An Assisted Photography Framework to Help Visually Impaired Users Properly Aim a Camera

MARYNEL VÁZQUEZ and AARON STEINFELD, Carnegie Mellon University

We propose an assisted photography framework to help visually impaired users properly aim a camera and evaluate our implementation in the context of documenting public transportation accessibility. Our framework integrates user interaction during the image capturing process to help users take better pictures in real time. We use an image composition model to evaluate picture quality and suggest providing audiovisual feedback to improve users' aiming position. With our particular framework implementation, blind participants were able to take pictures of similar quality to those taken by low vision participants without assistance. Likewise, our system helped low vision participants take pictures as good as those taken by fully sighted users. Our results also show a positive trend in favor of spoken directions to assist visually impaired users in comparison to tone and silent feedback. Positive usefulness ratings provided by full vision users further suggest that assisted photography has universal appeal.

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

Additional Key Words and Phrases: Photography, visually impaired, universal design, accessibility, transit

ACM Reference Format:
Marynel Vázquez and Aaron Steinfeld. 2014. An assisted photography framework to help visually impaired users properly aim a camera. ACM Trans. Comput.-Hum. Interact. 21, 5, Article 25 (November 2014), 29 pages.
DOI: http://dx.doi.org/10.1145/2651380

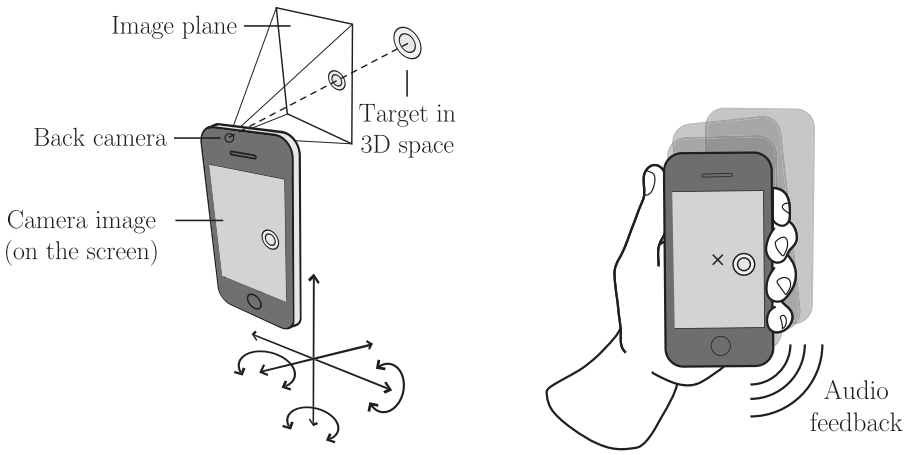
1. INTRODUCTION

Many people in the visually impaired community want to photograph people, events and, objects, just like fully sighted users [Jayant et al. 2011]. Some would also like to use cameras to obtain visual information, like the denomination of currency [Liu 2008] or whether their clothes match [Burton et al. 2012]. However, these users generally have difficulty properly aiming a camera, as does anyone lacking the visual reference provided by a viewfinder. This problem translates to an increased likelihood for capturing pictures with undesirable compositions—in other words, blind users might easily crop faces in a photograph by accident. These compositions reduce the value of the pictures and, ultimately, may impede understanding of the photographer's intent and message.

We address this need for assisted photography in the visually impaired community with an emphasis on camera aiming and show that such assistance can be beneficial in the context of documenting accessibility barriers related to public transportation.

This work is supported by grant number H133E080019 from the U.S. Department of Education through the National Institute on Disability and Rehabilitation Research.
Authors' address: M. Vázquez and A. Steinfeld, Robotics Institute, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213; emails: {marynel, steinfeld}@cmu.edu.

DOI: http://dx.doi.org/10.1145/2651380



(a) Image target (b) Audio feedback

Fig. 1. An illustrative description of our implementation. (a) The location of the 3D target captured by the user. The location of the target in the camera image depends on the relative position of the phone in space. (b) Real-time audio feedback provided by the system to help improve camera aiming. Visual feedback is provided with respect to the center of the image (cross mark).

Our general assumption is that users can spatially localize what they wish to capture and roughly aim the camera in the proper direction. As a result, the initial view of a scene provided by the user is very informative, and small camera motions suffice for obtaining a good image composition. Ways of relaxing this assumption are further discussed in Section 3.1.

The main differences between the assisted photography framework presented in this work and the way conventional cameras work are as follows:

(1) We evaluate image quality based on a real-time image composition model while users take pictures.
(2) We try to steer users to orient the camera in such a way that improves image composition.
(3) We automatically select the final picture taken by users.

A composition model is a set of guides or principles that aid in the placement of visual elements within the image. This model is the key in our formulation because it allows for objective image quality evaluation. Although constraining users to a composition model may appear to limit creative potential, many photographers use composition rules in practice. These rules can be computationally modeled [Datta et al. 2006; Banerjee and Evans 2007; Sung et al. 2012] and therefore incorporated into our methodology.

We evaluated our implementation of the proposed framework with a camera phone application, as in Figure 1. We designed the system in the context of documenting accessibility barriers related to public transportation and chose a centering image composition model for this purpose because it naturally highlights evidence. We focused our efforts on the documentation scenario because pictures are an attractive reporting method for riders [Steinfeld et al. 2010a], and they serve as persuasive evidence for promoting changes in transit accessibility [Steinfeld et al. 2010b]. Besides supporting assisted photography, it is our hope that our approach will enable visually impaired riders to document accessibility barriers through pictures, which in turn can lead to

better communication with the transit authorities. A different computational model, such as the rule of thirds, could be used if a more artistic photograph were desired.[1]

The present article expands our previous work [Vázquez and Steinfeld 2011a, 2011b, 2012] as follows:

—The proposed assisted photography framework is introduced as a guideline for reasoning about real-time systems that help users aim a camera. We illustrate how existing photography systems, or parts of them, fit into our general model.
—New results from our experiment are presented in the context of documenting transit accessibility barriers using our implementation of the proposed framework.
—Recommendations for future system developers are provided, based on the lessons learned from our experience.

An important novel aspect of our empirical evaluation is considering fully blind, low-vision, and fully sighted users from a universal design perspective. Section 5 analyzes pictures taken with and without assistance by these users, both from an objective perspective and a subjective point of view. We address questions such as follows: Is the desired target of the image in the composition? Can third-party observers identify this target as the main subject of the composition? Although our system implementation leverages contextual information from the transit domain, the lessons learned from this experience are applicable to other assisted photography methods.

2. RELATED WORK

The following subsections provide an overview of work related to helping users aim a camera, as well as reference relevant computer vision approaches pertaining image composition evaluation and region of interest (ROI) selection.

2.1. Helping Users Aim a Camera

The process of pointing the camera in the right direction was described as *focalization* in Jayant [2010]. This process generally relies on transforming visual information seen from the camera into another useful representation, which can be done with the help of humans or completely automatically.

Human-driven approaches to help aim the camera rely on human knowledge rather than computational analyses of the image content. The tele-assistance system for shopping by Kutiyanawala et al. [2011] is an example. This system was designed to establish verbal communication between a sighted guide and a visually impaired user who carries the camera. The user transmits images of a shelf in a store to the remote sighted guide, who uses this data to help pick out target products. The guide assists in aligning the camera toward targets and reads nutritional facts from the image to the user.

To the best of our knowledge, VizWiz was the first crowd-based assisted photography system for blind people [Bigham et al. 2010a]. The system was designed to answer visual questions about pictures using Amazon's Mechanical Turk, such as "Do you see the picnic tables across the parking lot?" Questions were answered in about 30 seconds, with warnings on dark and blurry images. Mitigating poor images was important since they reduced the number of good answers provided by Mechanical Turk workers.

VizWiz::LocateIt, a subsystem of VizWiz, was designed to help blind people locate arbitrary items in their environment [Bigham et al. 2010b]. This human-assisted subsystem provided audible feedback to users about how much they needed to turn the

[1]According to the centering rule, the main elements of the composition should be placed in the middle of the photo. In contrast, the rule of thirds suggests to divide the picture into nine equal parts by two horizontal and two vertical lines. The main elements of the composition should then be placed along these lines or their intersections [Bohn 2006].

camera toward a target object. Feedback modes included tone and clicking sounds, as well as a voice that announced a number between one and four indicating how far from the target the camera was aimed. For the evaluation, researchers acted as confederates and responded within about 10 seconds. Participants liked the clicking sound when 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.

Computer vision enables automated approaches for helping to aim cameras, thereby mitigating issues with human assistance latency and availability. Different to our assisted photography application, most of these computational approaches use face detectors or reference images to identify an object's position within a picture. For example, Bae et al. [2010] helped users aim a camera to match an existing photograph from the same view point, simplifying the task of rephotography. The system operated in real time, on a computer connected to a camera, using an initial reference image for input. This image was used to compute a relative viewpoint difference with respect to the current view of the camera. Users were informed of this difference through arrows displayed on the computer screen.

Headshot [Schwarz 2011] is a Windows Phone 7 application designed to help sighted users take a picture of their face with the back camera of the phone. Since the screen and camera are on opposite sides of the phone, fully sighted users cannot obtain a visual representation of the camera image. Thus, sighted users are similar to people with visually impairments when using the application. Headshot detects the face of the user and provides audio feedback toward a manually predefined location in the image. On reaching good positioning, the system provides a spoken warning ("say cheese") and then takes a picture.

The EasySnap framing application [White et al. 2010; Jayant et al. 2011] also used image processing to help visually impaired users aim a camera phone. In one mode, it detected faces and announced their size and position within the screen. In a second mode, it described how much and which part of the current view of the camera was occupied by an initial, close-up view of an object. In a third mode, it detected the contour of a document and tried to steer users' aiming position toward centering this document within the picture. Results from a study about the effectiveness of EasySnap to help visually impaired users in the first and second case revealed that most participants thought that the system helped their photography and found it easy to use. Third-party observers agreed that 58.5% of the pictures taken with EasySnap feedback were better framed than those without. Neutral ratings in both conditions were obtained in 12% of pictures, and the remaining 29.5% were better without feedback. As far as we know, no experimental results to date have been provided on the Document mode of EasySnap.

