Helping Visually Impaired Users Properly Aim a Camera

Marynel Vázquez
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA
marynel@cmu.edu

Aaron Steinfeld
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA
steinfeld@cmu.edu

ABSTRACT

We evaluate three interaction modes to assist visually impaired users during the camera aiming process: speech, tone, and silent feedback. Our main assumption is that users are able to spatially localize what they want to photograph, and roughly aim the camera in the appropriate direction. Thus, small camera motions are sufficient for obtaining a good composition. Results in the context of documenting accessibility barriers related to public transportation show that audio feedback is valuable. Visually impaired users were not affected by audio feedback in terms of social comfort. Furthermore, we observed trends in favor of speech over tone, including higher ratings for ease of use. This study reinforces earlier work that suggests users who are blind or low vision find assisted photography appealing and useful.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User interfaces – Input devices and strategies, Interaction styles

General Terms

Design, Experimentation, Human Factors

Keywords

Photography, Visually Impaired, Accessibility, Transit

1. INTRODUCTION

The goal of this work is to enable assisted photography for people who would normally have trouble taking a picture due to a visual impairment. There is evidence that people who are blind and low vision desire the ability to photograph people, events and objects, just like sighted users [11]. Furthermore, there is a desire to use cameras to obtain visual information, like the denomination of currency [14]. However, there is a basic barrier in the first step of the photography process. It is difficult to take a picture when one cannot see what is shown in the viewfinder.

Properly aiming the camera is crucial when taking a picture. Besides aesthetics, aiming is important because poor image compositions can make pictures hard to understand, thereby reducing their value. For example, cropped faces are a common result of improper camera aiming, and strongly discourage people with visual impairments from photographing other people. Likewise, a badly aimed picture of an accessibility barrier may not capture adequate information to properly document the barrier.

To the best of our knowledge, little research has explored different interaction modes to help visually impaired users properly aim a camera. Survey data suggests that spoken directions are the preferred type of guidance cue, with respect to audio tones, and vibrations [11]. Systems that rely on spoken information to help users aim the camera include the native iOS5 camera application for the iPhone platform with VoiceOver activated, VizWiz::LocateIt [3], and EasySnap [11]. The former uses face recognition to inform about faces in the view of the camera. The middle uses voice to inform about proximity to an object. The latter provides spoken information about the position of the camera with respect to an initial view. However, each of these systems is limited and has characteristics which can bias results. The iOS5 implementation only works for faces and provides limited feedback on where to aim the camera. VizWiz::LocateIt requires human assistance and may impose a delay of at least 10 seconds per round of feedback. EasySnap in *people mode* is similar to the iOS camera application, and in *object mode* requires users to first take a picture of the object up close. This can be problematic and hard to attain for larger objects, where close proximity could be dangerous.

In this work, we implemented and evaluated three interaction modes to assist visually impaired users during the camera aiming process: speech, tone, and silent feedback. We assume users are able to spatially localize what they want to photograph, and roughly aim the camera in the appropriate direction. Therefore, small camera motions are sufficient for obtaining a good composition.

We are particularly interested in the following research questions:

1. Is audio feedback valuable when users roughly know the direction in which to aim the camera?

2. Is speech-based feedback preferred over methods with more abstract guidance?

3. How do subjective factors (e.g., overall preference, perceived social comfort, and ease of use, etc) change for these interaction modes?

ASSETS'12, October 22–24, 2012, Boulder, Colorado, USA.
Copyright 2012 ACM 978-1-4503-1321-6/12/10 ...$15.00.

The first question is important because the proposed interaction modes rely on users roughly aiming the camera in the direction of what they want to capture. Therefore, users may feel audio feedback is unnecessary and prefer the silent mode, which has reduced sound contamination on environmental awareness. The other questions seek to identify how the different modes impact preference and acceptance.

We present findings in the context of documenting accessibility barriers related to public transportation. This scenario is motivating because pictures serve as persuasive evidence for promoting changes in transit accessibility [18]. In this context, good composition means a centering model: image subjects, or the main area of interest in an photo, should be framed in the middle. Centering naturally highlights visual evidence for documentation purposes, and increases the chances of including relevant context in images. Alternative composition models, such as the rule of thirds, might be preferred in other cases.

2. RELATED WORK

The process of pointing the camera in the right direction, also known as *focalization* [10], is important when designing camera-based assistive technologies for the visually impaired community. In general, the key to assisting low vision and blind users aim the camera is to transform visual information into another useful representation. Computational approaches to reach this goal can be grouped in two categories: human-driven, and fully automated methods.

Human-driven approaches to help aim the camera rely on human-based knowledge, more than on computing to understand image content. The tele-assistance system for shopping by Kutiyanawala et al. [13] is an example. It was designed to establish communication between a sighted guide and a visually impaired user who carries a camera. The user transmits images of a shelf in a store to the sighted guide through this system, and then the guide uses this data to help pick out target products. The guide further assists in aligning the camera towards targets, and reads nutritional facts from the image to the user. Verbal communication between the sighted guide and the user is key in this process.

