Improving Public Transit Accessibility for Blind Riders by Crowdsourcing Bus Stop Landmark Locations with Google Street View: An Extended Analysis

KOTARO HARA, University of Maryland, College Park
SHIRI AZENKOT, Cornell Tech
MEGAN CAMPBELL and CYNTHIA L. BENNETT, University of Washington
VICKI LE, SEAN PANNELLA, and ROBERT MOORE, University of Maryland, College Park
KELLY MINCKLER and ROCHELLE H. NG, University of Washington
JON E. FROEHLICH, University of Maryland, College Park

Low-vision and blind bus riders often rely on known physical landmarks to help locate and verify bus stop locations (e.g., by searching for an expected shelter, bench, or newspaper bin). However, there are currently few, if any, methods to determine this information *a priori* via computational tools or services. In this article, we introduce and evaluate a new scalable method for collecting bus stop location and landmark descriptions by combining online crowdsourcing and Google Street View (GSV). We conduct and report on three studies: (i) a formative interview study of 18 people with visual impairments to inform the design of our crowdsourcing tool, (ii) a comparative study examining differences between physical bus stop audit data and audits conducted virtually with GSV, and (iii) an online study of 153 crowd workers on Amazon Mechanical Turk to examine the feasibility of crowdsourcing bus stop audits using our custom tool with GSV. Our findings reemphasize the importance of landmarks in nonvisual navigation, demonstrate that GSV is a viable bus stop audit dataset, and show that minimally trained crowd workers can find and identify bus stop landmarks with 82.5% accuracy across 150 bus stop locations (87.3% with simple quality control).

Categories and Subject Descriptors: H.5 [**Information Interfaces and Presentation**]: User Interfaces; K.4.2 [**Social Issues**]: Assistive Tech for Persons with Disabilities

General Terms: Measurement, Design, Experimentation, Human Factors

Additional Key Words and Phrases: Crowdsourcing accessibility, accessible bus stops, Google Street View, Mechanical Turk, low-vision and blind users, remote data collection, bus stop auditing

ACM Reference Format:
Kotaro Hara, Shiri Azenkot, Megan Campbell, Cynthia L. Bennett, Vicki Le, Sean Pannella, Robert Moore, Kelly Minckler, Rochelle H. Ng, and Jon E. Froehlich. 2015. Improving public transit accessibility for blind riders by crowdsourcing bus stop landmark locations with Google Street View: An extended analysis. ACM Trans. Access. Comput. 6, 2, Article 5 (March 2015), 23 pages.
DOI: http://dx.doi.org/10.1145/2717513

This work is supported by the National Science Foundation, under grant IIS-1302338, a Google Faculty Research Award, funding from AT&T Research Labs, and an IBM Ph.D. Fellowship.
A preliminary version of this article appeared in the Proceedings of the 15th ACM SIGACCESS International Conference on Computers and Accessibility, October 21–23, 2013.

Authors' addresses: K. Hara, V. Le, S. Pannella, R. Moore, and J. E. Froehlich, Department of Computer Science, University of Maryland, College Park, 20742 MD; S. Azenkot, Jacobs Technion-Cornell Institute, Cornell Tech, 111 8th Ave. New York NY 10011; M. Campbell, C. L. Bennett, K. Minckler, and R. H. Ng, Computer Science Department, University of Washington, 185 W Stevens Way NE, Seattle, WA 98195.

DOI: http://dx.doi.org/10.1145/2717513

1. INTRODUCTION

For people who are blind or have low vision, public transportation is vital for independent travel [American Foundation for the Blind 2013; Blind Citizens Australia 1994; National Federation of the Blind 1986; Marston and Golledge 2003]. In previous formative work, we interviewed six blind adults about accessibility challenges in using public transportation [Azenkot et al. 2011]. We found that while buses were frequently a preferred mode of transit for our participants, determining the exact location of a bus stop was a major challenge [ibid., p. 3249]. Strategies for finding bus stops included asking other pedestrians for information (if available) or locating known landmarks such as bus stop signs, shelters, or other physical objects (e.g., benches).

In this article, we focus specifically on the role of landmarks in helping blind and low-vision people find and identify bus stop locations. While some transit agencies provide brief descriptions of their bus stops online (e.g., Metro Transit–King County DOT [2013]), this information often lacks detail or is inaccessible to visually impaired riders—if available at all. Similar to our previous interview findings [Azenkot et al. 2011], the American Foundation for the Blind (AFB) notes that locating bus stops is a significant access barrier often because the bus stops are not clearly marked with nonvisual indicators or are placed inconsistently off roadways [AFB 2013]. The challenge of locating a bus stop is exacerbated when traveling to an unfamiliar location where both the bus stop placement and the position and type of surrounding landmarks are not known to the traveler *a priori*.

In this article, we introduce and evaluate a new method for collecting bus stop location and landmark descriptions using online crowdsourcing and Google Street View (GSV). Using a custom tool that we built called *Bus Stop CSI (Crowdsourcing Streetview Inspections),* crowd workers virtually navigate to and label bus stop signs and surrounding landmarks in GSV (e.g., Figure 1). This new approach is highly scalable in comparison to previous bus stop crowdsourcing work (e.g.*, GoBraille* [Azenkot et al. 2011] and *StopInfo* [Prasain 2011; Campbell et al. 2014]), which required users to describe bus stops *in situ* using a mobile device. While this article focuses largely on data collection methods, we envision future work that integrates this data into transit agency websites and location-aware mobile transit tools such as *OneBusAway* [Ferris et al. 2010] and *StopInfo* [Campbell 2014]. For example, imagine a smartphone application that uses GPS and text-to-speech to automatically describe nearby and upcoming landmarks as a blind pedestrian navigates toward a bus stop.

We report on three studies, beginning with an interview study (Study 1) of 18 people with visual impairments (seven with no functional vision) to inform the design of our crowdsourcing tool. These interviews extend our aforementioned formative work [Azenkot et al. 2011] and further emphasize the importance of nonvisual landmarks in helping blind/low-vision travelers find and verify a bus stop location. We then transition to describing two studies of GSV: a comparative study (Study 2) examining differences between physical bus stop audit data and audits conducted virtually with GSV, and an online study (Study 3) using Amazon Mechanical Turk (MTurk) designed to examine the feasibility of crowdsourcing bus stop audits using our tool.

In Study 2, we found a high correlation between our physical bus stop audit data and GSV images across four field sites in the Washington, DC and Seattle metropolitan areas. This finding provides initial support for using GSV as a viable bus stop audit method. In Study 3, 153 MTurk crowd workers (turkers) labeled 150 bus stops using GSV via our custom tool. Overall, our results show that an individual turker is able to find and correctly label a bus stop and surrounding landmarks (e.g., benches, trash cans) with 82.5% accuracy. This increases to 87.3% with simple seven-turker majority vote for quality control. While not perfect, these results point to the feasibility of using

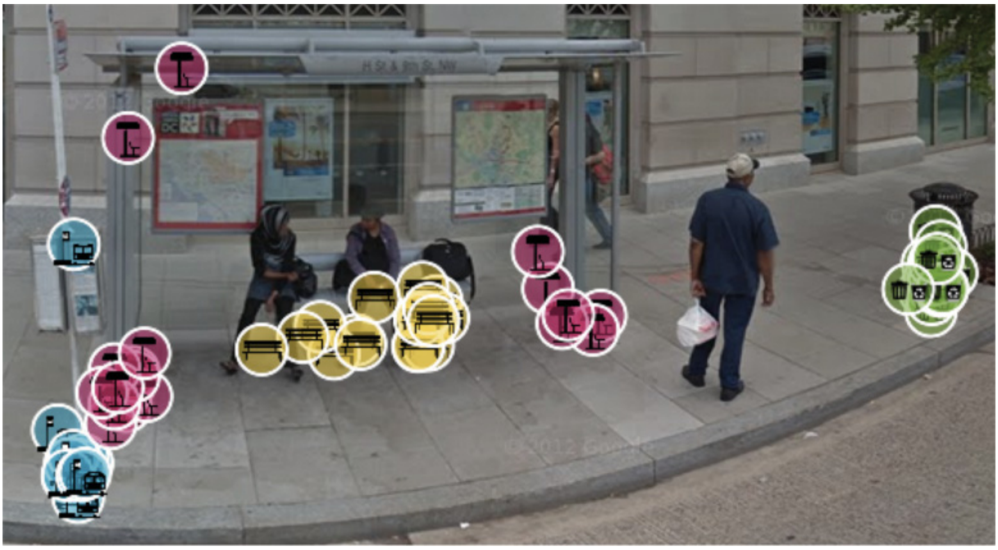


Fig. 1. We built a system that allowed crowd workers to collect detailed information about bus stops using Google Street View (GSV) to improve accessibility. Crowd workers were asked to label the landmarks in a GSV image. The image shows actual labels from crowd workers (Study 3). From left to right: blue circular icon = *bus stop sign*, magenta = *bus stop shelter*, yellow = *bench*, green = *trash/recycling can*.

GSV and crowdsourcing to remotely gather detailed bus stop descriptions. Future work should focus on crowd worker training, quality control to increase accuracy, and methods to address false-negative labeling errors discussed in the following.

In summary, the contributions of this article are threefold, involving both formative and summative findings: (i) our interview study adds to the existing literature on how blind and low-vision persons use bus transit, with a specific focus on navigating to and identifying bus stops; (ii) our comparative physical versus virtual bus audit study is the first of its kind for bus stop auditing and establishes that GSV is a viable data source for remotely collecting descriptions of bus stop features and surrounding landmarks; and, finally, (iii) our custom tool (Bus Stop CSI) and online crowdsourcing study shows that minimally trained crowd workers can find and describe bus stops using GSV with reasonable accuracy (>82% without quality control).

1.1. Authors' Note

This journal article is an invited TACCESS submission from the ACM ASSETS 2013 conference proceedings. We offer three main extensions from the conference counterpart [Hara, Azenkot, et al. 2013]. First, we include additional formative study results from our interviews with people with visual impairments. Second, we offer a more detailed description of the interface we developed to train turkers how to perform our labeling tasks. Finally, we provide an expanded analysis of turker labeling performance and attempt to uncover what characteristics make a bus stop hard or easy to label.

2. RELATED WORK

Using public transit requires navigating a wealth of visual information from maps and schedules to bus stop markings and bus route signs. This reliance on visual information makes using public transit difficult for people with severe visual impairments [AFB 2013; Marston and Golledge 2003]. With bus transit specifically, blind or low-vision people can struggle with determining route and schedule information, purchasing fare, finding the correct bus stop location, getting on the appropriate bus, and getting off at the right stop [Yoo et al. 2010; Golledge et al. 1997; Azenkot et al. 2011; AFB 2013].