The PortraitFramer application by the same authors [Jayant et al. 2011] informed users on how many faces were within the camera image. Visually impaired users could explore the touchscreen panel of the phone to feel the position of faces through vibration and pitch cues. This information could 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 VoiceOver 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 and size for some scenarios. Common phrases that the system speaks include "no faces," "one face," "small 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 for visually impaired users also try to provide cues with respect to camera aiming. For example, Liu's currency reader [Liu 2008] 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.

The mobile application by Tekin and Coughlan [2010] tries to automatically direct users toward centering product barcodes in images. Users hold the camera about 10 to 15 cm 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 this element in the picture. 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 blind and low vision users is also relevant when studying camera aiming. The indoor navigation system by Hub et al. [2004] 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. [2008] also 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.

2.2. Image Composition Evaluation

Image composition models have been traditionally studied in the context of aesthetic photography. Many books (e.g., David Präkel [2006] and Bohn [2006]) describe composition principles to help amateur photographers capture the world as professionals. Some of these principles have been applied to 3D rendering applications [Gooch et al. 2001] or have been modeled computationally to automate image composition evaluation. For example, Banerjee and Evans [2007] seek to automate the composition of images with one main subject, based on its position according to the rule of thirds and how prominent it is in the picture. Datta et al. [2006] consider visual features such as the location of the main element of the image, the distribution of brightness, texture, and others to automatically infer the aesthetic quality of photos. Recently, Sung et al. [2012] proposed an interactive optimization method for photo composition on a mobile platform. The latter system uses several composition rules to evaluate image aesthetics based on the main subject, which is manually indicated by users. This system is proof that a variety of image composition models can be evaluated in real time, besides the one considered in our own implementation.

Autonomous camera control systems are popular in robotics, where motion commands are less noisy than human actions and human interaction may be infrequent or hard to obtain in real time. Similar to our assisted photography approach, these systems work under a particular image composition model. Dixon et al. [2003] and Kim et al. [2010], for example, describe implementations of robot photographers designed to capture pictures of people, with framing strategies similar to those described previously. Likewise, Desnoyer and Wettergreen [2010] worked toward aesthetically aware autonomous agents.

2.3. Region of Interest

The image composition model that we implemented for our study relies on visual saliency for estimating which part of an image is relevant for the documentation task. Regions that are visually salient tend to be considered as information carriers that deliver the photographer's intention and capture part of the observers' attention as a whole [Chen et al. 2002].

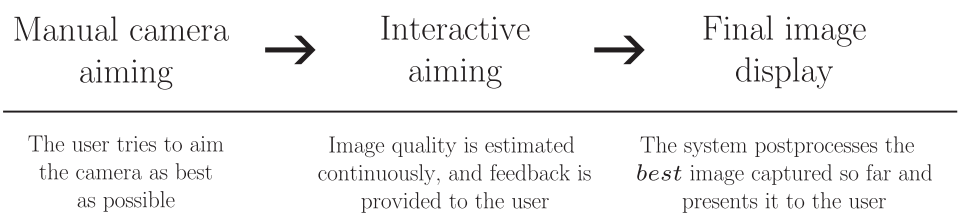


Fig. 2. Assisted photography framework.

Luo et al. [2003] posit that a good composition is the most influential attribute for image emphasis selection. Their system evaluated composition using saliency maps and then used this information to estimate the image of an event that received the most attention.

In our documentation domain, we focus on stimulus-driven visual attention for finding the regions of interest in street pictures, since transit elements tend to be salient and of high contrast. The number, type, and combination strategy of features used for estimating saliency in images is a problem of its own [Frintrop et al. 2010]. Different situations call for different solutions, and so we chose to test several methods with a variety of street photographs [Walther and Koch 2006; Hou and Zhang 2007; Guo et al. 2008; Bian and Zhang 2008; Achanta et al. 2009]. Interested readers should refer to our previous work for more details [Vázquez and Steinfeld 2011a].

3. ASSISTED PHOTOGRAPHY FRAMEWORK

This section describes our proposed assisted photography framework, characterized by the integration of real-time user interaction during the image capturing process. This framework was designed to help visually impaired users improve the way in which they aim the camera and thus the composition of the images taken. In keeping with the spirit of universal design, we believe that such a system is also useful to fully sighted users.

The framework is composed of three sequential components, as depicted in Figure 2. The first component prepares the system while the user initially aims the camera. The second component provides real-time user feedback to refine the position of the camera and improve image quality. Finally, the third component optionally postprocesses the final best image that was captured throughout the whole interaction and displays it to the user.

3.1. Manual Camera Aiming

The first component of the framework observes while the user aims the camera as best as possible and waits until he or she signals readiness to take a picture. The signal can be input through typical gestures, such as a tap gesture on the view finder or screen of a camera phone. A tap is advantageous because it does not require precise input from the user, although a shutter button could similarly be used in a more traditional camera system.

In general, we require users to be able to localize the desired target in space and roughly aim the camera in its direction before advancing to the interactive aiming phase. This means that the initial view of the scene is very informative about what the photographer is trying to capture, and small camera motions will be adequate to improve photo composition.

The preceding requirement is very important from a computational perspective, because it saves us from reasoning about what is left out of the initial view of the scene. Likewise, it may allow us to avoid the complexity of specifically identifying the

target being photographed and rather focus on roughly estimating its location in the image, as explained in Section 4. The fact that small camera motions suffice to improve image composition is also important for user adoption, since users should be able to take a picture with a small amount of effort and in a short period of time.

If we constraint the target being photographed to object classes that can be quickly and reliably identified with current computer vision techniques, then it is possible to relax the preceding requirement. For example, if the assisted photography system is meant to help users take pictures of other people, then it could focus on detecting faces [Viola and Jones 2001] from the time the user roughly aims the camera in the direction that he or she thinks is best. If the system effectively detects a face in the current view of the camera, then it can use this information to suggest how to better frame the person in the picture. Otherwise, it can inform the user that a person was not found in the image and wait for a face to appear or stop further processing.

3.2. Interactive Aiming

The interactive aiming component of the proposed framework iteratively estimates whether the view of the scene can be improved based on an image composition model. This component steers the user toward better aiming positions and saves the *best* image captured during this process. The interactive aiming phase automatically stops whenever the current view of the scene is good enough, based on the composition model. Alternatively, this phase ends either when the user has consumed too much time trying to improve camera aiming or image quality estimation fails. The latter may happen when image quality depends on the position of the target in the image, such as when using a centering model, but the system cannot estimate the target's location.

A computational model of image composition is the key for objectively estimating image quality and providing user feedback. The particular choice of composition model and the features considered for its evaluation depend on the type of pictures that we expect users to take. For example, we chose a centering model for the documentation task described in Section 4 because it naturally highlights evidence and increases the chance of including relevant context in images. Nonetheless, a rule of thirds model may be more appropriate for cases in which we want to help users capture aesthetically appealing images [Präkel 2006; Bohn 2006] or a model that considers target size if we are photographing faces of people [Jayant et al. 2011].

User feedback can be provided through a variety of output modes during the interactive aiming phase, including audio feedback. The latter is particularly attractive for advising changes in camera aiming position because the stimuli can be very expressive and quick to process. From a universal design perspective, we also suggest displaying the current view of the scene to the user and render visual cues to help aim the camera. Figure 1 shows an example.

It is also possible to use vibrations and respond to tactile user input during the interactive aiming phase. This interaction would be similar to how PortraitFramer [Jayant et al. 2011] or the iOS5 camera application inform about faces. However, we fear that vibrations might deviate users' aiming position due to the added device motion. Requiring tactile user input might also be complicated and prohibitively difficult, since users would need to hold the camera steady with one hand and explore the tactile display with the other. Users who need to use a white cane or a guide dog tend to not have the freedom to use two hands when interacting with a camera.

3.3. Final Image Display

The last component of the proposed framework optionally postprocesses the best image captured during the interactive aiming phase and displays it to the user. Image post-processing can be performed to check image quality and, for example, alert the user

Table I. Comparison of Several Interactive Photography Applications Based on the Components of the Proposed Assisted Photography Framework

	PF	iOS5	ES-P	ES-O	ES-D	HS	Ours
Designed for visually impaired users	✓	✓	✓	✓	✓	—	✓
Manual aiming							
Waits for user input to start processing the view from the camera	✓	—	—	✓	—	✓	✓
Interactive aiming							
Continuously estimates the location of the expected target on the image	✓	✓	✓	✓	✓	✓	✓
Continuously informs the user about the location and/or size of the target	✓	✓	✓	✓	—	—	—
Automatically evaluates image quality based on a composition model	—	—	—	—	✓	✓	✓
Informs the user about how to improve camera aiming	?	—	—	—	✓	✓	✓
Automatically selects the *best* image and stops the capturing process	—	—	—	—	?	✓	✓
May select a picture, other than the last one captured, as the *best* image	—	—	—	—	—	—	✓
Final image display							
Alerts users if the final image has poor quality	?	—	✓	✓	✓	—	—
Automatically enhances the final image	—	—	—	—	—	—	✓

Note: "PF" stands for PortraitFramer [Jayant et al. 2011]. "iOS5" refers to the iOS5 camera application. "ES-P" and "ES-O" stand for EasySnap in People and Object mode [Jayant et al. 2011], whereas "ES-D" stands for EasySnap in Document mode [White et al. 2010]. "HS" is Headshot [Julia Schwarz 2011]. "Ours" is our assisted photography application. A question mark (?) in a cell indicates that the authors do not specify the presence or absence of the corresponding feature, although the system may have it.

if the image does not have proper exposure or sharpness levels [Jayant et al. 2011]. Postprocessing operations can also be carried to enhance the final image. For example, in our implementation, we adjusted for camera rotation. One could also opt to apply image filters, crop the final best image to focus on an ROI [Suh et al. 2003], perform a content-aware resizing operation [Avidan and Shamir 2007], or another image improvement process. Note that this step may be unnecessary if the final image is deemed adequate by the system, and displaying an image may not be necessary for some users and applications.

