Title

Towards Mobile OCR: How To Take a Good Picture of a Document Without Sight

Authors

Cutter, Michael
Manduchi, Roberto

Publication Date

2015-09-01

Peer reviewed

Towards Mobile OCR: How To Take a Good Picture of a Document Without Sight

Michael Cutter
University of California, Santa Cruz
mcutter@soe.ucsc.edu

Roberto Manduchi
University of California, Santa Cruz
manduchi@soe.ucsc.edu

ABSTRACT

The advent of mobile OCR (optical character recognition) applications on regular smartphones holds great promise for enabling blind people to access printed information. Unfortunately, these systems suffer from a problem: in order for OCR output to be meaningful, a well-framed image of the document needs to be taken, something that is difficult to do without sight. This contribution presents an experimental investigation of how blind people position and orient a camera phone while acquiring document images. We developed experimental software to investigate if verbal guidance aids in the acquisition of OCR-readable images without sight. We report on our participant's feedback and performance before and after assistance from our software.

Categories and Subject Descriptors

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

General Terms

Design, Experimentation, Human Factors

Keywords

Visual Impairment, Optical Character Recognition, Document Processing

1. INTRODUCTION

There is increasing interest in mobile applications that can allow a blind person to access printed information such as restaurant menus, bills, signs on a door, etc. The ever increasing computational power of modern smartphones, combined with high quality on-board cameras, is enabling the development of OCR-based, low-cost applications that have great potential for benefiting the blind community. The fact that these software systems run on mainstream platforms (Android and iOS), rather than on customized devices, is an important bonus, since the latter are often expensive, lack support, and, like many assistive technology tools, are sometimes not well accepted due to the associated "stigma". However, as utilities for consumer mobile devices such as the Voice Over feature on iPhones proliferate, we can expect more wide spread adoption for accessibility applications.

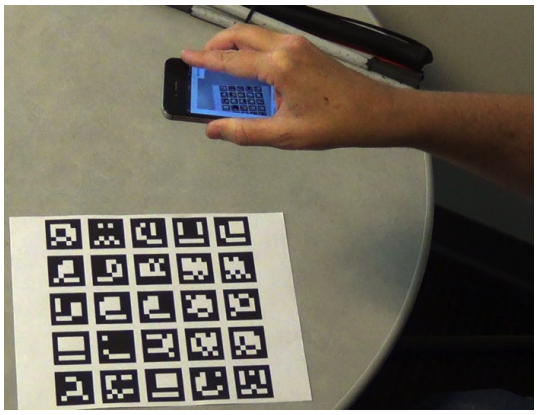

Figure 1: A participant positioning an iPhone over a document printed with ArUco fiducials.

Unfortunately, even the best OCR algorithm fails if the text in the image is cropped, the image has low resolution, is blurred, or badly lit. For sighted users, this is not a problem: one just needs to look at the scene through the screen, moving and orienting the phone until the desired text document is correctly framed and exposed before taking a shot. Even the best systems on the market can only provide post-facto indirect confirmation that a picture was actually readable – normally after producing garbled or incomplete text. The latency associated with OCR processing only makes things worse: one may have to wait for several minutes just to find out that the image was not OCR-readable and another snapshot of the document needs to be taken.

This contribution presents an experimental study with 12 blind participants. We first investigate how they hold and position the camera during image acquisition. Then they use our software that provides feedback while he or she tries to take an OCR-readable picture of a document. This experimental tool offers two modalities of usage. The first modality provides real-time confirmation when the user has reached a *compliant pose*, that is, when he or she has moved the camera to a position from which an OCR-readable image of the whole document can be taken. The second modality utters spoken directions to the user about where to move the phone next in order to increase the likelihood of reaching a compliant pose. After trying out our software we measure how they hold the camera again without assistance from our system.

This experimental study addresses two important research issues, concerning: (1) the ability of blind people to correctly position a smartphone in order to obtain an OCR-readable picture of a document; (2) the potential for increasing the success of this task by means of system-generated feedback. These results will hopefully inspire more research into mechanisms that could enable more efficient use of mobile OCR applications, and thus allow better access to printed information for blind users.

2. RELATED WORK

Document scanners coupled with OCR and text-to-speech have been used successfully by many blind people to access printed text [19]. Kane et al. [9] developed an augmented reality digital desk assistive environment which allows blind people to interact with complex paper documents. Their acquisition technology is a mounted desktop camera, which captures a live stream of images. The largest contour in the image is assumed to be the document and processed by optical character recognition. One of their user interface contributions is an "edge menu", inspired by the author's previous work [10]. The edge menu displays an alphabetical list of detected words. By clicking on an word on the list translational guidance is spoken to the image coordinate where the word was detected.

In recent years, a number of mobile OCR applications have been introduced to the market, to enable quick text access "on the go". The KNFB Mobile reader [2] and Blindsight's Text Detective [3] iPhone app are perhaps the best known such systems. The KNFB reader, which runs on the Nokia N82 phone, generates an optional "field of view report" via synthetic speech a few seconds after a picture has been taken of a document. This report contains information about the angle of the camera relative to the page and about whether all corners are visible or some text is cut off. By carefully holding the phone in position after the first picture has been taken, the user may be able to re-position the camera, if needed, so as to take a better framed picture. In practice, after taking a snapshot with KNFB, one has to wait for OCR to be completed before realizing that the shot was not compliant. Since multiple shots are normally needed, the whole process may be intolerably slow (possibly several minutes). However, KNFB just released an iOS version of their application which might remedy some of the latency issues. Unlike the KNFB Reader, Text Detective lets the user move the phone over the document, processing images continuously as they are taken by the phone's camera. As soon as an image is found containing text-like patterns, the phone vibrates briefly and the OCR process (which takes a few seconds) is started. This "opportunistic" approach is made possible by a fast text detection algorithm that is used to select promising images to be passed to the more computationally intensive OCR. However, their text spotter does not measure compliance: it will take a picture as soon as some text-like pattern is seen, possibly resulting in truncated lines etc. Only after OCR processing will the user find out that the shot was not compliant and that the hovering operation needs to be restarted. Often multiple hovering-OCR iterations are necessary, resulting in a long acquisition time. A similar opportunistic strategy is taken by an iPhone app named Prizmo [1], which processes each input image to find the edges of a rectangular document.