Most transit tools designed to assist blind and low-vision bus riders focus on two issues: helping identify the correct bus to board when waiting at a bus stop [Banâtre et al. 2004; Noor et al. 2009] or providing alerts for an upcoming stop while riding the bus [Jacob et al. 2011; Kostiainen et al. 2011]. We are interested in addressing a prerequisite challenge: helping visually impaired riders find and verify bus stop locations through the use of physical landmarks and detailed bus stop descriptions (e.g., the presence of benches and bus shelters). In a survey of 55 persons with visual impairments, 85% reported difficulties in finding public transit pickup points such as bus stops [Golledge et al. 1997]. Recent work has emphasized the importance of physical landmarks in helping low-vision and blind users navigate to public transit [Guentert 2011; Azenkot et al. 2011; Campbell et al. 2014]. Landmarks can only be used for navigation, however, when their location and spatial context (e.g., proximity to other physical objects) is known. Typically, this information is not captured or shared via traditional navigation tools (e.g., online maps).

Most relevant to our work is the GoBraille project [Azenkot et al. 2011] and its follow-up StopFinder [Prasain 2011], which was later renamed to StopInfo [Campbell 2014]. Both projects emphasize *in situ* mobile crowdsourcing to collect and present data about bus stops and surrounding landmarks to aid blind travelers (ibid., p. 323). Their *in situ* crowdsourcing approach takes advantage of the traveler's "downtime" while waiting for a bus: users fill out a simple form describing the bus stop (e.g., its location, relative direction, and encountered landmarks). While the reliance on blind users for bus stop data provides insights that are important to that community (e.g., nonvisual perceptions of a landmark), the approach has issues of critical mass and data scarcity. While our aim is similar, the approach we present here is unique: crowdsourcing data collection online using GSV where anyone at any time can contribute.

The use of omnidirectional streetscape imagery such as that found in GSV, Microsoft Bing Maps, and some Nokia Maps has become increasingly popular as a virtual audit technique in fields from urban informatics to public health research [Badland et al. 2010; Rundle et al. 2011; Clarke et al. 2010; Guy and Truong 2012; Hara et al. 2012; Hara, Le, et al. 2013]. Reported benefits over physical audits include time savings and the ability to monitor and analyze multiple cities from a central location [Badland et al. 2010; Rundle et al. 2011]. As an emerging area of research, most work thus far has focused on examining agreement between virtual (e.g., GSV) and physical field audit data (e.g., Badland et al. [2010], Rundle et al. [2011], Clarke et al. [2010], and Guy and Truong [2012]). Important for our work here, high levels of agreement have been found for measures including pedestrian safety, traffic and parking, and pedestrian infrastructure. To our knowledge, however, no one has specifically looked at the concordance between physical and virtual audit data for *bus stops* and their surrounding environment (which is the focus of Study 2).

With regard to crowdsourcing for accessibility, Bigham and colleagues argue that current technological infrastructure provides unprecedented access to large sources of human power that can be harnessed to address accessibility challenges [Bigham et al. 2011]. Recent examples of such crowdsourcing systems include VisWiz [Bigham et al. 2010] and Legion:Scribe [Lasecki et al. 2012]. More relevant to our work is Tiramisu [Steinfeld et al. 2011], a mobile crowdsourcing tool developed via universal design to help gather and disseminate information about bus arrival time and capacity. Our approach is complementary but does not rely on mobile crowdsourcing or continuous, active use by crowd workers to provide benefits. Finally, in the last decade, a growing number of crowdsourcing systems dedicated to geographic content have emerged (e.g., Wikimapia, OpenStreetMap, and Cyclopath [Panciera et al. 2010]). Interestingly, past work has found that user-contributed map data quality is high even when compared to proprietary systems (e.g., Flanagin and Metzger [2008] and Haklay [2010]). Though we currently rely on paid labor via MTurk, we plan to explore community-sourcing and volunteer contribution.

3. STUDY 1: FORMATIVE INTERVIEWS

In 2010,[1] we conducted formative interviews with six blind adults to learn about the challenges faced by visually impaired persons when using public transit [Azenkot et al. 2011]. Here, we extend this previous work by covering a wider variety of transit systems and involving a greater diversity of visually impaired participants. In addition, we specifically investigate the role of nonvisual landmarks in bus stop navigation, which is the primary focus of this article.

3.1. Interview and Analysis Method

We recruited 18 participants (10 male) with visual impairments from the United States and Canada with an average age of 52.1 (SD = 12.0; range = 24–67). Eleven participants could not easily read street signs due to their visual impairment. Of these, seven had no functional vision. As bus transit systems differ across population densities, we sought participants from different neighborhood types: eight participants lived in urban areas, seven suburban, and three in small towns. Participants were recruited via mailing lists affiliated with blindness organizations and were paid $15. The recruitment email stated that we were investigating public transit accessibility and that participants must be blind or low vision.

We conducted semistructured phone-based interviews, which lasted approximately 40 minutes. We asked participants about patterns of public transit use, challenges experienced therein, and coping and mitigation strategies. We then described a hypothetical smartphone application that provided the location and description of bus stops and surrounding landmarks during the user's journey (e.g., via GPS tracking and text-to-speech). We asked participants to assess the importance of various landmarks for this software application. We recorded, transcribed, and coded the interviews using an open coding methodology. While our interviews covered a broad range of subjects related to transit accessibility, in the following we primarily concentrate on findings related to locating bus stops.

3.2. Bus Stop Related Interview Findings

For most participants, public transit was critical for daily mobility. One woman, for example, stated that the lack of accessible public transit "*played into her decision*" to retire. Other forms of transit mentioned included walking, rides from family members, and paratransit. Experiences with paratransit service varied, however. In Seattle and Washington, DC, paratransit includes shared rides that must be reserved in advance. However, the service is often running behind schedule and passengers endure long routes to accommodate all riders. One participant expressed his preference for the fixed-route bus system as opposed to his local paratransit service because he can "*gain more independence*" and the fixed-route system is "*cheaper [for tax payers] and more sustaining.*"

Similar to prior work [Yoo et al. 2010; Azenkot et al. 2011; Golledge et al. 1997], participants (both blind and low-vision) described challenges when using public transit including finding bus stops, knowing which bus to board, and when to disembark. Most transit agencies in the United States require their bus drivers to announce upcoming

[1] Interviews were done in 2010, but the findings were published in 2011.

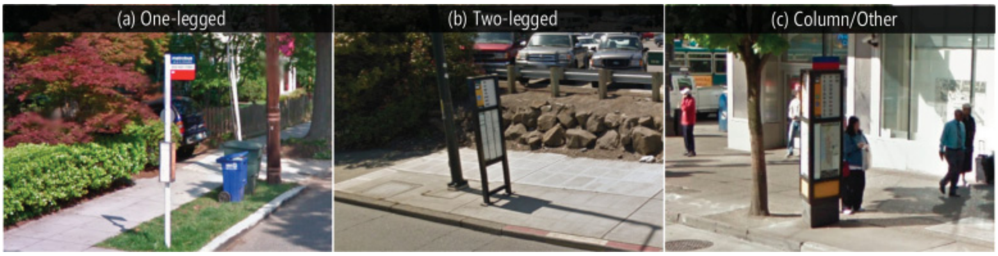


Fig. 2. Three different bus stop sign designs: (a) one-legged, (b) two-legged, and (c) column/other. Some interviewees mentioned that knowing the shape of a bus stop sign *a priori* is helpful for locating a stop. These examples are pulled from our Study 2 and 3 areas in Washington, DC (left) and Seattle ((a) and (b)).

stops (e.g., U.S. DOT [2007]); modern systems often make these announcements automatically. Similarly, some newer bus systems announce their bus route number and direction automatically when stopping to let passengers board. Not a single participant, however, found such announcements reliable, most often because of issues with audibility and accuracy (e.g., the announcement is out of sync with the vehicle's path).

When waiting at a stop, many participants found it difficult and stressful to know which bus to board when multiple buses arrived at a stop at the same time, since they could not hear the announcements for all buses. Nearly all participants relied heavily on asking bus drivers for information. One low-vision participant felt that he was at the "*discretion*" and "*mercy*" of the bus driver.

Most relevant to this article, half of the participants experienced difficulty finding the exact location of bus stops when traveling (three blind, six low-vision). Difficulties included determining the specific location of a bus stop (e.g., near-side of intersection, halfway down the block), obtaining accessible information sources, and knowing which landmarks and businesses indicate a proximal bus stop. Because bus stop designs and placement can vary widely within a city—from stops with a myriad of physical landmarks (e.g., shelters, benches, trash cans, and newspaper boxes) to stops with only a pole—one participant said with frustration:

> *There's really no rhyme or reason of where they put bus stops. And there's no way to . . . tell where a bus stop [is], 'cause you don't ever know where the pole is, or how it's marked, or . . . anything like that.* (P3, age = 63, blind)

For this participant, the main reason he did not use public transit was because of the challenges he faced in finding bus stops. Another participant noted that some stops in his city were hard to find because they had no nonvisual landmarks, only painted curbs. Many noted that consistent stop locations and landmarks would significantly help them overcome this accessibility challenge (Figures 2 and 3). For both blind and low-vision participants, finding an unfamiliar stop took a lot of time and, as one participant explained, required adjusting expectations to reduce stress:

> *I think also just not to worry about it so much. Just not stress out about it. Just know that it will be new and it will take a little more time to figure it out.* (P14, age = 55, blind)

To find bus stops, participants mentioned using walking directions from transit trip planners (if available in an accessible form), calling the transit agency,[2] or asking a sighted person questions about the stop's location. Ten participants (53%; six blind,

[2] In our prior work, one participant noted poor experience with calling transit agencies because they could not adequately explain bus stop locations over the phone (perhaps because the agency itself did not store sufficiently detailed descriptions about their bus stops in their database) [Azenkot et al. 2011].

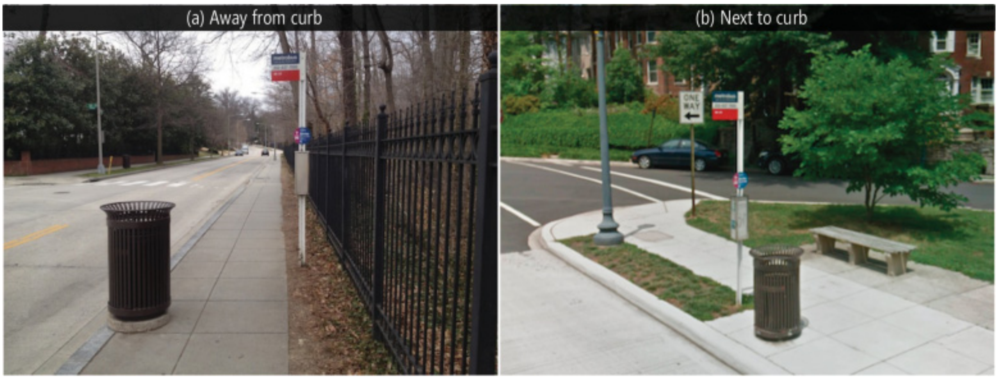


Fig. 3. The position of bus stop signs relative to the curb is inconsistent within and across geographic regions (both photos are from Washington, DC). Some of our interviewees expressed interest in knowing whether a bus stop sign is placed (a) away from the curb or (b) next to the curb. Occasionally, bus stop signs are attached to existing structures rather than their own poles, further obscuring their nonvisual accessibility.

four low-vision) reported asking pedestrians or other transit riders for information—a strategy only available when others are present (i.e., more difficult at night or in more rural areas). Some participants used orientation or mobility instructors to help guide them to routine bus stops. Once participants reached the vicinity of the stop, they commonly searched for landmarks. For example, if a person uses a cane, she or he can hear an echo from a shelter when walking by.

