Remote Sighted Assistants for Indoor Location Sensing of Visually Impaired Pedestrians

PAYMON RAFIAN and GORDON E. LEGGE, University of Minnesota-Twin Cities

Because indoor navigation is difficult for people with visual impairment, there is a need for the development of assistive technology. Indoor location sensing, the ability to identify a pedestrian's location and orientation, is a key component of such technology. We tested the accuracy of a potential crowdsourcing-based indoor location sensing method. Normally sighted subjects were asked to identify the location and facing direction of photos taken by a pedestrian in a building. The subjects had available a floor plan and a small number of representative photos from key locations within the floor plan. Subjects were able to provide accurate location estimates (median location accuracy 3.87ft). This finding indicates that normally sighted subjects, with minimal training, using a simple graphical representation of a floor plan, can provide accurate location estimates based on a single, suitable photo taken by a pedestrian. We conclude that indoor localization is possible using remote, crowdsourced, human assistance. This method has the potential to be used for the location-sensing component of an indoor navigation aid for people with visual impairment.

CCS Concepts: • **Human-centered computing** → **Accessibility**; • **Social and professional topics** → *People with disabilities*

Additional Key Words and Phrases: Wayfinding, crowdsourcing, indoor navigation, visual impairment, blind, low vision, location sensing

ACM Reference Format:

Paymon Rafian and Gordon E. Legge. 2017. Remote sighted assistants for indoor locations sensing of visually impaired pedestrians. ACM Trans. Appl. Percept. 14, 3, Article 19 (July 2017), 14 pages.
DOI: http://dx.doi.org/10.1145/3047408

1. INTRODUCTION

Pedestrian mobility has two major components. Obstacle avoidance is the ability to safely move through a space without running into objects. Dog guides and white canes are often used by visually impaired pedestrians for obstacle avoidance. The second component is wayfinding, also called spatial navigation. Wayfinding is the ability to plan and execute a route to a desired destination, which requires learning building layouts and updating one's location along a route. This article is concerned with assistive technology for indoor wayfinding.

A variety of assistive technologies have been developed for visually impaired wayfinding (for a review, see Giudice and Legge [2008]). In particular, GPS technology has been exploited for speech-based navigation for visually impaired wayfinding outdoors. Roentgen et al. [2011] provided an evaluation

This work is supported by research funds from the Department of Psychology at the University of Minnesota.
Authors' address: University of Minnesota-Twin Cities, 75 East River Road, Minneapolis, MN 55455; emails: {rafia003; legge}@umn.edu.

2017 ACM 1544-3558/2017/07-ART19
DOI: http://dx.doi.org/10.1145/3047408

of four commercially available systems. More recently, iPhone apps, designed specifically for visually impaired users—such as Ariadne GPS [2017], Seeing Eye GPS [2017], and BlindSquare [2017]—have proven to be quite successful.

But indoors, away from windows, GPS signals are not reliable, nor do they provide adequate spatial resolution for finding rooms and other key locations. A survey of electronic travel aids for the blind identified nine commercially available navigational aids, but none were designed for indoor usage [Roentgen et al. 2009]. A subsequent review confirmed that a viable indoor navigation system for people with visual impairments has not yet been developed [Gallagher et al. 2012]. One reason is the difficulty of locating pedestrians indoors. Many attempts have been made to create successful indoor location-sensing methods, but to our knowledge none have reached the convenience and accessibility needed for widespread use by visually impaired pedestrians. For discussion of the limitations of other indoor location-sensing systems, see Legge et al. [2013].

An indoor navigation system typically has three components: a method of pedestrian location sensing, a digital map, and an interface passing information to the pedestrian. Location sensing involves identifying the current location and heading of a pedestrian. A digital map ties a pedestrian's location and orientation information to functionally significant points of interest in the layout. Finally, an interface communicates meaningful information to the pedestrian, such as route directions, in an accessible format. This article proposes a method of indoor location sensing that could potentially take advantage of crowdsourcing.

Crowdsourcing refers to many people, working individually and remotely, performing a specific task. Crowdsourcing has already been used in assistive technology for visually impaired users. One application for crowdsourcing is smartphone apps for object recognition. VizWiz [2017] and TapTapSee [2017] are two examples. VizWiz allows a visually impaired person to take a picture of any object, ask a question about the picture (e.g., "What does this sign say?"), and have one of many "web workers" describe the object [Bigham et al. 2010]. TapTapSee has the user take a picture of an object, which the app identifies out loud to the user [Holton 2013]. With the app BeMyEyes, a remote sighted volunteer is shown a view from a visually impaired user's iPhone camera and answers questions via audio about objects in the camera's field of view [Be My Eyes 2017].

Successful use of crowdsourcing for object recognition prompted us to ask whether crowdsourcing could be used for accurate, indoor location sensing of visually impaired pedestrians. We conceive of a system in which a visually impaired pedestrian inside a building would take a picture of his/her surroundings using a cellphone app. The image would be transmitted to the "crowd" where a web worker would receive the pedestrian's image and use it along with a floor plan and set of sample images to identify the pedestrian's location and orientation. This information would be sent back to the pedestrian and used in conjunction with a digital map of the space and an accessible interface (e.g., synthetic speech) to provide the user with information about nearby points of interest or route instructions. For example, Legge et al. [2013] showed that blind and low-vision subjects could find nearby points of interest (such as a room on the same corridor) and navigate to remote locations in the floor plan with the use of a tablet-based digital map and synthetic-speech interface.

Our goal is to determine whether accurate location estimates can be obtained from a single picture taken by a visually impaired pedestrian.