None of these smartphone OCR applications ensure that a blind user will be able to take a well-framed image of the document. In order to help a person take a good picture of a document, the use of mechanical stands has been proposed (e.g. the Optical Scan Stand tool that is available for the Galaxy Core Advance handset). These devices may be very useful for fixed-size documents, but do not allow the user to reduce or increase the distance to the document, which is often necessary to account for small font size or large document size.

Shilkrot et al. [11] created a wearable device to support reading on the go. Their system is worn on the finger and like Text Detective reads small blocks of text. They explored continuous tone and haptic feedback to alert the user that they have reached the end of a textblock; or have veered too far from the textline. There is also a commercial product worn as glasses called OrCam [17] that provides real time OCR by users pointing their head and finger at the block of text they wish to be read. However, neither of these approaches ensure that the user has captured the entire document and both require that the person buy a dedicated piece of technology.

The difficulty of taking good pictures without sight represents a hurdle not only for mobile OCR, but also for other applications of camera-based information access, as well as for recreational photography. For example, Bigham et al. [18] used simple computer vision techniques along with crowdsourcing to help a blind user point the camera correctly to an object (for example, to better identify it or to get closer to it). Brady et al [6] analyzed the type of objects blind people take photos of in a crowd sourcing answer seeking scenario. Their analyis also includes photo quality assessment. They found that 46% of the questions asked by their recent power users regarded 'reading'. The use of remote sight operators, who can look at the image taken by a blind person and provide advice on how to orient the camera to take a better picture, was also considered by Kutiyanawala et al. [21] in a tele-assistance system for shopping. TapTapSee [15] is another popular app that uses crowdsourcing for text reading from an image taken by an iPhone. Zhong et al. [14] developed a key-frame selection algorithm and combined it with a cloud based visual search engine to help blind people identify objects continuously.

EasySnap and the next iteration, PortraitFramer, are mobile applications developed by Jayant et al. [8], that give feedback to a blind photographer about the scene light, or about the presence and location in the picture of an object or of a person. The use of real-time feedback to help a blind person photo document transit accessibility was also studied by Vazquez and Steinfeld [12]. In this scenario, there is no clearly defined "target" (e.g., a face) that could be used to guide framing. Instead, a general–purpose saliency map is used to select a region of interest. A camera-based system for barcode access, equipped with a guidance mechanism that suggests how to move the camera in order to precisely center a detected barcode, was developed by Tekin and Coughlan [22]. The process of taking a precisely framed picture of a document for OCR processing could potentially be facilitated by stitching together multiple pictures, each containing a partial view of the document, into a panoramic image (or mosaic) of the whole document, as suggested by Zandifar et al. [13].

Experiments with sighted, blindfolded participants using a system similar to the one discussed in this contribution were conducted in a study by Cutter and Manduchi [7]. This study used a naive guidance algorithm, and included experiments that were meant only to validate the feasibility of such an approach. In fact, sighted people are likely to develop, through daily experience with vision-mediated camera handling, mechanisms and skills that are very different from those available to blind people, and thus cannot, even when blindfolded, be considered representative of blind users for the tasks considered in the experiments. With respect to the pre-

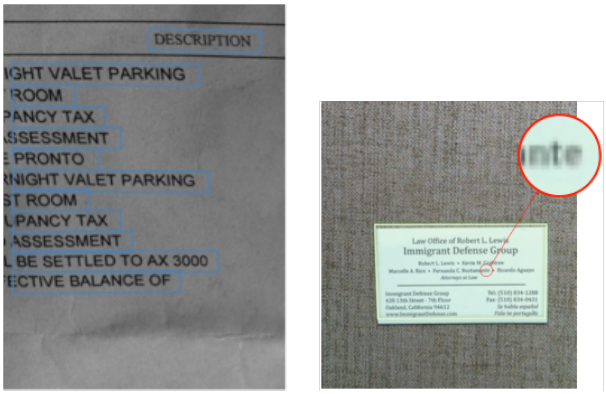

Figure 2: Non-OCR-compliant images (detail). Left: Text lines were successfully identified by the TextDetective app (blue rectangles), but parts of the lines are not visible. Right: The business card was correctly framed (yellow rectangle) by the Prizmo app, but the resolution is too low for OCR (see zoomed-in inset).

liminary study in [7], we re-designed the guidance algorithm, the experiments, and the evaluation criteria, and only considered blind participants.

3. METHOD

3.1 Overview and Rationale

Our goal in this work was to shed light on the process by which a blind person can operate a hand-held camera (embedded in a smartphone) to access text data printed on a document. We assume that the user can rely on OCR software capable of decoding printed text provided that: (1) the entirety of the page is visible, and (2) text is imaged with a certain minimum size. Furthermore, we assume that the OCR software is able to decode text at any orientation and even with noticeable perspective distortion (due e.g. to camera slant), as these factors can be corrected by proper image processing. In these conditions, text access can be obtained as long as the user is able to take a proper (*compliant*) picture of the document, in a way that is precisely formalized in a later section.

The main questions driving our investigation are:

1. *How difficult is it to take a compliant picture of a document without sight?* To the best of our knowledge, there are no published studies about the ability of blind people to maneuver a camera in order to take a readable picture of a document. Our research seeks to establish a baseline against which any proposed assistive technology for mobile OCR can be compared.

2. *Could this process be facilitated by proper system-generated feedback?* We considered two different approaches to provide feedback to the user. In the first approach, the system continuously takes images (frames) and analyzes each image to verify whether the imaged document is readable; as soon as a compliant (readable) image is taken, the user is notified and the process is stopped. In the second approach, the system additionally provides instructions to the user about where to move the phone to increase the likelihood of a compliant picture being taken.