When asked about which landmarks at bus stops are most important to navigation, participants identified shelters and benches as the most helpful, followed by trash cans, newspaper bins, and grass shoulders. Blind people could find such landmarks by walking with a cane or a guide dog and low-vision people could see these landmarks when walking close to them. A few blind participants also mentioned knowing the shape of the bus stop pole (e.g., thin vs. thick, two-column vs. one). One participant emphatically stated that all landmark information would be of critical importance. Some participants expressed more fine-grained information as important, such as types of sidewalk surface material, position of a bus stop sign relative to the curb (Figure 3), and the presence of a grass shoulder nearby the stop. Five participants (three blind, two low-vision) also mentioned the importance of knowing nearby businesses because of their distinct sounds and smells.

> *I look for landmarks . . . like a bus shelter at a certain place . . . or if there's a hedge, like bushes in front of a certain place and right by those bushes there's a newspaper rack or something like that then I know that it's my stop. If it's in front of a coffee shop . . . if there's a hotdog stand there, then I know that the bus stop is in front of the hot dog stand, you smell it . . . Noises too, you know different sounds.* (P14, age = 55, blind)

Though participants relied on various technologies for planning a trip on public transit, only five participants (26%; three blind, two low-vision) used smartphone applications for such tasks. These applications provided either real-time or scheduled arrival information, and helped participants determine which bus to board. GPS-based tools (e.g., iMove[3] and Sendero GPS LookAround[4]) were used to identify an address of one's current location; useful information to identify if he or she is on the correct street where a bus stop is located. None of the participants used technology tools to "tag" the

[3] iMove: https://itunes.apple.com/us/app/imove/id593874954?mt=8.
[4] Sendero LookAround: https://itunes.apple.com/us/app/sendero-gps-lookaround/id386831856?mt=8.

specific location of a stop, perhaps because they did not know how to do so. One participant who did not own a smartphone said she would likely buy one if she believed there would be good navigation applications for blind people. Another participant expressed his distress with unreliable technology tools.

> *One of the trip planners gave me accurate info and the other one gave me an inaccessible map or text-based directions so I found that I have to use them both in tandem which makes it more complicated and confusing. I kind of preferred to call the 800 number for the [transit agency] customer service.* (P18, age = 31, low-vision)

This quote highlights not only the importance of making navigation technology tools accessible for blind and low-vision users but also that the data they rely on is up to date and accurate.

3.3. Study 1 Summary

In summary, although our first interview study was conducted 3 years ago [Azenkot et al. 2011], the major challenges of blind and low-vision public transit riders remain the same. This is despite the technological improvements in navigation tools, smartphone applications, and accessible bus systems (e.g., automated announcements). Both blind and low-vision participants expressed similar challenges in public transit use. Seven blind and eight low-vision participants (out of 15 total) said that having information about landmarks would enable them to use transit more easily (even five participants who could sometimes read street signs). Descriptions of the shape and location of bus stop poles, shelters, and benches as well as information indicating their presence seemed most beneficial.

4. STUDY 2: PHYSICAL VERSUS GSV AUDITS

To assess the viability of using GSV to audit bus stops, we needed to first establish that the bus stops captured in the GSV image dataset do not differ significantly from current reality (e.g., because of image age). Thus, in Study 2, we conducted both in-person bus stop audits and GSV-based audits across the same four target geographic areas and compared the results. An audit here means logging the existence of landmarks at bus stops using a predefined codebook (described in Section 4.2). While the primary aim of this study was to explore what differences, if any, would exist between the GSV and physical bus stop audit data, we had two secondary aims. First, to investigate the feasibility and difficulty of the audit task itself (e.g., can members of our research team agree amongst themselves on the application of audit measures across various bus stop scenes?). Second, to produce a high-quality *ground truth* dataset that could be used to assess crowd worker performance in Study 3.

Our bus stop audit sites included four neighborhoods in the Washington, DC and Seattle, WA metropolitan areas (Figure 4). As bus stop designs differ across cities and neighborhoods, we selected a range of densities (e.g., downtown vs. suburban) and neighborhood types (e.g., residential vs. commercial). Additionally, we emphasized areas that have high demand for public transit, including schools, major department stores, convention centers, and museums. These same areas are also used in our crowdsourcing audit study (Study 3).

4.1. Collecting Physical Audit Data

Two separate research teams physically visited the bus stop locations: one team in Seattle and the other in Washington, DC. Teams walked or biked down each street in the predefined study area. They carried smartphones with GPS to help navigate to and track bus stops. An online spreadsheet prefilled with bus stop locations (e.g., Baltimore and Campus Dr.) and a Google Map URL allowed the researchers to track their position

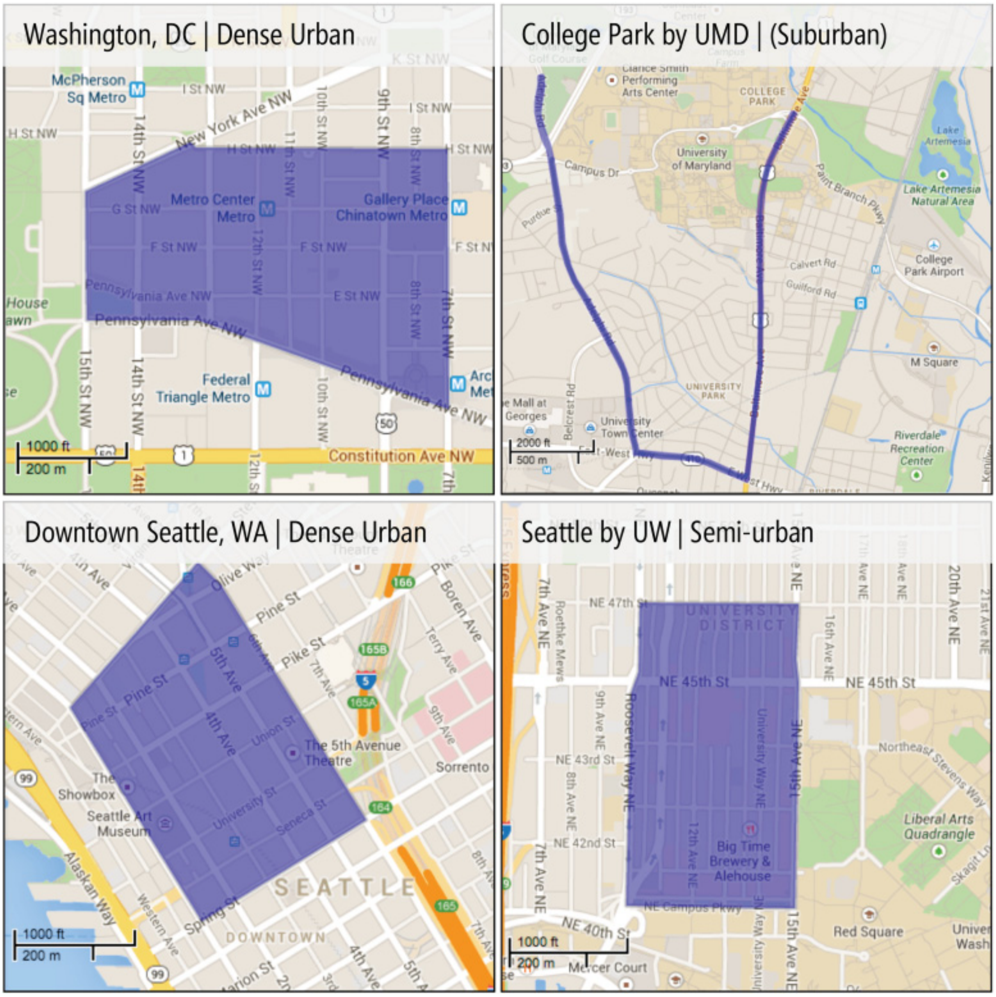


Fig. 4. The four audit areas used in Studies 2 and 3 spanning a range of neighborhood types in Washington DC and Seattle, WA. In all, we surveyed 179 bus stops across 42.2 linear kilometers. Each field site took approximately 1 day to physically survey (∼6h). See also Table I.

and the target bus stop in real time on an interactive map. Visited stops were marked in the spreadsheet and linked to a unique index for later analysis.

At each bus stop location, we took 7–10 geo-time-stamped pictures from varying angles—roughly 360 degrees around the bus stop from the sidewalk and street (far more angles than GSV)—and analyzed them *post hoc*. We were careful to capture clear images without occlusion problems. This photographic approach had two primary advantages: it created an image dataset analogous to GSV, which allowed us to apply a similar auditing methodology to both, and it allowed us to examine the image dataset multiple times without returning to the field site.

4.2. Auditing Methodology

While we had two separate teams photograph bus stops during the in-person field site visits, we used one single team of three researchers to independently audit (code)

Table I. The Four Areas Surveyed in Our Physical and Virtual (GSV) Audits
Total linear kilometers surveyed represents unidirectional surveying distance except for the *, which is bidirectional because of wider streets separated by a median (i.e., auditors walked/biked one side of road and then the other). See also Figure 4.

	Washington, DC	College Park by UMD	Downtown Seattle	Seattle by UW	Overall
Description of audit area	Dense urban	Suburban (next to Univ. of Maryland)	Dense urban	Semiurban (next to Univ. of Washington)	N/A
Total linear km surveyed	11.2	11.9*	8.0	11.1	42.2
# of bus stops found in physical audit	82	36	35	26	179
# of bus stops found in physical audit but missing from Google Maps API	21	4	3	1	29
Avg. GSV data age (SD)	1.9y (0.3)	1.0y (0.7)	1.9y (0.3)	2.1y (1.1)	1.75y (0.7)

both the physical and GSV image datasets. This reduced confounds due to different auditors. Three researchers who also conducted field visits in the Washington, DC area were responsible for the coding. Although bus stop auditing may seem like an objective process, it is, in fact, *subjective* and requires following a qualitative coding methodology. For example, one auditor may simply miss seeing a particular object in a scene or may misperceive or mislabel one object as another. By following the iterative coding method from Hruschka et al. [2004], our aim was to produce two high-quality audit datasets—one for each image dataset: physical and GSV—that could then be compared.