3.4. Framework Instantiations and Other Related Systems

Table I compares several interactive photography applications based on the components of the proposed framework and whether the approach was designed for users with visual impairments. The full description of our assisted photography application is postponed until Section 4.1. Unless necessary for comparison, additional information about our system is deferred to later sections.

Although most of the photography applications listed in Table I were already described in Section 2.1, it is worth discussing a few additional aspects of their implementation. For example, Headshot [Schwarz 2011] fits very well in our assisted photography framework, although it was not designed for visually impaired users and thus may pose problems for this target audience. Headshot considers a picture to be badly composed until the face of a person is positioned in a manually set location in the image. Visually impaired people, however, will have trouble choosing this location because it may depend on the desired background of the picture, and reasoning about this is hard without visual ques and out of physical reach. Alternatively, this system could use image composition models that do not require manual input when the user

is visually impaired. For example, the system could evaluate image composition based on the size of the face in the image and its position (e.g., according to the rule of thirds as in Dixon et al. [2003] and Kim et al. [2010]).

Other applications compared in Table I can be considered partial instantiations of our framework. For example, Easy Snap in Document mode [White et al. 2010] helps users center a book, newspaper, or banknote in the image, as well properly orient the camera. The application provides audio feedback as suggested in the description of the interactive aiming component of the proposed framework. However, the authors do not explicitly indicate whether the user or the system decides when to stop processing new views of the scene, and whether the system displays the final image. In addition, to the best of our knowledge, the final picture taken is the last frame that was captured by the camera, not the best one captured. The other modes in which EasySnap operates [Jayant et al. 2011] let the user decide when to stop processing frames.

Our framework can also be seen as a real-time, automated time-shifting method to help visually impaired users. *Time shifting* is a recent feature of camera phones, such as the BlackBerry Z10 or Samsung Galaxy Note II, that records a sequence of frames when the user takes a picture. After recording, the user can select a favorite frame as the final image. Our proposal is similar in that we process a sequence of frames during the interactive aiming phase, but rather than asking the user to select the best image at the end, we propose to do it automatically and in an online fashion. Note that time shifting does not help users aim the camera better.

Apple also has submitted patents pertaining to automatic frame selection when users request capturing an image [James et al. 2010; Brunner et al. 2012]. These patents describe systems that automatically select a frame from a collection that was captured prior or simultaneously to the request and present this frame in response to the user. The selection of the frame is described in terms of detected device movement using a motion-sensing component or image contrast. However, assisted camera aiming is not discussed in these patents.

4. EMPIRICAL EVALUATION

We implemented the proposed framework and conducted an experiment to evaluate our system in the context of documenting accessibility barriers in public transit settings. We chose this context because we believe that we can improve the communication between riders and transit authorities by helping to collect visual evidence of problems. In turn, this can help authorities appropriately solve the issues that matter to the community [Steinfeld et al. 2010b].

4.1. Assisted Photography Application

We implemented the proposed assisted photography framework on an iPhone 4S, as in Figure 1(b), because of its versatility and high levels of adoption by our main target users. We constrained image orientation to portrait mode because this simplified training users on how to take pictures with our application, as well as our experimental analysis. Nonetheless, landscape compositions could be implemented in the same manner as portrait arrangements. We could make the system trivially identify and adapt to these orientations based on data from the accelerometer in the device.

The problem of estimating image quality in a documentation context is difficult, but it is 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. Although 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 that riders are able to localize the barrier that they want to document in space and roughly

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

Fig. 3. Automatically proposed view on simulation test.

aim a camera in the proper direction, as assumed in Section 3.1. This means that only small camera motions are needed to balance photo composition and correct unwanted camera orientation.

Image quality assessment is performed by our system using a centering image composition model with a strong preference for detailed and sharp images. This composition model dictates that the target of the picture should be centered in the middle of the composition. If two images are composed similarly, our system selects the image with less blur.

We believe that centering the target is appropriate for the documentation context because it naturally highlights evidence for documentation purposes and increases the chance of including relevant context in images. Consider Figure 3 as an example. Our system initially estimates that the target is in the top-right part of the image and suggests positioning this region in the middle of the picture. As a result, the bus stop sign becomes the main element of the composition, and the image includes a wider variety of surrounding context.

The following paragraphs describe our implementation of the assisted photography framework for the documentation task. Some of the information provided below was summarized in Table I, and more details are given in our electronic Appendix A. This appendix describes how we implemented the main processing routine of our application from a systems perspective and details our choice of parameters.[2]

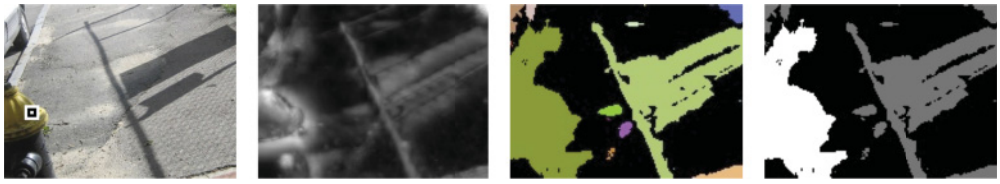
4.1.1. Manual Aiming. Our system waits for the user to roughly aim the camera toward the target of interest before entering the interactive aiming phase. The system displays the camera view on the phone screen during this period and waits for a tap gesture from the user.

Our system behaves just like a regular point-and-shoot camera up to when the manual aiming phase ends at the tap gesture. For this reason, we can analyze the first image processed when the interactive aiming phase starts as if it had been taken with a conventional camera.

4.1.2. Interactive Aiming. Our system estimates a region of interest (ROI) in the first image captured after the tap gesture and suggests this region as the new image center. The ROI is expected to contain the target being photographed or at least a significant portion of it.

Our technique to estimate the ROI is based on visual saliency and can be described as a method to avoid leaving out information that is expected to be most relevant. We opted for visual saliency because we can compute it quickly from low-level visual features and, more importantly, because the transit domain is strongly composed by conspicuous elements. For example, bus stop signs are generally designed to be easy

[2]Code is available at https://github.com/marynelv/assisted-photography.



(a) Original image (b) Saliency map (c) Potential ROI (d) Selected region

Fig. 4. ROI selection process. Figure 4(c) shows all potential regions of interest in different colors, and Figure 4(d) shows the selected ROI in white. The suggested image center (weighted mean of the ROI) is depicted in Figure 4(a) as a rectangle.

to identify, with bright colors and contrasting elements (Figure 3). Similarly, hydrants (Figure 4(a)), route signs, traffic cones, and other transit elements are salient in their context. Recognition capabilities could be added to the system to make it more robust, but this could limit the ability to satisfy our real-time operation constraints.

Rationale, algorithm details, and evaluation of our ROI estimation approach can be found in Vázquez and Steinfeld [2011a]. The following procedure briefly summarizes the steps of this method for completeness:

(1) Compute a saliency map from the image, which encodes local conspicuity in a manner akin to Itti and Koch [2001].
(2) Threshold the saliency map to generate candidate regions of interest.
(3) Select the most meaningful candidate region based on its size and the amount of saliency that it contains.

The weighted center of the ROI is suggested as the new center for the image, using saliency for the weights (Figure 4).[3] The suggested center is biased toward the most salient point in the ROI, 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 toward small salient regions that are less likely to be a good composition subject. For example, the point of maximum saliency in Figure 4(a) is a tiny portion of green grass located in the top-right corner of the picture.

If the ROI is not centered in the first image after the tap, then users are given the opportunity to improve image composition by slowly changing their aiming direction. The system processes every frame received thereafter as fast as possible and keeps track of the best image captured so far, based on the position of the ROI in the composition and blur. We track the ROI in successive frames through a standard template matching algorithm [Baker and Matthews 2004] and estimate blur using a non-reference blur metric [Crete et al. 2007]. This metric does not require a template image to estimate blur, which makes it particularly well suited for blur estimation in dynamic environments.

The application operates in one of three feedback modes during this interactive aiming phase:

—*Speech-based feedback:* Spoken words provide information about the relative orientation of the suggested center with respect to the middle of the composition, as well as the distance between the two. The system repeatedly speaks “up,” “down,” “left,” or “right” to indicate orientation toward a better composition. Words are spoken with

[3]We frequently abuse terminology and use *region of interest* and *suggested center* (computed from the ROI) interchangeably. Centering the ROI in the middle of the composition is undefined when this region is asymmetric. Thus, readers should keep in mind that centering the ROI is, effectively, centering the weighted mean of this region in the middle of the composition.

different pitch, indicating how close the suggested center is to the middle of the image. 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. As before, higher pitch means that the user is closer to the recommended position. No orientation information is provided.
—*Silent feedback:* The system lets the user capture the scene continuously without providing any audible guidance.

The underlying operation of the system is the same with all feedback modes. Although silent feedback does not seem appropriate for visually impaired users, we believe that 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. We sometimes call this *paparazzi* mode since it would be effective when holding a camera above a crowd. Similar to other modes, silent feedback requires real-time operation to track the ROI as the camera moves and to update the information displayed to the user on the screen of the device. With respect to the latter, the system renders the current view of the camera on the screen for users who can see the display. In addition, the system draws an overlay marker to indicate the location of the suggested center.

Our application alerts the end of the interactive aiming phase by playing an audio clip. This happens when the ROI has been centered, tracking fails, or the user spends more than 1 minute attempting to improve the image. The first case is the ideal situation, in which users were able to compose the image as proposed by the system. The second case occurs when the ROI cannot be tracked from one frame to the next, such as if the ROI ends up moving outside of the image or template matching fails due to extreme blur from fast camera motion. The third case happens when users exceed our time limit for improving image composition. The latter usually occurs when a user does not change the aiming direction significantly and holds the camera phone still.

4.1.3. Final Image Display. Our system shows the user the best image captured during interactive aiming as the final image. If the device was held vertically when this picture was taken (i.e., the device was not tilted too much), then our system also processes the image before displaying it on the screen to correct for excessive camera roll. We detect when the phone is held straight using the accelerometer values registered when the picture was taken and correct camera roll by rotating the picture in plane. We believe that this correction facilitates image understanding in the documentation context, because it makes vertical elements appear vertical in the composition rather than with a different orientation. Figure 5 shows an example.