To the best of our knowledge, VizWiz was the first crowd-based assisted photography system for blind people [3]. The system was designed to answer visual questions about pictures using Amazon's Mechanical Turk, like "Do you see the picnic tables across the parking lot?". Questions were answered in about 30 seconds, with best times reached with the help of warnings on dark and blurry images. Mitigating poor images was important since they reduced the number of good answers provided by MTurk workers.

VizWiz::LocateIt, a subsystem of VizWiz, was further designed to help blind people locate arbitrary items in their environment [2]. This subsystem provided audible feedback to the user about how much he or she needs to turn the camera in the direction of a target object. Feedback modes included tone and clicking sounds, as well as a voice that announced a number between one and four to indicate how far the user is from the target. Researchers answered requests from users in about 10 seconds for the purpose of evaluating the subsystem, instead of using Mechanical Turk workers. Participants liked the clicking sound to aid in finding a cereal box, and some suggested vibration, verbal instructions, and other familiar sounds as alternatives. No detailed comparison on the perception of feedback modes was provided.

Richardson also explored the use of Mechanical Turk workers to collect information about images [15]. His Descriptive Camera works like a normal camera, in the sense that users aim at what they want to capture. But, instead of producing an image, it outputs a text description of the scene, as provided by a Mechanical Turk worker. In about 6 minutes, the system provides descriptions such as, "This is a faded picture of a dilapidated building. It seems to be run down and in the need of repairs."

Computer vision enables automated approaches for helping aim cameras. The EasySnap framing application [11] relies on image processing to help users aim the camera towards people or particular objects. In the first case, it detects faces, and announces their size and position within the screen. In the second case, it describes how much and which part of the current view of the camera is occupied by an initial, close-up view of an object. Results from a study about the effectiveness of EasySnap to help visually impaired users revealed that most participants thought that the system helped their photography and found it easy to use. Third party observers agreed that 58.5% percent of the pictures taken with EasySnap feedback were better framed than those without, while 12% obtained neutral ratings between the two conditions. The remaining 29.5% were better without feedback.

The PortraitFramer application by the same authors [11] further informs about how many faces are in the camera's sight. Visually impaired users can explore the touchscreen panel of the phone to feel the position of faces through vibration and pitch cues. This information information can then be used to position people in photographs as desired.

Apple's camera application for the iPhone works in a similar manner to PortraitFramer. The release of the iOS5 mobile operating system updated the camera application with face recognition capabilities natively integrated with Apple's built-in speech-access technology. The camera application announces the number of faces in the current view of the camera, as well as a simple descriptor of face position for some scenarios. Common phrases that the system speaks up include "no faces", "one face" and "face centered". Moreover, the system plays a failure tone when users touch the screen outside of a region containing a face, thus providing a physical reference on how well a face is centered.

Other automated, camera-based applications outside of the photography domain also try to provide cues with respect to camera aiming. For example, Liu's currency reader [14] does not actively encourage a particular camera motion, but does provide real time response on whether a bill is readable within the image. This binary feedback is useful for identifying and learning good aiming positions.

Likewise, the mobile application by Tekin and Coughlan [19] tries to automatically direct users towards centering product barcodes in images. Users hold the camera about 10 to 15cm from a product, and then slowly scan likely barcode locations. The system is silent until it finds sufficient evidence for a barcode, and then provides audio feedback for centering. Guidance is provided through four distinct tone or verbal sounds that indicate left, right, up or down camera motions. Initial results published by the authors do not provide insight on particular audio feedback preferences.

Work on camera-based navigation for visually impaired users is also relevant when studying camera aiming. The indoor navigation system with object identification by Hub, Diepstraten and Ertl [8] answers inquiries concerning object features in front of the camera. The authors use a text-to-speech engine to identify objects, and provide additional spatial information. The system by Deville et al. [5] guides the focus of attention of blind people as they navigate. Rather than speech, these authors use spatial sounds generated from color features to indicate noteworthy parts of the scene.

3. METHOD

We conducted an experiment to study different interaction modes to steer users towards proper camera aiming positions. We framed this study in the context of documenting accessibility barriers related to public transportation. Our motivation in this scenario is twofold: rich multimedia documentation of problems serves as persuasive evidence for promoting changes in transit accessibility [18]; and previous research suggests photos are an attractive reporting method for riders [17]. Besides supporting assisted photography, we hope our findings encourage problem documentation through pictures between the visually impaired community. Empowering these riders to collect visual evidence of problems can lead to better communication between riders and transit authorities. Thus, there is a higher chance issues will get solved faster and more appropriately.