To address these questions, we developed the necessary experimental software tools and designed experiments. We decided to emulate an "ideal" OCR software and feedback mechanism by means of an image processing system based on augmented reality (AR) markers. Rather than dealing with a regular printed document, our participants interacted with a sheet of paper on which a number of AR markers (*fiducials*) were printed. Based on the image taken by the iPhone camera of these fiducials, the system quickly and robustly identifies its own position and orientation (collectively called *pose*) with respect to the document. This information is sufficient to establish whether the image of a "real" document of known size taken from the same camera pose would be OCR-readable (i.e., the pose is compliant), and to provide feedback and guidance to the user. This almost-Wizard-of-OZ mechanism allows us to abstract from the actual OCR software employed and to concentrate on the user interaction component of the system, under the assumption of an "ideal" image processing software. Using this tool, we can ascertain whether feedback mechanisms have potential for improving the user experience with mobile OCR without sight, which would justify further research in this direction; additionally, this system allows us to investigate the most promising strategies to present feedback to the user.

3.2 Population

Twelve blind participants (four females and eight males) were recruited through announcements on newsletters and word of mouth. All but one participant had at most some residual light perception. The participant who had some residual vision left had acuity of 20/3800 in one eye; the other eye had no vision (prosthetic). In order to remove any possibility that the little residual vision could bias results, this participant was blindfolded during the test. The participants were of age between 18 and 65, with a median age of 53. Of these participants, two were congenitally blind, two became blind at age three, and all others lost their sight after the age of ten. Two of the participants had lost their sight less than five years prior to the experiment. Seven participants were regular iPhone users, and four participants had tried mobile OCR systems before (but were not regular users of this technology).

3.3 The Compliant Pose Space of a Document

A *compliant* picture of a document is a picture that contains all of the text in the document, at enough resolution that it can be read by OCR. More precisely, a picture of a letter-sized (8.5" by 11") document is considered compliant for the sake of this study if: (1) all four corners of the printable area are visible, where in our case the printable area has top and bottom margins of 1.5" and left and right margins of 0.5"; and (2) a small letter placed anywhere in the printable area is seen in the picture at enough resolution that it can be read accurately by OCR. A "small letter" could be, for example, a lowercase 'x' character typed in 12 point Arial font, which has height of 4.23 mm. By "accurately readable by OCR" , we mean that the height of the letter in the image should be of at least 12 pixels [13]. This is based on the readability constraint discussed in [7] calculated at 8MP photo resolution of the iPhone. Thus, a compliant image of a document is such that the whole content can be read via OCR. Note that we define compliance only in geometric terms: factors such as bad illumination or blur certainly contribute to the quality of OCR reading, but are neither considered in this definition nor in this study.

We define *compliant pose* as the pose (3-D location + camera orientation, with respect to a reference system fixed with the document) of a camera that takes a compliant picture. Note that the compliance

of a pose depends on the camera's optical/imaging characteristics (intrinsic parameters [20]). For example, a pose that is compliant using a wide field-of-view lens may be non-compliant using a longer lens (because the document may not be seen in its entirety in the second case). Likewise, a compliant pose for a narrow field-of-view lens may be non-compliant for a shorter lens due to reduced angular resolution.

For a given camera, the set of all compliant poses form the *compliant pose space*. The compliant pose space of a document can be computed based on geometry. In addition, given a non-compliant pose, one could predict whether moving the camera in a certain direction and rotating it around a certain axis will result in a compliant pose. This information may be used in a guidance mechanism to provide hints to the user about how to move the camera in order to take an OCR-readable image. Of course, this assumes that the camera pose can be somehow computed – a difficult problem in itself. Several techniques are available for image-based pose estimation, ranging from stereo triangulation (when a system with two cameras is available), to structure from motion/SLAM, to methods that use fiducials printed on the page at known locations.

In our study, we used printed fiducials for camera pose estimation. In fact, in our experiments we give away completely with textual information, and use a document containing solely well-calibrated fiducials instead (see Fig. 1). This approach is justified by the fact that the goal of this investigation is to study the mechanisms that can facilitate reaching a compliant pose and thus obtaining an OCR-readable image of the document. In this way, we are able to separate the *technical* difficulties of pose estimation from the *human factors* that pertain to holding a camera and taking a compliant picture.

3.4 Interaction Modalities

We considered three different interaction modalities in our study. Each modality represents a mechanism by which the user may try to take a compliant picture of a document using a smartphone. The three considered modalities are described below.

3.4.1 Snapshot

In the *snapshot* modality, the user simply takes a snapshot of the document (e.g. by pressing a button or tapping the screen), from a position and orientation that, in his or her judgment, results in a compliant picture. No feedback is provided by the system, except to confirm (via synthetic speech) that a snapshot action was registered.

3.4.2 Hovering: Just Confirmation

In this case, the user moves the camera over the document ("hovering") while the system takes and processes pictures continuously. As soon as a compliant picture is detected, the system notifies the user and the process is stopped. The user is not required to take any action (such as pressing a button) besides moving the camera around the position that he or she expects to be the most appropriate for a compliant picture.

3.4.3 Hovering: Guidance

This represents a more interactive version of the "hovering" modality. The system continuously takes and processes pictures, and in addition produces hints (in the form of short synthetic speech sentences) advising the user about where to move the camera next in order to increase the likelihood of reaching a compliant pose.

3.5 Apparatus

3.5.1 Pose Estimation

The application developed for this experiment runs on an iPhone 4S (with a 4:3 aspect ratio and video resolution of 640x480). To compute the camera pose from a picture of the printed fiducials, we use the ArUco [4] Augmented Reality library, implemented with OpenCV [5]. A letter-size sheet is printed with ArUco's fiducial patterns in known locations (see Fig. 1). The software detects the location of the fiducials in the camera's field of view and computes the pose of the camera (previously calibrated off-line). Only a single fiducial is necessary for pose estimation, but accuracy is increased when multiple fiducially are seen. The software is able to process 20 images per second on average, although in practice the effective frame rate is smaller due to other concurrent processing on the phone. Given the camera's pose (computed with respect to a reference system centered at the paper sheet), one can obtain the homography (perspective transformation [20]) that maps points in the paper sheet to pixels. This information is used to compute compliance of the current pose, based on the criteria discussed above (visibility of all corners of the document's printable area, minimum resolution). Note that pose compliance detection (along with proper user confirmation) is all that is needed for the *hovering: just confirmation* modality. The *guidance* modality requires further processing and a more complex user interface, as explained below.

