Combining Crowdsourcing and Google Street View to Identify Street-level Accessibility Problems

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

| (a) *Object in Path* | (b) *No Curb Ramp* | (c) *Surface Problem* | (d) *Prematurely Ending Sidewalk* |

Figure 1: In this paper, we propose and investigate the use of crowdsourcing to find, label, and assess sidewalk accessibility problems in Google Streetview (GSV) imagery. The GSV images and annotations above are from our experiments with Mechanical Turk crowd workers.

ABSTRACT

Poorly maintained sidewalks, missing curb ramps, and other obstacles pose considerable accessibility challenges; however, there are currently few, if any, mechanisms to determine accessible areas of a city *a priori*. In this paper, we investigate the feasibility of using untrained crowd workers from Amazon Mechanical Turk (turkers) to find, label, and assess sidewalk accessibility problems in Google Street View imagery. We report on two studies: Study 1 examines the feasibility of this labeling task with six dedicated labelers including three wheelchair users; Study 2 investigates the comparative performance of turkers. In all, we collected 13,379 labels and 19,189 verification labels from a total of 402 turkers. We show that turkers are capable of determining the presence of an accessibility problem with 81% accuracy. With simple quality control methods, this number increases to 93%. Our work demonstrates a promising new, highly scalable method for acquiring knowledge about sidewalk accessibility.

Author Keywords

Crowdsourcing accessibility; accessible urban navigation; Google Street View; Mechanical Turk; image labeling

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI)

INTRODUCTION

According to the most recent US Census (2010), roughly 30.6 million individuals have physical disabilities that affect their ambulatory activities [32]. Of these, nearly half report using an assistive aid such as a wheelchair (3.6 million) or a cane, crutches, or walker (11.6 million) [32]. Despite aggressive civil rights legislation for Americans with disabilities (*e.g.,* [3, 21]), many city streets, sidewalks, and businesses in the US remain inaccessible [24].

The problem is not just that sidewalk accessibility fundamentally affects *where* and *how* people travel in cities but also that there are few, if any, mechanisms to determine accessible areas of a city *a priori*. Indeed, in a recent report, the National Council on Disability noted that they could not find comprehensive information on the "degree to which sidewalks are accessible" across the US [24]. Traditionally, sidewalk assessment has been conducted via in-person street audits [17,29], which are labor intensive and costly [25], or via citizen call-in reports, which are done on a reactive basis. As an alternative, we propose the use of crowdsourcing to locate and assess sidewalk accessibility problems *proactively* by labeling Google Street View (GSV) imagery (Figure 1).

We report on two studies in particular: a feasibility study (Study 1) and an online crowdsourcing study using Amazon Mechanical Turk (Study 2). Because labeling sidewalk accessibility problems is a subjective and potentially ambiguous task, Study 1 investigates the viability of the labeling sidewalk problems amongst two groups of diligent and motivated labelers: three members of our research team and three "sidewalk accessibility experts"—in this case,

wheelchair users. We use the results of this study to: (i) show that the labeling approach is reliable, with high intra- and inter-labeler agreement within and across the two groups; (ii) acquire an understanding of baseline performance—that is, what does good labeling performance look like? (iii) provide validated ground truth labels that can be used to evaluate crowd worker performance.

For Study 2, we investigate the potential of using crowd workers on Mechanical Turk (turkers) to perform this labeling task. We evaluate performance at two levels of labeling accuracy: *image level*, which tests for the presence or absence of the correct label in an image, and *pixel level*, which examines the pixel-level accuracies of the labels provided (as in Figure 1). We show that, when compared to ground truth, turkers are capable of determining that an accessibility problem *exists* in an image with 80.6% accuracy (binary classification) and determining the correct *problem type* with 78.3% accuracy (multiclass classification). Using a simple majority voting scheme with three turkers, this accuracy jumps to 86.9% and 83.8% respectively. We also examine the effect of two quality control mechanisms on performance: statistical filtering and multilevel review (see [22]). Our findings suggest that crowdsourcing both the labeling task and the verification task leads to a better quality result. We also demonstrate the performance/cost tradeoffs therein.

The primary contributions of this paper are threefold: (i) the first step toward a scalable approach for combining crowdsourcing and existing online map imagery to identify perceived accessibility issues, (ii) measures for assessing turker performance in applying accessibility labels, and (iii) strategies for improving overall data quality. 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 assessment, and as a mechanism for city governments and citizens to report on and learn about the health of their community's sidewalks (*e.g.,* through accessibility scores similar to *walkscore.com*).

BACKGROUND AND RELATED WORK
We provide background on sidewalk accessibility and sidewalk audit methods, in addition to related work on crowdsourcing and image labeling.

Factors Affecting Street-Level Accessibility
The US Department of Transportation [33] and the US Access Board [20] describe common problems that inhibit pedestrian access, including: (i) no place to walk—paths are either non-existent or not well-connected to destinations such as schools and transit; (ii) poor walking surfaces; (iii) blocked pathways, either temporarily (*e.g.,* by a vehicle) or permanently (*e.g.,* by a utility pole); (iv) difficult street crossings (*e.g.,* long walkway with no median, no curb ramps); (v) narrow sidewalks: wheelchair and scooter users require a wider path than ambulatory pedestrians, with most guidelines suggesting at least 60 inches.

Existing Sidewalk Audit Methods
In the US, state and federal departments conduct and encourage road safety audits that can also include walkability and pedestrian access (*e.g.,* [17,34]). Less formally, community organizations organize "Walk Audits" to find and assess deficiencies such as missing sidewalks, curb ramps, and/or dangerous street crossings (*e.g.,* [29]). Participatory reporting of accessibility problems has also been accomplished through applications that allow citizens to report non-emergency neighborhood issues to local government agencies (*SeeClickFix.com*) or to share information on wheelchair accessibility of businesses (*Wheelmap.org*); however, these applications do not support remote, virtual inquiry and have not been shown to scalably collect data on accessible public rights-of-way.

Street and neighborhood audits are also conducted by researchers in public health, sociology, and urban planning with the goal of studying the built environment and its impact on human behavior (*e.g.,* [12,17,27,30]). Since physical audits are often time-consuming and expensive [25], some studies have explored more efficient methods including recording video while driving for later review [27], or using satellite imagery and other map tools (*e.g.,* [19,30]). Omnidirectional streetscape imagery such as that in GSV has recently been used to perform *virtual* audits [5,11,25]. Reported benefits include time-savings and the ability to monitor and analyze multiple cities from a central location [5,25]. As an emerging area of research, work thus far has focused on the robustness and reliability of such approaches. Most importantly for our work, high levels of concordance have been reported between GSV *vs.* physical audit data for measures including pedestrian safety, traffic and parking, and pedestrian infrastructure [5,11,25].