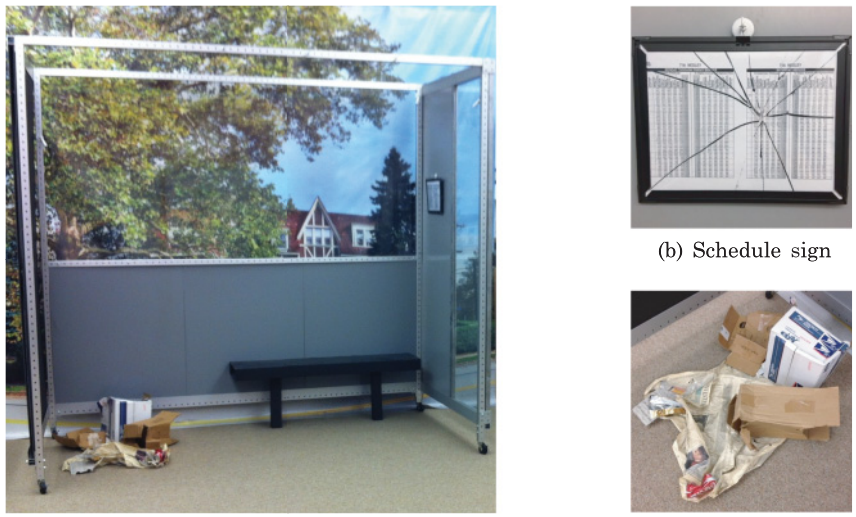
4.2. Participants

There were three groups of six participants each: full vision or corrected to full vision (F), low vision (L), and blind (B). Although 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 the presence of 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. During recruitment, the participants were informed that they would be completing surveys and documenting items in our laboratory. All participants were paid and fully consented.

(a) Camera roll (b) Captured image (c) Corrected image

Fig. 5. Automatic camera orientation adjustment of our assisted photography application. The system rotates the best image captured during the interactive aiming phase to correct for unwanted camera roll. This correction only happens when the phone is mostly held vertically—that is, the phone is not pointed significantly upward or downward. The corrected image is the final picture displayed to the user.



(a) Bus shelter (b) Schedule sign (c) Ground obstacles

Fig. 6. Simulated bus shelter used for our study. The schedule sign and the obstacles documented by participants were inside the shelter.

4.3. Experimental Setup

We used a simulated bus shelter inside our laboratory for the study (Figure 6). This shelter 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. The shelter was comparable in dimensions and layout to real shelters in the Pittsburgh area and was used to limit bias from lighting conditions, bystanders, and inclement weather.

We used a within-subjects design and counterbalanced the three interaction modes (speech, tone, and silent) using a three-level Latin square. Conditions were tested with two documentation tasks: a damaged and nonaccessible schedule sign (shoulder height on side wall near glass) and ground obstacles inside the shelter (back-left corner). Although the schedule sign might seem inappropriate for blind users, it can be of their interest when it contains tactile material, such as content in Braille. Likewise, it

can contain machine-readable optical labels, such as QR codes, with more information about the bus stop.

Participants were asked to take three practice pictures during the beginning of each condition to become familiarized with the feedback modes. These pictures were taken at a table in the laboratory, and their content included common objects (e.g., a plastic container, magazines). After practice, participants were asked to take six 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, whereas the rest started with the ground clutter. The duration of the experiment varied depending on the speed in which participants completed the tasks.

Participants were asked to imagine that they were waiting for a bus and to 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 toward the schedule or the obstacles, as we did not want to induce bias for particular camera angles.

Although 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 were and would try to take pictures from afar without having confirmed the location of the target. To make the study more realistic, we asked the participants to physically find the problems before documenting them.

The application started recording data when users tapped the screen up until they were done taking a picture. Logged measures included the length of the interactive aiming phase, the distance from the suggested center to the middle of the composition, device acceleration, and other useful metrics.[4] After each feedback mode, participants completed an identical postcondition survey with questions about ease of use, usefulness, and social comfort on a seven-point scale. See Vázquez and Steinfeld [2012] for more details.

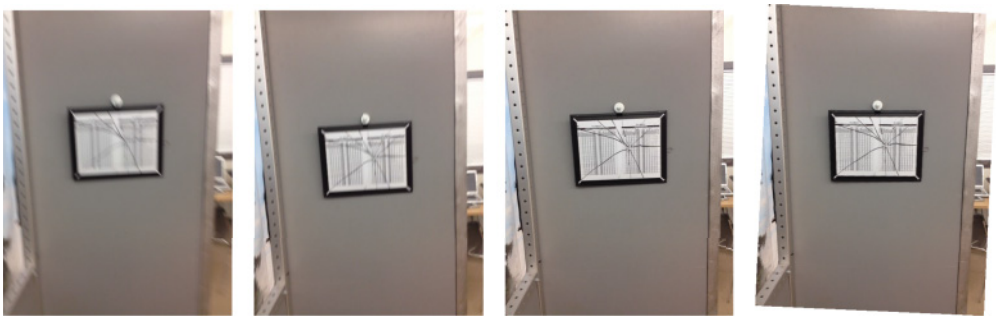
Participants also completed a pretest survey covering demographics, disability, and technology attitudes, as well as a posttest survey covering experiences and preferences. The former was a subset of questions developed by the Quality of Life Technologies Engineering Research Center [Beach et al. 2009]. These were selected to provide the option to draw inferences to the thousands of samples collected by the survey developers and related transportation studies (e.g., Beyene et al. [2009]). The latter included questions on transit complaint filing, technology use, and seven-point scale ratings for feedback mode preference.

4.4. Measures

The results presented in this article are focused on the data collected from the 324 trials in which participants used our application to take a picture (18 participants $\times$ 6 pictures $\times$ 3 feedback modes). Survey results and camera aiming statistics from Vázquez and Steinfeld [2012] are summarized in this article to complement our analysis of the following:

—The ROI selection process in our experimental environment
—Third-party target identification

[4]An average of 16 frames per second were processed in our experiments, with added background logging processes for data analysis.



(a) First I_i (b) Closest I_c (c) Best I_b (d) Rotated best I_b^R

Fig. 7. Example images considered in our third-party evaluation. The differences between the images are better observed digitally with zoom. Refer to Section 4.4 for more details.

—Third-party image understandability
—Perceptible image quality based on blur.

The effectiveness of our ROI estimation approach was measured by counting the number of times the suggested center (inside the ROI) fell into the target being photographed. This count considered the first images processed during the interactive aiming phase and ignored those pictures where the target was not in the composition. The latter are cases in which our main assumption was not true: users did not aim the device roughly in the right direction during the manual aiming phase.

A subset of 867 pictures logged during the experiment were evaluated by third-party observers through Amazon's Mechanical Turk. A total of five surveys were collected per image, and \$0.04 was paid for each complete survey pertaining target identification, image understandability, and blur. For a given trial, we considered:

—The first image processed when the interactive aiming phase started (I_i)
—The image with the suggested center closest to the middle of the composition (I_c)
—The image considered to be the best one by our system (I_b)
—The rotated best image (I_b^R).

Note that I_i, I_c, and I_b may all be the same for a trial, because it is possible that the initial image provided by the user is well composed for our image composition model. Similarly, it is possible for I_i to be different than the best I_b but for I_b to still be equal to I_c. In addition, some trials may not have a rotated best image I_b^R, since our system only compensated for unwanted camera roll if the phone was held mostly straight up. Figure 7 shows examples of these pictures.

The instructions provided to the workers introduced the task with an image similar to Figure 6, where the schedule sign and the trash were identified. The instructions then directed the workers to answer a short survey about a picture taken to document either barrier and provided examples on how to fill this survey and rate image blur. Appendix B provides for more details.

The survey was presented next to the image being evaluated and without explicitly indicating whether the trash or the schedule sign were the intended target. Survey questions were asked as follows:

(1) Rate the following statement (seven-point scale): *I can easily identify the main subject of the picture.*
(2) What do you think is the main subject of the picture? (option 1) trash, (2) schedule sign, (3) other, (4) don't know.

(3) In which part of the image is the main subject? (option 1) In the center, (2) in the left side, (3) in the right side, (4) in the top part, (5) in the bottom part, (6) in the bottom-right part, (7) in the bottom-left part, (8) in the top-right part, (9) in the top-left part, (10) the main subject is not in the image.
(4) Rate your perception of blur in the whole image: (1) imperceptible, (2) perceptible but not annoying, (3) slightly annoying, (4) annoying, (5) very annoying.
(5) Rate the following statement (seven-point scale): *I am confused about what the photographer tried to capture because the picture is hard to understand.*

The first and last questions were rated from 1 to 7, from "strongly disagree" to "strongly agree," and were used to measure image understandability. The second question was open in the sense that workers could select "other" (option 3) and indicate what they thought was the main subject of the composition in their own words. For cases where "other" was selected and the workers indicated a main subject that aligned with the actual target problem being photographed, we considered as if the target had been selected. For example, when a worker indicated that the main subject was the "broken glass" that covered the schedule sign, we counted his response as if he had selected "schedule sign" (option 2) in the second question of the survey.

The responses to the second question were used to analyze target identification, the third to check if workers were paying attention to the task, and the fourth to examine the effect of incorporating blur into our image quality evaluation process. We used a standard procedure for evaluating perceptual image quality based on blur [Crete et al. 2007] and averaged blur ratings per image as a mean opinion score (MOS).