To evaluate this idea, we tested normally sighted subjects playing the role of an off-site "web worker". The subjects attempted to identify a remote pedestrian's location and orientation from a single image sent from the pedestrian's smartphone camera. The subject had access to an interface composed of a floor plan and a set of sample camera images from the same floor plan (Figure 1). Our primary question was to ask how accurately subjects could determine the pedestrian's location and heading. Secondary

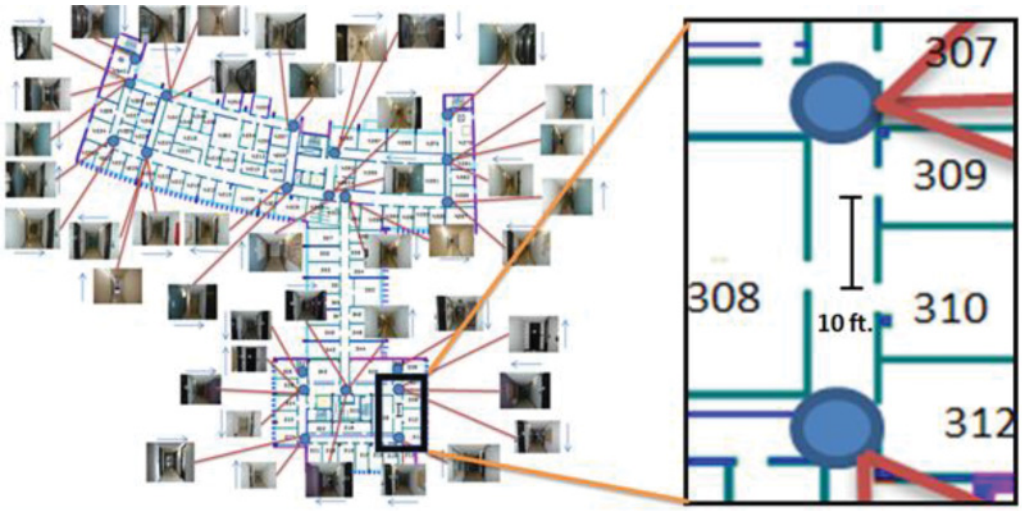

Fig. 1. One of three interface displays. The interface was used by subjects to identify the location and orientation of each image from Experiments 1 and 2. Blue circles show hallway intersections including dead ends. Red lines connect representative images to corresponding hallway intersections. Blue arrows indicate representative image facing directions. The inlay map shows 10ft on the interface.

questions included the time required for subjects to make location-sensing judgments, and the effect of the content of the images transmitted by the pedestrian.

In the first of two experiments, normally sighted subjects were shown forty test images one at a time from two floors in our campus building. The subject indicated the camera's location by clicking the mouse on a screen rendering of the floor plan. The forty images were taken by one of the authors to explore the impact of specific camera locations and facing directions.

The promising results of the first experiment prompted a follow-up. The second experiment was similar to the first except the test photos were taken by ten visually impaired pedestrians. The goal of the second experiment was to determine whether the visually impaired pedestrians could take camera photos of sufficient quality and content for good crowdsourced location sensing.

This article does not provide the specifics for the development of an entire indoor navigation aid. Our goal was to determine whether indoor localization is possible using a single user image and crowdsourced, human assistance.

2. METHODS

2.1 Materials and Apparatus

2.1.1 *Interface for Locating a Pedestrian.* Our interface design is composed of a floor plan showing room numbers and a set of representative images. Figure 1 shows a sample interface display. Representative images were taken at hallway intersections with one photo facing down each connecting hallway. Therefore, an L-junction has two representative images; a T-junction has three, and so on. Dead ends were included and had one representative image. Table I shows the number of representative images of each type for each floor. Hallway intersections were chosen as representative image locations because such pictures give information about hallway length and often the geometry of an adjacent hallway intersection. This, combined with geometrical information provided by the floor plan has been shown to be predictive of a person's ability to make judgments about a building's layout [Legge

Table I. Number of Representative Images by Floor

Floor	Total Images	L-junction Images	T-junction Images	Dead end Images
Second	22	10	9	3
Third	45	12	29	4
Practice	9	2	6	1

et al. 2003]. The interface was created with Microsoft PowerPoint 2010 on Windows 8.1. A floor plan was inserted and a thumbnail of each representative image was placed near the hallway intersection to which it belonged. Each representative image on the floor plan was hyperlinked to a separate slide that contained an entire screen-sized version of the representative image. In presentation mode, any thumbnail representative image could be clicked to view a full-sized version of that image. Separate interface displays were created for testing on two floors of a campus building. A third interface display featuring a section of another floor was used to introduce subjects to the use of the interface. The two tested floors differed in complexity. The third floor was more complex with 14 corridors, 20 junctions, and 102 rooms while the second floor had 8 corridors, 11 junctions, and 77 rooms.

The coordinates (x, y) of a subject's location estimate on the PowerPoint slide of the floor plan were captured using the following method. To make a location estimate, the subject activated a pen tool by pressing control-p. After marking a location on the floor plan, the subject exited PowerPoint's presentation mode by pressing the escape key. The coordinates of this location (measured in inches from the top left corner of the slide) were then obtained by right-clicking on the pen marking and selecting "position" and then "size and position" from the context menu.

A subject's location response was converted from inches on the interface to feet in the real world by measuring a distance of 10 feet in the building and associating this distance with interface coordinates for the corresponding locations.