Finally, most relevant to our work is the recent *CrossingGuard* paper by Guy and Truong [14], which focuses on navigation aids for visually impaired pedestrians and includes a small-scale study of GSV with three turkers. While similar in spirit, Guy and Truong focus exclusively on intersections for the visually impaired while we examine sidewalks for people with mobility impairments. More importantly, we ask turkers to mark *perceived* accessibility problems (*i.e,* perform judgments of accessibility) while [14] asks turkers to "check-off" the existence of traffic objects (*e.g.,* left turn signal, stop sign). Lastly, we ask turkers to directly label pixels, which helps establish an important initial baseline for collecting streetview accessibility training examples for computer vision.

Crowdsourcing and Image Labeling
Our image labeling task is analogous to that commonly performed in computer vision research for image segmentation, and object detection and recognition [6,26, 28,35,1]. Since manually building a large dataset of annotated images for training computer vision algorithms is expensive and time consuming [26], a number of web-based image labeling tools have been developed to capitalize on the large user population accessible over the

Figure 2: Labeling GSV images is a three step process consisting of *marking* the location of the sidewalk problem in the image, *categorizing* the problem into one of five types, and *assessing* the problem's severity. Here, the utility pole is labeled *Object in Path* and rated 5 (*Not Passable*).

Internet (*e.g.,* [1,2,26,28]). These tools differ in the level of information acquired about each image and their userbase.

For example, in von Ahn *et al.'s* work, textual labels are provided for images through a clever collaborative game-with-a-purpose, where users provide captions to describe objects in an image [1] or draw bounding boxes around specific items [2]. *LabelMe* [26] provides even more granular segmentation by allowing users to draw polygonal-outlines around objects, which are publically viewable and editable. Finally, to our knowledge, Sorokin and Forsyth [28] were the first to experiment with "outsourcing" image labeling to Mechanical Turk. In a series of experiments, they showed that a large number of high quality image annotations could be acquired relatively cheaply and quickly. Others have successfully used Mechanical Turk for a variety of purposes including document editing [7], graphical perception experiments [16], and near real-time assistance with visual problems for blind people [8].

ANNOTATION INTERFACE AND DATASET

To collect geo-labeled data on sidewalk accessibility problems in GSV images, we created an interactive online labeling tool in JavaScript, PHP and MySQL (Figure 2). We also created a verification interface (Figure 3) where users could accept or reject previously collected labels. Below, we describe the annotation interface and the primary dataset used in our studies. We return to the verification interface in the Study 2 section.

For the annotation interface, labeling is a three-step process consisting of *marking* the location of the problem (if one exists)*, categorizing* it into one of five types*,* and *assessing* its severity. For the first step, the user draws an outline around the perceived accessibility problem in the image (similar to LabelMe [26]). A pop-up menu then appears with five problem categories: *Curb Ramp Missing, Object in Path, Surface Problem, Prematurely Ending Sidewalk,* and *Other*. After a problem 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 ("not passable") and 1 is least severe ("passable"). The label is then complete. After all identified sidewalk problems have been labeled in a given image, the user can select "submit labels" and another image is loaded. Images with no apparent sidewalk problems can be marked as such with a button labeled "There are no accessibility problems in this image." Users can also skip images and record their reason

(*e.g.,* "image too blurry", "sidewalk not visible"). Other labeling techniques were explored in early prototypes [15].

The test dataset used in the labeling interface consists of 229 images 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 229 images, 179 contained one or more of the aforementioned problem categories; 50 had no visible sidewalk accessibility issues and were used, in part, to evaluate *false positive* labeling activity. Based on our majority-vote ground truth data (described later), we determined the following composition: 67 images with *Surface Problems,* 66 images with *Object in Path*, 50 with *Prematurely Ending Sidewalk*, and 47 with *Curb Ramp Missing*. This count is not mutually exclusive—48 images in total included more than one problem type. The label *Other* was used 0.5% of the time in Study 1 and 0.6% in Study 2 and is thus ignored in our analyses. As of September 2012, the average age of the images is 3.1 years old (*SD*=0.8 years). We return to the potential issue of image age in the discussion.

EVALUATING ANNOTATION CORRECTNESS

In this section, we provide an overview of the correctness measures used in our two studies. Because this is a new area of research, we introduce and explore a range of metrics—many of which have different levels of relevancy across application contexts (*e.g.,* calculating the accessibility score of a neighborhood *vs.* collecting training data for a computer vision algorithm).

Defining Levels of Annotation Correctness

Assessing annotation correctness in images is complex. To guide our analysis, we derived two spectra that vary according to the type and granularity of data extracted from each label: the localization spectrum and the specificity spectrum. The *localization* spectrum describes the positioning of the label in the image, which includes two discrete levels of granularity: *image level* and *pixel level*. For image level, we simply check for the absence or presence of a label anywhere within the image. Pixel level is more precise, examining individual pixels highlighted by the label outline. Our pixel-level analysis is analogous to image segmentation in computer vision and, indeed, our evaluation methods are informed from work in this space.

The *specificity* spectrum, in contrast, varies based on the amount of descriptive information evaluated for each label.

Figure 3: The verification interface used to experiment with crowdsourcing validation of turker labels—only one label is validated at a time in batches of 20. (a) A correctly labeled *No Curb Ramp* problem; (b) A false positive *Object in Path* label (the utility pole is located in the grass and not in the sidewalk); (c) A false negative example: The cars should have been marked as *Object in Path*.

At the finest level of granularity, we check for matches based on the five label categories as well as corresponding severity ratings: *Object in Path, Prematurely Ending Sidewalk, Surface Problem, Curb Ramp Missing*, and *No Problem* (indicating the user had clicked "no accessibility problems found"). Note that *Curb Ramp Missing* and *No Problem* were exempt from severity ratings. At the next level of granularity, we only examine problem types, ignoring severity ratings; we refer to this level as *multiclass*. Finally, at the coarsest level of granularity we group all problem categories into a *binary* classification of problem vs. no problem.