5. RESULTS

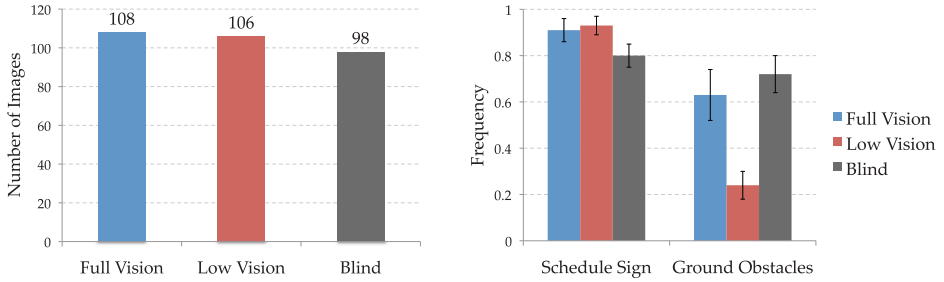
This section analyzes the data collected during the study and mostly presents new results on the content of the pictures captured by the participants. Statistical test assumptions (e.g., normality) were verified as part of our analysis. We encourage readers to refer to Vázquez and Steinfeld [2012] for more details on survey results and camera aiming statistics, and to Vázquez and Steinfeld [2011a] for a more complete evaluation of our ROI estimation approach.

5.1. Demographics

A total of 18 participants were recruited for the study. Participants were categorized into full vision (F), low vision (L), and blind (B) groups based on the information they provided when completing our demographics survey, and on how well they could see and read the screen of the iPhone used to take pictures. Only one participant self-categorized himself as being in a different group than that to which he was assigned. He said that he was low vision but did not mention using any vision aid and was able to read and write as well as the full vision participants. Therefore, we assigned him to the (F) group for the purposes of this study.

The average ages per group were 24, 56, and 55 years for the (F), (L), and (B) groups, with standard deviations of 6.7, 11.8, and 12.1, respectively. The percentage of women completing the experiment for (F), (L), and (B) was 50%, 50%, and 83%, respectively. Visually impaired participants reported using white canes (58%), guide dogs (25%), magnifiers on glasses (25%), tinted glasses (25%), handheld telescopes (17%), and other devices to get around. One blind participant reported that she wore hearing aids.

All visually impaired participants had a cell phone, and 66.7% of these devices had a camera. In the full vision group, six out of six participants said that they take photos, whereas three and one in the low vision and blind groups said that they take photos. In terms of device usage, 25% of the participants in the (L) and (B) groups said that they take pictures with a phone, whereas only 33% of the low vision participants reported



(a) The target was in the initial image (b) Frequency at which the proposed center was in the target

Fig. 8. (a) The number of times the desired target was in the view of the camera when the interactive aiming phase started. (b) The frequency at which the suggested center fell in the desired target when the interactive aiming phase started.

using a regular camera. Three totally blind participants said that they had never taken a picture before.

Only one participant in the fully sighted group said that he had filed a complaint about a transit problem, whereas five people in the (L) group and six in the (B) group reported having filed complaints. Phone calls were the common way of reporting problems for visually impaired participants.

5.2. Initial Aiming

There were 12 trials, 4% of 324, in which participants failed to roughly orient the camera in the right direction during the initial aiming phase. These are cases in which the ground obstacles or the schedule sign were not in the picture from which the ROI was estimated. Figure 8(a) shows the distribution of the remaining trials in which participants roughly aimed in the proper direction and captured at least part of the desired target when the interactive aiming phase started.

The initial image (I_i) was selected as the best image of the trial 25% of the time. Our application considered the initial aiming position provided by the user was good enough in these cases, or that any image processed afterward did not significantly improve image composition. We performed a restricted maximum likelihood (REML) analysis [Patterson and Thompson 1975; Stroup 2012] to evaluate the frequency (out of one) at which this happened on the effects of participant Group (full vision, low vision, and blind), image Target (ground obstacles and schedule sign), and Participant as random effect and nested by Group. This analysis resulted in a significant difference only for Target ($F[1, 106] = 7.54$, $p = 0.007$). The Student's post hoc t-test showed that the frequency at which I_i was selected as the best image was significantly higher for the sign ($M = 0.31$, $SE = 0.31$) than for the ground obstacles ($M = 0.19$, $SE = 0.23$). In other words, the initial aiming position provided by the participants when capturing the schedule sign tended to be better than the initial position provided when capturing the ground obstacles, according to our image composition model.

5.3. Suggested Center

We computed the frequency (over one) at which the suggested center was selected by our system inside the intended target. We considered the three pictures taken per participant Group, feedback Mode, and Location to compute this frequency. On average, our system successfully suggested a new center inside the target 70% of the time ($N = 108$, $SE = 0.04$), considering as false cases the 12 initial images for which participants failed to roughly aim the camera in the proper direction.

Table II. Distribution of the 113 Images in Which at Least One Mechanical Turk Worker Did Not Identify the Intended Target as the Main Subject of the Composition

	Silent	Tone	Speech	Total (per Group)
Full Vision	1 (1%)	9 (8%)	8 (7%)	18 (16%)
Low Vision	11 (10%)	9 (8%)	4 (3%)	24 (21%)
Blind	26 (23%)	26 (23%)	19 (17%)	71 (63%)
Total (per Mode)	38 (34%)	44 (39%)	31 (27%)	

We conducted a REML analysis for the frequency at which the suggested center fell in the target. Group and image Target were considered as fixed effects, whereas Participant was a random effect nested within Group. We found significant differences for Target ($F[1, 106] = 59.28$, $p < 0.001$) and the interaction between Group and Target ($F[2, 105] = 16.01$, $p < 0.001$). The Student's post hoc test for the former showed that the frequency at which the suggested center fell in the schedule sign ($M = 0.88$, $SE = 0.03$) was significantly higher than for the ground obstacles ($M = 0.53$, $SE = 0.06$). Figure 8(b) shows the average results per Group and Target.

The fact that our ROI estimation approach was not as successful with ground obstacles as with the schedule sign is strongly related to our reliance on visual saliency. The schedule sign pops out visually in its context more than the trash. Plus, participants tended to take pictures closer to the sign than the trash, especially those in the (F) group. Naturally, when more context is added to the picture, the less relevant a particular target becomes. As mentioned previously, object recognition could be added to our system to improve the suggestion of a new target to the user. However, this is out of the scope of the present work.

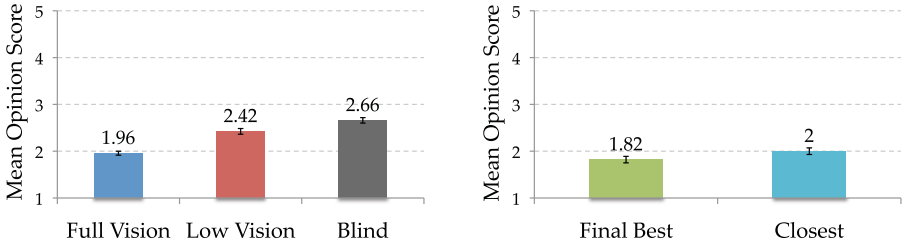
Although these objective results are an indication of the success of our approach at suggesting an appropriate image center, they do not tell the whole story. These results do not consider the overall appearance of the images. In some cases, for example, users may have taken a picture very close to the target and our system may have succeeded at suggesting a new center inside the schedule sign or the trash. However, the image may have been taken so close that it turned out to be difficult to understand. Likewise, in many cases, our system may have suggested a new center that was not contained in the target but still helped improve the composition. For this reason, we suggest considering these results alongside the following third-party image evaluation.

5.4. Image Subject Identification

Third-party observers were surveyed about the main subject of the compositions, as explained in Section 4.4. We found that all five workers who evaluated 87% of these pictures agreed that the main subject was the intended barrier captured by the photographer. The remaining 113 pictures did not have consensus across the workers and were mostly taken by blind participants. Table II details these results.

The lack of consensus for the (F) group with respect to the main subject seemed high considering that fully sighted participants had no trouble roughly aiming the camera in the proper direction. Further inspection of the data led us to realize that the results of Table II for the (F) group corresponded to pictures taken by a single full vision participant. This participant tended to take pictures far from the targets, as compared to the other participants, and these longer shots complicated our ROI estimation process. Furthermore, these shots also generated confusion between third-party raters with respect to the main subject of the pictures.

We computed the percentage of third-party evaluators who did identify the intended target as the main subject in the set of pictures without consensus. A REML analysis with participant Group, feedback Mode, image Type (I_i, I_c, I_b, and I_b^R), and intended Target as main effects, as well as Participant as random effect nested by Group, revealed



(a) MOS per Group (b) MOS for I_b (Final Best) and I_c (Closest)

Fig. 9. MOS for the images that were rated through Amazon's Mechanical Turk. A lower MOS is better.

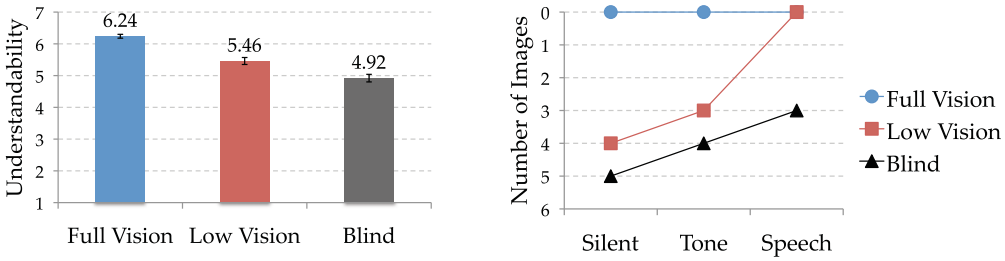
significant differences for Target ($F[1, 111] = 7.08$, $p = 0.009$). The average percentage of workers who identified the trash as the main subject ($M = 0.51$, $SE = 0.06$) was significantly higher than the average for the schedule sign ($M = 0.42$, $SE = 0.04$). No other effects or interactions were significant, but Group was close. On average, 63% of the evaluators identified the intended target as the subject in the pictures taken by fully sighted participants ($SE = 0.05$), 53% in the pictures taken by low vision participants ($SE = 0.07$), and 38% in the remaining cases ($SE = 0.04$).

5.5. Perceptible Image Quality Based on Blur

We checked for consistency between the five blur ratings collected per image through Mechanical Turk (Cronbach's alpha 0.89) and averaged these responses to create an MOS for blur.