The subject also estimated the facing direction of the camera view by manually indicating on the interface the predicted facing direction of the pedestrian along the hallway.

The number of representative images for each interface display. Due to the close proximity of two junctions on the third-floor interface display, one T-junction representative image was not included.

2.1.2 *Test Images.* In Experiment 1, four types of test images were used to identify whether certain image categories yielded better location sensing performance than others (Figure 2). Test images were photos taken by a normally sighted researcher using a Nikon D80 camera with 18–135mm lens (1514 $\times$ 1013 resolution). Similar locations on each floor were chosen for test image locations in order to keep the difficulty of test images between floors as similar as possible. The location of each test image was recorded for future comparison with subject location responses.

The four types of test images were:

(1) *Junction Corridor*. Photo taken at a hallway intersection facing straight down a connecting hallway. That is, similar to the representative images on the interface.
(2) *Junction Featural*. Photo taken at a hallway intersection pointing at an object such as a door sign, a change in floor material, unique wall color, or other stable feature in the surroundings that might facilitate location sensing.
(3) *Non-junction Corridor*. Photo taken anywhere between two hallway intersections facing straight down the hallway.
(4) *Non-junction Featural*. Photo taken anywhere between two hallway intersections pointing at an informative object.

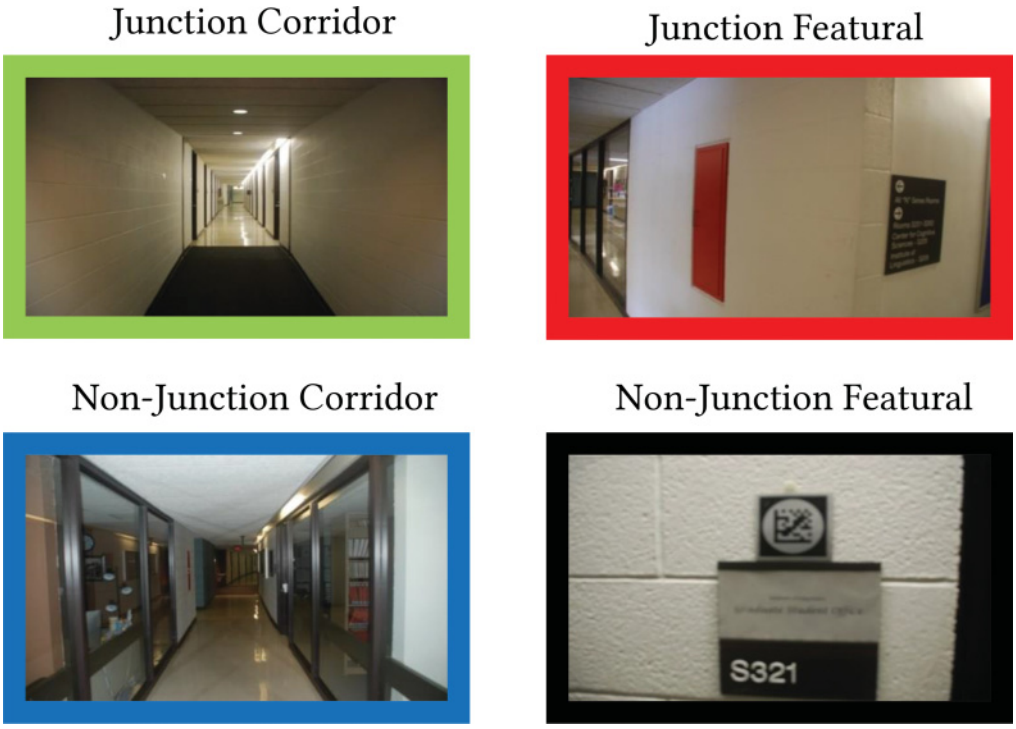

Fig. 2. A sample test image of each type: junction corridor, junction featural, non-junction corridor, and non-junction featural (see text). For all junction corridor test images and many junction featural test images, it cannot be determined from the image alone that it was taken at a junction. In this figure, however, the junction featural test image can be identified as having been taken at a junction.

Twenty test images (five of each of the four types) were collected from each of the second and third floors. Test images from each floor were tested in separate blocks. Each image was inserted into a PowerPoint presentation corresponding to the floor on which it was taken. Test images were pseudo-randomized so that the four types of test images were evenly spread throughout the PowerPoint presentation. The result was a pseudo-randomized PowerPoint presentation containing twenty images from each floor. We called these presentations *test image configurations*.

In Experiment 2, ten visually impaired participants (five blind and five with low vision) took ten pictures from each of the same two floors used in the first experiment. Photos from an eleventh subject were excluded because the subject's finger covered the camera's lens for the majority of the images. Figure 4 shows a sample of six images taken by visually impaired subjects. Images were selected to illustrate variations in quality and informativeness of the features in all pictures taken by visually impaired subjects. Five of the subjects took pictures with their own iPhones, and five subjects used one of two Android cellphone cameras of 13 or 16 megapixels provided by an experimenter. Half of the images from each floor were taken at hallway intersections and the other half were taken at locations between hallway intersections. Subjects were instructed to hold the cellphone at eye level, take pictures only containing aspects of the hallways, and avoid temporary hallway features. A temporary feature of the building is any part that is not likely to be stable over time (e.g., a chair outside an office door or another pedestrian). Subjects were also told to take pictures that they thought might help a normally sighted web worker perform location sensing, such as door signs. The cell phones automatically

adjusted for variances in lighting. Pictures were taken in either landscape or portrait orientation depending on what was most comfortable for the subject. These images were compiled into four configurations of 50 images each. Of the 50 images in each configuration, 25 were junction pictures and 25 were non-junction pictures. The number of images of each type (junction or non-junction) was split as equally as possible between low-vision and blind subjects. Configurations were pseudo-randomized so pictures taken by a particular subject and the location of images in the building were spread throughout the configurations.