As the first work in the area, these dimensions of analysis are important for understanding crowd worker performance across various measures of correctness. Identifying an appropriate level of correctness may depend on the specific application context. For example, because of the focal length and camera angles used in GSV imagery, simply identifying that an accessibility problem exists in an image (*i.e.,* image-level, binary classification) localizes that problem to a fairly small geographic area: a specific street side and sidewalk within a city block. This level of geographic precision may be sufficient for calculating accessibility scores or even informing accessibility-aware routing algorithms. Binary classification—whether at the image level or the pixel level—also helps mitigate the subjectivity involved in selecting a label type for a problem (*e.g.,* some persons may perceive a problem as *Object in Path* while others may see it as a *Surface Problem*). In other cases, however, more specific correctness measures may be needed. Training computer vision algorithms to segment and, perhaps, automatically identify and recognize obstacles, would require pixel-level, multiclass granularity.

Image-Level Correctness Measures

For image-level analysis, we computed two different correctness measures: a straightforward accuracy measure and a more sophisticated measure involving precision and recall. For *accuracy*, we compare ground truth labels with turker labels for a given image and calculate the percentage correct. For example, if ground truth labels indicate that three problem types exist in an image: *No Curb Ramp, Object in Path*, and a *Surface Problem*, but a turker only labels *No Curb Ramp*, then the resulting accuracy score would be 50% (1 out of 3 problems identified correctly and 1 correct for *not* providing *Sidewalk Ending*). Though easy

to understand, this accuracy measure does not uncover more nuanced information about *why* an accuracy score is obtained (*e.g.,* because of false positives or false negatives).

As a result, we incorporated a second set of correctness measures, which extend from work in information retrieval: *precision*, *recall*, and an amalgamation of the two, *f-measure*. All three measures return a value between 0 and 1, where 1 is better:

$$Precision = \frac{\# \ of \ True \ Pos \ Labels}{\# \ of \ True \ Pos \ Labels + \# \ of \ False \ Pos \ Labels} \quad \text{(Eq. 1)}$$

$$Recall = \frac{\# \ of \ True \ Pos \ Labels}{\# \ of \ True \ Pos \ Labels + \# \ of \ False \ Neg \ Labels} \quad \text{(Eq. 2)}$$

$$F\text{-}measure = 2 * \frac{Precision * Recall}{Precision + Recall} \quad \text{(Eq. 3)}$$

True positive here is defined as providing the correct label on an image, *false positive* is providing a label for a problem that does not actually exist in the image, and *false negative* is *not* providing a label for a problem that *does exist* in the image. In this way, precision measures the accuracy of the labels actually provided (*i.e.,* a fraction expressing the ratio of correct labels over *all labels provided*) while recall measures the comprehensiveness of the correct labels provided (*i.e.,* a fraction expressing the ratio of correct labels over all *possible correct labels*). For example, a precision score of 1.0 means that every label the turker added was correct but they could have missed labels. A recall score of 1.0 means that the turker's labels include all of the actual problems in the image but could also include non-problems. Given that algorithms can be tuned to maximize precision while sacrificing recall and vice versa, the f-measure provides a single joint metric that encapsulates both. We use accuracy, precision, recall, and f-measure to describe our image level results.

Pixel-Level Correctness Measures

Pixel-level correctness relates to image segmentation work in computer vision. Zhang [1] provides a review of methods for evaluating image segmentation quality, two of which are relevant here: the *goodness method*, which examines segmentation based on human judgment and the *empirical discrepancy method*, which programmatically calculates the difference between test segmentations and "ground truth" segmentations for a given image. The goodness method can be advantageous in that it does not require ground truth; however, it is labor intensive because it relies on human judgment to perceive quality. Though judging the quality of

segmentations can also be crowdsourced, partly mitigating the labor concern (*e.g.,* [7]), the quality of the judgment itself remains an issue.

Thus, we also explored two empirical discrepancy methods: *overlap* (or area of intersection) [2,31] and, again, precision/recall combined with f-measure [9,10], which is similar to that explained above though applied at the pixel level rather than the image level. For our first discrepancy method, overlap is defined as:

$$Overlap(A, B) = \frac{Area(A \cap B)}{Area(A \cup B)} \qquad \text{(Eq. 4)}$$

where *A* and *B* are the pixel outlines. Note that if the outline *A* is perfectly equal to the outline *B*, then *Overlap(A,B)=1*. If *A* and *B* are disjoint, then *Overlap(A,B)=0*. Although this metric is easy to understand, similar to the straightforward accuracy measure for image-level analysis, it fails to capture nuances in correctness. Thus, for our second discrepancy metric we define precision, recall, and f-measure at the pixel level. From the image segmentation literature [6], *precision* is defined as the probability that a generated outline-fill pixel area correctly highlights the target object and *recall* is the probability that a true outline-fill pixel is detected. Thus, in order to calculate precision and recall at the pixel level, we need to compute three different pixel counts for each image:

1. **True positive pixels:** number of overlapping pixels between the ground truth segmentation and the test segmentation;
2. **False positive pixels**: number of pixels in the test segmentation *not* in the ground truth segmentation;
3. **False negative pixels:** number of pixels in the ground truth segmentation *not* in the test segmentation.

Precision and recall can then be computed by the following formulae (f-measure is the same as *Eq. 3* above):

$$Precision = \frac{\# \, of \, True \, Pos \, Pixels}{\# \, of \, True \, Pos \, Pixels + \# \, of \, False \, Pos \, Pixels} \qquad \text{(Eq. 5)}$$

$$Recall = \frac{\# \, of \, True \, Pos \, Pixels}{\# \, of \, True \, Pos \, Pixels + \# \, of \, False \, Neg \, Pixels} \qquad \text{(Eq. 6)}$$

Before calculating pixel-level correctness for any of the measures, we flatten all labels with equivalent type into the same layer and treat them as a single set of pixels. This allows us to more easily perform pixel-by-pixel comparison between ground truth labels and test labels marked with the same problem type.

STUDY 1: ASSESSING FEASIBILITY

Labeling accessibility problems perceived in streetscape images is a subjective process. As such, our first study focused on demonstrating that informed and well-motivated labelers could complete the labeling task and produce consistent results. We had two additional goals: (i) to produce a vetted ground truth dataset that could be used to calculate turker performance in Study 2, and (ii) to help contextualize Study 2 results (*i.e.,* what does "good" performance look like?).

We collected independently-labeled data from two groups: three members of our research team and three wheelchair users (who served as "sidewalk accessibility experts"). We then computed intra- and inter-annotator agreement scores for within and between each group respectively. We explore agreement at both the image level and the pixel level across binary and multiclass classification.