3.5.2 Guidance

The goal of the *guidance* mechanism is to give clear instructions as to where to move the camera to reach a compliant pose. This algorithm produces a *correction vector* that takes the camera to a compliant pose if the same orientation is maintained. The correction vector links the current camera position with the closest point in the *compliant segment* (see Fig. 3), which is the set of points on a line through the center of the sheet, parallel to the optical axis of the camera, such that each point in the segment is a compliant camera location under the current orientation. The compliant segment for a given camera orientation is defined by two endpoints, $\mathbf{p}_1$ and $\mathbf{p}_2$, where $\mathbf{p}_2$ is higher (with respect to the document) than $\mathbf{p}_1$.

However, if the slant of the camera with respect to the sheet normal is too large (*non–compliant orientation*), the compliant segment for the current camera orientation may contain no points, meaning that, in order to reach a compliant pose, the camera needs to be rotated.

Correction information is communicated to the user through synthetic speech. Synthetic speech capabilities are provided by the Flite [16] library. Each short sentence contains directions along at most two Cartesian axes, and precisely those in need of the largest correction (e.g., "Move up 5 and forward 3" or "Move left 4"). We felt that specifying three vector coordinates (e.g., "Move up 5, forward 3 and left 8") would generate exceedingly long sentences and possibly become confusing. Units are expressed in centimeters, and the reference system is fixed with respect to the paper sheet (not the user). This could create a conflict if the user construes the direction as if in reference to his or her body; however, we noted that most participants kept the paper sheet aligned with their body, reducing the risk of conflicting frames of reference. Note that the camera pose is monitored in real time, and directions are produced continuously (with a minimum gap of 1 second between two sentences).

If a non–compliant orientation is detected, the system utters the sentence "Reset orientation", which prompts the user to re-orient the phone, ideally bringing the phone parallel to the document.

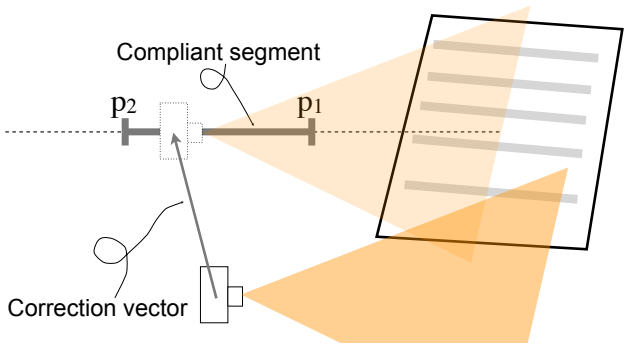

Figure 3: A simple guidance example. The current camera pose (shown in solid line) is not compliant, because part of the document is outside of its field of view. If the camera is moved by the correction vector, it will reach a position in the compliant segment. If the orientation is kept constant, any position on the compliant segment is compliant.

Upon detection of a compliant pose, the system utters the sentence "Pose compliant", terminating the trial.

Our strategy for determining the correction vector was inspired by a similar algorithm originally proposed by Cutter and Manduchi [7]. Their algorithm does not consider camera orientation: it always produces a correction vector that would bring the camera to a compliant pose *under the assumption that the sheet is seen front-to-parallel*. With the system used in their study, the heights of $\mathbf{p}_1$ and $\mathbf{p}_2$ are of 28 cm and 42 cm, respectively and centered at the orgin. In practice, this means that the correction vector is potentially incorrect as soon as the iPhone is not held parallel to the sheet (i.e. at non-null *off-axis angles*). As shown in Fig. 7, off-axis angles of 10 degrees or more are to be expected, which highlights the need for explicit orientation reasoning as in the new algorithm proposed here. With our algorithm $\mathbf{p}_1$ and $\mathbf{p}_2$ and set dynamically given the current orientation and position.

3.6 Procedure

Participant were given an introduction to the goals of the experiment and to its procedures. They were informed that, in order to take a "good" (compliant) picture of the document, the camera should be at a height of between approximately one foot and one and a half feet over the document, with the iPhone level (horizontal) and well aligned with the document. Each participant was asked to sit on a chair in front of a small desk, and invited to adjust the height of the chair to ensure that he or she was able to raise his or her iPhone-holding hand comfortably at least 40 cm above the desktop. Participants were informed that they could stand up during the experiment, if they felt that this would increase their comfort, and that they could use either or both hands to hold the phone. Most participants decided to sit for the duration of the experiment, although three participants decided to stand for all or some of the trials. Several of the participants experimented with multiple positions of the phone holding hand throughout the experiment.

After this preliminary phase, each participant performed the experiment, structured as an ordered sequence of sessions: Pre-intervention, Intervention, and Post-intervention. Each session was comprised of 12 identical trials; participants were informed that the first three trials of each session were to be considered practice trials. At the beginning of each trial, the paper sheet was slightly moved and rotated on the desktop, and the iPhone was placed flat (the camera facing downwards) over the document's left corner closest to the participant. In this way the participants frame of reference was reset; each trial simulates a fresh document scanning scenario. Each participant was assigned a Group ID (0 or 1), such that the IDs were evenly distributed across participants.

3.6.1 Pre-Intervention

The goal of each trial was to take a compliant picture of the document using the *snapshot* modality described earlier. The participant was asked to pick up the iPhone, and position it where he or she thought a good picture of the document could be taken. Once they were confident of the position they took a picture by pressing either of the two small volume buttons on the side of the iPhone. Participants were free to re-position the document on the desktop if they wanted to, and could take as much time as they wanted before taking the snapshot.

Several participants found the action of pressing one of the volume buttons difficult to execute, especially when holding the phone with one hand, although others found it very natural. Two participants expressed concern about the possibility that while reaching with a finger for these buttons, the phone may be inadvertently moved, generating blur or resulting in the picture taken from an incorrect location; however, this didn't seem to be the case, and all snapshots taken this way were correctly processed by the system.