To begin the auditing process, an initial codebook was derived for each bus stop landmark: (i) bus stop signs, (ii) bus stop shelters, (iii) benches, (iv) trash or recycling cans, (v) mailbox and newspaper bins, and (vi) traffic signs and other poles. These landmarks were selected based on the findings from our interviews as well as from bus stop design guidelines (e.g., Easter Seals [2008]). The codebook provided detailed definitions of each along with visual examples. We also defined the audit area around a bus stop as 20ft (∼6.1m) in either direction from the bus stop sign (from [Intercity Transit 2010]). Note, however, that as our audits were performed via visual inspection of images (for both the physical and GSV datasets), auditors could only estimate distances.

During auditing, count data was entered into a preformatted spreadsheet tracking the number of each landmark at each bus stop. As prescribed by Hruschka et al. [2004], each auditor began by independently coding a small subset of data—in this case, 15 DC and five College Park bus stop locations. Afterwards, the auditors came together to discuss and modify problematic codes. With the updated codebook, the entire DC and College Park physical image dataset was audited (including the original 20 locations, which were reaudited) followed by the GSV dataset. We conducted a similar iterative coding process for the two Seattle audit areas. The codebook descriptions were updated to reflect Seattle bus stop designs.

The GSV audits differed from the physical image dataset audits in two ways: first, the auditors used a GSV interface where they could control camera angle and location rather than browse through a set of static images; second, the auditors rated the overall difficulty of auditing each location on a 5-point Likert scale, where 1 = very easy to assess and 5 = very hard to assess. These ratings will be used later in Study 3 to investigate whether crowdsourcing audit accuracy changes based on rated difficulty.

4.3. Interrater Agreement on Audit Data

Before comparing the physical audit data to the GSV audit data (Section 4.4), we needed to first calculate interrater agreement *between researchers* for each individual dataset. For this, we applied the Krippendorff's Alpha (α) statistical measure (see

Table II. Krippendorff's Alpha Interrater Agreement Scores between Three Researchers on Both the Physical Audit and GSV Audit Image Datasets

Following the iterative coding methodology of Hruschka et al. [2004], a second audit pass was conducted with an updated codebook for low-agreement scores—in our case, $\alpha < 0.800$. *Excluded categories for second coding pass because original agreement $\alpha = 0.800$.

	Physical Audit Image Dataset		GSV Audit Image Dataset	
Bus Stop Landmark	**First Pass (α)**	**Second Pass (α)**	**First Pass (α)**	**Second Pass (α)**
Bus stop sign	0.937	0.972	0.761	0.916
Bus stop shelter*	0.991	0.991	0.955	0.955
Bench*	0.940	0.940	0.870	0.870
Trash/Recycling can	0.876	0.886	0.793	0.946
Mailbox/Newspaper bin	0.886	0.957	0.768	0.938
Traffic signs/ Other poles	0.683	0.866	0.685	0.874
Overall	0 909	0 944	0 850	0 930

Krippendorff [2003]). Although we have previously used Fleiss' kappa to compute interrater agreement on streetscape audit tasks [Hara, Le, et al. 2013], this statistical measure cannot be applied to count data, which is what we have here. Our results are presented in Table II (first pass columns). The overall α score between researchers was 0.909 for the physical audit dataset and 0.850 for the GSV audit dataset.

Similar to most statistical measures for interrater agreement, there is no universally accepted threshold for determining high agreement with the Krippendorff's Alpha measure. However, Krippendorff [2003] suggests that agreement scores of $\alpha \geq 0.800$ are generally considered reliable, while data below $\alpha < 0.667$ should be discarded or recoded (p. 241). Though none of our α scores fell below 0.667 for either dataset, some categories had $\alpha < 0.800$. One primary source of disagreement involved differing perceptions of what geographic area constituted a bus stop (recall the 20ft perimeter). For some bus stop locations, traffic signs, poles, and other landmarks extended just beyond or just within the prescribed bus stop range. These edge cases were difficult to assess and contributed to the lower α score. Note also that the GSV agreement scores were lower on average than the physical audit dataset often because of inferior-quality images (e.g., the GSV privacy protection algorithm misidentified some bus stop signs as vehicle license plates and blurred them out; see Google [2013]).

To alleviate such disagreements as recommended by Hruschka et al. [2004], the three auditors discussed low agreement codes (any $\alpha < 0.800$) and updated the codebook once again. The auditors then took a second full independent pass on both the physical and GSV audit datasets but focused only on those bus stop landmarks that previously had an α score <0.800. The updated results are in Table II (second pass columns). On this second pass, the overall agreement increased from 0.909 to 0.944 for the physical audit dataset and from 0.850 to 0.930 for the GSV dataset. Importantly, all α scores were now ≥ 0.800 thereby completing our iterative coding scheme. The summary of the total number of each landmark and the number of bus stops with each landmark are shown in Table III.

4.4. Comparing Physical versus GSV Audit Data

The high agreement scores within the physical and GSV datasets provides evidence that the audit data is consistent and of good quality. Consequently, we can move toward examining the key research question of Study 2: How does the physical audit dataset compare to the GSV dataset? To investigate this question, two more small procedural steps are required: first, we need to amalgamate the three-auditor count data into a single count set for both datasets and then we need to decide upon some mathematical approach to compare them. For the amalgamation method, we take the median of the three auditor counts for each bus stop landmark at each bus stop location. For example,

Table III. Summary of Landmarks Found during Physical Audit and GSV Audit in 179 Bus Stop Locations
In each cell, values in parentheses indicate counts in the subset 150 of the 179 bus stops that were available in the Google Transit dataset. The numbers are the median of three researchers' count data.

	Bus Stop Sign		Bus Stop Shelter		Bench		Trash / Recycling		Mailbox / News. Bins		Traffic Signs / Other Poles	
	Physic al	GSV	Physical	GSV	Physical	GSV	Physical	GSV	Physical	GSV	Physical	GSV
# of the landmarks found in total	167 (140)	152 (128)	102 (88)	98 (84)	133 (112)	121 (101)	100 (81)	95 (86)	69 (57)	56 (47)	162 (132)	153 (126)
# bus stops with a corresponding landmark type	166 (139)	151 (127)	87 (73)	70 (67)	95 (78)	86 (69)	74 (59)	65 (58)	29 (24)	24 (21)	109 (89)	105 (86)

Table IV. Following Rundle et al. [Rundle et al. 2011], we Performed a Spearman Rank Correlation between the Physical and GSV Bus Stop Landmark Count Audit Datasets. For all coefficients (ρ), $p < 0.001$.

Physical vs. GSV Audit Data	Bus Stop Sign	Bus Stop Shelter	Bench	Trash / Recycling	Mailbox / News. Bins	Traffic Signs / Other Poles
Coefficient (ρ)	0.612	0.877	0.875	0.715	0.776	0.811

if R1 found one traffic sign at a specific bus stop location, R2 found four traffic signs, and R3 found five, then the median count between them would be four. This approach allowed us to transform the three count datasets into one for both the physical and GSV audit data. For the comparison method, similar to Rundle et al. [2011], we calculate a Spearman rank correlation between the two count sets (physical and GSV).

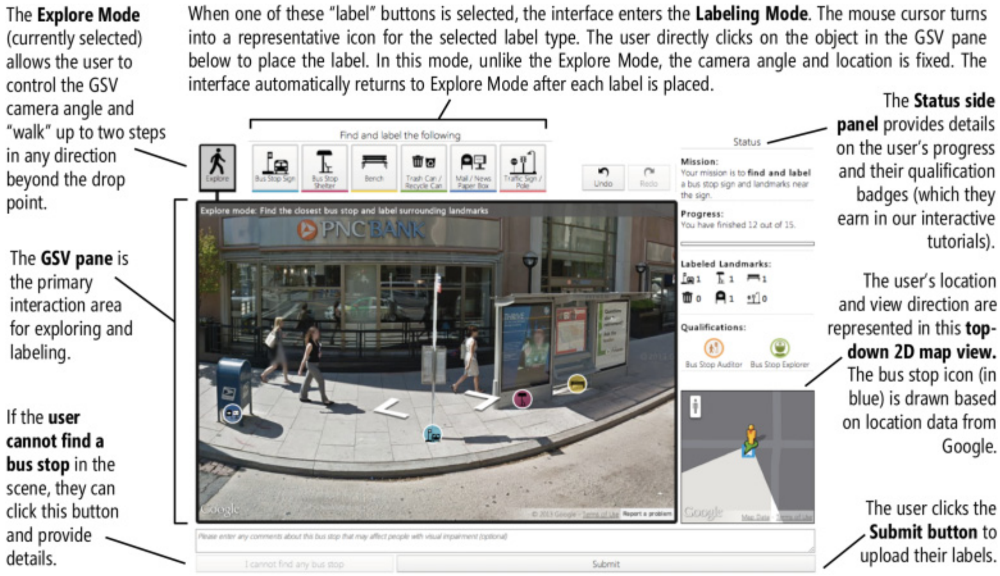
Our results are presented in Table IV; all are statistically significant at $p < 0.001$. Using the definition of high correlation by Rundle et al. [2011], all of our landmark coefficients (ρ) are highly correlated ($\rho > 0.60$) between the physical and GSV datasets. The two highest are for bus stop infrastructure: *Bus Stop Shelters* ($\rho = 0.88$) and *Benches* ($\rho = 0.88$). The two lowest are *Bus Stop Signs* ($\rho = 0.61$), which are sometimes difficult to see in GSV, and *Trash/Recycling Cans* ($\rho = 0.72$), which are likely to be the most transient landmark type (e.g., they may move a lot over time). It is important to note that during the physical audit, we encountered 29 bus stops that were *not* in Google Transit's bus stop location dataset (21 of which were in downtown Washington, DC); see Table I. This Google Transit dataset is independent of the GSV images. Only *three* of these bus stops, however, were also missing in GSV (due to outdated images). The preceding correlation results are for all 179 physical audit locations with zeros filled in for the three missing bus stops in the GSV datasets.

4.5. Study 2 Summary

In summary, Study 2 demonstrates that bus stop auditing is a subjective process but, more importantly, that the GSV audit dataset is highly correlated with the physical audit dataset. This indicates that despite instances of GSV image ages being over 2 years old, GSV is a viable data source for gathering up-to-date information on bus stop locations and surrounding landmarks.