Collecting Wheelchair User Ground Truth Data

Three wheelchair users were recruited via listservs and word-of-mouth: two males with spinal cord injury (tetraplegia) and one male with cerebral palsy. All three used motorized wheelchairs; one also used a manual wheelchair but rarely. Each wheelchair user took part in a single labeling session at our research lab. Participants were asked to label the images based on their own experiences and were instructed that not all images contained accessibility problems. They were also asked to "think-aloud" during labeling so that we could better understand the rationale behind their labeling decisions.

The sessions lasted for 2-3 hours and included a short, post-labeling interview where we asked about the participant's personal experiences with sidewalk/street accessibility and about potential improvements to our labeling tool. In consideration of participant time and potential fatigue, only a subset of the total 229 image dataset was labeled: 75 in total. These images were selected randomly from each of the four problem categories (*4 categories x 15 images = 60*) plus an additional 15 randomly selected "no problem" images. Participants were compensated $25-35 depending on session length. Below, we report on evaluating agreement between the researchers, the wheelchair users, and the researchers *compared* to the wheelchair users. For the latter calculation, we compare *majority vote* data from each group so *N*=2 rather than *N*=6. We describe both image-level and pixel-level performance.

Evaluating Image-Level Agreement and Performance

We computed inter-rater agreement on labels at the image level using Fleiss' kappa [18], which attempts to account for agreement expected by chance. As this was an image-level analysis, 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. Multiple labels of the *same type* were compressed into a single "binary presence" indicator for that label. For example, if three individual *Surface Problems* were labeled in an image, for our analysis, we only considered the fact that a *Surface Problem* was detected and not how many occurrences there were exactly. This helped control for different annotator tendencies—some who would provide one large label to cover contiguous problem areas and others who would provide separate labels. Results are shown in Table 1 for both binary and multiclass classification (*N* represents the number of annotators and *I* the number of images, Table 2 uses the same notation).

Three key results emerge: first, both the researchers and the wheelchair users had moderate to substantial levels of agreement [18], which indicates that the labeling task, at least at the image-level, is feasible and that the labels are

Image-Level Label Specificity	Label	Researchers (N=3, I=229)	Wheelchair Users (N=3, I=75)	Researchers vs. Wheelchair Users (N=2 groups, I=75)
Binary Classification	No Problem vs. Problem	0.81	0.68	0.79
Multiclass Classification	No Curb Ramp	0.81	0.82	0.83
	Object in Path	0.56	0.55	0.62
	Sidewalk Ending	0.86	0.71	0.78
	Surface Problem	0.62	0.40	0.74
	Overall	0.69	0.62	0.74

Table 1: Fleiss' kappa annotator agreement scores for image-level analysis between the researchers, the wheelchair users, and the researchers compared to the wheelchair users (this lattermost comparison is based on majority vote data within each group).

Pixel-Level Label Specificity	Correctness Measure	Image Comparisons	Researchers (N=3, I=229)	Wheelchair Users (N=3, I=75)	Researchers vs. Wheelchair Users (N=2 groups, I=75)
Binary Classification	Area Overlap	Same	0.31 (0.21)	0.26 (0.22)	0.27 (0.21)
		Different	0.02 (0.05)	0.01 (0.04)	0.01 (0.04)
	F-Measure	Same	0.43 (0.25)	0.37 (0.26)	0.38 (0.26)
		Different	0.03 (0.08)	0.02 (0.06)	0.03 (0.07)
Multiclass Classification	Area Overlap	Same	0.27 (0.21)	0.22 (0.22)	0.23 (0.21)
		Different	0.01 (0.03)	0.00 (0.02)	0.00 (0.02)
	F-Measure	Same	0.38 (0.26)	0.32 (0.27)	0.33 (0.27)
		Different	0.01 (0.05)	0.01 (0.04)	0.01 (0.04)

Table 2: The results of our pixel level agreement analysis (based on [23]) between the researchers, wheelchair users, and researchers compared to wheelchair users. Similar to Table 1, the rightmost column is majority vote data. Cell format: average (stdev).

fairly consistent across labelers; second, and just as importantly, the third column in Table 1 shows high agreement *between* the majority vote data of the research team and the wheelchair users, which indicates that the accessibility problems identified by the research team are consistent with "experts"; and, finally, the multiclass agreement results show that *Object in Path* and *Surface Problem* have more disagreement than *No Curb Ramp* and *Sidewalk Ending*. This is likely because *Object in Path* and *Surface Problems* are often less salient in images and because they are occasionally substituted for one another (*e.g.,* some labelers perceive a problem as *Object in Path* while others as a *Surface Problem*).

Evaluating Pixel-Level Agreement and Performance
Calculating pixel-level agreement is more challenging. Because no widespread standards exist for evaluating pixel-level agreement for human labelers, we followed the process prescribed by Martin *et al.* [23]. We verify the labeling process by showing that pixel-level label overlap and f-measure scores are higher between labelers on the *same* image than across *different* images. These scores will later act as a baseline for defining good pixel-level performance when evaluating turker labels. To compare between the *same* images, 678 comparisons are required (3 annotators x 229 images). For *different* images, 156,636 comparisons are required (3 annotators x (229 x 229 – 229)). Because the wheelchair users only labeled 75 of the 229 images, their comparison count is correspondingly lower (225 for same, 16,650 for different). We ignore images for which all annotators labeled *No Problems Found* (as no pixel labels exist in these images). Our results are shown in Table 2.

From these results, we conclude that our pixel level annotations across labelers are reasonably consistent, although less so than for image level. Unsurprisingly, agreement is higher for binary classification than for multiclass, though not substantially. This indicates that a major source of disagreement is not the label type (*e.g., Object in Path* vs. *Surface Problem*) but rather the pixels highlighted by the outline shape. We emphasize, however, that pixel outlines for even the same object across labelers will rarely agree perfectly; the key then, is to determine what level of overlap and f-measure scores are acceptable and good. Our results suggest that overlap scores of 0.31

and 0.27 and f-measure scores of 0.43 and 0.38 for binary and multiclass classification respectively are indicative of what a motivated and diligent annotator can achieve. We emphasize that even 10-15% overlap agreement at the pixel level would be sufficient to confidently localize problems in images and highlight these areas in accessible map routing interfaces. This level of consistency, however, may not be sufficient for training computer vision. We return to this point in the discussion.

PRODUCING GROUND TRUTH DATASETS
Finally, now that we have shown the feasibility of the labeling task and found reasonably high consistency amongst labelers, we can use these Study 1 labels to produce a ground truth dataset for evaluating turker performance. We consolidate the labeling data from the three researchers into four unified ground truth datasets: binary and multiclass at both the image and the pixel level

Consolidating Image-Level Labels: To combine image-level labels across the three labelers, we simply create a majority-vote "ground truth" dataset. Any *image* that received a label from at least two of the three researchers was assigned that label as "ground truth."