3.6.2 Intervention

The goal of these trials was to move the iPhone over the document so as to reach a compliant pose using one of the *hovering* modalities described earlier. Participants in Group 0 used the *hovering: guidance* modality, while participants in Group 1 used the *hovering: just confirmation* modality. The starting procedure at each trial was the same as for the pre-intervention trials. A time-out period T_{to} of 150 seconds was set for each trial: if a compliant pose was not reached within the time-out period, the trial was terminated.

3.6.3 Post-Intervention

This session was identical to the Pre-intervention session. All participants used the *snapshot* modality to try to take compliant pictures of the document. These trials were meant to investigate whether experience with a hovering modality in the Intervention trials could increase the user's awareness of the compliant space, and thus facilitate taking a compliant snapshot of a document without system assistance. At the end of the three sessions, participants were asked to answer a short questionnaire, described in detail in the Results section.

The experiments described in [7] also consider similar interaction modalities to those considered here, albeit under different names. However, the experiment design in [7] and in the study presented here are very different. Participants in the experiments of [7] all underwent the same sequence (Snapshot; Hovering:Just Confirmation; Hovering:Guidance). This design does not allow one to evaluate whether experience with a hovering modality can increase one's skill at taking compliant snapshots without system assistance (which is the reason for the Pre- and Post-Intervention phases of the new design). In addition, the experiment design from [7] did not balance the order of the hovering modalities, resulting in a potentially biased analysis.

3.7 Metrics

3.7.1 Accuracy

Each *snapshot* trial can be characterized by a binary variable (*success*) that is equal to 1 if the snapshot resulted in a compliant picture, 0 otherwise. The *success rate* (*SR*) represents the average success value over all trials in a session.

We also derive a "softer" measure of accuracy (*proportion legible*) defined as the number of equivalent 12-point characters in the printable area that are OCR-readable from the image, divided by the total number of characters in the printable area, assuming the the printable area is filled with 12-point characters in an ordered grid. (This grid is designed based on standard inter-character and inter-line spacing.) Note that proportion legible = 1 implies success = 1; the opposite is not true. The proportion legible metric gives an indication of the document area that can be accessed by OCR. Note, however, that this does not translate directly into "readable portion of a document": if, for example, the right half of a text column is outside of the camera's field of view, the whole column is not "readable" (even though individual words in the left half can be decoded by OCR). A more useful metric, which we will consider in future work, would take the document structure into account. For each session, we computed the *median proportion legible* over all trials in the session.

3.7.2 Time

For the *hovering* trials, we measure the time from the beginning of the trial until a compliant pose is reached (*time-to-completion,* T_c). If a compliant pose is not reached before the time-out period T_{to}, we simply set $T_c = T_{to}$.

4. RESULTS

4.1 Snapshot Modality

4.1.1 General Results

Figs. 4 and 5 show the results, in terms of success rates and proportion legible, for the pre- and post-intervention trials using the *snapshot* modality. From these plots, it results clear that, while some participants were quite proficient at this task, others had serious difficulties. In particular, seven participants could not take a single compliant picture in the pre-intervention trials; three of them could not take any compliant picture in the post-intervention trials either.

To investigate the main causes of failure, we need to consider all conditions that can result in a non-compliant pose. The space of poses PS can be divided into four disjoints sets:

PS1: Poses that can be made compliant by simply re-positioning the camera (orientation unchanged) but not by simply re-orienting the camera (position unchanged).

PS2: Poses that can be made compliant by simply re-orienting the camera (position unchanged) but not by simply re-positioning the camera (orientation unchanged).

PS3: Poses that can be made compliant by simply re-orienting the camera or re-positioning the camera.

PS4: Poses that can be made compliant only by re-orienting and re-positioning the camera.

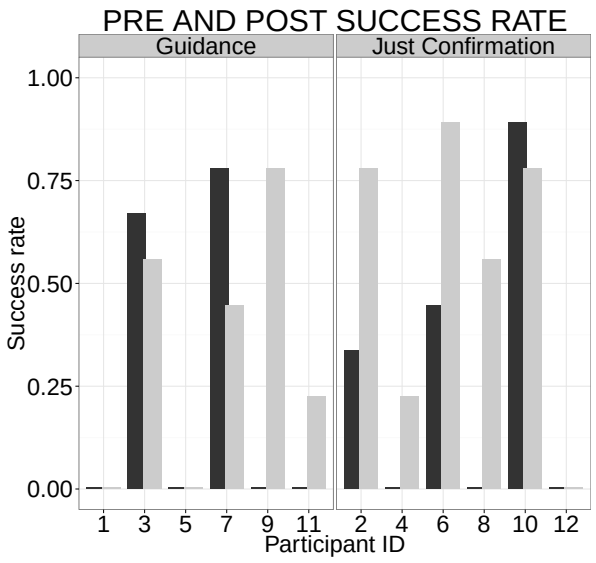

Figure 4: Success rate for all participants in the snapshot-type trials. Left: Group 0 (Guidance). Right: Group 1 (Just confirmation). Black: pre-intervention. Gray: post-intervention.

We analyzed the poses of the non-compliant snapshots, in order to obtain proportion of occurrence of the different types of poses above. This is expressed as probabilities (see Tab. 1.)

Pr(PS1)	Pr(PS2)	Pr(PS3)	Pr(PS4)
0.35	0.1	0.49	0.06

Table 1: The probability distribution of non-compliant poses across the four conditions considered.

This data suggests that in most cases ($\Pr(PS1) + \Pr(PS3) = 0.84$) a simple re-positioning of the camera would have led to a compliant snapshot. In a smaller proportion of cases ($\Pr(PS2) + \Pr(PS3) = 0.59$), a compliant pose would have been reached by simply re-orienting the phone. The more serious situation of a pose requiring both orientation and position adjustment occurs only 6% of the time.