We found a significant correlation between the MOS and the automatic blur metric of Crete et al. [2007] that we used in our system ($r(824) = 0.56$, $p < 0.001$). This correlation was computed considering all images for which our system estimated blur in real time and were evaluated by third-party observers.

To further examine the MOS obtained from Mechanical Turk, we conducted a REML analysis with participant Group, feedback Mode, intended Target, and image Type (I_i, I_c, I_b, and I_b^R) as main effects, along with Participant as random effect nested within Group. This analysis resulted in significant differences for Group ($F[2, 864] = 7.33$, $p = 0.006$), Target ($F[1, 865] = 5.55$, $p = 0.02$), and Type ($F[3, 863] = 23.20$, $p < 0.001$). The average MOS tended to be lower (better) for the fully sighted group, followed by that of the low vision group (Figure 9(a)). The Tukey-HSD post hoc showed that the difference in MOS between the (F) and (B) groups was significant in this respect. In addition, the pictures of the ground obstacles were significantly less blurred ($M = 2.2$, $SE = 0.05$) than those of the schedule sign ($M = 2.45$, $SE = 0.04$), and the images I_c ($M = 2.00$, $SE = 0.07$) and I_b ($M = 2.10$, $SE = 0.06$) were significantly less blurred than the initial images I_i ($M = 2.64$, $SE = 0.06$). The difference between I_i and I_b^R ($M = 2.34$, $SE = 0.07$) was not significant, suggesting that the rotation applied to some of the best pictures induced noticeable blur artifacts. Last, we also found significant interactions between Group and Mode ($F[4, 862] = 2.64$, $p = 0.03$), Group and Target ($F[2, 864] = 6.67$, $p = 0.001$), and Group and Type ($F[6, 860] = 6.29$, $p < 0.001$). In the first case, the post hoc showed that the pictures taken by blind participants with speech feedback were significantly more blurred than those taken by fully sighted users. In the second case, the pictures of the ground obstacles taken by full vision participants were significantly less blurred ($M = 1.69$, $SE = 0.05$) than the rest, except for the pictures of the schedule sign taken by low vision participants ($M = 2.41$, $SE = 0.08$). In the third case, the post hoc analysis showed that the initial pictures I_i taken by low



(a) Understandability for initial images

(b) I_i has ROI closest to image center

Fig. 10. Results for the initial images (I_i) for which the intended target was identified as the main subject by third-party raters. (a) Understandability for I_i per participant Group. (b) The number of images in which the initial image did not have the ROI centered in the composition but had the smallest distance between the ROI and the center compared to following frames.

vision participants were significantly more blurred than all other photos taken by this group.

There were cases in which the best image of a trial (I_b) did not have the minimum distance between the ROI and the middle of the composition, compared to all other images processed during the interactive aiming phase. This is because we evaluated image quality based on blur regardless of the distance between the ROI and the image center. To check if this extra consideration was useful, we performed an additional REML analysis of MOS on the effects of participant Group, image Target, feedback Mode, and image Type (I_c and I_b), with Participant as random effect nested within Group. The analysis excluded all cases where $I_c = I_b$—that is, where the best image was the picture that had the suggested center closest to the middle of the composition. We found significant differences for image Type ($F[1, 264] = 4.97$, $p = 0.027$). The average MOS for the selected best image I_b was significantly lower (better) than the one for I_c, as shown in Figure 9(b).

5.6. Image Understandability

We created an Understandability index by averaging the seven-point scale responses obtained for *I can easily identify the main subject of the picture* and *I am confused about what the photographer tried to capture because the picture is hard to understand* (reversed). Cronbach's alpha was 0.89, above the nominal 0.7 threshold for question reliability.

Understandability was used to compare the images in the set with subject consensus, where all Mechanical Turk raters agreed that the main subject was the intended target. Analyses were performed as follows.

5.6.1. Initial Images. Average Understandability ratings for the first images I_i in the group with consensus were positive in general ($M = 5.6$, $SE = 0.07$, $N = 274$). To check for differences between these images, we conducted a REML analysis on Understandability with Participant as random effect nested within Group, and with participant Group, image Target, feedback Mode, and Final Best Image as main effects. The latter corresponds to whether or not an image was selected as the final best (I_b) for its trial. Only significant differences per Group were found between initial images ($F[2, 271] = 13.09$, $p < 0.001$), as shown in Figure 10(a). The Tukey HSD post hoc showed that Understandability was significantly higher for the initial images in the (F) group compared to the (L) and (B) groups. Although the difference was not significant between the latter two, there was an upward trend in favor of the (L) group. These

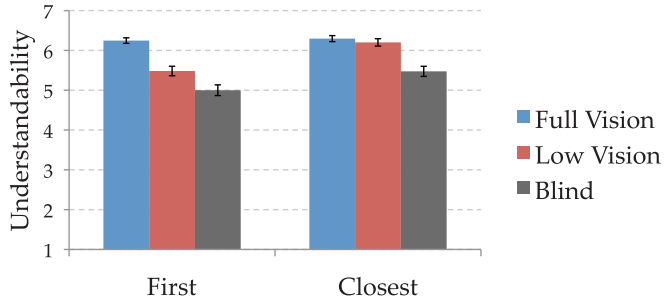


Fig. 11. Understandability for first images (I_i) and those with the suggested center closest to the middle of the composition (I_c). Third-party raters agreed that the main subject of these pictures was the target.

results align with camera-aiming statistics discussed in Vázquez and Steinfeld [2012], including the fact that the suggested center was significantly farther away from the middle of the first images captured by blind participants compared to those captured by the (F) and (L) groups. No other effects or interactions were significant in terms of Understandability.

Given the previous results, we decided to further analyze the 60 first images that were selected as the best (I_b) in the group with consensus. We found that 40% of this set were I_i pictures with the ROI centered in the middle of the composition. This percentage represents 24 trials in which the interactive aiming phase ended immediately. Of these trials, 16 belonged to the (F) group (67%), 6 to the (L) group (25%), and 2 to the (B) group (8%).

A subset of 19 initial images with subject consensus did not have the suggested center in the middle of the composition but had the smallest distance between these two points compared to all following frames in their trial. Most of these images were taken by the (B) group, whereas none were taken by participants in the (F) group, as shown in Figure 10(b). Similar results were observed for the average percentage of time that users increased the distance from the suggested center to the middle [Vázquez and Steinfeld 2012].

5.6.2. Centering the Suggested Center. There were 223 trials in which the the initial image (I_i) was different from the picture with the suggested center closest to the middle of the composition (I_c), and where both pictures had third-party consensus with respect to the main subject. We conducted a REML analysis to check for differences in Understandability between these images, considering participant Group, feedback Mode, image Target, and image Type (I_i and I_c) as main effects, as well as Participant as random effect nested within Group. We found significant differences for Group ($F[2, 443] = 9.14$, $p = 0.003$). Average Understandability ratings were significantly higher for images in the (F) group ($M = 6.27$, $SE = 0.05$, $N = 170$) compared to those in the (B) group ($M = 5.24$, $SE = 0.09$, $N = 116$). We also found significant differences for image Type ($F[1, 444] = 32.08$, $p < 0.001$), with average Understandability ratings of 5.64 ($SE = 0.07$) for I_i and 6.04 ($SE = 0.06$) for I_c. As depicted in Figure 11, the interaction between Group and Initial image was significant as well ($F[2, 443] = 7.73$, $p < 0.001$). Centering the suggested center brought Understandability ratings for the (B) group up to the levels of the (L) group without assistance, and those of the (L) group up to the levels of the (F) group. The interaction between Mode and Target was slightly significant ($F[2, 443] = 3.02$, $p = 0.049$), and a Tukey HSD post hoc did not reveal any pairwise significant differences.

5.6.3. Final Image Selection. To evaluate the effect of our image composition model, we compared third-party Understandability for the initial image processed (I_i) and the final best (I_b). We excluded those cases where $I_i = I_b$, and where at least one of these images did not have third-party subject consensus. A REML analysis with Participant as random effect nested within Group and participant Group, feedback Mode, image Target, and image Type (I_i or I_b) as main effects showed significant differences. The results for Group ($F[2, 415] = 10.51$, $p = 0.001$) were similar to previous findings: group (F) was significantly better than group (B) in terms of Understandability, and there was a trend in favor of group (L) in comparison to group (B). With respect to Type, $F[1, 416] = 35.99$ ($p < 0.001$), the best images had significantly higher average Understandability, $M = 6.06$ ($SE = 0.06$), than the initial images, $M = 5.61$ ($SE = 0.07$). The interaction between Group and Type was also significant ($F[2, 415] = 6.13$, $p = 0.002$). A Tukey HSD post hoc showed that blind participants reached the ratings of the low vision group after the interactive aiming phase. In a similar manner, low vision participants reached the Understandability ratings of the full vision group by using our system. These results are strongly aligned to those obtained for I_i versus I_c, as presented in Figure 11. In addition, the interaction between Group and Mode was marginally significant ($F[4, 413] = 2.47$, $p = 0.044$). The post hoc revealed that Understandability ratings were significantly lower when group (B) took pictures in silent and tone modes in comparison to group (F).

Since the differences between I_i and I_b were very similar to those obtained between I_i and I_c, we did another REML analysis on Understandability to compare I_b versus I_c. We considered those images with third-party subject consensus and discarded trials with $I_b = I_c$—that is, where the final best image had the ROI closest to the middle of the composition. We tested for differences on participant Group, feedback Mode, image Target, and image Type (I_b or I_c), with Participant as random effect nested within Group but only found significant differences for Group ($F[2, 207] = 8.51$, $p = 0.005$). No novelties were observed in this case.

In addition, we inspected the differences in Understandability between all final best images (I_b) with third-party subject consensus ($N = 222$). A REML analysis with Participant as random effect nested by Group and participant Group, feedback Mode, and image Target as main effects resulted in significant differences for Group ($F[2, 219] = 5.39$, $p = 0.02$). As before, the post hoc showed that Understandability was significantly higher for I_b taken by group (F) than by group (B). The effect of Target and the interaction between Group and Target were close to significant, with similar trends to those presented previously.