Consolidating Pixel-Level Labels: Combining labels from the three researchers at the pixel level is less straightforward. The consolidation algorithm will directly impact the results obtained from our correctness measures. For example, if we combine highlighted pixel areas across all three researchers (union), then turker precision is likely to go up but recall is likely to go down. If, instead, we take the intersection across all three labelers, the ground truth pixel area will shrink substantially, which will likely increase turker recall but reduce precision. Consequently, we decided to, again, adopt a majority vote approach. To produce the majority vote pixel-level dataset, we look for labels from at least two of the three researchers that overlap by 15% of their unioned area. The value of 15% was chosen because it is the lower-quartile cutoff using researcher overlap data. For binary classification, the label type was ignored—thus, any labels that overlapped by 15% or more were combined. For multiclass, the labels had to be of the same type.
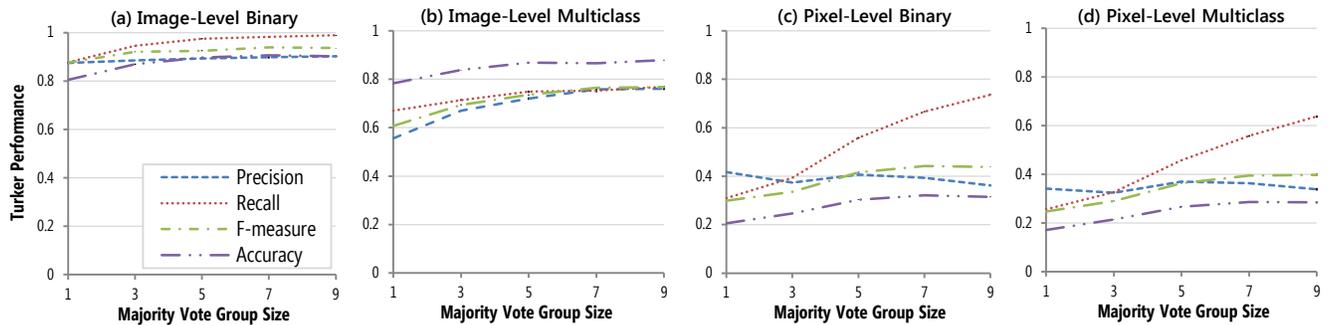
Figure 4: Binary and multiclass performance at the image- and pixel-levels with varying majority vote group sizes. Each graph point is based on multiple permutations of the majority vote group size across all 229 images. Standard error bars are in black (barely visible due to low variance).

STUDY 2: CROWD WORKER PERFORMANCE

To investigate the potential of using untrained crowd workers to label accessibility problems, we posted our task to Mechanical Turk during the summer of 2012. Each "hit" required labeling 1-10 images for 1-5 cents (0.5 to 5 cents per image). Each turker new to the task was required to watch at least half of a 3-minute instructional video, after which the labeling interface automatically appeared. Note: one task encompasses labeling one image.

We first describe high-level results before performing a more detailed analysis covering labeler count *vs.* accuracy, two quality control evaluations, and the best and worst performing images. For the analysis below, we do not consider severity ratings. Instead, we leave this for future work. However, given that we found a high rate of false positives amongst the turker data, we did examine the effect of removing labels that received a severity rating of a 1 (*Passable*) or a 2 (*Fairly Passable*). Our findings did not change significantly as a result.

High-Level Results

In all, we hired 185 distinct turkers who completed 7,517 image labeling tasks and provided a total of 13,379 labels. Turkers completed an average of 40.6 tasks (*SD*=61.2); 20 turkers labeled only 1 image and 10 turkers labeled all 229. The median image labeling time was 33.3s (*SD*=89.0s) and the average number of labels per image was 1.79 (*SD*=1.27). When compared with our ground truth dataset, overall turker accuracy at the *image* level was 80.6% for binary classification and 78.3% for multiclass classification. At the pixel level, average area overlap was 20.6% and 17.0% for binary and multiclass, respectively. These numbers are reasonably close to the values of 27% and 23% that we saw for wheelchair users *vs.* researchers.

Accuracy as a Function of Turkers per Image

Collecting multiple annotations per image helps account for the natural variability of human performance and reduces the influence of occasional errors; however, it also requires more workers [28]. Here, we explore accuracy as a function of turkers per image. We expect that accuracy should improve as the number of turkers increases, but the question then, is by how much? To evaluate the impact of the number of turkers on accuracy, we collected labels from 28 or more turkers for *each* of our 229 images. We compare our majority vote ground truth data with majority vote data

Image-Level Label Specificity	Label	Maj Vote Size: 1	Maj Vote Size: 3	Maj Vote Size: 5	Maj Vote Size: 7	Maj Vote Size: 9
Binary	No Prob vs. Prob	80.6 (0.1)	86.9 (0.3)	89.7 (0.2)	90.6 (0.2)	90.2 (0.2)
Multiclass	No Curb Ramp	78.6 (0.1)	86.0 (0.1)	90.2 (0.3)	91.6 (0.2)	93.7 (0.3)
	Object in Path	73.0 (0.1)	78.1 (0.2)	81.3 (0.3)	82.2 (0.1)	83.4 (0.2)
	Sidewalk Ending	84.7 (0.1)	88.3 (0.1)	88.5 (0.4)	89.5 (0.4)	89.8 (0.3)
	Surface Problem	77.0 (0.1)	82.1 (0.2)	84.9 (0.3)	85.9 (0.4)	88.4 (0.3)
	Overall	78.3 (0.0)	83.8 (0.1)	86.8 (0.2)	86.6 (0.2)	87.9 (0.1)

Table 3: Binary and multiclass label type accuracy at the image level across five majority vote group sizes. Cell format: avg% (stderr %).

across five turker group sizes: 1, 3, 5, 7, and 9. Because we have 28 turkers per image, we can run the analysis multiple times for each group size, average the results, and calculate error margins (which produces a more accurate portral of expected future performance for each group size). For example, when we set the majority vote group size to three, we randomly permute nine groups of three turkers. In each group, we calculate the majority vote answer for a given image in the dataset and compare it with ground truth. This process is repeated across all images and the five group sizes, where (*X=majority vote group size, Y=number of groups*): (1,28), (3, 9), (5,5), (7, 4), (9, 3). To compute the majority vote *answer* for each group size, we use the same label consolidation process as that used for the researcher majority vote labels.