2.1.3 *Subjects.* Twenty-four normally sighted young adults (13 females and 11 males) participated as subjects in Experiment 1. All subjects self-reported normal vision or corrected to normal vision. A few subjects had visited the building before, and three said they had previously visited at least one of the floors used for testing.

In the second experiment, ten visually impaired subjects (five blind and five with low vision) were recruited to take photos. For the purpose of this study, blind subjects were people lacking any visual information for navigation. Low-vision subjects were able to obtain some visual information to aid in navigation, but none had sufficient acuity to read building signage except at extremely close distances. The acuities of the five low-vision subjects ranged from 20/250 to less than 20/1000 with an average of 20/582. Six normally sighted subjects who had taken part in the first experiment were recruited to perform the same location-sensing task in Experiment 2 using the images taken by the visually impaired pedestrians.

All subjects read and signed consent forms (consent forms were read to subjects with visual impairment) and were compensated for their participation. The research procedures and informed consent were approved by the University of Minnesota's IRB.

3. EXPERIMENTS

3.1 Experiment 1

Twenty-four normally sighted subjects identified the location and facing direction of a series of camera images from two floors of a campus building. Testing lasted 1-1/2 to 3 hours. Informed consent was administered and subjects were told of the study's purpose. Subjects were given an explanation of the interface's features and used the practice interface display to perform five practice trials with the same procedure used for testing. After each practice trial, subjects were told the correct response and how, using the interface, the correct response could be determined. For actual testing, subjects performed location sensing for each test image of the second and third floor. Figure 3 shows the subject's view during testing. At the conclusion of testing, the subject went through an informal debriefing. Four subjects did not complete testing due to time constraints, but their data are included in our analysis.

The twenty test images from the second and third floors were tested in separate blocks, with the order of floors alternating across subjects. For each floor, the test images were sequenced from 1 to 20. To reduce order effects, consecutive subjects began testing on the first, fifth, tenth, or fifteenth image and then proceeded through the sequence. Before starting each floor, subjects were given up to 5 minutes to explore the interface display for that floor. Most subjects used this time to study the floor plan and use the zoom feature to view each representative image.

Subjects communicated a location and orientation judgment using the method described in Section 2.1.1. The experimenter recorded the location coordinates and orientation associated with the subject's response.

A stopwatch was used to measure the time from the onset of the test image to the subject's report on the test image's orientation, which we refer to as response time. This response time included the time for the subject to mark the judged location on the slide with the pen tool.

Fig. 3. A subject's view during testing. Two displays screens were used. One display showed an interface corresponding to one floor (15.6in HD LED screen with 1366 × 768 resolution). The other display showed a test image from that floor (23in and 1920 × 1080 resolution). The subject was able to continuously refer to both the test image and the interface during testing. The subject's task was to determine the camera location and facing direction of the test image.

3.2 Experiment 2

The six normally sighted subjects were asked to identify the location and orientation of fifty images taken by low-vision and blind subjects. Testing lasted about 2 hours. Subjects gave informed consent and then performed the same five practice trials as in Experiment 1. Before testing on each floor, subjects were given up to 5 minutes to reacquaint themselves with the interface display. During testing, subjects made location and orientation responses for each image in one of the four configurations described in Section 2.1.2, and response time was collected as it was in Experiment 1. In the informal debriefing at the end of testing, subjects were asked to compare their experiences in Experiments 1 and 2.

4. RESULTS

4.1 Experiment 1

4.1.1 *Location Accuracy.* Our main goal was to determine whether crowdsourcing could be used for accurate location sensing of pedestrians in indoor spaces. A subject's location response was considered correct if the location fell within the same hallway as the actual test image, and the subject correctly identified the facing direction of the test image. The 24 subjects in Experiment 1 gave correct location responses on 97.07% of the trials (894 of 921) and correct orientation responses on 92.18% of trials (849 of 921). Considering only correct location responses, median subject location accuracy was 3.87ft (IQR from 1.64 to 7.74).

Adopting a more stringent definition for a correct location response (that the response must be within 10ft of the actual test image location), 80.24% (739 of 921) of location responses and 78.18% (720 of 921) of orientation responses were correct.

4.1.2 *Response Time.* A secondary goal was to determine the time required by a subject to locate a pedestrian based on the single test image. The median response time for correct trials was 49.88 seconds (IQR from 29.19 to 87). Response times improve during testing. For the first floor in which subjects were tested, the median time per trial was 55.01 seconds. On the second tested floor, the median time per trial was 44.45 seconds. Complexity also affected response time. Median response time on the more complex floor was 52.37 seconds compared to 45.16 seconds on the less complex floor.

Table II. Spatial Accuracy and Response Time for the Four Image Types

Image Type	Percent of Trials with Correct Responses	Median Location Accuracy (ft)	Median Response Time (sec)
Junction Corridor	98.25	2.36 [1.03 to 7.34]	43.52 [27.28 to 70.00]
Junction Featural	97.79	2.08 [1.16 to 3.37]	36.63 [23.03 to 59.43]
Non-junction Corridor	97.39	6.57 [4.62 to 8.90]	64.00 [37.51 to 110.9]
Non-junction Featural	94.49	4.52 [2.03 to 9.89]	61.64 [33.04 to 116.5]

For each type of test image, the percent of subject responses with correct location judgments, the median location accuracy, and the median response time are shown. Images taken at hallway junctions had better median location accuracy and response time than non-junction images. Ranges in brackets are the upper and lower bounds of the IQR.

