can improve aggregate crowd performance on image segmentation tasks. Future work may investigate ways to better understand how tool diversity generalizes to other domains, which would introduce a novel, powerful, and complementary crowdsourcing approach.

Conclusion and Future Work

Our collective observations open opportunities and directions with pursuing a deeper understanding of how multiple tools influence the aggregate performance on diverse crowdsourcing tasks. Our future work includes applying tool diversity to additional crowdsourcing tasks, such as behavior coding (Lasecki et al. 2014) or activitiy recognition (Lasecki et al. 2013) to demonstrate its generality. We would also like to investigate the effect of financial incentives on tool diversity, which has not been discussed in depth in our study. Tool diversity may also help seamlessly integrate computer and human input to train machine learning algorithms.

Acknowledgments

This work was supported in part by MCity at the University of Michigan. We would also like the thank Stephanie O'Keefe for her input on this work.

References

Bearman, A.; Russakovsky, O.; Ferrari, V.; and Fei-Fei, L. 2016. Whats the point: Semantic segmentation with point supervision. In *ECCV 2016*.

Bell, S.; Upchurch, P.; Snavely, N.; and Bala, K. 2013. Opensurfaces: A richly annotated catalog of surface appearance. In *ACM TOG 2013*.

Carlier, A.; Charvillat, V.; Salvador, A.; Giro-i Nieto, X.; and Marques, O. 2014. Click'n'cut: Crowdsourced interactive segmentation with object candidates. In *CrowdMM 2014*.

Gouravajhala, S.; Song, J. Y.; Yim, J.; Fok, R.; Huang, Y.; Yang, F.; Wang, K.; An, Y.; and Lasecki, W. S. 2017. Towards hybrid intelligence for robotics. In *CI 2017*.

Gurari, D.; Sameki, M.; and Betke, M. 2016. Investigating the influence of data familiarity to improve the design of a crowdsourcing image annotation system. In *HCOMP 2016*.

Lasecki, W. S.; Murray, K. I.; White, S.; Miller, R. C.; and Bigham, J. P. 2011. Real-time crowd control of existing interfaces. In *UIST 2011*.

Lasecki, W. S.; Song, Y. C.; Kautz, H.; and Bigham, J. P. 2013. Real-time crowd labeling for deployable activity recognition. In *CSCW 2013*.

Lasecki, W. S.; Gordon, M.; Koutra, D.; Jung, M. F.; Dow, S. P.; and Bigham, J. P. 2014. Glance: Rapidly coding behavioral video with the crowd. In *UIST 2014*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV 2014*.

Mao, A.; Kamar, E.; Chen, Y.; Horvitz, E.; Schwamb, M. E.; Lintott, C. J.; and Smith, A. M. 2013. Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing. In *HCOMP 2013*.

Russell, B. C.; Torralba, A.; Murphy, K. P.; and Freeman, W. T. 2008. Labelme: a database and web-based tool for image annotation. In *IJCV 2008*.

Zhong, Y.; Lasecki, W. S.; Brady, E.; and Bigham, J. P. 2015. Regionspeak: Quick comprehensive spatial descriptions of complex images for blind users. In *CHI 2015*.