We conducted this analysis at the image and pixel levels for binary and multiclass classification across our multiple correctness measures. Results are shown in Figure 4 (image and pixel level) and Table 3 (image level only). As expected, performance improves with turker count but these gains diminish in magnitude as group size grows. For example, at the image level, binary accuracy improves from 80.6% to 86.9% with 3 turkers and to 89.7% with 5 turkers but only to 90.2% with 9 turkers. For image-level multiclass, we see a similar trend. At the pixel level, the binary area overlap measure improves from 20.6% to 30.3% with 5 turkers but only to 31.4% with 9 turkers. Again, multiclass performance is similar (see Figure 4d). Even though group sizes beyond 5 continue to improve results at both the image and pixel level, this benefit may not be worth the additional cost.

Note that for the pixel level, the recall score rises dramatically in comparison to other metrics. This is because the consolidated majority vote pixel area tends to grow with turker count (with more pixels labeled, recall will go up).
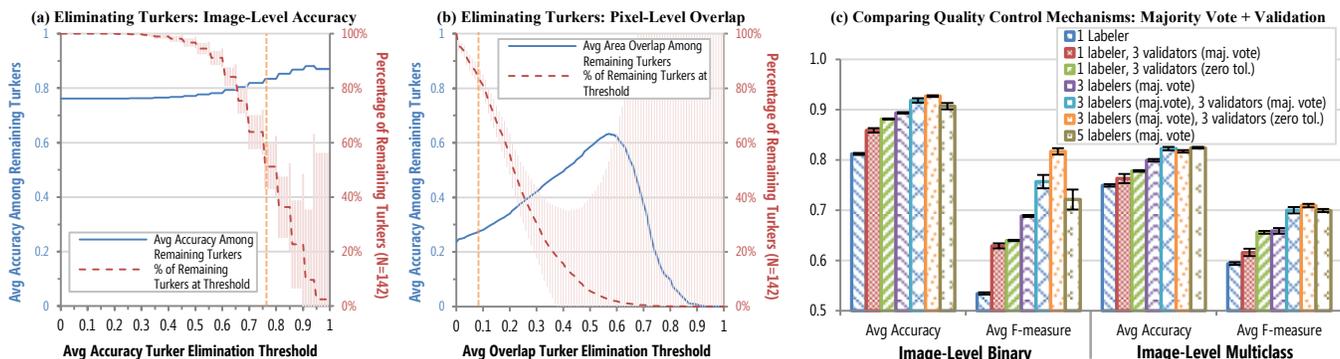
Figure 5: (a and b) Show the effect of increasingly aggressive turker elimination thresholds at the image- and pixel-levels based on average multiclass performance of 5 images. Error bars are standard deviation (for blue) and standard error (for red). As the threshold increases, fewer turkers remain and uncertainty increases. (c) Compares the effectiveness of various quality control mechanisms on performance at the image level.

Different consolidation processes will produce different results. Finally, similar to Study 1, *Sidewalk Ending* and *No Curb Ramp* labels performed better than *Object in Path* and *Surface Problem* labels (Table 3).

Quality Control Mechanisms

We explore two quality control approaches: filtering *turkers* based on a fixed threshold of acceptable performance and filtering *labels* based on crowdsourced validations collected through our verification interface. In both cases, we perform our analyses offline, which allows us to simulate performance with a range of quality control mechanisms.

Statistical Filtering: For the first approach, we explored the effect of eliminating turkers based on their average multiclass performance at both the image and pixel level. The goal was to uncover effective performance thresholds for eliminating poor quality turkers. We assign measure of errors to image-level and pixel-level correctness by using a Monte Carlo-based resampling approach called Bootstrap [13]. We first eliminate all turkers from our dataset who had completed fewer than five tasks. We then take samples of the remaining 142 turkers with replacement. For each sampled turker we randomly select five tasks that s/he completed to measure their average multiclass accuracy (for image level) or multiclass overlap (for pixel level). We shift our elimination threshold by increments of 0.01 and reject turkers if their average performance is lower than this threshold. At each increment, we also calculate overall performance across *all tasks* among the remaining turkers. We repeat this process independently at the image and pixel levels *N*=1000 times to calculate error bars.

Results are shown in Figure 5 (a and b). In both figures, we see overall performance steadily increase as poor performing turkers get eliminated. However, the threshold where elimination takes effect differs between the two mechanisms due to differences in difficulty. For example, to achieve the same accuracy level as we would expect from majority vote with 3 turkers (0.84), the average performance elimination threshold needs to be 0.76 (marked in orange in the graph). At that threshold, image-level multiclass accuracy amongst the remaining turkers goes up to 0.84, but at a cost of eliminating 51.2% of our workforce. For pixel-level data, to achieve a score similar

to the average area overlap between researcher labels (0.27), the elimination threshold needs to be set to 0.08, which increases the overlap score from 0.24 to 0.27 but reduces our workforce by 15% (again, orange line in graph). Thus, as expected, our results show accuracy gains with increasingly aggressive elimination thresholds; however, these accuracy gains come at a cost of reducing the effective worker pool. We expect that future systems can use these results to identify poor performing turkers *proactively* during data collection via ground truth seed images (*e.g.,* see [22]), and either offer additional training, or, in the extreme case, rejecting the work outright and blacklisting the turker. The threshold used depends on the accuracy needs of the application.

The Verification Interface: For the second quality control approach, we use our verification interface (Figure 3) to subjectively validate labels via crowdsourcing. Here, turkers validate existing labels rather than provide new ones. We ensured that the same turker did not label and validate the same image. As the validation task is simpler than the labeling task, we batched 20 validations into a single hit at a cost of 5 cents. We collected three or more validations per label across 75 images (the same subset used by the wheelchair users in Study 1). In all, we collected 19,189 validations from 273 turkers. Whereas the median time to label an image was 35.2s, the median time to validate a label was 10.5s. Thus, collecting validations is quicker and cheaper than collecting new labels.