4.1.3 *Effect of the Type of Test Image.* A third goal of this study was to identify the type of test image that yields the best performance for location sensing. Table II shows results of location-sensing accuracy and response time for each test image type. Results were only calculated with trials that had correct location responses.

The percent of trials with correct location responses was greater than 94% for all image types. Images taken at hallway junctions yielded more accurate location estimates than those taken mid-hallway, with median location accuracies of 2.36ft and 2.08ft for junction corridor and junction featural test images, respectively. Response times were also faster for the junction images, with median response times of 43.52 seconds for junction corridor images and 36.63 seconds for junction featural images.

In post-testing debriefing, all subjects reported that test images containing doors signs, such as the non-junction featural test image in Figure 3, were easiest for location identification. Considering only test images containing readable door signs, 98% (196 of 200) of location responses were correct and 95% (190 of 200) of location responses were within ten feet of the actual test image location. These test images also had a median trial duration of 27.37 seconds (IQR from 19.5 to 39.91) and a median location accuracy of 2.09ft (IQR from 1.22 to 4.01).

Finally, we note that there were substantial individual differences in performance across our 24 subjects. Percent of correct judgments ranged from 80% to 100% across subjects. Median location accuracies ranged from 2.09ft to 7.34ft and median response times ranged from 30.84 seconds to 97.50 seconds. There was no correlation between response time and location accuracy.

4.2 Experiment 2

In Experiment 2, six of the subjects from Experiment 1 returned and made location estimates based on the photos taken by the ten visually impaired subjects.

4.2.1 *Location Accuracy.* Subjects gave correct location responses 91% of trials (273 of 300) and correct orientation responses on 86.33% of trials (259 of 300 trials). One subject identified the correct location on 80% of trials (40 of 50) and correct orientation on only 58% of trials (29 of 50). Ranges for the other five subjects were 86% to 98% for location accuracy and 92% to 98% for orientation accuracy. Median location accuracy for the six subjects in Experiment 2 was 3.56ft (IQR from 1.69 to 6.85), compared with their median accuracy in Experiment 1 of 3.77ft (IQR from 1.73 to 7.5). Median location accuracy for images taken by blind subjects was 3.65ft (IQR from 1.64 to 8.78) and 3.32ft (IQR from 1.8 to 6.44) for images taken by low vision subjects.

4.2.2 *Response Time.* The median response time dropped from 49.86 seconds (IQR from 29 to 85) in Experiment 1 to 25.12 seconds (IQR from 15.78 to 41.85) in Experiment 2.

Fig. 4. Six sample images taken by visually impaired subjects. The top three images were taken by subjects with low vision. The bottom three were taken by blind subjects. Images were selected to represent the range in quality and inclusion of informative features in all pictures taken by visually impaired subjects. Images regress in quality and usefulness from left to right. A few images were either extremely blurry or did not contain any informative features (e.g., blank, white wall).

4.2.3 *Image Type.* Images taken at hallway junctions had a median location accuracy of 3.32ft (IQR from 1.22 to 5.71). Non-junction images had a median location accuracy of 3.66ft (IQR from 1.93 to 9.76). The median response time was 23.37 seconds (IQR from 16.06 to 42.85) for junction images and 27.41 seconds (IQR from 15.76 to 45.5) for non-junction images.

Similar to Experiment 1, test images from corridor junctions yielded faster responses with greater spatial accuracy.

5. DISCUSSION

Our primary goal was to determine the accuracy of indoor location sensing based on single test images taken with a camera from within a building. We tested normally sighted subjects as surrogates for a pool of web workers in a potential crowd-sourced application. These subjects had available to them an interface showing a conventional floor plan and a set of representative camera images taken in the corresponding building.

In the first experiment, the test images were collected by a normally sighted researcher. The subjects had a median location accuracy of less than four feet for correct location responses. In the second experiment, the test images were photos taken by ten visually impaired pedestrians. Subjects in Experiment 2 were even more accurate with median location responses of just over 3.5ft. In both experiments, the error in location estimates was less than the distance between doors in most buildings.

Coupled with suitable route instructions from a speech interface, this level of accuracy would usually be adequate to guide a visually impaired pedestrian to a destination room door in an office building.

We also wanted to know how quickly subjects could perform location sensing. The median time per trial in Experiment 1 of just under 50 seconds should be considered an upper bound. Improvements to the interface should significantly lower this figure. For example, implementing a zoom feature that is activated by hovering over a representative image would allow subjects to scan representative images more quickly. Also, the time taken by subjects to mark a location estimate with the procedure described in Section 2.1.1 is included in the response time. From informal observation, this part of the data collection procedure added about 5 seconds per trial. An automated system for recording location estimates on the interface could eliminate this overhead time. Though our response-time measurements did not account for the time required to send location information to the user, the object recognition app TapTapSee is advertised as being able to provide its service to users in about 7 to 10 seconds [TapTapSee 2017]. Taking into account the time required to perform the object recognition task, this suggests that transmission time to and from the user is not a limiting factor.