3.1 Assisted Photography Application

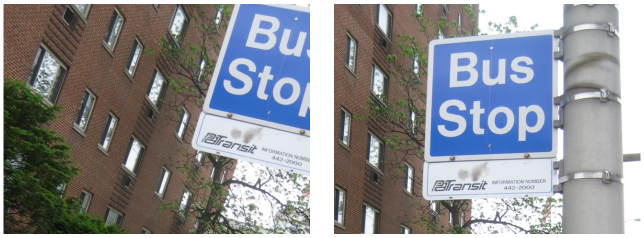
We created an interactive application for the iPhone platform to assist visually impaired users during photographic documentation of transit accessibility. We chose this mobile platform because of its versatility, screen reader capabilities, and high levels of adoption between our main target users.

The problem of taking a "good" picture in this context is difficult, but dramatically simplified by the task characteristics. First, aesthetics are not an issue for problem documentation, thereby mitigating a significant challenge. Second, we do not need to know what the barrier is – we only need to know where it is. While being able to automatically annotate barriers might be useful for documentation, it is not essential. This mitigates the need for object recognition. Third, we can assume the rider is able to localize the barrier in space and roughly aim a camera at the target. This means only small camera motions are needed to balance photo composition and correct unwanted camera orientation.

Consider Figure 1a as an example. We can deduce from the initial view of the scene that the area of interest in the picture is related to the stop sign. Thus, one way of improving the image would be to aim the camera towards the upper-right region of the initial view, bringing the sign to the center of the picture. A centering image composition model helps in this context because it naturally highlights evidence, and increases the chances of including relevant context in pictures. Figure 1b shows the suggested view, as automatically proposed by our system in a simulation.

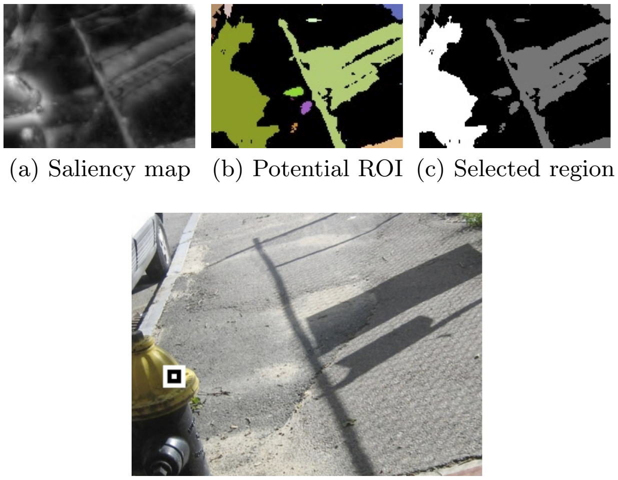
3.1.1 *Region of interest selection*

Our system automatically selects a region of interest (ROI) in pictures, and suggests it as the main subject of the composition for documentation purposes. Our technique can be described as a method to avoid leaving out information that is expected to be most relevant. This strategy was designed for the transit domain without explicit knowledge of object models, and leverages the fact that this domain is strongly composed by conspicuous elements. Complete rationale, algorithm details, and evaluation of this approach can be found in [20].



(a) Initial camera view (b) Suggested final view

Figure 1: Automatically proposed view on simulation test



(a) Saliency map (b) Potential ROI (c) Selected region

(d) Original image and suggested center

Figure 2: Automatic ROI selection process, and suggested image center (rectangle)

Our system constructs a model of visual attention in an image employing a modified version of saliency maps, as defined by Itti and Koch [9]. These maps highlight visual stimuli that are intrinsically salient in their context, which tends to be the case for transit elements in street pictures.

Possible regions of interest are generated by thresholding the saliency map of an image. These regions are later ranked based on their size and saliency, and the one with highest score is selected as the ROI. Figures 2a, 2b and 2c depict this process.

3.1.2 *Image composition assessment*

Our system suggests the weighted center of the ROI as the new center for an image, using saliency for the weights. The suggested center is biased towards the most salient point in the ROI, as shown in Figure 2d, which may not be the most salient point in the image. If we chose the most salient point in the image directly, then our proposed center would be driven towards small salient regions that are less likely to be a good composition subject. The point of maximum saliency in Figure 2d is a tiny portion of green grass, for example, which is located in the top-right corner of the picture.

Our system considers the image to have a good composition when the weighted mean of the ROI is near the geometric center of the picture. If this is not the case, then the system enters in an interactive mode to try to help users frame the ROI during problem documentation.

3.1.3 Interaction Modes

After an initial aiming direction is set, users slowly move the phone to improve image composition, based on the location of the center suggested by the system. Every frame received from the camera is processed as fast as possible to track the position of the region of interest, and provide real-time feedback during this phase.[1] Tracking is accomplished through a standard Lucas-Kanade template matching algorithm [1].

Our mobile application operates in one of three feedback modes while the user tries to frame the ROI:

Speech-based feedback: Spoken words provide information about the relative orientation of the suggested center with respect to the middle, as well as the distance between the two. The system repeatedly speaks "up", "down", "left" or "right" to indicate orientation, depending on whether the suggested center is located in the upper part of the image, the lower part, etc. Words are spoken with different pitch as a cue on how close the suggested center is to the middle. Higher pitch means closer.

