Characterizing Image Segmentation Behavior of the Crowd

Mehrnoosh Sameki, Danna Gurari, and Margrit Betke; Boston University Computer Science Department

1. INTRODUCTION

Crowdsourcing platforms empower individuals and businesses to rapidly gather large amounts of human input. The challenge arises how to trust the results obtained from crowdsourced laypersons. Our work is a contribution to the growing exploration of how to successfully leverage large groups of crowdsourced laypersons to collect high-quality image annotation results.

Across many fields, methods have been proposed to control the quality of results and minimize human errors. Among them, two main categories can be observed [Allahbakhsh et al. 2013]. The first category of methods is relevant for the design of the crowdsourcing task. By requiring the appropriate qualifications of potential workers [Biewald and Pelt 2010], applying the appropriate visual clues [Sampath et al. 2014] and comprehensive instructions to avoid ambiguity, and showing good and bad examples of the desired result [Kulkarni et al. 2012], requesters can boost their chance of obtaining high-quality responses from the crowd. The second category of quality-control methods can be applied at runtime. The more commonly studied techniques in this category include domain expert review and the use of redundancy, where multiple crowd results are combined into a single final result [Ipeirotis et al. 2010]. A less researched area is how to leverage knowledge about user behavior to evaluate the final result [Rzeszotarski and Kittur 2011].

Collective intelligence has been used to solve labor-intensive computer vision tasks. The focus of our work is the task of delineating boundaries of objects in images (segmentation), which is a critical step for many computer vision tasks, including collecting information in regions of interests, matching objects in different images (registration), following objects over time (tracking), and differentiating between different types of objects (classification).

Image content can affect user behavior and the quality of acquired annotations – a user may require more time to draw complicated object boundaries, e.g., the boundary of a flower versus a box. One may take advantage of this observation in order to reduce the cost of collecting drawings of object boundaries [Carlier et al. 2014; Vijayanarasimhan and Grauman 2009]. Another informative characteristic of user behavior is the number of clicks a user makes to demarcate the boundary. The idea is that more complicated shapes require more points to capture this shape in detail. It has been suggested that the number of user clicks relates to the accuracy one may expect from crowd workers when asking them to draw boundaries of objects [Sorokin and Forsyth 2008]. Drawing tools can also have an effect.

We conducted a study to examine the drawing behavior of crowdsourced workers when given familiar and unfamiliar image content, shown in everyday and biomedical images. We analyzed the connection between accurate task completion and task completion time. We also compared the number of clicks workers performed to draw a boundary with the accuracy with which this boundary was drawn. We found it very valuable to broaden the analysis of crowd behavior by introducing unfamiliar data. We revealed problems that seem to have been overlooked so far and may require rethinking or extending generally held assumptions about crowd behavior. In particular, we revealed surprising differences in user behavior based on familiarity of image content. Our results provide insights into how crowdsourcing may be used for the task of image segmentation.

Collective Intelligence 2015.

2. EXPERIMENTAL METHODOLOGY

Our crowdsourcing experiment focused on the image segmentation task [Gurari et al. 2014]. To collect crowdsourced drawings of object boundaries, we used the freely-available online annotation tool LabelMe [Russell et al. 2008]. We connected it to Amazon Mechanical Turk, an Internet marketplace from which we recruited workers who had previously completed at least 100 tasks (HITs) and maintained at least a 92% approval rating. We accepted all submitted HITs. We provided workers with detailed instructions on the steps of the drawing tasks and showed them pictures of good and bad annotations. We obtained task completion times from Mechanical Turk.

To evaluate drawing quality, we used the region overlap ratio, a standard evaluation metric, to measure how closely a segmentation matches the "true" (gold standard) object boundary. The overlap ratio counts the number of pixels common to both the annotated and true regions that are in the combination of regions (i.e., $\frac{|A \cap B|}{|A \cup B|}$, where A represents the set of pixels in the true region and B represents the set of pixels in the annotated region). Scores range from 0 to 1 with higher scores reflecting greater similarity and so better accuracy.

We used two freely available benchmark image libraries that contain two types of images: "everyday images" of familiar objects, e.g., birds and buildings [Alpert et al. 2007], and biomedical images of relatively unfamiliar content, e.g., muscle cells and cross sections of the heart aorta [Gurari et al. 2014] (see figures). In addition to raw image data, the libraries contain, per image, multiple versions of object boundaries drawn by experts (scientists in the case of the biomedical images). To obtain a single gold standard for the "true" object boundaries, we combined these expert-drawn boundaries using the pixel-level majority vote (assign a pixel to belong to the object when at least half of the expert annotations assign the pixel to the object).

