A Feasibility Study of Crowdsourcing and Google Street View to Determine Sidewalk Accessibility

Kotaro Hara, Victoria Le, and Jon Froehlich
Human-Computer Interaction Lab
Computer Science Department, University of Maryland
College Park, MD 20742
{kotaro, jonf}@cs.umd.edu; vnle@umd.edu

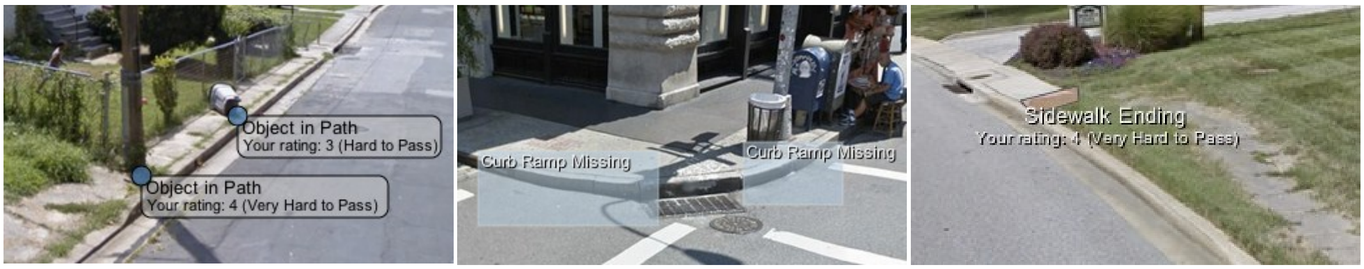

Figure 1. Using crowdsourcing and Google Street View images, we examined the efficacy of three different labeling interfaces on task performance to locate and assess sidewalk accessibility problems: (a) *Point*, (b) *Rectangle*, and (c) *Outline*. Actual labels from our study shown.

ABSTRACT
We explore the feasibility of using crowd workers from Amazon Mechanical Turk to identify and rank sidewalk accessibility issues from a manually curated database of 100 Google Street View images. We examine the effect of three different interactive labeling interfaces (*Point, Rectangle, and Outline*) on task accuracy and duration. We close the paper by discussing limitations and opportunities for future work.

Categories and Subject Descriptors
K.4.2 [**Computer and Society**]: Social Issues-Assistive technologies for persons with disabilities

Keywords
Crowdsourcing accessibility, Google Street View, accessible urban navigation, Mechanical Turk

1. INTRODUCTION
The availability and quality of sidewalks can significantly impact *how* and *where* people travel in urban environments. Sidewalks with surface cracks, buckled concrete, missing curb ramps, or other issues can pose considerable accessibility challenges to those with mobility or vision impairments [2,3]. Traditionally, sidewalk quality assessment has been conducted via in-person street audits, which is labor intensive and costly, or via citizen call-in reports, which are done on a reactive basis. As an alternative, we are investigating the use of crowdsourcing to locate and assess sidewalk accessibility problems *proactively* by labeling online map imagery via an interactive tool that we built.

In this paper, we specifically explore the feasibility of using crowd workers from Amazon Mechanical Turk (mturk.com), an online labor market, to label accessibility issues found in a manually curated database of 100 Google Street View (GSV) images. We examine the effect of three different interactive labeling interfaces (Figure 1) on task accuracy and duration. As the first study of its kind, our goals are to, first, investigate the viability of reappropriating online map imagery to determine sidewalk accessibility via crowd sourced workers and, second, to uncover potential strengths and weaknesses of this approach. We believe that our approach could be used as a lightweight method to bootstrap accessibility-aware urban navigation routing algorithms, to gather training labels for computer vision-based sidewalk accessibility assessment techniques, and/or as a mechanism for city governments and citizens alike to report on and learn about the health of their community's sidewalks.

ASSETS'12, October 22-24, 2012, Boulder, Colorado, USA.
ACM 978-1-4503-1321-6/12/10.