Tone-based feedback: The pitch of a looping tone indicates distance from the suggested center to the middle of the image. Higher pitch means closer as before. No orientation information is provided.

Silent feedback: The system lets the user capture the scene continuously, without providing any audible guidance.

In all three modes, the collected image is one where the ROI is closest to the center. For this reason, we have nicknamed the silent mode as *paparazzi* mode. A user can simply wave the phone in slow motion and the most centered frame will be selected. This mode is still interesting because it does not reduce surrounding awareness through noise pollution, and allows users to take pictures without attracting others' attention. Similar to the other modes, it requires real-time operation to track the ROI as the camera moves, and does alert when enough data has been collected.

We also tried vowel-like sounds proposed by Harada, Takagi and Asakawa [7] to represent radial directions during the pilot phase of our study. We soon realized that the limited time users had for familiarization with the system was not enough to learn the mapping of these sounds. However, we believe these sounds are promising for providing orientation information when users have the opportunity for longer practice times.

3.1.4 User Interface

The user interface of our application is very simple. When the application starts, the camera view is shown on the full screen. Once roughly aimed, users hold still and tap the touchscreen anywhere. The system quickly suggests a new image center based on the estimated ROI in the initial image, and draws a circle over this point to indicate its location. An "X" mark also appears on the middle of the image as a reference for those who can see the screen. The system plays a short tone afterwards to let users know they can begin moving the camera slowly to center the ROI. One of the feedback modes described previously guides (or not) the user towards the ROI.

A trial finishes in several ways. Ideally, the user will steer the ROI into the center of the image, given a small margin for error. In this case, it saves the last frame as the best image captured. The system fails and stops early, when the ROI exits the image, or camera motion induces extreme blur and tracking fails. Upon finishing, the system plays a sound and shows the best image captured during the aiming process.

3.1.5 Other implementation details

Many final images taken with our system were blurry during preliminary testing. These images showed low spatial detail and had reduced edge sharpness, in comparison to the initial image users tried to capture. This was discouraging for documenting accessibility barriers, so we decided to add blur estimation capabilities to our system. Our hope was that this would help reduce the number of times significantly blurred images were selected as the best captured.

We chose the no-reference blur metric by Crete et al. [4] for our system. The metric is not computationally intensive, and had better agreement with human ratings of blur than other methods found in the literature [12, 6, 16]. The evaluation was performed on 100 images depicting Pittsburgh's public transportation system, which were captured in the wild by team members using our assisted photography application. Figure 3 shows objective blur ratings obtained with [4] versus subjective opinions. More detailed results on blur estimation are out of the scope of this paper.

We altered the frame evaluation criteria of the application when we incorporated blur detection. The final implementation examines the final set of frames and tries to pick the best combination of close proximity to the center and low blur. Note that if the initial image is very blurry, subsequent best frames may be blurry as well. The system does not deal with focus or exposure, though this would be a nice addition.

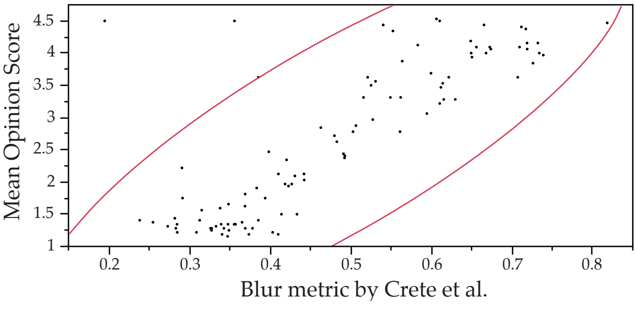


Figure 3: Strongly correlated subjective and objective [4] blur scores ($r = 0.8258$, $p = 0.0001$) on public transportation images. Mean Opinion Scores were computed as the average subjective ratings collected in a 5-point scale

[1] An average of 16 frames per second are processed with added background logging processes for future data analysis.

3.2 Participants

During recruitment, the participants were informed they would be completing surveys and documenting items in our laboratory. All participants were paid volunteers and fully consented. There were three groups of six participants each: full vision or corrected to full vision (F), low vision (L), and blind (B). While the first group may seem unnecessary, universal design practices recommend testing systems for broad appeal. The second group included participants with a wide range of visual impairments, none of whom could easily read the screen of an iPhone. The third group was limited to participants who could only perceive light or were totally blind.

Participants were recruited from local universities and the general public using contacts in local organizations, and community email lists. Participants were required to be 18 years of age or older, fluent in English, and not affiliated with the project.