We also expect that providing web workers with training and additional practice would yield faster response times. Evidence for this was the improvement in median response times during testing in Experiment 1 (55.01 seconds on the first tested floor to 44.45 seconds on the second tested floor). Despite these improvements, a response time of 44.45 seconds is probably too slow for practical application as a location sensing method. But in Experiment 2, where the images were provided by visually impaired pedestrians, the response time dropped to 25.12 seconds. The improvement in response time from the first to the second experiment likely occurred because visually impaired subjects were told to take pictures of informative features of the building while the collection of test images in Experiment 1 aimed to capture a wide range of images, some of which turned out to be difficult for location sensing. Additionally, subjects participating in Experiment 2 were familiar with both the set-up of the experiment (e.g., interface design) and the layout of the floors from participation in Experiment 1.

In one experiment, over a one-day period, web workers responded to 15 VizWiz user images in an average of 27.0 seconds [Bigham et al. 2010]. A response time close to this value would be more acceptable for an indoor location sensing system. In Experiment 2, the median response time was 25.12 seconds (down from 49.88 seconds in Experiment 1), which suggests that location sensing with this method can be fast enough for practical use and that more experience with the location sensing task may result in significant improvements in median response time. Still, we do acknowledge that this median response time was reached only when subjects had previous familiarity with the floor plan. Training of paid or volunteer web workers may be necessary to obtain response times near this level. Floorplan complexity also affected response time, and it's likely that the nature and distribution of distinctive features would as well.

In addition to individual improvement through testing, there was substantial individual variability in performance among the twenty-four subjects in Experiment 1. Median response times ranged from 30.85 to 97.50 seconds, and median location accuracy ranged from 2.09 to 7.34ft. These variations may indicate that certain individuals have inherently better spatial abilities in performing our task. The differences might also be due to differences in task strategy. Subjects that had high location accuracy as well as short response times were able to quickly verify their location judgments. Less successful subjects either did not spend enough time verifying their responses and made many mistakes or spent too long and had high response times. For example, one subject with a median location accuracy of 2.92ft and a median response time of 97.5 seconds quickly narrowed an image location to a short stretch of hallway, but often spent over a minute deciding on the exact location. Higher performing subjects seemed to be able to place little weight on temporary aspects of the building when making location estimates while subjects with higher median response times were delayed by small details in

the building that had changed between the time when representative images and test images were taken. Again, there was no correlation between response time and location accuracy.

A third goal of this research was to identify the type of pedestrian image that resulted in the highest location accuracy and shortest time required for location sensing. In both experiments, junction test images yielded better location accuracy and shorter response times than photos taken in mid-hallway locations. While pictures taken anywhere might prove useful, visually impaired pedestrians could be told that photos taken at intersections are likely to yield better performance. It should be noted, however, that this finding may be related to our use of intersection photos as the representative images used by our subjects. Had we also included mid-hallway representative photos, the performance difference between the two types of images might have disappeared. Overall, the median location accuracy of images from both experiments did fall within the 10-foot benchmark described in Section 4.1.1. This level of accuracy among all test image types allows for flexibility in the type of pictures that can be taken by pedestrians.

The key question in Experiment 2 was to determine if the pictures taken by visually impaired pedestrians would have adequate quality to yield the high performance levels observed with the test images in Experiment 1. Recall that the test images in Experiment 1 were photos taken by a normally sighted experimenter. The six subjects who participated in Experiments 1 and 2 commented informally that both groups of images were similar in quality (e.g., levels of blurriness). They also noted that visually impaired pedestrians were able to capture informative aspects of the building in their camera images. Only rarely was there a complete lack of useful information (e.g., an image containing only a white wall). The cellphones used functioned adequately across the range of lighting conditions in the test spaces. Among pictures taken by visually impaired pedestrians, there was only a small difference in median location error (3.32ft for low vision and 3.65ft for blind) suggesting this location-sensing method would be useful to users with no vision. The location accuracy of the six subjects in Experiment 2 was slightly better than their performance in Experiment 1. As with response time, this was likely due to a combination of practice, increased familiarity with the interface, and experience with both floor layouts. Also, visually impaired pedestrians were able to provide photos with enough informative features to allow for good location sensing.

If this method of location sensing were to be implemented using crowdsourcing, it could likely be done at a low cost to the user as demonstrated by the VizWiz object recognition app (about $0.07 per question) [Bigham et al. 2010]. Also, web workers who participate in social service related crowdsourcing work have been shown to have altruistic rather than monetary motivations [Takagi et al. 2013]. This may allow for a consistent population of web workers. Encouragement comes from the BeMyEyes app, which hasn't had difficulty maintaining unpaid web workers. The app has a large ratio of signed-up web workers to visually impaired users at over 13:1. But reliance on a base of remote volunteers could be limited by (1) round-the-clock availability of volunteers, and (2) the spatial skill or training level of the volunteers. If a consistent, high-performing population of volunteers were to be impractical, web workers would need to be paid. This adds an ongoing cost that is not present in other indoor localization methods.

Implementation of this method for location-sensing also relies on the availability of floor plans linked to suitable representative images. Collection of these images and rendering the floor plans would require significant overhead effort and periodic updating.

There is also the technical problem of making sure a web worker receives the correct floor and interface information associated with a pedestrian's current location. Even if web workers can effectively locate a pedestrian, we must figure out how location and orientation information can be sent back to the pedestrian in a useful format. In the case of an indoor navigation system, users would likely receive route instructions through synthetic speech (i.e., "walk 10 steps then turn right") similar to the

system proposed by Legge et al. [2013] who showed that given a known location, subjects can use a digital map and properly structured verbal instructions to locate nearby points of interest or follow a series of waypoints to a remote destination in the floor plan. In the event that a user is provided an inaccurate location, the user would submit another photo after realizing the error. Web workers would also be trained to provide a location only when they are highly confident the location is accurate. Otherwise, the user would be instructed to take another photo (e.g., "location not found, please send a more informative photo").

