Predicting Quality of Crowdsourced Image Segmentations from Crowd Behavior

Mehrnoosh Sameki, Danna Gurari, Margrit Betke
Boston University

Introduction

Crowdsourcing is leveraged to rapidly and inexpensively collect annotations, but concerns have been raised about the quality of the results. We explore how to predict the quality of a crowd worker's annotation in the absence of additional human input. We do this by studying whether the interactive behavior of a worker correlates with annotation quality.

The crowd task we use as our experimental tool is drawing the boundary of an object in an image (segmentation). For this task, a user clicks a series of points that the software connects sequentially with straight lines to create a closed polygon (**Figure 1**). Workers have complete freedom in drawing and decide themselves how many clicks are needed to accurately capture details of diverse object shapes (e.g., house versus dog). Moreover, workers may need to allocate extra time to decide how to separate the more complicated object boundaries from other objects and/or the background.

Researchers have suggested that the quality of crowdsourced segmentations can be inferred by two behavioral cues: 1) the time a worker spends to draw a boundary (Vijayanarasimhan and et al 2009) and 2) the number of clicks a user makes to demarcate the boundary (Russell and et al 2008). However, no previous work has quantitatively validated this claim. In addition to examining the validity of this claim, we explore how to predict the quality of a crowd-submitted segmentation using a collection of behavioral cues. Our analysis reveals that the predictive power of learned models differs for different types of data, in particular, familiar everyday images vs. unfamiliar biomedical images.

Methods

To collect crowd-drawn segmentations, we configured the open source code of the image annotation tool LabelMe (Russell and et al 2008), with Amazon Mechanical Turk (AMT). Before an internet worker on AMT can accept our Human Intelligent Task (HIT), he/she is shown five instruction steps in English, followed by pictures exemplifying accurate and inaccurate annotations.

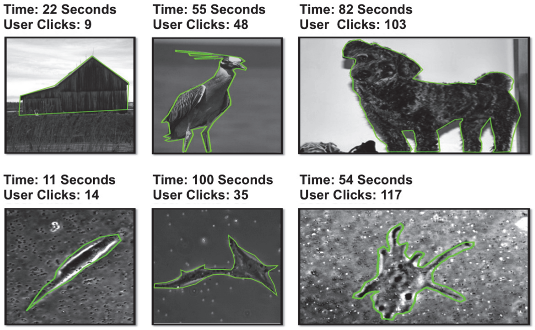

Figure 1: Given a crowd drawing, can we predict its quality based on worker behavioral clues such as time per drawing and number of user clicks?

We measure the quality of crowd segmentations by measuring the similarity of each crowd segmentation to a gold standard segmentation. We adopt as our similarity metric the widely used Jaccard index which measures how closely two segmentations resemble each other at the pixel level. It computes the number of pixels common to two segmentations, normalized by the number of pixels in both segmentations.

We propose three predictive features that may be indicative of segmentation quality:

- **Time Per Task**: We collect this metric from an AMT report that indicates the lapsed time for each completed HIT.
- **Number of User Clicks**: We count the number of points that the user places on the boundary of the object.
- **Time-Per-User-Click**: We normalize against the effect of boundary complexity by computing $(\frac{TimeperTask}{NumberofUserClicks})$.

Inspired by the observation that segmentation quality can gradually change from nearly perfect to seemingly meaningless, we chose a regression model to capture the continuous nature of segmentation quality variability. Specifically, we performed supervised learning by training a multiple linear regression model to learn whether crowd behavior can be predictive of the quality of crowd-drawn segmentations.

Image Library (# segmentations):	**All** (2025)		**BU-BIL** (1525)		**Weizmann** (500)	
Evaluation Metric:	CC	MAE	CC	MAE	CC	MAE
1. Time	-0.09	0.1	0.06	0.09	-0.01	0.15
2. Number of User Clicks	-0.07	0.1	0.16	0.08	0.22	0.14
3. Time & Number of User Clicks	-0.07	0.1	-0.07	0.09	-0.08	0.15
4. Time/Number of User Clicks	0.21	0.1	-0.01	0.09	0.33	0.14
5. Time/Number of User Clicks & Time & Number of User Clicks	**0.27**	**0.1**	**0.17**	**0.08**	**0.4**	**0.13**

Table 1: Evaluation and comparison of our prediction models using five different sets of predictive features.