3.3 Experimental Setup

We used a real-size, simulated bus shelter inside our laboratory for the study (Figure 4). This included a bench, a tempered glass panel on the upstream side of the shelter, a place to mount route information signs, and a bus stop sign. This shelter is comparable in dimensions and layout to real shelters in the Pittsburgh area. We opted for a simulated shelter in order to limit bias from lighting conditions, bystanders, and inclement weather.

We used a within–subject design, and counterbalanced the three interaction modes (Speech, Tone and Silent) using a 3-level Latin Square. Conditions were tested with two documentation tasks: a damaged and non-accessible schedule sign (shoulder height on side wall near glass), and ground obstacles inside the shelter (back left corner). Participants were asked to take 3 practice pictures during the beginning of each condition to get familiarized with the feedback modes. These pictures were taken at a table in the laboratory, and their content included common objects (e.g., plastic container, magazines, etc.). After practice, participants were asked to take 6 trial pictures per condition, alternating between the schedule and the obstacles. Half of the participants per group started with the schedule as initial documentation task, while the rest started with the ground clutter. The duration of the experiment varied depending on the speed in which participants completed the tasks.

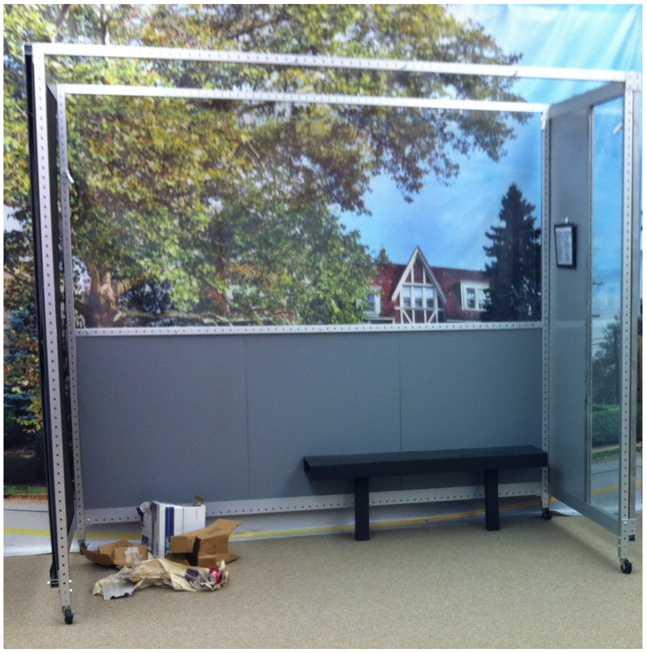


Figure 4: Simulated bus shelter used during the experiment. The schedule sign and the obstacles documented by participants are inside the shelter

The application started recording data when users tapped the screen, up until they were done taking a picture. The following information was collected per trial image:

- Time since the participant tapped the screen and the system presented the best image (Trial Time)
- Distance from the suggested center to the middle of the first processed image (Initial Distance)
- Distance from the suggested center to the middle of the best image presented to the user (Best Image Distance)
- Whether the user brought the suggested center to the middle, or the application stopped because tracking failed (Reached Middle)
- Percentage of the time that users increased the distance from the suggested center to the middle (Moving Away)
- Average device acceleration (Acceleration)

Participants were asked to imagine they were waiting for a bus and document the aforementioned issues using our assisted photography application. They were free to take pictures from where they thought was best for documentation. We did not guide participants towards the schedule or the obstacles, since we did not want to induce bias for particular camera angles.

While the shelter closely mimicked a real shelter, we worried that participants with visual impairments would not be able to find the schedule or the obstacles quickly during the first trial. This initial learning phase could bias the results, so we gave participants a tour of the shelter at the beginning of the study. We removed the ground clutter to allow participants to navigate freely, and familiarize as they would in a real situation. There was also concern that visually impaired participants would get a sense of where the schedule and the obstacles where, and would try to take pictures from afar without having confirmed the location of the targets. To make the experiment more realistic, we asked these participants to physically find the problems before documenting.

Participants completed a pre-test survey covering demographics, disability, and technology attitudes and a post-test survey covering experiences and preferences. The latter included questions on transit complaint filing, technology use, and 7-point scale ratings for feedback mode preference.

Within the study, each participant completed an identical post-condition survey (Table 1) after each condition. This survey was developed by Steinfeld et al. [17] to study modality preference for rider reports on transit accessibility problems, and was previously validated with wheeled mobility device users. Participants were not shown the index labels.

Table 1: Post-condition survey (7-point scale from strongly disagree to strongly agree; R means reversed for analysis)

#	Question	Ease of Use	Usefulness	Social Comfort
1	Learning to use this method was easy.	×		
2	Becoming skillful with this method was easy.	×		
3	I had no problem physically using this method.	×		
4	Using this method would improve my performance in reporting observations.		×	
5	Using this method for reporting observations would increase my productivity.		×	
6	I feel this method is too slow for everyday use. R		×	
7	I felt uncomfortable using this method when people were around in public. R			×
8	When I use this method, I feel like other people are looking at me. R			×
9	Using this method in front of strangers embarrasses me. R			×
10	I like the idea of using this method.		×	
11	I would have done as good a job without using this method. R		×	
12	Carrying items to do this method on daily trip is such a hassle to me. R		×	
	Cronbach's α:	0.849	0.833	0.828

4. RESULTS

As implied by the research questions in the Introduction, this paper is mostly focused on survey results. Complete analysis of the actual content of the data collected by the participants is deferred to future publications.