Fig. 6 shows the location of the camera at the time of the snapshot for compliant poses (black dots) and non-compliant poses (grey dots). (Remember that locations higher than 42 cm and lower than 28 cm with respect to the document are non-compliant.) The plot suggests that in many cases, non-compliance was due to the participant keeping the phone too close to the document (the difference in height means between compliant and non-compliant poses is significant at $p < 0.001$). Fig. 7 shows the histogram of *off-axis angles* (defined as the angle between the camera's optical axis and the normal to the document) at the time of the snapshot. (Note that the off-axis angle, by itself, does not determine compliance: if the camera is located to the side of the document, a moderately large off-axis angle may be required for compliance.) This histogram shows that, on average, non-compliant poses were characterized by a larger off-axis angle than compliant poses (the difference in means is significant at $p < 0.001$).

The median time to take a snapshot (over all trials in a pre- or post-intervention session) ranged from 5.6 sec. to 39.3 sec., with a mean

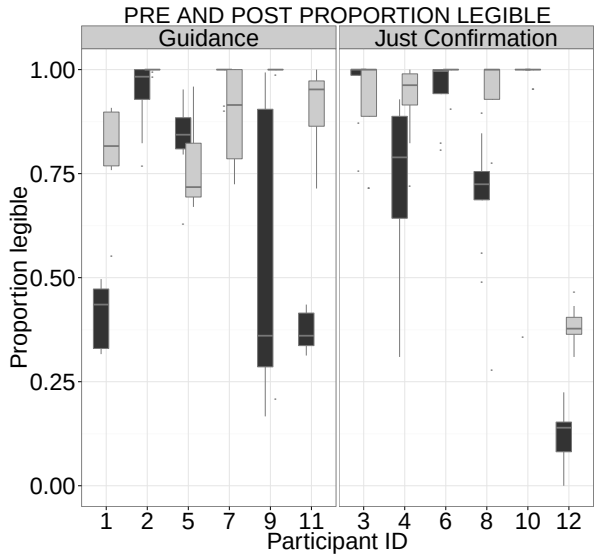

Figure 5: Proportion legible for all participants in the snapshot-type trials, shown as box plots. Left: Group 0 (Guidance). Right: Group 1 (Just confirmation). Black: pre-intervention. Gray: post-intervention.

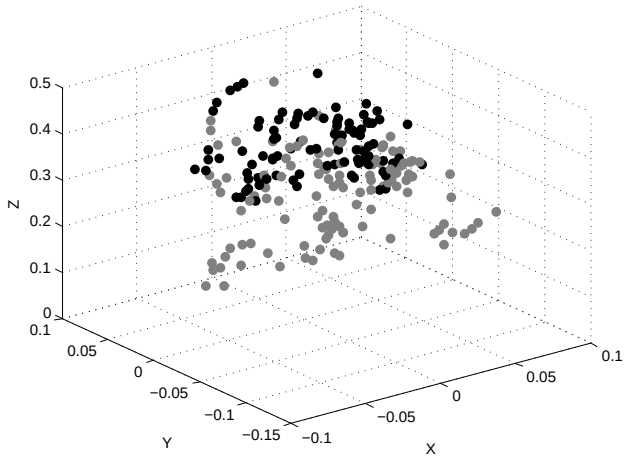

Figure 6: 3-D locations of camera pose in the pre- and post-intervention trials, with respect to a reference system centered at the center of the paper sheet (units are in meters). Black: compliant pose. Gray: non-compliant pose.

of 12.4 sec.

4.1.2 Pre- and Post-Intervention Comparison

We compared the success rate and median proportion legible for pre- and post-intervention sessions using a standard 2×2 mixed factorial design model. Note from Fig. 4 that among those participants who were able to take compliant pictures in the post-intervention trials, two in Group 0 and four in Group 1 improved their success rate after the intervention session, while two in Group 0 and one in Group 1 worsened their performance. The difference in mean success rate between pre- and post-intervention and across groups was not found to be significant at $\alpha = 0.05$.

The difference in mean between the pre- and post-treatment median proportion legible is significant at $p = 0.04$ (mean equal to 0.72 for pre-treatment, 0.89 for post-treatment). However, the main effect of intervention type (*guidance* vs. *just confirmation*) was not found to be significant at $\alpha = 0.05$. No significant difference was found between the means of camera height, horizontal offset (distance to the line perpendicular to and centered at the sheet), or off-axis angle at the time snapshots were taken for the pre- and post-intervention trials. However, for the participants that were not able to take a single compliant snapshot in the pre-intervention trials (participants 1,4,5,8,9,11,12; see Fig. 4), we noted that the median (across trials) of the horizontal offset decreased from 5.5 cm to 3.7 cm (paired one-sided t-test; $p = 0.03$). This may help explain why all but one of these participants performed better (in terms of proportion legible) in the post-intervention trials.

4.2 Hovering Modalities

4.2.1 Time-to-Completion

Fig. 8 shows a box plot of the logarithm of the time-to-completion values for all hovering-type trials (Intervention session). The median (over all trials) time-to-completion ranged from 3.6 sec to 48.1 sec, with an average value of 13.9 sec. We notice that one participant in Group 1 (ID=12) took much longer to complete the hovering trials than the other participants; the mean value of the time-to-completion medians with this participant removed drops to 10.8 sec. Multiple-sample repeated measurements ANOVA analysis did not find a significant difference in the mean time-to-completion between participants in Group 0 (*guidance*) and Group 1 (*just confirmation*).

4.3 Participant Surveys

At the end of the experiment, each participant was asked to complete a short survey. Participants were asked to comment on a number of statements using a five-point Likert scale (with 'strongly disagree' represented by '1' and 'strongly agree' represented by '5'). The statements, reported verbatim below along with the median response, differed slightly across the two participant groups.