2. LABELING STREET VIEW IMAGES
To collect geo-labeled data on sidewalk accessibility problems in GSV images, we created an interactive online labeling tool in Javascript, PHP and MySQL, which works across browsers. Labeling GSV images is a three step process consisting of *marking* the location of the sidewalk problem, *categorizing* the problem into one of five types, and *assessing* the problem's severity. For the first step, we created three different marking interfaces: (i) *Point*: a point-and-click interface; (ii) *Rectangle*: a click-and-drag interface; and (iii) *Outline*: a path-drawing interface. We expected that the *Point* interface would be the quickest labeling technique but that the *Outline* interface would provide the finest pixel granularity of marking data (and thereby serve, for example, as better training data for a future semi-automatic labeling tool using computer vision).

Once a problem has been marked, a pop-up menu appears with four specific problem categories: *Curb Ramp Missing, Object in Path, Prematurely Ending Sidewalk,* and *Surface Problem*. We also included a fifth label for *Other*. These categories are based on sidewalk design guidelines from the US Department of Transportation website [3] and the US Access Board [2]. Finally, after a category has been selected, a five-point Likert scale appears asking the user to rate the severity of the problem where 5 is most severe indicating "not passable" and a 1 is least severe indicating "passable." If more than one problem exists in the image, this process is repeated. After all identified sidewalk problems have been labeled, the user can select "submit labels" and another image is loaded. Images with no apparent sidewalk problem can be marked as such by clicking on a button labeled "There are no accessibility problems in this image." Users can also choose to skip an image and record their reason (*e.g.,* image too blurry, sidewalk not visible).

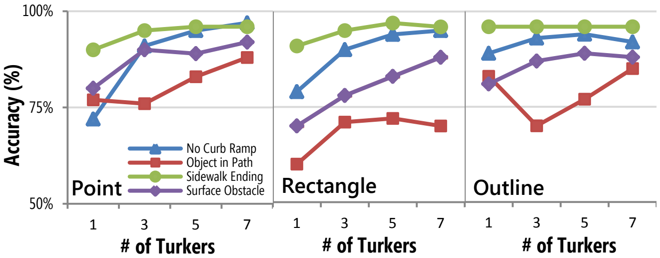

Figure 2. The number of turkers per image vs. accuracy for each of the three labeling interfaces. Note that the y-axis begins at 50%.

3. STUDY METHOD

To investigate the feasibility of using crowd workers for this task, we posted our three labeling interfaces (*Point, Rectangle,* and *Outline*) to Amazon Mechanical Turk. Crowd workers ("turkers") could complete "hits" with all three interfaces but would see each image at most once. Before beginning the labeling task with a particular interface, turkers were required to watch the first half of a three-minute instructional video. Three videos were used, one for each condition, which differed only in the description and presentation of the corresponding labeling interface. After 50% of the video was shown, the labeling interface would automatically appear (thus, turkers were not forced to watch the entire video).

Each labeling interface pulled images from the same test dataset, which consisted of 100 GSV images. These images were manually scraped by the research team using GSV of urban neighborhoods in Los Angeles, Baltimore, Washington DC, and New York City. We attempted to collect a balanced dataset. Of the 100 images, 81 contained one or more of the aforementioned problem categories. The remaining 19 images had no visible sidewalk accessibility issues and were used, in part, to evaluate *false positive* labeling activity.

To evaluate turker performance, we created baseline label data by having each of the three authors independently label all 100 images in each of the three interfaces. Inter-rater agreement was computed on these labels at the *image* level using Fleiss's kappa for each interface. More specifically, we tested for agreement based on the absence or presence of a label in an image and not on the label's particular pixel location or severity rating. We found moderate to substantial agreement [1] (ranging from 0.48 to 0.96). From these labels, we created a majority-vote "ground truth" dataset. Any image that received a label from two of the three authors was assigned that label as "ground truth" (Table 1).

	No Curb Ramp	Object in Path	Sidewalk Ending	Surface Problem
Point	34	27	10	29
Rectangle	34	27	11	28
Outline	34	26	10	29

Table 1. Frequency of labels at the *image* level in our ground truth dataset based on a "majority vote" from three trained labelers.