4.1 Demographics

A total of 18 participants were recruited for the study. The average age per group was 24, 56, and 55 for (F), (L), and (B), with standard deviations of 6.7, 11.8 and 12.1. The percentage of women that completed the experiment was 50%, 50%, and 83%, respectively. One blind participant indicated wearing hearing aids.

Visually impaired participants reported using white canes (58%), guide dogs (25%), magnifiers on glasses (25%), tinted glasses (25%), and hand-held telescopes (17%), between other devices to get around. All these participants had a cellphone, and 66.7% of these devices had a camera.

All participants in the full vision or corrected to full vision group take photos, while 3 and 1 in the low vision and blind groups do. Three totally blind participants said that they had never taken a picture before the experiment. In terms of device usage, 25% of the participants in the (L) and (B) groups said they take pictures with a phone, and only 33% of the low vision participants use a regular camera.

Only one participant in the fully sighted group said that he had filed a complaint about a transit problem, while 5 people in the low vision and 6 in the blind group indicated having filed complaints. Phone was the common way of reporting problems between visually impaired participants.

4.2 Camera Aiming Statistics

A repeated measures ANOVA on Group and Mode was used to analyze log data recorded by the application. Participants took significantly longer to take pictures in Silent mode than in Speech mode, $F(3) = 5.07$ (p $= 0.0068$). The interaction between effects showed that there was a significant difference between the two modes for low vision participants.

The difference in Initial Distance between groups and modes was significant, with $F(3) = 8.42$ (p $= 0.0035$) and $F(3) = 3.56$ (p $= 0.0297$), respectively. The Tukey post-hoc showed that blind participants started off target significantly more than others. Interestingly, Initial Distance with the tone-based feedback was significantly greater than with speech, even though audio feedback was only provided after initial distances were logged.

There were significant differences in Group on Best Image Distance, $F(3) = 6.26$ (p $= 0.0106$), and Reached Middle, $F(3) = 13.86$ (p $= 0.0004$). Fully sighted participants were able to bring the suggested center significantly closer to the middle with respect to blind participants. Moreover, participants in (F) and (L) reached the middle significantly more times than those in (B).

Differences in Mode on Best Image Distance and Reached Middle were significant as well, with $F(3) = 4.99$ (p $=$ 0.0074) and $F(3) = 10.42$ (p < 0.0001), respectively. Post-hoc analyses showed that when users used Speech, their distances from the suggested center to the middle of the best image were significantly smaller than those obtained in with the other modes. Likewise, participants reached more the middle with Speech.

The interaction between Group and Mode was also significant for Distance and Reached Middle, $F(3) = 2.80$ (p $= 0.0261$) and $F(3) = 3.45$ (p $= 0.009$). The post-hoc analysis revealed that Speech gets (B) participants into the final distance and success range of the (L) group. Furthermore, Speech gets (L) participants into the success range of the (F) group for reaching the middle (Figure 5).

The analysis also indicated significant differences in Group and Mode on Moving Away, with $F(3) = 12.37$ (p $= 0.0007$) and $F(3) = 9.78$ (p < 0.0001). Participants in the full vision group moved away from the target less time than the rest, which is not surprising since they can see the view finder of the camera and will notice when they are not making progress towards centering the target. Trials with Speech feedback had significantly lower percentages of time moving away with respect to other modes.

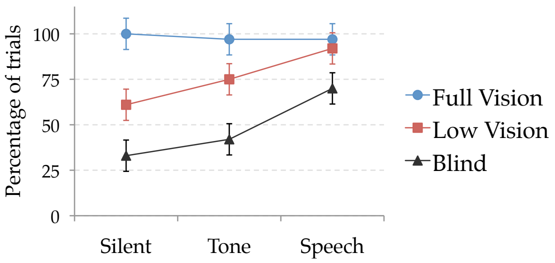


Figure 5: Percentage of trials the target was centered

The average magnitude of device acceleration was significantly different only between groups, $F(3) = 4.65$ (p = 0.0268). Participants in (F) moved the camera significantly slower than those in (B), which was expected because full-vision participants can easily take advantage of the visual information provided through the screen.

4.3 Post-condition ratings

Unless otherwise noted, comparisons were analyzed using a full factorial ANOVA with participant Group and feedback Mode as main effects, followed by a Tukey HSD post-hoc where appropriate. For the purposes of analysis, responses to each question within each post-condition survey category (Ease of Use, Usefulness, and Social Comfort) were flipped to align positive/negative direction, with higher as better, and averaged as a group. Index groups all surpassed the 0.7 reliability acceptance threshold used in the literature (Table 1). ANOVA analyses did not reveal any Ordering effects.