Questions for Group 0 (*Hovering: Guidance*)	**Median response**
I feel that, after interacting with the system, I am now able to take better pictures of the document by myself.	**4**
It was easy to follow the directions from the system.	**5**
The directions from the system helped me take better pictures of the document.	**4**
If the guidance system were available as an application, I would be interested in using it.	**5**

Questions for Group 1 (*Hovering: Just Confirmation*)	**Median response**
The system helped me take better pictures of the document.	**4**
It was easy to follow the directions from the system.	**5**
If this system were available as an app, I would be interested in using it.	**5**

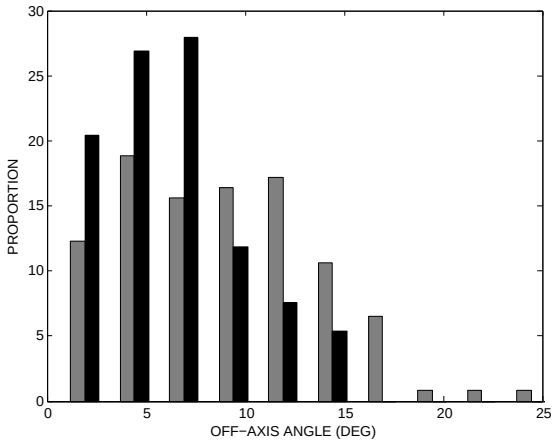

Figure 7: Histogram of off-axis angles for compliant (black) and non–compliant (gray) terminal poses in the pre- and post-intervention trials.

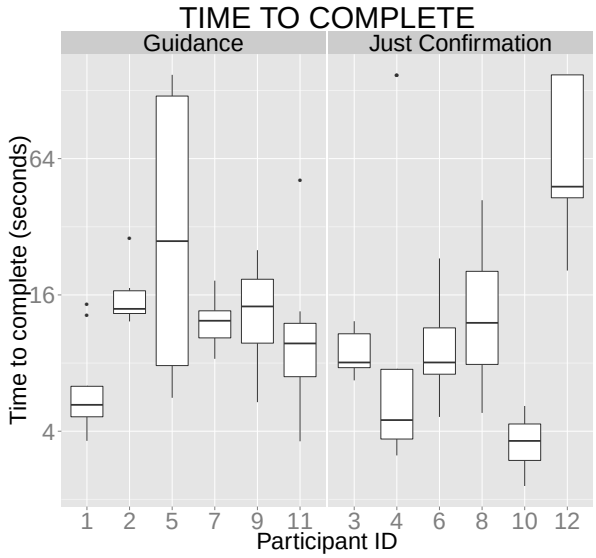

Figure 8: Time-to-completion values for all participants shown as a box plot on logarithmic scale. Left: Group 0 (Hovering: Guidance). Right: Group 1 (Hovering: Just Confirmation).

5. DISCUSSION

Participants exhibited a wide diversity of skill taking compliant snapshots without help from the system (Figs. 4 and 5). By observing the participants during the experiment, it was clear that some were much more "methodical" than others in the way they moved the phone to take a snapshot. Interestingly, as shown by Fig. 6, participants tended to take snapshots at a short distance from the document: the maximum recorded height of a snapshot was 44 cm, which is slightly above the maximum compliant height (42 cm). As mentioned earlier, participants were informed that the correct height was approximately between one foot and one and a half feet, but it seems that they preferred to err on the lower end. Of course, since no feedback was provided in the pre-intervention session, participants did not have a means to correct what could be a biased perception of the camera height. However, this tendency did not change even after the Intervention phase, in which participants had a chance to experiment first-hand the range of compliant heights.

Can the proprioception skills that are necessary to correctly position a camera be taught? We note that during the trials performed as part of the pre and post-test, we observed no trend of improvement between the first and the second half of the trials. This makes sense since there is no feedback during the snapshot trials. However, for many participants we observed improvement between the the pre and post-test. In addition, our quantitative results with the experimental system, along with the outcomes from the participant surveys, supports this observation. However, these results do not provide a clear indication of what exactly was learned through the Intervention phase.

As mentioned above, participants in the post-test trials continued to take snapshots from a relatively low height, something that undoubtedly contributed to a fair portion of failures. However, anecdotally a participant in the guidance group said after several trials of the intervention "ahah now i've got it". Similar "aha" moments occurred for other participants during the intervention; at which point the subsequent intervention trials were quickly completed.

We were surprised by the discovery that both the *guidance* and the *just confirmation* intervention modalities produce comparable results. We carefully designed a complex guidance modality, and expected that it would help the user reach a compliant pose faster. This expectation was supported by preliminary results using a similar system with sighted blindfold participants presented in [7]. Although as discussed earlier, the experimental design and the chosen metrics in [7] may have been inappropriate for this type of analysis.

Why is it, then, that the guidance modality, with its rich system feedback, did not prove superior to the just confirmation modality in terms of time-to-completion in the present study? We believe that the reason for this lies in the sub-optimal design of the user interface used in these prior experiments. Upon careful analysis of the videos collected during the trials, we determined two main pitfalls of the current design:

Lack of explicit orientation guidance. As shown in Fig. 7, non-OCR-complaint images are often associated with excessive off-axis angles. Our original guidance system gave directions in terms of translation but not of orientation; this was a deliberate choice in order to keep the complexity of directions low. Participants were advised to keep the iPhone horizontal; only upon detection of a large off-axis angle was a synthetic speech warning produced. However, most participants found it difficult to re-orient the phone correctly (horizontally), resulting in the off-axis warning being re-issued several times before the orientation of the iPhone was properly adjusted. When this happened, the whole process was slowed down, which generated frustration among some participants. We now believe that some form of orientation correction guidance would be very beneficial. Indeed, as discussed earlier, in 59% of the non-compliant snapshot cases, a simple camera re-orientation would have been sufficient to make the pose compliant, and in 6% of the cases this correction would in fact have been necessary.