5. OUR BUS STOP LABELING TOOL

Shifting now to preparations for our third study—to allow crowd workers to examine and describe bus stops and surrounding landmarks in GSV—we created an interactive online labeling tool called *Bus Stop CSI* in JavaScript, PHP, and MySQL. Unlike previous crowdsourcing GSV work that uses static imagery to collect labels (e.g., Guy and Truong [2012], Hara et al. [2012], and Hara, Le, et al. [2013]), our labeling interface is fully interactive and allows the crowd worker to move about and control the camera view in the 360-degree GSV panoramic space (see Figure 5). Although this interactive



The **Explore Mode** (currently selected) allows the user to control the GSV camera angle and "walk" up to two steps in any direction beyond the drop point.

When one of these "label" buttons is selected, the interface enters the **Labeling Mode**. The mouse cursor turns into a representative icon for the selected label type. The user directly clicks on the object in the GSV pane below to place the label. In this mode, unlike the Explore Mode, the camera angle and location is fixed. The interface automatically returns to Explore Mode after each label is placed.

The **Status side panel** provides details on the user's progress and their qualification badges (which they earn in our interactive tutorials).

The **GSV pane** is the primary interaction area for exploring and labeling.

The user's location and view direction are represented in this **top-down 2D map view.** The bus stop icon (in blue) is drawn based on location data from Google.

If the **user cannot find a bus stop** in the scene, they can click this button and provide details.

The user clicks the **Submit button** to upload their labels.

Mission: Your mission is to **find and label** a bus stop sign and landmarks near the sign.

Progress: You have finished 12 out of 15.

Explore mode: Find the closest bus stop and label surrounding landmarks

Please enter any comments about this bus stop that may affect people with visual impairment (optional)

Fig. 5. The Bus Stop CSI Interface. We use the Google Maps Transit API to determine drop locations nearby bus stops. Crowd workers use the Explorer Mode to move around and look for the target bus stop (indicated by the blue icon in the 2D-map view) and the Labeling Mode to label any of the six bus stop landmark types. Clicking the Submit button uploads the labels (in this case, a mailbox, bus stop sign, shelter, and bench). The worker is then transported to a new location unless the HIT is complete (14–16 bus stop locations are included in each HIT).

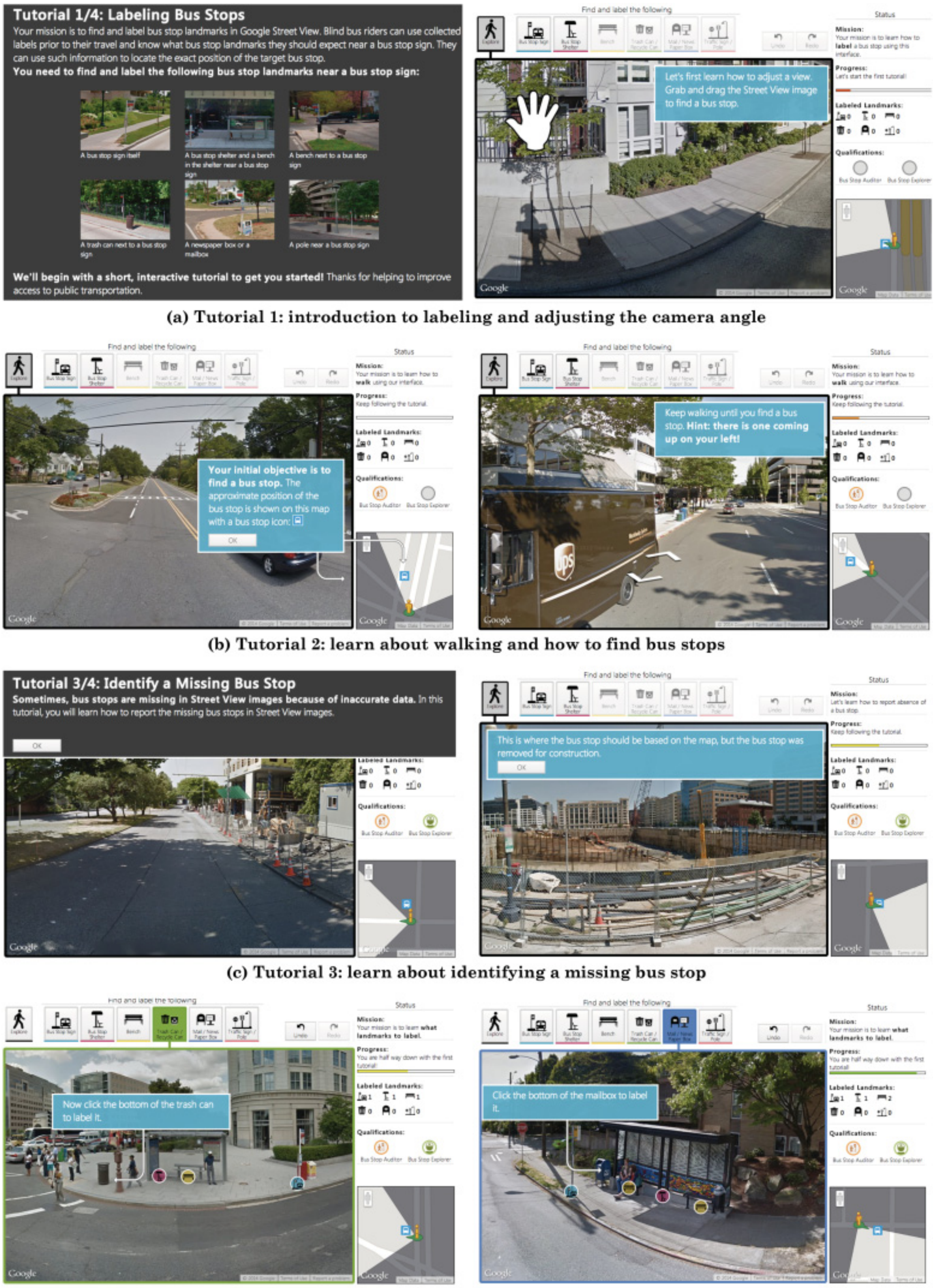
freedom increases task complexity, the benefits are twofold: first, the crowd worker can "walk" in GSV to find the target bus stop; second, the crowd worker can shift their view to find an optimal labeling perspective (e.g., a camera view that avoids occlusions). As we deployed our tool on MTurk, the following description is written for that context.

When a turker accepts our HIT (a bundle of labeling tasks) for the first time, they are greeted by a four-stage interactive tutorial (Figure 6). Each stage is dedicated to progressively teaching the turker about some new interaction or labeling feature in our tool:

—Tutorial 1: This tutorial teaches a turker how to control the camera angle in the GSV pane and find a bus stop.
—Tutorial 2: We teach a turker that sometimes he or she needs to "walk" to find a bus stop in GSV, because bus stops could be too far away to identify.
—Tutorial 3: A turker is taught that sometimes a target bus stop icon may show up in the 2D-map view but may not actually exist in the GSV pane.
—Tutorial 4: In this tutorial, we ask turkers to label different types of landmarks, letting them once again review types of landmarks that they have to label.

Turkers must successfully complete one tutorial stage before moving on to the next. Because the bus stop signs and landmarks differ in look and feel across cities, we created separate interactive tutorials for Washington, DC and Seattle (eight tutorials in total; four for each metropolitan area). If a turker was trained in one city, they were required to retrain in the other city.

Once the tutorials are successfully completed, we query the Google Maps API to drop the turker close to a bus stop in the audit area and the labeling task begins. Bus Stop CSI has two modes of interaction: the *Explorer Mode* and the *Labeling Mode*. In the



Tutorial 1/4: Labeling Bus Stops

Your mission is to find and label bus stop landmarks in Google Street View. Blind bus riders can use collected labels prior to their travel and know what bus stop landmarks they should expect near a bus stop sign. They can use such information to locate the exact position of the target bus stop.

You need to find and label the following bus stop landmarks near a bus stop sign:

A bus stop sign itself

A bus stop shelter and a bench in the shelter near a bus stop sign

A bench next to a bus stop sign

A trash can next to a bus stop sign

A newspaper box or a mailbox

A pole near a bus stop sign

We'll begin with a short, interactive tutorial to get you started! Thanks for helping to improve access to public transportation.

Mission: Your mission is to learn how to **label** a bus stop using this interface.

Progress: Let's start the first tutorial!

Let's first learn how to adjust a view. Grab and drag the Street View image to find a bus stop.

(a) Tutorial 1: introduction to labeling and adjusting the camera angle

Mission: Your mission is to learn how to **walk** using our interface.

Progress: Keep following the tutorial.

Your initial objective is to find a bus stop. The approximate position of the bus stop is shown on this map with a bus stop icon:

OK

Mission: Your mission is to learn how to **walk** using our interface.

Progress: Keep following the tutorial.

Keep walking until you find a bus stop. **Hint: there is one coming up on your left!**

(b) Tutorial 2: learn about walking and how to find bus stops

Tutorial 3/4: Identify a Missing Bus Stop

Sometimes, bus stops are missing in Street View images because of inaccurate data. In this tutorial, you will learn how to report the missing bus stops in Street View images.

OK

Mission: Let's learn how to report absence of a bus stop.

Progress: Keep following the tutorial.

This is where the bus stop should be based on the map, but the bus stop was removed for construction.

OK

(c) Tutorial 3: learn about identifying a missing bus stop

Mission: Your mission is to learn **what landmarks to label.**

Progress: You are half way down with the first tutorial!

Now click the bottom of the trash can to label it.

Mission: Your mission is to learn **what landmarks to label.**

Progress: You are half way down with the first tutorial!

Click the bottom of the mailbox to label it.

(d) Tutorial 4: reinforce prior tutorial learnings and have turker label a complex scene

Fig. 6. The four interactive tutorials. We created two sets of interactive tutorials, one for Washington, DC (left column) and one for Seattle (right column). Turkers had to complete all four tutorials successfully before working on a task in earnest.

Explorer Mode, the user interacts in the GSV pane using the traditional Street View inputs. Walking is controlled by clicking the arrow widgets (<, >, ∨, and ∧). Horizontal and vertical panning in the 360-degree view is controlled by clicking and dragging the mouse across the image. When the user is first dropped into a scene, he or she is defaulted into Explorer Mode. When the user clicks on one of the six labeling buttons, the interface switches automatically to the Labeling Mode. Here, mouse interactions no longer control movement and camera view. Instead, the cursor changes to represent the currently selected label. The user can then apply the selected label by clicking on the appropriate landmark in the GSV pane. Our tool automatically tracks the camera angle and repositions the applied labels in their correct location as the view changes—in this way, the labels appear to "stick" to their associated landmark. Turkers cannot see previously entered labels by other workers.