Ease of Use ratings for our application were positive in general (first column of Table 2). There was a significant difference on Ease of Use between participant groups, $F(3) = 6.61$ (p = 0.0030). Full vision or corrected to full vision participants gave statistically significant higher ratings for Ease of Use, with respect to the rest. No other effects or interactions were significant for Group and Mode, but a slight upward trend was observed for Speech.

We realized after running the experiment that there is potential for a small bias in the Ease of Use metric, because the success sound feedback only told participants that a trial had ended, and not whether it had ended successfully. We averaged log statistics per Mode, and checked if there were inconsistencies or unexpected results with respect to Ease of Use. We found that there were significant negative correlations between Ease of Use and Trial Time ($r = -0.4673$, $p = 0.0004$), and between Ease of Use and Moving Away ($r = -0.5381$, $p < 0.0001$). There was also reasonable, significant positive correlation between Ease of Use and Reached Middle ($r = 0.5591$, $p < 0.0001$).

There was a significant difference in Group on Usefulness, $F(3) = 3.57$ (p = 0.0363), and Social Comfort, $F(3) = 5.67$ (p = 0.0064). A post-hoc analysis on the former revealed that full vision participants reported significantly reduced Usefulness as compared to low vision participants (second column of Table 2). A post-hoc on the latter result showed that full vision participants gave significantly reduced Social Comfort ratings than low vision participants (third column of Table 2). Even though the interaction between Group and Mode was not significant, we noticed a trend that suggests that Social Comfort is not affected by audio feedback in the case of people with visual impairments.

4.4 Post-test ratings

A full factorial ANOVA showed significant differences in Mode on post-test preference ratings, $F(3) = 3.32$ (p = 0.0453). Speech mode ratings where significantly higher at the end of the experiment, than those collected for Silent mode. Even though differences in preference per Group were not significant, there were differences in the interaction between Group and Mode, $F(5) = 13.85$ (p < .0001). Visually impaired participants ended up preferring audio feedback over Silent mode, while participants in the full vision group did the contrary (Figure 6).

Table 2: Average ratings on Ease of Use, Usefulness and Social Comfort per group. Standard deviation is shown between parenthesis

	Ease of Use	**Usefulness**	**Social Comfort**
Full Vision	6.76 (0.42)	4.69 (0.88)	4.07 (1.62)
Low Vision	5.83 (1.53)	5.69 (1.39)	5.69 (1.59)
Blind	5.46 (1.13)	5.01 (1.12)	4.61 (1.33)

4.5 Other Findings

Even though the Speech mode was preferred in many cases, we were able to notice some difficulty with the spoken sounds when the phone was held in an orientation other than straight up. For illustrative purposes, consider the case when the system says “up” to indicate that the suggested center is in the upper part of the image. If the user is holding the phone straight up vertically and is aiming the camera to the front, then it is natural to translate the device upwards to bring the center to the middle of the picture. Nonetheless, if the phone is aimed downwards, e.g., towards the ground, then the user should move the phone forward to frame the suggested center in the middle. This dichotomy was a problem for several blind participants, who ended up translating the phone upwards and not forward when aiming downwards. It was hard for them to understand why it was taking so long to center the target in these cases.

Qualitative data, mostly in the form of interviews and comments, were captured during this study. Only one blind participant expressed no interest at all in photography. She was totally blind, and said that she would only do it if there was a way she could feel images, e.g., feel the shape of buildings and big spaces captured in pictures. All other visually impaired users indicated they like (or would like) to take pictures of events, people, and objects.

A low vision participant was a photographer who has been losing his sight progressively. He cleaned the iPhone camera prior to use, and was very concerned about taking the “best” picture for documentation purposes. *“What do you think tells the best story?”* – he kept repeating to himself. Throughout the experiment he got very excited with the system because it was suggesting centers close to the middle. In other words, the application tended to agree that his aiming was appropriate for documentation.

Multiple visually impaired participants used the application to take a picture of their guide dog, and requested a copy for their personal use. Other participants with visual impairments suggested using the system for documenting potholes, which they considered extremely dangerous.

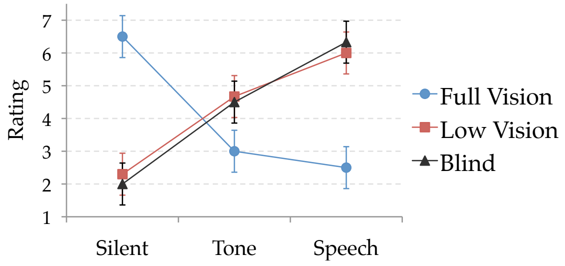


Figure 6: Post-test preference ratings by Mode and User

5. DISCUSSION

Is audio feedback valuable when users roughly know the direction in which to aim the camera?
Yes. Audio feedback helped steering users towards centering targets in pictures, and visually impaired users indicated preference for both Speech and Tone modes, versus Silent mode. Objective data showed that when they interacted with the system in Speech mode, their performance tended to be better (e.g., faster aiming time, more centering, etc).