Disruptive guidance modality. The synthetic speech directions produced by the system contained precise metric indication of where to move the phone next. Ideally, the user would move the phone ex-

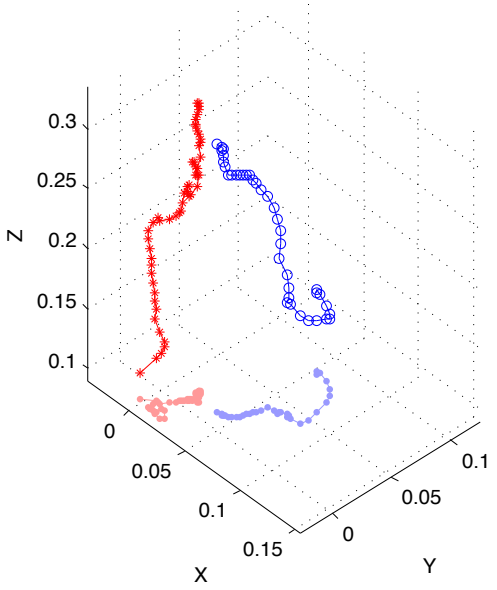

Figure 9: The paths represent camera locations during two trials, using the hovering:just confirmation modality (red) and the hovering:guidance modality (blue). Units are in meters. The projection of the paths on the horizontal plane are shown with faded color. Circular blue marks and red asterisks are placed at constant time periods of 0.1 s. Only the portion of the path after a certain time lag is shown as measurements cannot be taken when the camera is too close to the document. This lag was of 6.8 s for the path marked in red and of 3.9 s for the path marked in blue.

actly as directed, ending at a compliant pose. In fact, this was rarely the case, due to the difficulty of moving the phone precisely as directed. This resulted in participants following a discrete sequence of movements; after each movement, they would pause and wait for the system to produce the next direction. In contrast, participants in the group that did not use the guidance system moved the phone in continuous motion; this allowed for a larger portion of space to be explored in the same amount of time. The difference in behavior for the two hovering modalities can be noticed in Fig. 9. The path marked in blue (*hovering:guidance*) is characterized by non-uniform velocity and several abrupt turns in response to a direction, whereas the path marked in red (*hovering:just confirmation*) shows a more uniform motion. In future work we will explore different types of acoustic interface that require less information processing by the user and encourage smooth trajectories.

6. CONCLUSIONS

We have presented an experimental study that investigated modalities to help a blind person take better pictures of a document faster through the use of image processing software. The overarching goal of this project is to facilitate the use of mobile OCR for printed text access.

The proposed mechanisms have been implemented using special printed fiducials, and could not be used directly with regular printed documents. This investigation explores the "best case scenario" of a perfectly functioning device; similar functionalities on regular printed documents are not out of reach.

Camera orientation can be computed from the device accelerometers and by measuring orientation of detected parallel text lines. By detecting the endpoints of text lines, one can make inferences about whether the text is fully visible (e.g. a line ending at the edge of the image is likely truncated) or, if not, where the camera should be moved for better visibility. Readability of characters can be computed by a fast text spotter (e.g. if characters in a line cannot be spotted, the camera is too far). Localization features could be approximately inferred by computer vision algorithms with heuristics about the visual structure of typical documents. These vision-based algorithms can obtain functionalities similar (albeit less accurate) to using fiducials with real-world documents.

Acknowledgments

Research reported in this publication was supported by the National Eye Institute of the National Institutes of Health under award number 1R21EY025077-01 The authors would like to thank Corinne Olafsen who assisted during the experiments.

7. REFERENCES

[1] Prizmo. `http://www.creaceed.com/prizmo`. Accessed: 2013-11-09.

[2] Knfb reader mobile. `knfbReadingTechnology, Inc`, 2008. http://www.knfbreader.com/.

[3] Text detective (blindsight inc.). `http://blindsight.com/textdetective`, 2011.

[4] Aruco: a minimal library for augmented reality applications based on opencv. Universidad D Cordoba, 2012. http://www.uco.es/investiga/grupos/ava/node/26.

[5] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[6] E. Brady, M. R. Morris, Y. Zhong, S. White, and J. P. Bigham. Visual challenges in the everyday lives of blind people. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2117–2126, New York, NY, USA, 2013. ACM.

[7] M. P. Cutter and R. Manduchi. Real time camera phone guidance for compliant document image acquisition without sight. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 408–412. IEEE, 2013.

[8] C. Jayant, H. Ji, S. White, and J. P. Bigham. Supporting blind photography. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, ASSETS '11, pages 203–210, New York, NY, USA, 2011. ACM.

[9] S. K. Kane, B. Frey, and J. O. Wobbrock. Access lens: A gesture-based screen reader for real-world documents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 347–350, New York, NY, USA, 2013. ACM.

[10] S. K. Kane, M. R. Morris, A. Z. Perkins, D. Wigdor, R. E. Ladner, and J. O. Wobbrock. Access overlays: Improving non-visual access to large touch screens for blind users. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pages 273–282, New York, NY, USA, 2011. ACM.

[11] R. Shilkrot, J. Huber, C. Liu, P. Maes, and S. C. Nanayakkara. Fingerreader: A wearable device to support text reading on the go. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '14, pages 2359–2364, New York, NY, USA, 2014. ACM.

[12] M. Vázquez and A. Steinfeld. Helping visually impaired users properly aim a camera. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*, ASSETS '12, pages 95–102, New York, NY, USA, 2012. ACM.

[13] A. Zandifar and A. Chahine. A video based interface to textual information for the visually impaired. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, ICMI '02, pages 325–, Washington, DC, USA, 2002. IEEE Computer Society.

[14] Y. Zhong, P. J. Garrigues, and J. P. Bigham. Real time object scanning using a mobile phone and cloud-based visual search engine. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '13, pages 20:1–20:8, New York, NY, USA, 2013. ACM.

[15] TapTapSee. www.taptapseeapp.com.

[16] A. Black and K. Lenzo Flite: a small fast run-time synthesis engine In *SSW4-2001*, paper 204. 2001

[17] ORCAM. www.orcam.com

[18] J. Bigham, C. Jayant, A. Miller, B. White, and T. Yeh. Vizwiz::locateit — enabling blind people to locate objects in their environment. In *Proc. Workshop on Computer Vision Applications for the Visually Impaired*, 2010.

[19] J. Coughlan and R. Manduchi. Camera-based access to visual information. In R. Manduchi and S. Kurniawan, editors, *Assistive Technology for Blindness and Low Vision*. CRC Press, 2013.

[20] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

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

[22] E. Tekin and J. Coughlan. A mobile phone application enabling visually impaired users to find and read product barcodes. In *Proc. International Conference on Computers Helping People with Special Needs (ICCHP '10)*, 2010.