In early pilot studies, we found that users would get disoriented by accidentally "looking" straight down (toward the street) or straight up (toward the sky) in the GSV pane. Thus, to simplify GSV interaction and to focus the view appropriately on street-level features, we reduced vertical panning to 20 degrees (0, −20). Other GSV adjustments included hiding the onscreen camera control and zooming widgets, disabling keyboard interactions (to prevent accidental movement), and hiding textual overlays (e.g., street names). In addition, we prevented users from moving more than two steps in any direction away from their initial drop point. This constraint prevented users from walking down streets in search of bus stops. In our dataset, a single GSV "step" translated to roughly 5–10m of real-world movement (GSV steps are smaller in denser areas).

6. STUDY 3: CROWDSOURCING LABELS

To investigate the potential of using minimally trained crowd workers to find and label bus stop landmarks, we posted our tool to MTurk in April 2013. In each HIT, turkers needed to label 14–16 bus stop locations. We paid $0.75 per HIT ($0.047–0.054 per labeling task); which was decided based on the task completion time in pilot studies (e.g., approximately $0.10 per minute). Although we used 179 bus stop locations in Study 2, here, we use a subset 150. This subset is necessary because, as previously mentioned, 29 bus stop locations do not show up in the Google Maps Transit API (see Table I). We use this API to automatically place turkers next to bus stops in our labeling tool. If the API is unaware of the bus stop, we cannot determine its location.

6.1. Assessing Accuracy

In order to assess turker performance, we needed ground truth data about which landmarks exist at each bus stop location. For this, we used the median count GSV dataset from Section 4.4. Recall that to produce this consolidated dataset, we calculated the median count of each landmark type from the three auditor datasets across every bus stop location. Here, we further transform these counts into binary presence indicators for each landmark type. In other words, our ground truth dataset is a 150 row (for bus stop locations) × 6 column (for landmark types) matrix where cells = 1 represent the presence of that landmark type at the specified bus stop and cells = 0 represent an absence. Although the *Bus Stop CSI* tool gathered raw landmark counts and relative location data on landmarks (e.g., a trash can is north of the bus stop sign), we did not evaluate this level of granularity here. Thus, our analysis focused only on whether crowd workers properly indicated the presence/absence of a landmark in a scene but without regard for multiple occurrences. We leave more sophisticated assessments for future work.

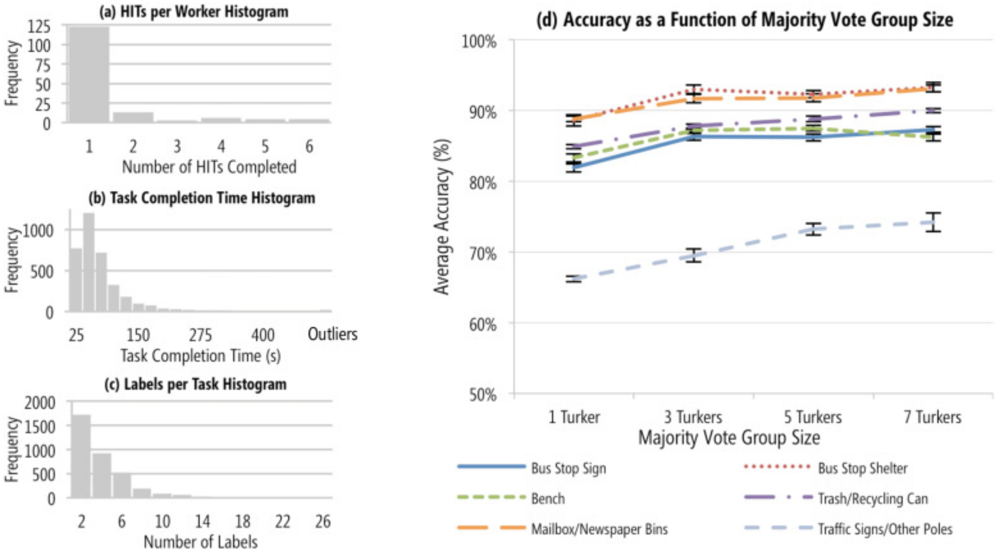


Fig. 7. (a) A histogram of the number of HITs completed per worker. The long-tail distribution shows that most workers (123/153) performed a single HIT and left. (b) A histogram of task completion time. Most tasks (93%) were completed within 150s. The rightmost bin includes outliers: 23 tasks with completion time >500s (max 7517s). (c) A histogram of label counts per task. Most turkers submitted under six labels, (d) Overall accuracy as a function of majority vote group size. Each graph point is based on multiple permutations of the majority vote group size across all 150 bus stop locations. Standard error bars are in black. Note: the *y* axis does not start at 0% (range: 50%–100%).

Table V. The Average Labeling Accuracy with One Turker Per Scene Across All 150 Bus Stops. Cell format: Average (Standard Error).

	Bus Stop Sign	Bus Stop Shelter	Bench	Trash/ Recycling	Mailbox/ News Bins	Traffic Signs / Other Poles	Overall
Avg. accuracy (*N*= 153 turkers)	81.9% (0.6)	88.6% (0.5)	83.3% (0.5)	84.9% (0.4)	88.8% (0.4)	66.2% (0.4)	82.5% (0.3)

6.2. High-Level Results

In total, 153 distinct turkers completed 226 HITS (3534 labeling tasks) and provided 11,130 bus stop landmark labels. On average, turkers completed 1.48 HITs (SD = 1.17), which is equivalent to 23.1 labeling tasks (SD = 19.0) (Figure 7(a)) The median labeling time per task was 44.7s (avg = 71.8s; SD = 213.1s) (Figure 7(b)) and the average number of labels per panoramic image was 3.15 (SD = 3.06) (Figure 7(c)). When compared with our ground truth dataset, overall turker accuracy was 82.5% (SD = 0.3%) for properly detecting the presence/absence of a landmark across the 150 bus stop locations.

When broken down by landmark type (Table V) the mailbox/newspaper bin landmark type followed by the bus stop shelter and bench had the highest accuracies at 88.8% (SE = 0.4%), 88.6% (SE = 0.5%), and 83.3% (SE = 0.5%), respectively. These all tended to be fairly salient landmark types in GSV. In contrast, the lowest scoring landmark type (Traffic Signs/Other Poles at 66.2%) was the most open-ended label (i.e., least defined) making it susceptible to confusion and misuse. This was particularly true given that our ground truth data had a constrained 20ft extent on either side of the bus stop sign meaning that potentially correct labels placed beyond that area could be flagged as incorrect. In the future, we plan to account for distance in our assessments.

Returning to the researcher-supplied difficulty ratings from Study 2, we found a significant difference ($p < 0.0001$) between turker performance on bus stop locations rated easy by our research team ($N = 116$) versus those rated medium-to-hard ($N = 34$). We compared overall accuracies in two groups using Welch's t-test. For the easy locations, our average per-turker accuracies were 84.5% (SE = 0.3%). This decreased to 74.3% (SE = 0.7%) for the hard locations, which suffered from occlusion, blurred images, and required more movement (including a scene where one virtual step leapt forward in a disorienting manner). We revisit the relationship between turker performance and bus stop scene in Section 6.5.

6.3. Accuracy as a Function of Majority Vote Size

Collecting multiple labels per bus stop location helps account for the natural variability of human performance and reduces the influence of occasional errors; however, it also requires more workers. Similar to Hara, Le et al. [2013] here we explore accuracy as a function of turkers per scene. We recruited 21 (or more) turkers for each of the 150 bus stop locations. We compared ground truth data with majority vote labels across four turker groups: 1, 3, 5, and 7. Because we have at least 21 turkers per bus stop location, we could compute accuracies multiple times for each group size, average the results, and calculate error margins. The overall goal here was to produce a more accurate portrayal of expected future performance for each group size. For example, when we set the majority vote group size to three, we randomly permuted seven groups of three turkers. For each group, we calculated the majority vote answer for a given bus stop location in the dataset and compared it with ground truth. This process was repeated across all locations and the five group sizes, where X = majority vote group size, Y = number of groups: (1,21), (3, 7), (5,4), and (7, 3). We used a similar evaluation technique in Hara, Le et al. [2013].

Overall, we found that accuracy does indeed increase with majority vote group size from 82.5% to 85.8% with three turkers and 87.3% with seven turkers. These gains, in general, diminish in magnitude as majority vote group size grows (Figure 7(d)). However, for the hardest landmark label type (*Traffic Signs Other Poles)*, we see a continued steady increase as the majority vote size grows—perhaps indicating the wisdom in the crowds for more challenging landmark types.

6.4. Individual Worker Performance

As reliable and performant workers are a critical component to any crowdsourcing system, in this subsection we analyze individual worker performance. Our goal here is to identify poor-performing workers and uncover patterns of behavior that may be automatically discovered and rectified in future versions of the Bus Stop CSI tool (e.g., by providing better feedback about performance in the user interface).

For each turker, we calculated a *per-worker average accuracy* metric by taking the average accuracy of all tasks he or she submitted. The results are shown in Figure 8(a). Individual worker accuracies varied between 58% and 98% (Avg = 81%, SD = 7%). From this data, we divided our 153 turkers into two groups: "poor performers" ($N = 21$) who had average accuracies one standard deviation below the mean (<74%) and all other turkers ($N = 132$)—the "other" group.

To better understand the behavior of this poor-performing group, we examined their false-positive and false-negative error behaviors and compared them to the "other" group. A false positive indicates that a turker provided a label for a landmark that does not actually exist in the scene (i.e., overlabeling). Conversely, a false negative means that the turker missed labeling a landmark that actually existed at the bus stop (i.e., underlabeling). Overall, the average false-positive and false-negative rate in the poor-performing group was 0.4 (SD = 0.3) and 1.5 (SD = 0.4), respectively, compared with

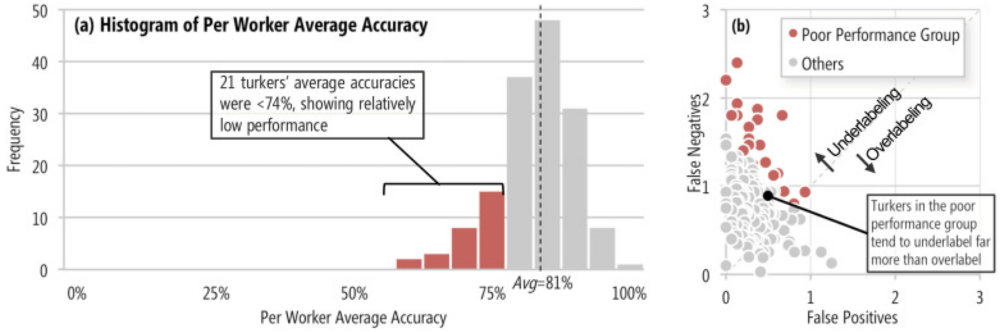


Fig. 8. (a) A histogram of per turker average accuracies of 153 turkers (bin size = 5%). (b) A scatter plot of the average false-positive vs. false-negative errors for each of the 153 turkers. Each point indicates a turker; the 21 "poorly performing" turkers are highlighted in red.