A variant of the method described here would be fully automated location sensing using computer image matching [Kawaji et al. 2010; Liu et al. 2010; Tomasi et al. 2013; Hile and Borriello 2008]. Given a sufficiently large number of representative images of an indoor layout, it is likely that a computer algorithm could achieve location sensing approaching the accuracy of our human subjects. In a hybrid system, a computer could sift through a set of representative images and provide a human web worker with a small number of possible matches for an incoming test image. A similar model combining the skills of humans and computers is that of the image description service WebInSight. WebInSight gives descriptions of web-based images for people with visual impairment from an existing database or calculates a description when one is not present using web context analysis and enhanced optical character recognition (OCR). If a description is not sufficient or if alternative text is not able to be automatically calculated, users can request an additional description from a web worker. This feature allows for lower costs as web workers are only used when necessary [Bigham et al. 2006]. Similarly, web workers may only be called upon to perform location sensing when the image matching system is unsure of the pedestrian's location.

The 20 indoor positioning systems reviewed by Liu et al. [2007] though not designed specifically for visually impaired users provide a comparison of location accuracy between our method and others. Mulloni et al. [2009] used camera phones and fiduciary markers to determine indoor localization. Localization accuracy could not be found for this method. One drawback of this system is that finding the fiduciary markers may be difficult for visually impaired users. Dutoit et al. [2016] were able to provide position and facing direction of a user in real-time using 3D mapping and inertial measurements. Speed for this method was high with localization updates made in an average of 214ms. Several fully automated image-matching localization systems have also been developed. Liu et al. [2010] created a system that compares image frames from user-captured video with those of a reference video of the indoor space. Results from testing showed that 90% of frames from the user video had median localization error of 1.0 meter. This system would require substantial user investment since a special camera is used as well as a backpack to carry the computer that performs localization. Tomasi et al. [2013] matched features from two user images to features from 360° panoramic images taken of the building. Thorough testing of localization ability has not been done, but performing a comparison with two user images reduced incorrect localizations significantly. Hile and Borriello [2008] were able to obtain localizations to 30cm within several seconds by matching pre-marked features from a floor plan with features extracted from a user image. Features, in this case, are corners of walls and doorway locations. Difficulties with this system occur on floors not consisting of just hallways and in cases where corners of walls cannot be distinguished. Benefits of image matching (especially speed) could be exploited in a system that couples our method with image matching.

Among systems specifically designed for location sensing of visually impaired pedestrians, Ran et al. [2004] were able to provide a location accuracy within 22cm by installing pilots (devices that transmit ultrasound signals) to receivers placed on each shoulder of the user. The flight time difference between ultrasound signals sent from two pilots is used to find the distance between the user and the pilots. However, Reflection and blockage of ultrasound signals resulted in dead-spots and installation of pilots is required. The San Francisco Airport has implemented a prototype indoor navigation aid for the

visually impaired using Bluetooth beacons with a reported location accuracy of within 5 meters, but it required installation of over 300 Bluetooth beacons [Iozzio 2014]. The required location accuracy will depend on the task in question; finding a room in an office building would usually require greater accuracy than finding a departure gate at an airport.

We have shown that a scheme relying on remote sighted assistance can provide accurate estimates of indoor spatial location and orientation to pedestrians. Improvements to the interface and training of subjects will likely reduce response times and improve location accuracy. We have also shown that pictures taken at hallway intersections facilitate location sensing. Finally, we have shown that visually impaired pedestrians are able to take cellphone pictures of sufficient quality to permit good localization judgments by remote sighted subjects. Our findings show the feasibility of crowdsourcing for indoor location-sensing and implementation in an indoor navigation system for visually impaired pedestrians.

ACKNOWLEDGMENTS

We thank Rachel Gage for recruiting subjects and experiment set-up, Yingchen He for help with data organization and suggestions on experiment protocol, and Bosco Tjan PhD for showing us how PowerPoint could be used to register accurate location sensing responses of subjects.

REFERENCES

Ariadne GPS. 2017. Ariadne GPS | Mobility and map exploration for all. Retrieved February 18, 2017 from http://www.ariadnegps.eu/.

Be My Eyes. 2017. Be My Eyes | Bringing Sight to the Blind and Visually Impaired. Retrieved February, 18. 2017 from http://www.bemyeyes.org/.

Jeffrey P. Bigham, Jayant Chandrika, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, Tom Yeh. 2010. VizWiz: Nearly real-time answers to visual questions. In *Procedings of the 23nd Annual. ACM Symposium on User Interface Software and Technology (UIST'10)*. ACM, New York, NY, 333–342. DOI:http://dx.doi.org/10.1145/1866029.1866080

Jeffrey P. Bigham, Ryan S. Kaminsky, Richard E. Ladner, Oscar M. Danielsson, and Gordon L. Hempton. 2006. WebInSight: Making Web Images Accessible. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility (Assets'06)*. ACM, New York, NY, 181–188. DOI:http://dx.doi.org/10.1145/1168987.1169018

Ryan C. Dutoit, Joel A. Hesch, Esha D. Nerurkar, and Stergios I. Roumeliotis. 2016. Consistent map-based 3D localization on mobile devices. *Comput. Res. Repos.* (2016).