Is speech-based feedback preferred? How do subjective factors change for these interaction modes?
Speech was preferred over Silent mode, but preferences were not significantly different between Speech and Tone mode. We noticed that visually impaired users were not affected by audio feedback in terms of Social Comfort, though this was not the case for the full vision group.

We observed trends in favor of Speech between the visually impaired community, including slightly higher ratings for Ease of Use. Subjective opinions on Ease of Use and Usefulness were supported by objective data that showed that orientation information (provided only by the Speech mode) seemed to help users center the target more easily.

6. FINAL REMARKS

This study reinforces earlier work that suggests that users who are blind or low vision find assisted photography appealing and useful. Furthermore, it appears there is overall acceptance of assisted photography, including users with full vision, due to the positive ratings of usefulness. The collected results suggest the participants with full vision do find value in the silent *paparazzi* mode, thereby suggesting assisted photography has universal appeal. However, it is clear that the interface may need to change when systems know the user is blind or low vision. The iOS5 camera application's altered behavior when VoiceOver is turned on is a good example of how this can be achieved.

7. ACKNOWLEDGMENTS

The Rehabilitation Engineering Research Center on Accessible Public Transportation (RERC-APT) is funded by grant number H133E080019 from the United States Department of Education through the National Institute on Disability and Rehabilitation Research.

8. REFERENCES

[1] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *Int'l J. Comput. Vision*, 56(3):221 – 255, March 2004.

[2] J. Bigham, C. Jayant, A. Miller, B. White, and T. Yeh. Vizwiz::locateit - enabling blind people to locate objects in their environment. In *Proc. CVPRW'10*, 2010.

[3] 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. Vizwiz: nearly real-time answers to visual questions. In *Proc. UIST'10*, 2010.

[4] F. Crete, T. Dolmiere, P. Ladret, and M. Nicolas. The blur effect: perception and estimation with a new no-reference perceptual blur metric. In *SPIE Conf. Series*, volume 6492, 2007.

[5] B. Deville, G. Bologna, M. Vinckenbosch, and T. Pun. Guiding the focus of attention of blind people with visual saliency. In *Proc. CVAVI'08*, 2008.

[6] R. Ferzli and L. Karam. A No-Reference Objective Image Sharpness Metric Based on the Notion of Just Noticeable Blur (JNB). *IEEE Trans. Image Process.*, 18(4):717 –728, april 2009.

[7] S. Harada, H. Takagi, and C. Asakawa. On the audio representation of radial direction. In *Proc. CHI'11*, 2011.

[8] A. Hub, J. Diepstraten, and T. Ertl. Design and development of an indoor navigation and object identification system for the blind. In *Proc. ASSETS'04*, 2004.

[9] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001.

[10] C. Jayant. Mobileaccessibility: camera focalization for blind and low-vision users on the go. *SIGACCESS Access. Comput.*, (96):37–40, 2010.

[11] C. Jayant, H. Ji, S. White, and J. P. Bigham. Supporting blind photography. In *Proc. ASSETS'11*, 2011.

[12] J. Ko and C. Kim. Low cost blur image detection and estimation for mobile devices. In *Proc. ICACT*, 2009.

[13] A. Kutiyanawala, V. Kulyukin, and J. Nicholson. Teleassistance in accessible shopping for the blind. In *Proc. ICOMP'11*, 2011.

[14] X. Liu. A camera phone based currency reader for the visually impaired. In *Proc. ASSETS'08*, 2008.

[15] Matt Richardson. Descriptive Camera Project. http://mattrichardson.com/Descriptive-Camera/. Last accessed May 2012.

[16] N. Narvekar and L. Karam. A No-Reference Image Blur Metric Based on the Cumulative Probability of Blur Detection (CPBD). *IEEE Trans. Image Process.*, 20(9):2678 –2683, Sept. 2011.

[17] A. Steinfeld, R. Aziz, L. Von Dehsen, S. Y. Park, J. Maisel, and E. Steinfeld. Modality preference for rider reports on transit accessibility problems. TRB 2010 Annual Meeting. Transportation Research Board, 2010.

[18] A. Steinfeld, J. Maisel, and E. Steinfeld. The value of citizen science to promote transit accessibility. In *First Intl. Symposium on Quality of Life Technology*, 2009.

[19] E. Tekin and J. M. Coughlan. A mobile phone application enabling visually impaired users to find and read product barcodes. In *Proc. ICCHP'10*, 2010.

[20] M. Vázquez and A. Steinfeld. An assisted photography method for street scenes. In *Proc. WACV'11*, 2011.