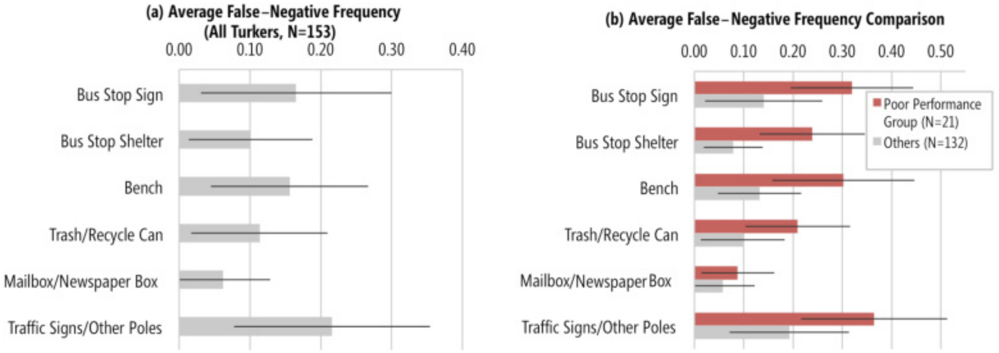


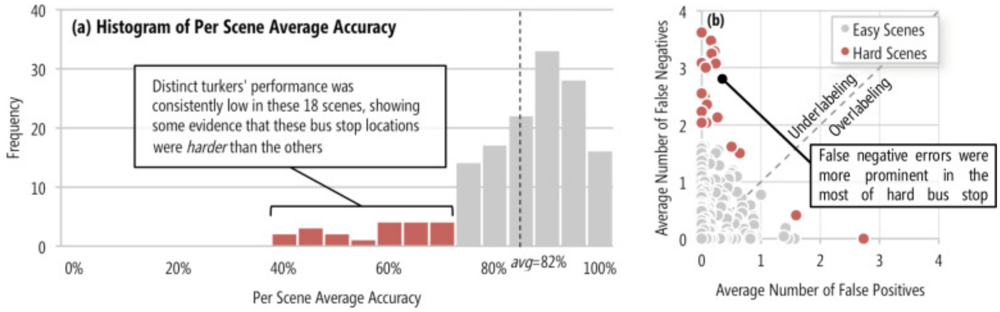
Fig. 9. (a) The average number of false-negative errors broken down per landmark type. (b) Comparison of average number of false-negative errors between the poor performance group and others. Error bars indicate standard deviation.

0.3 (SD = 0.2) and 0.7 (SD = 0.3) for the other group. While false-negative errors were more common than false positives for both groups, the poor performers had a higher number of both (Figure 8(b)). More work is needed to determine why false negatives are more prevalent. For example, it would be interesting to investigate whether false negatives are due to a misunderstanding of the task, problems with occlusion, or simply worker apathy.

Performance can be further broken down by landmark type (Figure 9). Traffic signs and other poles were missed most frequently, probably because they are the most open-ended landmark type and confusing label to use (as noted in the overall results). Mailboxes and newspaper boxes had the lowest false-negative scores, most likely because they did not appear in many bus stops that were used in our study, and, even when they appeared, they were easy to spot due to their visual salience. These trends occurred for turkers in the poor performance group and in the other group.

6.5. Scene Difficulty

Identifying bus stop landmarks is harder in some GSV images than in others. To this end, we evaluated turker performance in each scene by measuring *per-scene average accuracy*. We calculated this for each bus stop by taking an average of turkers' label accuracies. For example, for a bus stop with 21 distinct turker labels, a per-scene average accuracy is calculated by adding all turkers' accuracy scores and dividing it



Distinct turkers' performance was consistently low in these 18 scenes, showing some evidence that these bus stop locations were *harder* than the others

False negative errors were more prominent in the most of hard bus stop

Fig. 10. (a) A histogram of per scene average accuracies of 150 bus stop locations. A bin size is 5% (e.g., the bar above "80%" indicates the frequency of bus stop locations with accuracy between 75% and 80%). 18 scenes' accuracies fell below 69% (one standard deviation away from the mean). (b) A scatter plot that shows the types of errors made in 150 bus stop locations. Each point indicates a bus stop location. Data points for the 18 "hardest" scenes are colored red. The x and y axis indicates the average number of false-positive and false-negative errors, respectively.

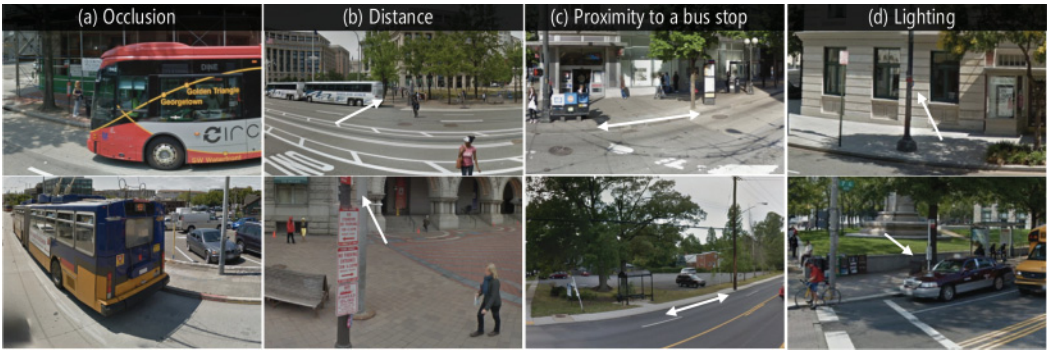
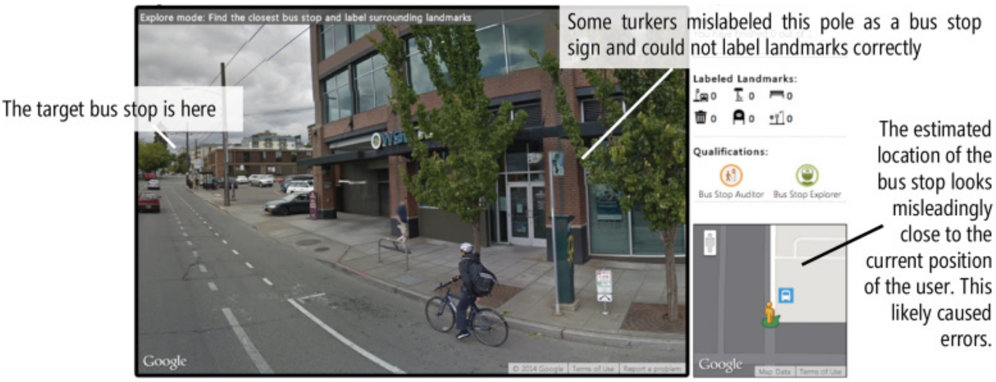


Fig. 11. (a) Occlusion: an obstacle such as a stopping bus can hide a part or all of the bus stop and its landmarks. (b) Distance: sometimes a bus stop is too far or too close to observe so turkers miss labeling them. (c) Proximity to the corresponding bus stop: it is sometimes ambiguous whether one should count a landmark as a part of a bus stop due to its distance from the bus stop. (d) Lighting: Shadows can cover landmarks and make it harder to find them (e.g., the bus stop sign in the first row and the trash can in the second row are not very visible). White arrows in the figures are used to highlight the positions of the bus stop signs and landmarks that are hard to see otherwise.

by 21. We looked into average numbers of errors (i.e., false-positive errors and false-negative errors) to clarify what types of mistakes were made.

We found that per-scene average accuracy varied from 39% to 100% (Avg = 82%, SD = 13%) across 150 bus stop scenes. Of all the scenes, the accuracy of 18 scenes fell under one standard deviation below the mean (<69%), indicating that distinct turkers consistently failed to provide accurate information about the presence/absence of landmarks in these scenes (Figure 10(a)). Notably, false-negative errors (i.e., failing to label existing landmarks) exceeded false-positive errors at most of these bus stops (Figure 10(b)). This suggests that the dominant cause of the low per-scene average accuracies is because turkers are underlabeling landmarks.

In these bus stop scenes, GSV panoramic images suffered from potential labeling difficulties, such as (i) occlusion, (ii) distance, (iii) ambiguity in landmark's proximity to a corresponding bus stop, (iv) lighting, and (v) misleading information on Google Maps pane (Figures 11 and 12):



The target bus stop is here

Some turkers mislabeled this pole as a bus stop sign and could not label landmarks correctly

The estimated location of the bus stop looks misleadingly close to the current position of the user. This likely caused errors.

Fig. 12. An inaccurate estimated location of the bus stop in the Google Maps pane could confuse turkers. In this picture, the actual bus stop is further down the street, but the bus stop icon on the Google Maps shows that the bus stop should be visible from the current position.

—*Occlusion*: Obstacles (e.g., a stopping bus) could block a view of a part of or an entire bus stop in GSV.
—*Distance*: A bus stop could be too far/close from views that are available in GSV (e.g., a wide street that makes observing one side of a road difficult).
—*Bus stop proximity*: It is sometimes ambiguous to judge if landmarks are close enough to a bus stop.
—*Lighting*: Bad lighting (e.g., due to shadow from trees) could affect difficulties of labeling.
—*Misleading information*: Bus stop latitude/longtitude coordinates in Google Transit data are not always precise. As a result, the estimated locations of bus stops shown in the Google maps pane could be faulty and misleading for users.

We speculate, for example, turkers struggled to find a bus stop and label landmarks when they needed to walk in a scene with occlusion. In such a scene, they either skipped tasks or labeled only easily visible landmarks and neglected things behind obstacles. We discuss potential design implications for the future design of the interface in the following.

6.6. Study 3 Summary

In summary, our current crowdsourcing experiments and analyses are the first results to demonstrate that minimally trained crowd workers can accurately find and label bus stop landmarks in GSV (>82%). We also provided insights on what traits poorly performing workers have and what makes scenes more difficult to label. Future work could explore more sophisticated analyses of worker labels including count and placement accuracy in each scene. In addition, more work is needed to establish the required accuracy level needed to provide value to transit agencies and navigation tools (e.g., with what data accuracy can people with visual impairment navigate themselves to a bus stop? What fallback strategies can be used in case there are errors in data?).