We collected five crowdsourced drawings for each of the 405 images in our data set. Each worker traced the boundary of the main object of the image by placing as many control points as he or she deemed necessary. The user interface connected the sequence of control points with straight lines. The worker finished tracing the boundary by clicking on the start point. We refer to the number of control points as the "number of user clicks." We used a pixel-level majority vote to combine the five crowd-collected outlines into a single outline (i.e., assign a pixel to belong to the object only if at least half of the workers assigned it to the object). We analyzed the accuracy of the computed outline by measuring its agreement with the expert-established gold standard outline using the overlap ratio metric.

3. RESULTS

In our experiment, 90 unique workers contributed 2,025 drawings. We here characterize crowd worker behavior by the relationship between the annotation quality and the number of crowd-drawn user clicks as well as the time it took to annotate. The average time it took workers to draw an object outline was 46 s, and the median was 30 s (25th percentile 21 s and 75th percentile 56 s). The median time it took workers to draw an outline in a biomedical image was 26 s (25th percentile 18 s and 75th percentile 46 s), and in a everyday image 32 s (25th percentile 48 s and 75th percentile 80 s).

The median number of user clicks was 27 (25th percentile 20 and 75th percentile 38), 27 for biomedical images (25th percentile 21 and 75th percentile 37 user clicks), and 29 for everyday images (25th percentile 17 and 75th percentile 43 user clicks).

3.1 Accuracy as a Function of Completion Time

To study the relationship between the time workers took to annotate images and the quality of their work, we arranged the completion time data into four quartiles, T1 for 0–25th percentiles, T2 for 25–50th, T3 for 50–75th, and T4 for 75–100th. We then analyzed the accuracy of annotation results for

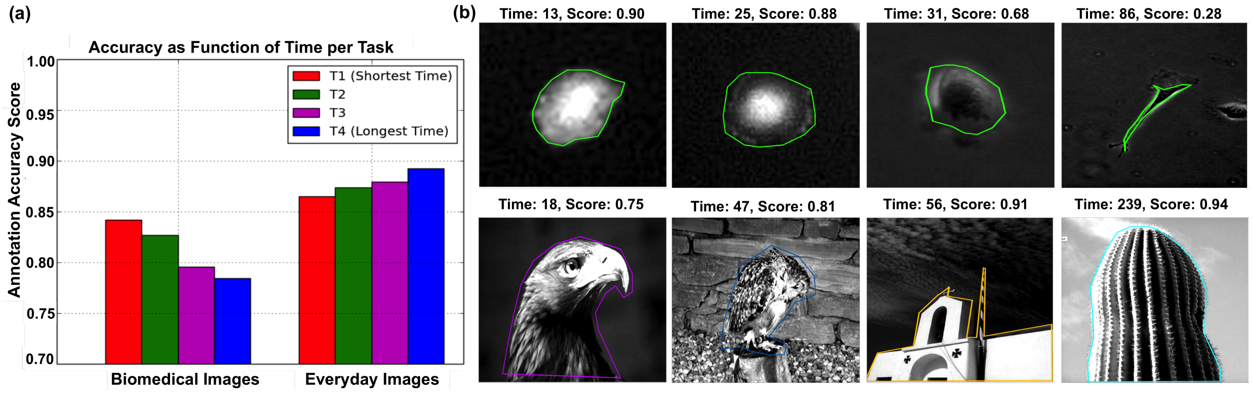

Fig. 1. (a) Annotation accuracy as a function of task completion time. The scores are binned into four quartiles (T1–T4) and reported for biomedical and everyday images separately. (b) Representative annotations showing one crowd drawing result from each of T1 T2, T3, and T4 for biomedical and everyday images. For each object, 17 user clicks were used.

biomedical and everyday images with respect to these quartile groupings (**Figure 1a**). When comparing annotation accuracy of biomedical versus everyday images, we observed an opposite pattern of change in accuracy scores with increasing completion time: As completion time per task increased, the image annotation quality worsened in biomedical images. In everyday images, however, the more time a worker spent, the more accurate the resulting image annotation became.

3.2 Accuracy as a Function of User Clicks

We studied the relationship between user accuracy and the number of user clicks (**Figure 2a**). As in the previous analysis, we divided the annotation results into four quartiles, P1 for the first quartile, P4 for the last. We also found opposite characteristics when comparing the annotation accuracy of biomedical and everyday images. For the biomedical data, the lowest accuracy scores were obtained for segmentations with the largest number of user clicks (P4). For the everyday images, however, more user clicks meant higher accuracy (P2–P4). The segmentations of everyday images obtained with the fewest user clicks (P1) had in the lowest accuracy scores and could be disregarded (or weighed less) in a voting procedure that yields a final annotation.

4. DISCUSSION AND CONCLUSIONS

We were surprised to find that more time led to more accurate results for annotating the familiar content in everyday images and less accurate results for the unfamiliar content of biomedical images. We visually inspected the images to infer why (**Figure 1b**). We hypothesize that users who spent more time did so in order to carefully place each control point along the object boundary. This boundary was understood for everyday objects but not understood and ambiguous for biomedical images (**Figure 2b**). We were also surprised to find that more user clicks, i.e., boundary control points, led to different observed trends of annotation quality with the familiar and unfamiliar image content. Visual inspection of the results reveals that the boundaries of the biomedical objects were so intricate that an increase in the number of boundary points generally was insufficient in improving segmentation accuracy (**Figure 2b**).