5.6.4. Camera Roll Compensation. The distribution of the rotated pictures I_b^R was unbalanced between targets. Only 11% of these pictures were taken when participants tried to photograph the ground obstacles, because users tended to tilt the phone downward in this case.

We conducted a REML analysis on Understandability for the final best images (I_b) and their rotated version (I_b^R), when our system adjusted for camera roll. We considered Participant as random effect nested by Group, as well as Group, feedback Mode and image Type (I_b and I_b^R) as main effects, but only Type resulted in a slight significant difference ($F[1, 202] = 4.19$, $p = 0.042$). The Student's post hoc showed that the final best images I_b had significantly higher Understandability ($M = 6.11$, $SE = 0.08$) than their rotated counterparts ($M = 5.93$, $SE = 0.08$).

We presented the rotated images on Mechanical Turk as in Figure 5(c) and believe that this had a slight negative effect on ratings, especially when camera roll was significant. It is possible that if we had cropped the rotated images to make them look vertical, like the rest, we may have obtained different results.

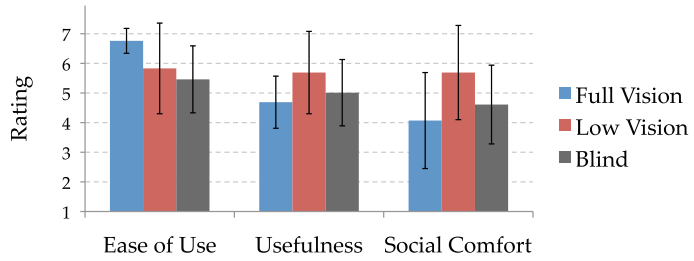


Fig. 12. Average ratings on Ease of Use, Usefulness, and Social Comfort per group.

5.7. Postcondition Ratings

We surveyed participants' opinion about our assisted photography application after they tried each of our feedback modes, as reported in Vázquez and Steinfeld [2012]. We summarize the survey data here for convenience. Postcondition survey responses were grouped into three categories: Ease of Use, Usefulness, and Social Comfort (Cronbach's alpha 0.849, 0.833, and 0.828, respectively). These responses 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. ANOVA analyses did not reveal any Ordering effects.

Ease of Use ratings for our application were positive in general, as depicted in Figure 12. There was a significant difference on Ease of Use between participant Groups ($F[3, 50] = 6.61$, $p = 0.003$). Full vision participants gave statistically significant higher ratings for Ease of Use with respect to the other groups. No other effects or interactions were significant for Group and Mode, but a slight upward trend was observed for speech.

There were significant differences in Group on Usefulness ($F[2, 51] = 3.57$, $p = 0.036$) and Social Comfort ($F[2, 51] = 5.67$, $p = 0.006$). The post hoc analyses revealed that full vision participants reported significantly reduced Usefulness and Social Comfort than low vision participants (Figure 12). Note that full vision participants still rated Usefulness and Social Comfort in the middle—not at the low end. Although the interaction between Group and Mode was not significant, we noticed a trend suggesting that Social Comfort is not affected by audio feedback for people with visual impairments.

5.8. Posttest Ratings

A full factorial ANOVA showed significant differences in Mode on posttest preference ratings ($F[2, 51] = 3.32$, $p = 0.045$). At the end of the study, speech mode ratings ($M = 4.9$, $SE = 0.5$) were significantly higher than silent mode ($M = 3.6$, $SE = 0.5$). Although differences in preference for Group were not significant, there were differences in the interaction between Group and Mode ($F[4, 50] = 13.85$, $p < 0.001$). Visually impaired participants preferred audio feedback over silent mode, whereas participants in the full vision group did the contrary (Figure 13).

5.9. Other Findings

Although 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), then it is natural to translate the device upward to bring the center to the middle of the picture. However, if the phone

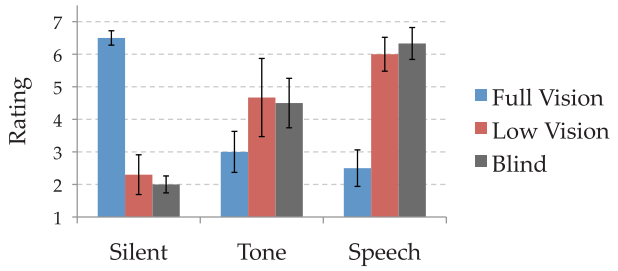


Fig. 13. Posttest preference ratings by feedback Mode and user Group.

is aimed downward, such as toward the ground, then the user should move the phone forward to frame the suggested center in the middle of the composition. This dichotomy was a problem for several blind participants when aiming downward, who ended up translating the phone upward rather than forward. It was hard for these participants 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, saying that she would only take pictures if there was a way that she could feel images (e.g., feel the shape of buildings and big spaces). All other visually impaired users indicated that they like, or would like, to take pictures of events, people, and objects.

Although speech feedback was generally preferred by the visually impaired community, some participants pointed out that this might change with extended use of our system. In particular, one blind participant said, "Now that I am not experienced, I prefer voice (speech mode), but once I learn, I might prefer the tone." He then compared the learning curve of our system with the Bop It! game by Hasbro Inc., implying that it is difficult to follow the instructions when starting to play, but after a while, the player becomes an expert and the game is easier to follow.

One 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. He kept repeating to himself, "What do you think tells the best story?" Throughout the experiment, he became excited with the system because it was suggesting centers close to the middle. In other words, the application tended to judge his initial camera aiming position as appropriate.

Multiple visually impaired participants used the application to take a picture of their guide dog, as shown in Figure 14, 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.

6. DISCUSSION

In general, our assisted photography application helped users take better pictures in real time. Third-party Understandability ratings for the final pictures taken were significantly better than for the first images processed by the system. This suggests that our approach is better than a traditional camera. Although our ROI selection process failed to suggest a new image center inside the desired target in some cases, average Understandability ratings were not reduced with the use of our system.

The Understandability results showed that our system helps blind participants achieve the level observed for the initial images in the low vision group. Likewise, our system elevates low vision participants to the level of unassisted, full vision participants. It is our hope that this boost in image quality with the use of our system enables

Fig. 14. Other pictures taken by the participants with our assisted photography application.

the development of other assisted photography applications that can positively impact the visually impaired community. Computer vision performance is heavily influenced by image quality, and this boost may permit previously infeasible applications.

We did not find significant differences between silent, tone, and speech feedback in terms of Understandability for the best images captured with our application. However, in general, audio feedback helped steer users toward centering the suggested center in the pictures more efficiently.

Participants had different preferences with respect to feedback modes. In particular, we observed trends in favor of speech mode for the visually impaired community. Subjective opinions on Ease of Use and Usefulness showed that orientation information, provided only by speech mode, seemed to help users center the ROI more easily. These results were supported by objective data, such as aiming time and how often the ROI was centered.

6.1. Possible System Improvements

Although speech mode tended to increase the performance of visually impaired users, we also noticed that our selection of spoken sounds were confusing in some situations. We used "up," "down," "left," and "right" to indicate how the device should be moved to center the ROI based on the location of the suggested center in the image. Instead, we should have provided these instructions based on how the device was held by the user.

Our assisted photography application did not know when participants were roughly aiming the camera in the proper direction or not, because it did not know the intended target. Even if it had known what participants were going to capture, it did not have object recognition capabilities to identify the target. Thus, the system could be improved by adding a few models of typical objects that users are expected to photograph. When the application starts, it could ask users what the target is and then decide if object recognition is appropriate for suggesting a new center based on the models in memory. If the system does not have a model for the target, then it could fall back to image saliency, as currently implemented, for estimating the ROI.

We believe that there is potential in building image mosaics from the best picture selected by our system and other images processed during the interactive aiming phase. Mosaics could add significant contextual information to the selected best image, which

may be useful in a documentation context. This feature, however, would require extension of the final image display phase, because composing the mosaic would likely take some time given computational limitations on phones. We would also need to address "ghosts" (i.e., elements that do not appear in the same position in all pictures used for the mosaic) and manage holes in the composition. The latter are parts of the mosaic that are empty, because no image covers these spatial regions.

Another way of making our system better would be to continue processing a few additional frames when the ROI is centered in the middle of the composition. This would allow the system to pick as best image a picture that was captured after the ROI was centered, not only before. The hope is that one of these extra images would have a similar composition to the one where the ROI was centered but would be significantly less blurred.

7. CONCLUSION

We presented an assisted photography framework aimed at helping people with visual impairments take pictures and described our implementation in the context of documenting accessibility barriers related to public transportation. Our results in this context reinforce earlier work suggesting that users who are blind or have low vision find assisted photography appealing and useful. Furthermore, it appears that there is overall acceptance of assisted photography, including users with full vision, due to positive usefulness ratings collected during our study.

Full vision participants seemed to find value in silent feedback mode, thereby suggesting that our assisted photography framework has universal appeal. However, it is clear that the interface of such a system may need to change when the user is blind or has low vision. The iOS5 camera application's altered behavior when VoiceOver is turned on is a good example of how this can be achieved.

Since our evaluation of the framework was focused on our particular implementation for the documentation scenario, more systematic evaluations are needed to verify its effectiveness in other settings. For example, the framework could be used to help users capture better images for optical character recognition or object labeling, which benefit from good image compositions. These systems would naturally require appropriate image quality evaluation mechanisms and corresponding user feedback modes.

We foresee multiple opportunities for assisted photography to become mainstream, especially as computer vision improves and processing power becomes cheaper and more accessible. In our opinion, the key to success with these systems is providing real-time feedback to users so that they can take better pictures in situ. It is also essential to acknowledge that users may have different preferences for how they want the system to behave. Therefore, it is important to provide a variety of options and features that users can decide to use when taking a picture based on their current needs. We believe that our proposed framework supports this general class of assisted photography systems.

ELECTRONIC APPENDIX

The electronic appendix for this article can be accessed in the ACM Digital Library.

REFERENCES

Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. 2009. Frequency-tuned salient region detection. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*. IEEE, Los Alamitos, CA, 1597–1604.