7. DISCUSSION AND CONCLUSION

While Study 1 extended upon our previous formative work [Azenkot et al. 2011], our findings reemphasized the significance of landmarks in aiding visually impaired navigation. For example, we found that benches and shelters were most helpful, which crowd workers correctly labeled 83.3% and 88.6% of the time, respectively, in Study 3—such a result demonstrates the interconnections between our studies. Study 2 showed that despite data age and occlusion problems, GSV could be used as a lightweight dataset for bus stop audits (even when compared to physical audit data). Finally, and perhaps most importantly, Study 3 showed that a minimally trained crowd worker could find and label bus stops in *Bus Stop CSI* with 82.5% accuracy, which jumps to 87.3% with a simple seven-turker majority vote scheme). The extended Study 3 results suggest that the dominant cause of labeling mistakes is false-negative errors (i.e., workers missed labeling landmarks). Addressing the underlabeling problem would further increase the overall accuracy. Taken together, these three studies advance the current literature and understanding of how information about bus stop landmarks could be potentially collected and used to guide low-vision and blind bus riders. With that said, our work is not without limitations. Here, we briefly discuss limitations that could affect the scalability and accuracy of our approach and opportunities for future work.

Inaccurate Bus Stop Locations. While our physical audit in Study 2 found 179 bus stops, 29 of these were missing from the Google Maps API. Because we rely on this same API in our *Bus Stop CSI* tool, these 29 bus stops could not be visited—even if they were visible in GSV (in this case, all but three were). Though this might be resolved in the future as Google or official transit organizations provide up-to-date bus stop location data, we could also proactively search for bus stops by asking turkers to sweep through streets in GSV.

Similarly, oftentimes we found that the exact locations of bus stops in the Google Maps API were inaccurate (e.g., wrong place on the block, wrong side of an intersection). This made our 2D-map pane confusing for some scenes—a worker would point the avatar toward the bus stop icon but would not see a bus stop in the GSV pane. We believe this led turkers to label artifacts that are not parts of bus stops or decide to skip and falsely report missing bus stops. This consequently increased false-positive and false-negative errors. Other data sources (e.g., OpenStreetMap) could likely be used to mitigate this problem. Other task assignment strategies should also be investigated. As current interface drops all turkers at a same place facing a same direction in GSV, many of them might have made same mistakes (e.g., incorrectly skipped a task). We would like to investigate the effect of dropping turkers at slightly different locations, which allows them to observe a same bus stop from various positions and camera angles with different occlusion/lighting conditions.

Image Age. While we observed high concordance between our GSV bus stop audit data and our physical audit data, the image age in GSV remains a concern. Although Google does not publicly specify a GSV update plan from city to city, Washington, DC has been updated at least three times in the last 4 years. In addition, Google updated 250,000 miles of road in early October 2012 (http://goo.gl/hMnM1).

Scene Difficulty. The following GSV-related problems made it more challenging to label bus stops: (i) *Distance:* most streets are driven once by a GSV car from a single car lane in one direction. This can create distant views of bus stops. (ii) *Occlusion*: bus stop landmarks are sometimes occluded by a parked bus or other obstacle. (iii) *Lighting*: shadows from trees and buildings can make bus stop landmarks hard to see. (iv) *Blur*: as previously mentioned, sometimes GSV misidentifies a bus stop sign as a license plate and blurs it out, which makes it harder to identify. One potential solution would be to integrate other streetscape imagery sources (e.g., Microsoft Streetside) to gain multiple simultaneous views of an area.

In the extended study, we also found that (v) landmarks' proximity to a bus stop could be a problem. Judging whether landmarks are close enough to a bus stop from a static image is hard. As a result, some turkers missed labeling some landmarks that are close enough to a bus stop, or overlabeled things that are far away from the bus stop. Future work should investigate the use of 3D point cloud data to estimate the physical placement of a bus stop and other landmarks. The interface should also provide more feedback and point out when turkers make false-negative mistakes using techniques such as ground truth seeding.

Though these factors likely caused difficulty in finding bus stop landmarks, future work should also investigate how often these problems occur with a larger dataset of bus stop scenes (e.g., how often lighting condition is bad in GSV). to better understand the extent of these problems. Future work could also investigate whether the performance difference is due to turker ability, bus stop scene difficulty, or tool/interface limitations

Selecting Bus Stop Landmarks. Our tool allowed crowd workers to label six landmark types but other landmark types could also be useful (e.g., grass, trees). For example, one turker left a comment saying, *"There is a tree very close to the bus stop sign."* Future work should examine other landmark types and continue performing user-centered design to see how these landmarks affect navigation.

ACKNOWLEDGMENTS

We thank Matt Mauriello and Leah Findlater for their feedback on earlier drafts. We also thank our ASSETS and TACCESS reviewers for their helpful critiques.

REFERENCES

American Foundation for the Blind. 2013. Accessible Mass Transit.

S. Azenkot, S. Prasain, A. Borning, E. Fortuna, R. E. Ladner, and J. O. Wobbrock. 2011. Enhancing independence and safety for blind and deaf-blind public transit riders. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'11)*. ACM, New York, NY, 3247–3256.

H. M. Badland, S. Opit, K. Witten, R. A. Kearns, and S. Mavoa. 2010. Can virtual streetscape audits reliably replace physical streetscape audits? *J. Urban Health: Bull. N. Y. Acad. Med.* 87, 6 (2010), 1007–1016.

M. Banâtre, P. Couderc, J. Pauty, and M. Becus. 2004. Ubibus: Ubiquitous Computing to Help Blind People in Public Transport. In *Proceedings of Mobile Human-Computer Interaction (MobileHCI'04)*. Lecture Notes in Computer Science. S. Brewster and M. Dunlop (Eds.). Vol. 3160, Springer, Berlin, 310–314.

J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, and T. Yeh. 2010. VizWiz: Nearly real-time answers to visual questions. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology* (*UIST'10)*. ACM, New York, NY, 333–342.

J. P. Bigham, R. E. Ladner, and Y. Borodin. 2011. The design of human-powered access technology. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'11)*. ACM, New York, NY, 3–10.

Blind Citizens Australia. 1994. BCA Public Transport Policy.

M. Campbell, C. Bennett, C. Bonnar, and A. Borning. 2014. Where's my bus stop?: Supporting independence of blind transit riders with stopinfo. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers and Accessibility* (*ASSETS'14)*. ACM, New York, NY, 11–18.

P. Clarke, J. Ailshire, R. Melendez, M. Bader, and J. Morenoff. 2010. Using Google Earth to conduct a neighborhood audit: Reliability of a virtual audit instrument. *Health Place* 16, 6 (2010), 1224–1229.

Easter Seals. 2008. *Toolkit for the Assessment of Bus Stop Accessibility and Safety*.

B. Ferris, K. Watkins, and A. Borning. 2010. OneBusAway: Results from providing real-time arrival information for public transit. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*. ACM, New York, NY, 1807–1816.

A. J. Flanagin and M. J. Metzger. 2008. The credibility of volunteered geographic information. *GeoJournal*, 72, 3–4 (2008), 137–148.

R. G. Golledge, J. R. Marston, and C. M. Costanzo. 1997. Attitudes of visually impaired persons toward the use of public transportation. *J. Vis. Impair. Blind.* 91, 5 (1997), 446–459.

Google. 2013. Google Street View Privacy.

M. Guentert. 2011. Improving public transit accessibility for blind riders: A train station navigation assistant. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'11)*. ACM, New York, NY, 317–318.

R. Guy and K. Truong. 2012. CrossingGuard: Exploring information content in navigation aids for visually impaired pedestrians. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'12)*. ACM, New York, NY, 405–414.

M. Haklay. 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environ. Plan. B: Plan. Des.* 37, 4 (2010), 682–703.

K. Hara, S. Azenkot, et al. 2013. Improving public transit accessibility for blind riders by crowdsourcing bus stop landmark locations with Google Street view. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility Technology*. 16:1–16:8.

K. Hara, V. Le, and J. Froehlich. 2012. A feasibility study of crowdsourcing and Google Street view to determine sidewalk accessibility. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'12), Poster Session*. ACM, New York, NY, 273–274.

K. Hara, V. Le, and J. Froehlich. 2013. Combining crowdsourcing and Google Street view to identify street-level accessibility problems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'13)*. ACM, New York, NY.

D. J. Hruschka, D. Schwartz, St. D. C. John, E. Picone-Decaro, R. A. Jenkins, and J. W. Carey. 2004. Reliability in coding open-ended data: Lessons learned from HIV behavioral research. *Field Meth.* 16, 3 (2004), 307–331.

Intercity Transit. 2010. *Bus Stop Specification Guidelines*.

R. Jacob, B. Shalaik, A. Winstanley, and P. Mooney. 2011. Haptic Feedback for Passengers Using Public Transport. In *Digital Information and Communication Technology and Its Applications*. H. Cherifi, J. Zain, and E. El-Qawasmeh (Eds.). Communications in Computer and Information Science. Springer, Berlin, 24–32.

J. Kostiainen, C. Erkut, and F. B. Piella. 2011. Design of an audio-based mobile journey planner application. In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments (MindTrek'11)*. ACM, New York, NY, 107–113.

K. H. Krippendorff. 2003. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Inc.

W. Lasecki et al., 2012. Real-time captioning by groups of non-experts. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology (UIST'12)*. ACM, New York, NY, 23–34.

J. R. Marston and R. G. Golledge. 2003. The hidden demand for participation in activities and travel by persons who are visually impaired. *J. Vis. Impair. Blind.* 97, 8 (2003), 475–488.

Metro Transit–King County Department of Transportation. 2013. Stop Information for E Thomas St and 16th Ave E.

M. Z. H. Noor, I. Ismail, and M. F. Saaid. 2009. Bus detection device for the blind using RFID application. In *Proceedings of the 5th International Colloquium on Signal Processing and Its Applications, 2009 (CSPA'09)*. 247–249.

K. Panciera, R. Priedhorsky, T. Erickson, and L. Terveen. 2010. Lurking? Cyclopaths?: A quantitative lifecycle analysis of user behavior in a geowiki. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*. ACM, New York, NY, 1917–1926.

S. Prasain. 2011. StopFinder: Improving the experience of blind public transit riders with crowdsourcing. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'11), Poster Session*. ACM, New York, NY, 323–324.

A. G. Rundle, M. D. M. Bader, C. A. Richards, K. M. Neckerman, and J. O. Teitler. 2011. Using Google Street View to audit neighborhood environments. *Am. J. Prev. Med.* 40, 1 (2011), 94–100.

A. Steinfeld, J. Zimmerman, A. Tomasic, D. Yoo, and R. Aziz. 2011. Mobile transit information from universal design and crowdsourcing. *Transp. Res. Rec.: J, Transp. Res. Board* 2217 (2011), 95–102.

The National Federation of the Blind. 1986. Resolutions adopted by the annual convention of the national federation of the blind. *The Braille Monitor* (October).

U.S. Department of Transportation, F. H. A., 2007. ADA Regulations, Part 37—Transportation Services for Individuals with Disabilities.

D. Yoo, J. Zimmerman, A. Steinfeld, and A. Tomasic. 2010. Understanding the space for co-design in riders' interactions with a transit service. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*. ACM, New York, NY, 1797–1806.

Received June 2014; revised November 2014; accepted January 2015