We performed a series of analyses with the validation data, using both majority vote validation and zero tolerance validation. For the latter, if any validator down-votes a label, that label is eliminated. We compare these results to no quality control (baseline), the use of majority vote labels, and a combination of majority vote labels plus subjective validation. Results are in Figure 5. As before, performance improves with additional turkers—either as labelers *or* as validators. The best performing quality control mechanism was 3 labelers (majority vote) plus 3 validators (zero tolerance) beating out 5 labelers (majority vote). This suggests that it is more cost effective to collect 3 labels *with* validation than 5 labels total per image, particularly given that validation requires less effort.
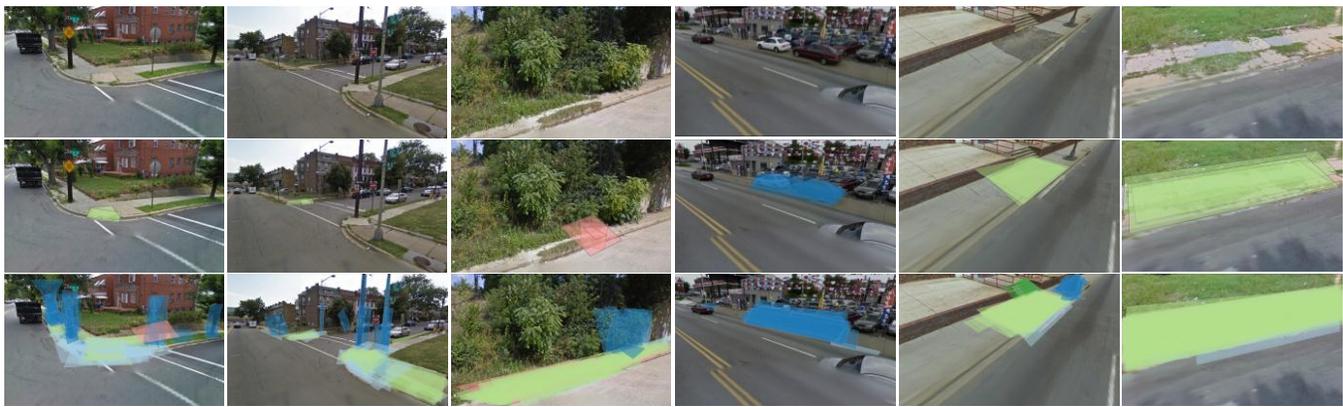
(a) 0.80, 0.59; 0.20, 0.0　　(b) 0.81, 0.70; 0.0 , 0.0　　(c) 0.86, 0.59; 0.15, 0.0　　(d) 0.74, 0.89; 0.52, 0.52　　(e) 0.97, 0.79; 0.58, 0.55　　(f) 0.94, 0.96; 0.63, 0.61

Figure 6: A selection of the bottom and top three performing images in our dataset based on *multiclass* pixel-level area overlap. Top row: original GSV image; middle row: majority vote ground truth from researchers using 15% overlap; bottom row: turker labels. Numbers show turker performance results for that image, from left to right: image-level binary, image-level multiclass; pixel-level binary, pixel-level multiclass.

Best and Worst Performing Images

Finally, to uncover what aspects of an image make it particularly easy or difficult to label, we sorted and visually inspected images in our dataset by multiclass pixel-level area overlap performance. Figure 6 shows a selection of the bottom and top performing images (left 3, right 3). For the worst performing images, there are many false positives: for example, utility poles and stop signs labeled as obstacles even though they are not in the sidewalk path. Figure 6c highlights two additional common problems: first, problem types can have ambiguous categories—in this case, the ground truth label indicates *Sidewalk Ending* while many turker labels selected *Surface Problem*; second, it is unclear *how much* of the problem area should be highlighted. For *Sidewalk Ending*, the ground truth labels highlight only the sidewalk termination point—some turkers, however, would label this section *and* any beyond it with no sidewalk (thereby greatly reducing their pixel-level scores). Future interfaces could detect these mistakes and provide active feedback to the turker on how to improve their labeling. In contrast, for the best performing images, the accessibility problems are, unsurprisingly, more salient and the camera angle provides a relatively close-angle shot.

DISCUSSION AND CONCLUSION

We have shown that untrained crowdworkers could find and label accessibility problems in GSV imagery. We also highlighted the effect of common quality-control techniques on performance accuracy. Here, we discuss limitations of our study and opportunities for future work.

Our prototype labeling system relied on a manually curated database of images selected by the research team. This approach was sufficient to demonstrate the feasibility of our idea but ignored important practical aspects such as locating the GSV camera in geographic space and selecting an optimal viewpoint. These challenges clearly need to be solved to produce a scalable approach. We are currently working on multiple solutions including: (i) a software agent that virtually "drives" through city streets in GSV and attempts to snapshot optimal angles of sidewalks and street intersections and (ii) deferring this complexity to the crowd

worker by allowing them to control camera angle and zoom level in an updated labeling interface.

An additional limitation relates to the GSV images themselves. Image quality can sometimes be poor due to lighting conditions, which can often be auto-corrected, or blurriness. More work is also needed to assess the degree to which sidewalk occlusion (*e.g.*, parked cars) is an issue in GSV images across different regions. Finally, GSV image age is also a potential problem (recall that images were 3.1 years old on average in our dataset). The following factors should mitigate this lattermost issue: (i) as noted earlier, virtual GSV neighborhood audits and physical audits have resulted in high concordance for pedestrian infrastructure data [5,11,25]; (ii) GSV is already being treated as a valuable resource by the accessibility community—*e.g.,* one of our mobility-impaired participants mentioned that he uses GSV to examine an area for traversability before leaving his house; (iii) GSV imagery is often updated as the GSV technology improves or simply to ensure accuracy *e.g.,* Google updated 250,000 miles of roads in early Oct 2012 (http://goo.gl/hMnM1). Moreover, all of the above GSV limitations may be potentially resolved through other data sources such as high-resolution top-down satellite or fly-over imagery [30], volunteer-contributed geo-located pictures (*e.g., SeeClickFix*), or government 311 databases.

While we captured important accessibility characteristics of sidewalks, other problems may exist. For example, the wheelchair users in Study 1 indicated that sidewalk narrowness can also reduce accessibility. We did not have a means of measuring sidewalk width or assess narrowness. Future work should look at the ability to calculate widths (*e.g.,* via computer vision-based mensuration), which could, perhaps, be reconstructed via the multiple camera angles offered by GSV or derived from the 3D-point cloud data that modern GSV cars collect (see [4]). While this 3D data is not yet publicly available, it could also be useful in object detection for automatically identifying problems.

We intend to integrate computer vision (CV) into our approach primarily for image triage, view selection,

mensuration, and semi-automatic object detection. Doing so will allow for further scalability, for example, where turkers verify automatically generated labels. However, while our current pixel-level results should be useful for *localizing* where problems exist in images, they may not be sufficient for training CV algorithms. To capture higher quality training data for CV, a future labeling tool should provide finer granularity outlines, feedback to turkers about their performance, proactive quality control, and better training.