Our work facilitates the determination of what to expect from crowd workers in terms of time and number of clicks for two challenging image sets that represent familiar and unfamiliar image content. We hope that this understanding will encourage individuals to create more generalized quality control

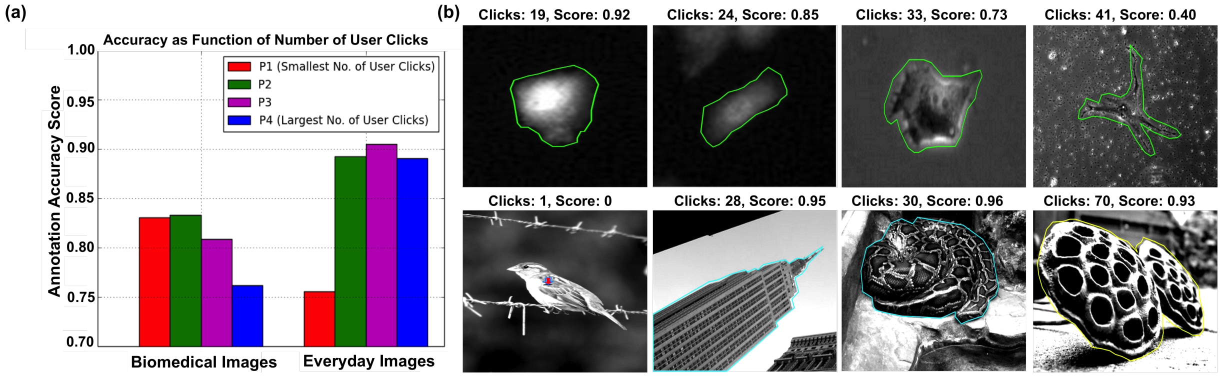

Fig. 2. (a) Annotation accuracy as a function of number of user clicks per task. The scores are binned into four quartiles (P1–P4) and reported for biomedical and everyday images separately. (b) Representative annotations showing one crowd drawing result from each of P1, P2, P3, and P4 for biomedical and everyday images. Annotation time per image varies from 42 to 56 seconds.

methods or web-based drawing systems for their crowdsourcing systems. In addition, we hope this work will encourage future interdisciplinary collaborations by highlighting how analyses on unfamiliar (biomedical) images revealed ways to improve our understanding about crowd behavior overall.

REFERENCES

M. Allahbakhsh, B. Benatallah, H. R. Motahari-Nezhad, A. Ignjatovic, and E. Bertino. 2013. Quality Control in Crowdsourcing Systems. *Internet Computing* 17, 2 (2013), 76–81.

S. Alpert, M. Galun, R. Basri, and A. Brandt. 2007. Image Segmentation by Probabilistic Bottom-Up Aggregation and Cue Integration. In *Proceedings of the Conferece on Computer Vision and Pattern Recognition (CVPR), 8 pages*.

L. Biewald and C. Van Pelt. 2010. Distributing a task to multiple workers over a network for completion while providing quality control. US Patent App. 12/817,946. (June 2010).

A. Carlier, V. Charvillat, A. Salvador, X. Giro i Nieto, and O. Marques. 2014. Click'n'Cut: Crowdsourced Interactive Segmentation with Object Candidates. In *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia*. ACM, 53–56.

D. Gurari, D. Theriault, M. Sameki, and M. Betke. 2014. How to use level set methods to accurately find boundaries of cells in biomedical images? Evaluation of six methods paired with automated and crowdsourced initial contours. In *Interactive Medical Image Computation Workshop (IMIC)*.

P. G. Ipeirotis, F. Provost, and J. Wang. 2010. Quality management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*. ACM, 64–67.

A. Kulkarni, M. Can, and B. Hartmann. 2012. Collaboratively crowdsourcing workflows with turkomatic. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. 1003–1012.

B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. 2008. LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision* 77, 1–3 (2008), 157–173.

J. M. Rzeszotarski and A. Kittur. 2011. Instrumenting the crowd: using implicit behavioral measures to predict task performance. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 13–22.

H. A. Sampath, R. Rajeshuni, and B. Indurkhya. 2014. Cognitively inspired task design to improve user performance on crowdsourcing platforms. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*. 3665–3674.

A. Sorokin and D. Forsyth. 2008. Utility data annotation with Amazon Mechanical Turk. *Urbana* 51, 61 (2008), 820.

S. Vijayanarasimhan and K. Grauman. 2009. What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In *Conference on Computer Vision and Pattern Recognition*. 2262–2269.