Shai Avidan and Ariel Shamir. 2007. Seam carving for content-aware image resizing. *ACM Transactions on Graphics* 26, 3, Article No. 10. DOI:http://dx.doi.org/10.1145/1276377.1276390

Soonmin Bae, Aseem Agarwala, and Frédo Durand. 2010. Computational rephotography. *ACM Transactions on Graphics* 29, 3, 1–15. DOI: http://dx.doi.org/10.1145/1805964.1805968

Simon Baker and Iain Matthews. 2004. Lucas-Kanade 20 years on: A unifying framework. *International Journal of Computer Vision* 56, 3, 221–255. DOI: http://dx.doi.org/10.1023/B:VISI.0000011205.11775.fd

Serene Banerjee and Brian L. Evans. 2007. In-camera automation of photographic composition rules. *IEEE Transactions on Image Processing* 16, 7, 1807–1820.

Scott Beach, Richard Schulz, Julie Downs, Judith Matthews, Bruce Barron, and Katherine Seelman. 2009. Disability, age, and informational privacy attitudes in quality of life technology applications: Results from a national Web survey. *ACM Transactions on Accessible Computing* 2, 1, 5.

Nahom Beyene, Rory Cooper, and Aaron Steinfeld. 2009. Driving status and the inner drive for community mobility and participation: A survey of people with disabilities and senior citizens from support groups in New Delhi, India. In *Proceedings of the Rehabilitation Engineering and Assistive Technology Society of North America (RESNA) Conference.*

Peng Bian and Liming Zhang. 2008. Biological plausibility of spectral domain approach for spatiotemporal visual saliency. In *Proceedings of the 15th International Conference on Advances in Neuro-Information Processing (ICONIP'08),* Springer-Verlag, Berlin, Heidelberg, 251–258.

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

Jeffrey P. Bigham, Chandrika Jayant, Andrew Miller, Brandyn White, and Tom Yeh. 2010b. VizWiz:: LocateIt—enabling blind people to locate objects in their environment. In *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).* IEEE, Los Alamitos, CA, 65–72.

Klaus Bohn. 2006. *50 Principles of Composition in Photography: A Practical Guide to Seeing Photographically through the Eyes of a Master Photographer.* CCB Publishing.

Ralph Brunner, Nikhil Bhogal, and James David Batson. 2012. Image Capturing Device Having Continuous Image Capture. Patent Publication No. US8289400 B2: Filed June 5, 2009, Published Oct. 16, 2012. Retrieved September 14, 2014, from http://www.google.com/patents/US8289400.

Michele A. Burton, Erin Brady, Robin Brewer, Callie Neylan, Jeffrey P. Bigham, and Amy Hurst. 2012. Crowdsourcing subjective fashion advice using VizWiz: Challenges and opportunities. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'12).* ACM, New York, NY, 135–142.

Liqun Chen, Xing Xie, Xin Fan, Wei-Ying Ma, Hong-Jiang Zhang, and Heqin Zhou. 2002. *A Visual Attention Model for Adapting Images on Small Displays.* Technical Report MSR-TR-2002-125. Microsoft Research.

Frederique Crete, Thierry Dolmiere, Patricia Ladret, and Marina Nicolas. 2007. The blur effect: Perception and estimation with a new no-reference perceptual blur metric. In *Proceedings of SPIE,* Vol. 6492. Human Vision and Electronic Imaging XII. 64920I–64920I–11. DOI: http://dx.doi.org/10.1117/12.702790

Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. 2006. Studying aesthetics in photographic images using a computational approach. In *Computer Vision—ECCV 2006.* Lecture Notes in Computer Science, Vol. 3953. Springer, 288–301.

David Präkel. 2006. *Basics Photography 01: Composition.* AVA Publishing.

Mark Desnoyer and David Wettergreen. 2010. Aesthetic image classification for autonomous agents. In *Proceedings of the 20th International Conference on Pattern Recognition (ICPR'10).* IEEE, Los Alamitos, CA, 3452–3455.

Benoit Deville, Guido Bologna, Michel Vinckenbosch, and Thierry Pun. 2008. Guiding the focus of attention of blind people with visual saliency. In *Proceedings of the Workshop on Computer Vision Applications for the Visually Impaired (CVAVI'08).* 1–13.

Michael Dixon, Cindy M. Grimm, and William D. Smart. 2003. *Picture Composition for a Robot Photographer.* Technical Report WUCSE-2003-52. Washington University in St. Louis, St. Louis, MO.

Simone Frintrop, Erich Rome, and Henrik I. Christensen. 2010. Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception* 7, 1, 1–39. DOI: http://dx.doi.org/10.1145/1658349.1658355

Bruce Gooch, Erik Reinhard, Chris Moulding, and Peter Shirley. 2001. Artistic composition for image creation. In *Proceedings of the 12th Eurographics Workshop on Rendering Techniques.* Springer-Verlag, Berlin, Heidelberg, 83–88.

Chenlei Guo, Qi Ma, and Liming Zhang. 2008. Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform. In *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08).* IEEE, Los Alamitos, CA, 1–8.

Xiaodi Hou and Liqing Zhang. 2007. Saliency detection: A spectral residual approach. In *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07)*. IEEE, Los Alamitos, CA, 1–8.

Andreas Hub, Joachim Diepstraten, and Thomas Ertl. 2004. Design and development of an indoor navigation and object identification system for the blind. In *Proceedings of the 6th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'04)*. ACM, New York, NY, 147–152. DOI:http://dx.doi.org/10.1145/1028630.1028657

Laurent Itti and Christof Koch. 2001. Computational modelling of visual attention. *Nature Reviews Neuroscience* 2, 3, 194–203.

Bryan James, Andrew Hodge, and Aram Lindahl. 2010. Camera Image Selection Based on Detected Device Movement. Patent Publication No. US20100309334 A1: Filed June 5, 2009, Published Dec. 9, 2010. Retrieved September 14, 2014, from http://www.google.com/patents/US20100309334.

Chandrika Jayant. 2010. MobileAccessibility: Camera focalization for blind and low-vision users on the go. *SIGACCESS Accessible Computing* 96, 37–40.

Chandrika Jayant, Hanjie Ji, Samuel White, and Jeffrey P. Bigham. 2011. Supporting blind photography. In *Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'11)*. ACM, New York, NY, 203–210.

Julia Schwarz. 2011. Headshot. Retrieved September 14, 2014, from http://juliaschwarz.net/appsandutilities/2012/05/04/headshot/.

Myung-Jin Kim, Tae-Hoon Song, Seung-Hun Jin, Soon Mook Jung, Gi-Hoon Go, Key-Ho Kwon, and Jae-Wook Jeon. 2010. Automatically available photographer robot for controlling composition and taking pictures. In *Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'10)*. IEEE, Los Alamitos, CA, 6010–6015.

Aliasgar Kutiyanawala, Vladimir Kulyukin, and John Nicholson. 2011. Teleassistance in accessible shopping for the blind. In *Proceedings of the 2011 International Conference on Internet Computing*. 18–21.

Xu Liu. 2008. A camera phone based currency reader for the visually impaired. In *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'08)*. ACM, New York, NY, 305–306.

Jiebo Luo, Amit Singhal, and Andreas Savakis. 2003. Efficient mobile imaging using emphasis image selection. In Proceedings of the PICS Conference. Society for Imaging Science and Technology, Springfield, VA, 355–359.

H. D. Patterson and R. Thompson. 1975. Maximum likelihood estimation of components of variance. In *Proceedings of the 8th International Biometric Conference*. 199–207.

Aaron Steinfeld, Rafae Dar Aziz, Lauren Von Dehsen, Sun Young Park, Jordana L. Maisel, and Edward Steinfeld. 2010a. Modality preference for rider reports on transit accessibility problems. In *Proceedings of the TRB 2010 Annual Meeting*. Transportation Research Board, Washington, DC.

Aaron Steinfeld, Rafae Dar Aziz, Lauren Von Dehsen, Sun Young Park, Jordana L. Maisel, and Edward Steinfeld. 2010b. The value and acceptance of citizen science to promote transit accessibility. *Technology and Disability* 22, 1–2, 73–81.

Walter W. Stroup. 2012. *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. Taylor & Francis.

Bongwon Suh, Haibin Ling, Benjamin B. Bederson, and David W. Jacobs. 2003. Automatic thumbnail cropping and its effectiveness. In *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology (UIST'03)*. ACM, New York, NY, 95–104. DOI:http://dx.doi.org/10.1145/964696.964707

Hachon Sung, Guntae Bae, Sunyoung Cho, and Hyeran Byun. 2012. Interactive optimization of photo composition with Gaussian mixture model on mobile platform. *Optical Engineering* 51, 1, 017001. DOI:http://dx.doi.org/10.1117/1.OE.51.1.017001

Ender Tekin and James M. Coughlan. 2010. A mobile phone application enabling visually impaired users to find and read product barcodes. In *Proceedings of the 12th International Conference on Computers Helping People with Special Needs (ICCHP'10)*. Springer-Verlag, Berlin, Heidelberg, 290–295.

Marynel Vázquez and Aaron Steinfeld. 2011a. An assisted photography method for street scenes. In *Proceedings of the 2011 IEEE Workshop on Applications of Computer Vision (WACV'11)*. IEEE, Los Alamitos, CA, 89–94.

Marynel Vázquez and Aaron Steinfeld. 2011b. Facilitating photographic documentation of accessibility in street scenes. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems (CHI EA'11)*. ACM, New York, NY, 1711–1716.

Marynel Vázquez and Aaron Steinfeld. 2012. Helping visually impaired users properly aim a camera. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility(ASSETS'12)*. ACM, New York, NY, 95–102.

Paul Viola and Michael Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'01)*, Vol. 1. IEEE, Los Alamitos, CA, I-511–I-518.

Dirk Walther and Christof Koch. 2006. Modeling attention to salient proto-objects. *Neural Networks* 19, 9, 1395–1407.

Samuel White, Hanjie Ji, and Jeffrey P. Bigham. 2010. EasySnap: Real-time audio feedback for blind photography. In *Adjunct Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology (UIST'10)*. ACM, New York, NY, 409–410.

Received June 2013; revised July 2014; accepted July 2014