For quality control, future applications will be using images where ground truth is *unknown*. Instead, "ground truth" seed images will need to be injected into the labeling dataset to actively measure turker performance (see [22]). Active monitoring will allow turkers to receive performance feedback, help assist them when they make common mistakes, and warn and, eventually, eliminate poor quality workers if they do not improve. Beyond turkers, we also plan to build a volunteer-based participatory website to both visualize our results and highlight areas that need data collection. In contrast to our current interface, we could allow for collaborative editing (*e.g.,* [26]) and experiment with incentivizing volunteers (*e.g.,* through gamification).

Our general approach of collecting useful, street-level information in a scalable manner from GSV images has application beyond sidewalks. We would like to expand our approach to assess the accessibility of building fronts, friction strips and stop lights at intersections (similar to [14] but with pixel-based labeling), and non-accessibility related topics such as tracking and labeling bike lanes in roadways.

REFERENCES

1. von Ahn, A. & Dabbish, L. (2004). Labeling images with a computer game. *CHI '04*, 319-326.
2. von Ahn, L., Liu, R., & Blum, M. (2006). Peekaboom: a game for locating objects in images. *CHI'06,* 55-64.
3. Americans with Disabilities Act of 1990, Pub. L. No. 101-336, 104 Stat. 328.
4. Anguelov, D., Dulong, C., *et al.* (2010). Google Street View: Capturing the World at Street Level. *IEEE Computer,* V.43, 32-38
5. Badland, H., Oplt, S., Witten, K., Kearns, R., & Mavoa, S. (2010). Can virtual streetscape audits reliably replace physical streetscape audits? *Journal of Urban Health*, 1007-1016.
6. Berkeley Segmentation Dataset and Benchmark (2007), http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/
7. Bernstein, M., Little, G., *et al.* (2010). Soylent: a word processor with a crowd inside. *UIST'10*, 313-322.
8. Bigham, J., Jayant, C., *et al.*, 2010. VizWiz: nearly real-time answers to visual questions. *UIST'10*, 333-342.
9. Chen, D., Bilgic, M., Getoor, L., and Jacobs, D. (2011). Dynamic Processing Allocation in Video. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 11, 2174-2187.
10. Cheng, C., Koschan, A., *et al.*(2012) Outdoor scene image segmentation based on background recognition and perceptual organization. *IEEE Trans Image Process*. 21(3):1007-19
11. Clarke, P., Ailshire, J., *et al.* (2010). Using Google Earth to conduct a neighborhood audit: Reliability of a virtual audit instrument. *Health & Place*, *16*(6), 1224-1229.
12. Clifton, K., Smith, A., & Rodriguez, D. (2007) The development and testing of an audit for the pedestrian environment. *Landscape and Urban Planning*; 80(1–2):95-110
13. Efron, B. & Tibshirani, R. (1994) An introduction to the Bootstrap. New York: Chapman & Hall.
14. Guy, R. & Truong, K. (2012). CrossingGuard: exploring information content in navigation aids for visually impaired pedestrians. *CHI'12*, 405-414.
15. Hara, K., Le, V., & Froehlich, J. (2012). A feasibility study of crowdsourcing and google street view to determine sidewalk accessibility. *ASSETS '12 Poster*. 273-274.
16. Heer, J. & Bostock, M. (2010). Crowdsourcing graphical perception: using mechanical turk to assess visualization design. *CHI'10*, 203-212.
17. Institute of Transportation Engineers, Pedestrian mobility and safety audit guide, http://www.ite.org/PedAudits/AuditGuide.pdf
18. Landis, R. & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*; 33:159 –74.
19. Leslie, E., Coffee, N., Frank, L., Owen, N., Bauman, A., Hugo, G. (2007) Walkability of local communities: Using geographic information systems to objectively assess relevant environmental attributes. *Health Place*. 3(1):111–22.
20. 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
21. Kinney *et al. vs.* Yerusalim & Hoskins. 93-1168 U.S. (1993). http://www.access-board.gov/PROWAC/yerusalim.htm
22. Quinn, A. & Bederson, B. 2011. Human computation: a survey and taxonomy of a growing field. *CHI'11*, 1403-1412.
23. Martin, D., Fowlkes, C., Tal, D., Malik, J. (2001) A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *ICCV'01*. 2: 416-23
24. National Council on Disability. (2007) The impact of the Americans with Disabilities Act: Assessing the progress toward achieving the goals of the ADA.
25. Rundle, A., Bader, M., Richards, A., Neckerman, K., & Teitler, J. (2011) Using Google Street View to audit neighborhood environments. *Am J Prev Med*; 40(1):94-100
26. Russell, B., Torralba, A., Murphy, K., & Freeman, W. (2008). LabelMe: A Database and Web-Based Tool for Image Annotation. *Int. J. Comput. Vision* 77, 1-3, 157-173.
27. Sampson, R. & Raudenbush, S. (1999) Systematic social observation of public spaces: a new look at disorder in urban neighborhoods. *Am J Social*; 105(3):603–51
28. Sorokin, A., Forsyth, D. (2008) Utility data annotation with Amazon Mechanical Turk. *CVPR'08*. 1-8.
29. StreetsWiki, Walk Audit, http://streetswiki.wikispaces.com/Walk+Audit
30. Taylor, B., Fernando, P., Bauman, A., Williamson, A., Craig, J., Redman, S. Measuring the quality of public open space using google earth. American Journal of Preventive Medicine 2011; 40(2):105-12
31. Timp, S. & Karssemeijer, N. (2004) A new 2D segmentation method based on dynamic programming applied to computer aided detection in mammography. *Medical Physics*. 31(5):958-72
32. U.S. Census Bureau (2010), Americans with Disabilities: 2010 Household Economic Studies, http://www.census.gov/prod/2012pubs/p70-131.pdf
33. U.S. Department of Transportation, Designing Sidewalks and Trails for Access, http://www.fhwa.dot.gov/environment/bicycle_pedestrian /publications/sidewalks/index.cfm
34. U.S. Department of Transportation, Road Safety Audits (RSA), http://safety.fhwa.dot.gov/rsa/
35. Yao, B., Yang, X., & Zhu, S. (2007). Introduction to a large-scale general purpose ground truth database: methodology, annotation tool and benchmarks. *EMMCVPR'07*, 169-183.
36. Zhang, Y. (2001) A review of recent evaluation methods for image segmentation. *ISSPA'01*. 1:148-51.