4. ANALYSIS AND RESULTS

We posted our task assignments to Mechanical Turk in batches of 20-30 over a one week period in June, 2012. In all, we hired 123 distinct workers who were paid three to five cents per labeled image. They worked on 2,235 assignments and provided a total of 4,309 labels (1.9 per image on average). As expected, the *Point* interface was the fastest with a median per-image labeling time of 32.9 seconds (*SD*=74.1) followed by *Outline* (41.5s, *SD*=67.6) and *Rectangle* (43.3s, *SD*=90.9). When compared with our ground truth dataset, overall turker accuracies at the *image* level were: 83.0% for *Point*, 82.6% for *Outline*, and 79.2% for *Rectangle*.

		No Curb Ramp	Object in Path	Sidewalk Ending	Surface Problem	Overall
Point	Precision	0.90	0.53	0.80	0.76	0.71
	Recall	0.82	0.93	0.73	0.93	0.87
Rectangle	Precision	0.85	0.48	0.80	0.59	0.63
	Recall	0.85	1.00	0.73	0.71	0.84
Outline	Precision	0.89	0.47	0.89	0.71	0.67
	Recall	0.91	0.93	0.73	0.89	0.89

Table 2. Precision and recall results for the three labeling interfaces based on majority vote data with three turkers compared to ground truth. "Object in path" is consistently the worst performing label.

We also explored accuracy as a function of the number of turkers per image and as a function of label type. To do this, we calculated four different turker-based majority vote datasets for each interface based on four different turker group sizes: 1, 3, 5, and 7. Group membership was determined based on the order of completion for each hit. The results are shown in Figure 2. Note that, again, we perform these comparisons at the *image* level rather than the individual label level and that we again ignore severity. These calculations are left for future work.

We did, however, employ an additional evaluation method by calculating the precision and recall rate of each interface, where:

$$Precision = \frac{True\ Pos}{True\ Pos + False\ Pos},\ Recall = \frac{True\ Pos}{True\ Pos + False\ Neg}$$

True positive here is defined as is providing the correct label on an image, *false positive* is providing a label for a problem that does not actually exist on the image, and *false negative* is *not* providing a label for a problem that *does exist* in the image. Our results are presented in Table 2. Both high precision and recall are preferred. The precision rate for *Object in Path* and *Surface Problems* are relatively low for all three interfaces. This indicates that turkers are making false positive decisions for those labels—that is, they tend to use these labels for things that are not actually problems.

5. DISCUSSION AND CONCLUSION

In this paper, we explored the feasibility of using crowd-sourced labor to label sidewalk accessibility problems from GSV images. We showed that untrained crowd workers can locate and identify sidewalk accessibility problems with relatively high accuracy (~80% on average). However, there is a clear problem with turkers *overlabeling* images (*i.e.,* we had a high false positive rate). In addition, there is a non-trivial number of bad quality workers—11 out of 123 had an error rate greater than 50%. In the future, we plan to explore automated methods of quality control to identify and expel poor quality workers programmatically. An additional limitation lies relates to using GSV as a data source: often times GSV images can be rather old (the average age of our images were 2.9 yrs) and some images are distorted due to sun glare or blurriness. Finally, sidewalks are not always visible in GSV. They can be blocked by cars, trees, guard rails or other obstacles. A future study emphasizing breadth is needed to determine the magnitude of this problem.

6. REFERENCES

[1] Landis R. J. and Koch G. G. 1977. The measurement of observer agreement for categorical data. Biometrics 1977; 33:159 –74.

[2] Public Rights-Of-Way Access Advisory Committee (PROWACC). 2007. Special Report: Accessible Public Rights-of-Way Planning and Design for Alterations, http://www.access-board.gov/prowac/alterations/guide.htm

[3] U.S. Department of Transportation, Designing Sidewalks and Trails for Access, http://www.fhwa.dot.gov/environment/bicycle_pedestrian/publications/sidewalks/index.cfm