Experiments and Results

We used the Weizmann single-object image library (Alpert and et al 2007) showing familiar objects such as animals, furniture, and buildings (100 images), and the BU-BIL library (Gurari and et al 2015) showing unfamiliar objects such as muscle cells and heart aorta (305 images). Both datasets are freely available and include expert-drawn annotations for use as a gold standard.

To capture in our training data the variability of segmentation behaviors that may arise due to differing worker skill, we collected five crowd-drawn segmentations per each image. We recruited crowdsourced workers from AMT, and accepted all workers who had previously completed 100 HITs and maintained at least a 92% approval rating. We paid workers $0.02 upon completion of each segmentation task.

We used the code from the freely-shared data mining system Weka (Hall and et al 2009) to analyze the predictive power of proposed prediction models. Our training dataset consisted of Jaccard scores calculated for 500 crowd-drawn segmentations for everyday images and 1525 crowd-drawn segmentations for biomedical images. We used 10-fold cross-validation to train and test prediction models. We evaluated each prediction model using Pearsons correlation coefficient (CC) and mean absolute error (MAE). CC indicates how strongly correlated predicted scores are to observed scores. MAE is a linear measure that indicates the average absolute size of prediction errors.

We analyzed five prediction models per dataset (i.e., Weizmann, BU-BIL) and across both datasets (All). We evaluated a prediction model for each of the three behavioral cues independently (**Table 1**, row 1, 2, 4) as well as in combinations (**Table 1**, row 3, 5). Overall, we found that crowd behavior is more strongly correlated with familiar everyday image content than unfamiliar biomedical content (**Table 1**).

Our findings highlight the validity of prior work which suggested that the quality of crowd segmentations may be inferred from a worker's number of clicks or time to annotate (Vijayanarasimhan and et al 2009; Russell and et al 2008). We found a weak linear correlation between time per task and segmentation accuracy (**Table 1**, row 1, All Images) as well as number of clicks and segmentation accuracy (**Table 1**, row 2, All Images).

Our findings illustrate how crowd behavior relates to different types of image content. While we found a slight correlation between the number of clicks and segmentation quality for both types of image content, we found a stronger correlation for everyday images than biomedical images (**Table 1**, row 2, BU-BIL and Weizmann). We hypothesize that the boundaries of the biomedical objects were so intricate that an increase in the number of boundary points did not necessarily imply improvement to segmentation accuracy.

Our findings demonstrate the effect of the combination of different behavioral metrics for prediction models. For both time-per-task and number-of-clicks, we observed weak correlations (**Table 1**, row 3). A model based on time-per-user-click led to stronger predictive power for all images. Interestingly, this prediction feature was better-suited for everyday content than for biomedical content (**Table 1**, row 4). Finally, the combination of time-per-task, number-of-clicks, and time-per-user-click as predictors led to an improved prediction model for all images as well as for the everyday and biomedical images individually (**Table 1**, row 5). However, we observed that the improvement of the prediction model for the biomedical images is minor from when only relying solely on the number of clicks.

Conclusions

We proposed an approach to predict the quality of crowd work directly from crowd behavior. Experiments revealed that time per number of clicks was the most effective cue for predicting segmentation quality. In addition, prediction models were more effective for estimating the segmentation quality from worker behavior for familiar everyday content than unfamiliar biomedical content.[1]

References

Alpert, S., and et al. 2007. Image segmentation by probabilistic bottom-up aggregation and cue integration. 8 pp.

Gurari, D., and et al. 2015. How to collect segmentations for biomedical images? A benchmark evaluating the performance of experts, crowdsourced non-experts, and algorithms. *WACV* 8 pp.

Hall, M., and et al. 2009. The weka data mining software: An update. *SIGKDD Explor. Newsl.* 11(1).

Russell, B. C., and et al. 2008. LabelMe: A database and web-based tool for image annotation. *IJCV* 77(1–3).

Vijayanarasimhan, S., and et al. 2009. What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In *CVPR*, 2262–2269.

[1] The authors acknowledge support by NSF, grant IIS RI 1421943.