Thomas Gallagher, Elyse Wise, Hoe Chee Yam, Binghao Li, Euan Ramsey- Stewart, Andrew G. Dempster, and Chris Rizos. 2012. Indoor navigation for the blind and vision impaired: Where are we and where are we going? *J. Locat. Based Serv. - Int. Conf. Indoor Position. Navig.* 8, 1 (2012), 54–73. DOI:http://dx.doi.org/10.1109/IPIN.2012.6418894

Nicholas A. Giudice and Gordon E. Legge. 2008. Blind navigation and the role of technology. *Eng. Handb. Smart Technol. Aging, Disabil. Indep*. (2008), 479–500. DOI:http://dx.doi.org/10.1002/9780470379424.ch25

Harlan Hile and Gaetano Borriello. 2008. Positioning and orientation in indoor environments using camera phones. *IEEE Comput. Graph. Appl.* 28, 4 (2008), 32–39. DOI:http://dx.doi.org/10.1109/MCG.2008.80

Corinne Iozzio. 2014. Indoor mapping lets the blind navigate airports. Retrieved February 18, 2017 from http://www.smithsonianmag.com/innovation/indoor-mapping-lets-blind-navigate-airports-180952292/?no-ist.

Hisato Kawaji, Koki Hatada, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2010. Image-based indoor positioning system: fast image matching using omnidirectional panoramic images. In *Proceedings of the 1st ACM International Workshop on Multimodal Pervasive Video Analysis (MPVA'10)*, ACM, New York, 1. DOI:http://dx.doi.org/10.1145/1878039.1878041

Gordon E. Legge, Paul J. Beckmann, Bosco S. Tjan, Gary Havey, Kevin Kramer, David Rolkosky, Rachel Gage, Muzi Chen, Sravan Puchakayala, and Aravindhan Rangarajan. 2013. Indoor navigation by people with visual impairment using a digital sign system. *PLoS One* 8, 10 (2013), 1–15. DOI:http://dx.doi.org/10.1371/journal.pone.0076783

Bill Holton. 2013. A review of the TapTapSee, CamFind, and Talking Goggles Object Identification Apps for the iPhone. (July 2013). Retrieved February, 18 2017 from http://www.afb.org/afbpress/pub.asp?DocID=aw140704.

Gordon E. Legge, Sarah J. Mason, Mark Brady, Nicholas Guidice, and Erick J. Schlicht. 2003. Maplets: Local geometrical components of human cognitive maps. *J. Vis.* 3, 9 (2003), 136. DOI:http://dx.doi.org/10.1167/3.9.136

Hui Liu, Houshang Darabi, Pat Banerjee, and Jing Liu. 2007. Survey of wireless indoor positioning techniques and systems. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* 37, 6 (2007), 1067–1080. DOI:http://dx.doi.org/10.1109/TSMCC.2007.905750

Jason J. Liu, Cody Phillips, and Kostas Daniilidis. 2010. Video-based localization without 3D mapping for the visually impaired. In *Proceeding of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops (CVPRW'10)*. IEEE, New York, NY, 23–30. DOI:http://dx.doi.org/10.1109/CVPRW.2010.5543581

Alessandro Mulloni, Daniel Wagner, Istvan Barakonyi, and Dieter Schmalstieg. 2009. Indoor positioning and navigation with camera phones. *IEEE Pervas. Comput.* 8, 2 (2009), 22–31. DOI:http://dx.doi.org/10.1109/MPRV.2009.30

Uta R. Roentgen, Gert Jan Gelderblom, and Luc P. de Witte. 2011. Users' evaluations of four electronic travel aids aimed at navigation for persons who are visually impaired. *J. Vis. Impair. Blind.* 105, 10 (Nov. 2011), 612–624.

Seeing Eye GPS. 2017. Sendero Group: The Seeing Eye GPS™ App for cell-enabled iOS devices. Retrieved February, 18. 2017 from http://www.senderogroup.com/products/shopseeingeyegps.html.

BlindSquare. 2017. Retrieved February, 18 2017 from http://blindsquare.com/.

Uta R. Roentgen, Gert Jan Gelderblom, Mathijs Soede, and Luc P. de Witte. 2009. The impact of electronic mobility devices for persons who are visually impaired: A systematic review of effects and effectiveness. *J. Vis. Impair. Blind.* 103, 11 (2009), 743–753.

Lisa Ran, Sumi Helal, and Steve Moore. 2004. Drishti: An integrated indoor/outdoor blind navigation system and service. In *Proceedings of the 2nd IEEE Annual Conference on Pervasive Computing and Communications (PerCom'04)*, New York, NY, 23–30. DOI:http://dx.doi.org/10.1109/PERCOM.2004.1276842

TapTapSee. 2017. TapTapSee - Blind and Visually Impaired Assistive Technology - powered by CloudSight.ai image recognition API. Retrieved February, 18 2017 from www.taptapseeapp.com.

Hironobu Takagi, Susumu Harada, Daisuke Sato, and Chieko Asakawa. 2013. Lessons learned from crowd accessibility services. Lecture Notes in Computer Science. Vol. 8117 (2013), 587–604. DOI:http://dx.doi.org/10.1007/978-3-642-40483-2_42

Matteo Tomasi, Paolo Anedda, Shrinivas Pundlik, Jin Zheng, and Gang Luo. 2013. Using smartphone sensor to improve image-based scene recognition for indoor localization. In *Proceedings of the 2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN'15)*. IEEE, New York, NY, 2007–2008.

VizWiz. 2017. VizWiz - Take a Picture, Speak a Question, and Get an Answer. Retrieved February, 18. 2017 from http://www.vizwiz.org/.

Received February 2016; revised December 2016; accepted December 2016