HHS Public Access

Author manuscript

Proc SIGCHI Conf Hum Factor Comput Syst. Author manuscript; available in PMC 2014 November 23.

Published in final edited form as:

Proc SIGCHI Conf Hum Factor Comput Syst. 2014 ; 2014: 3113–3122. doi:10.1145/2556288.2557328.

The Last Meter: Blind Visual Guidance to a Target

Roberto Manduchi and

Computer Engineering Dept., University of California, Santa Cruz, manduchi@soe.ucsc.edu

James M. Coughlan

Smith-Kettlewell Eye Research Inst., San Francisco, CA, coughlan@ski.org

Abstract

Smartphone apps can use object recognition software to provide information to blind or low vision users about objects in the visual environment. A crucial challenge for these users is aiming the camera properly to take a well-framed picture of the desired target object. We investigate the effects of two fundamental constraints of object recognition – frame rate and camera field of view – on a blind person's ability to use an object recognition smartphone app. The app was used by 18 blind participants to find visual targets beyond arm's reach and approach them to within 30 cm. While we expected that a faster frame rate or wider camera field of view should always improve search performance, our experimental results show that in many cases increasing the field of view does not help, and may even hurt, performance. These results have important implications for the design of object recognition systems for blind users.

Keywords

Assistive technology; Blindness; Wayfinding; Camera-based access to information

ACM Classification Keywords

H.5.2. Information interfaces and presentation: User Interfaces-- Input devices and strategies; Interaction styles

INTRODUCTION

A growing number of smartphone apps are now available that use the smartphone camera to provide information to a blind or low vision user about objects in his or her visual environment. Such apps use a combination of computer vision-based object recognition algorithms or crowd-sourcing techniques to perform tasks such as identifying grocery products, determining the denominations of paper currency, reading a sign posted on the wall or reading a printed document such as a restaurant menu. However, this technology

poses a fundamental challenge: how can a user with little or no vision take a well-framed picture of the desired target object? This process entails both *exploration* in search of a target, and, once the target has been detected, *guidance* to the target using feedback from the system. Our work concentrates specifically on the guidance phase, which is a crucial bottleneck in the overall search process but which has received little attention in past research.

Recent research [18,17] has explored various real-time guidance mechanisms that help a blind or low vision take well-framed pictures. Indeed, in our previous work on our smartphone based color marker detection system [7,4,13], we explored and tested a variety of user interface (UI) options before arriving at the UI used in our current system (see the "Apparatus" Section). Given such a mechanism (which is fixed in our current study), we explore the effects that *fundamental constraints imposed by the object recognition technology itself* have on the user's performance in acquiring well-framed pictures. Among the most important of these constraints are the video frame rate (the rate at which video frames are processed by recognition algorithms) and the camera field of view (FOV, which is determined by the camera optics). In many circumstances it is possible to trade off one constraint against another in the design of the object recognition system; however, little is known about the practical consequences of these trade-offs for visual search and framing by a blind or visually impaired person. For instance, frame rate can often be increased by down-sampling the video frames, but at the expense of limiting the maximum range at which the target can be resolved. Similarly, expanding the FOV (e.g., with a wide-angle or fisheye lens) has the potential to speed up the initial search for a target, but it also reduces the image resolution, and (as we show in this paper) may make it more difficult to localize the target from close-up.

We investigate the effects of these constraints using a fast and extremely reliable computer vision-based object recognition smartphone app, developed in-house, which was used by a total of 18 blind participants to find visual targets beyond arm's reach and to approach them to within a distance of approximately 30 cm, using continuous audio feedback from the app. Compared with the authors' initial supposition that either a faster frame rate or wider camera field of view should always improve search performance, the results of our statistical analysis of the experiment are more nuanced, showing that in many cases increasing the field of view does not help, and may even hurt, performance.

While our study used a specific type of visual target in a particular search task, we argue that the results of the study generalize to nearly any object recognition-based visual search task performed by a blind user. Specifically, any mobile object recognition task requiring the target to be sufficiently well resolved and fully contained within the camera's field of view will be subject to two fundamental system constraints, the frame rate and FOV. Thus, the main contribution of our work is to explore the effects that the fundamental constraints of object recognition technology have on search performance for blind users.

RELATED WORK

A number of technologies to support independent orientation and mobility for persons with visual impairment have been proposed and investigated by the research community [14]. Much less attention has been devoted to the topic considered in this contribution, that is, precise guidance to a target through continuous visual-based tracking. Document access via mobile OCR (such as the KNFB Mobile reader [1] and Blindsight's Text Detective [3]) presents similar problems: the user needs to take a close-up, well-framed picture of the document. Beyond OCR, other applications of visual information access include barcode reading. 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 [17]. Preliminary experiments with a guidance system using similar color markers as in the present study were reported in [13]. The study in [13] was mostly qualitative in nature; it highlighted the difference between *exploration* in search of a target, and, once the target has been detected, *guidance* to the target using feedback from the system. The present work concentrates on the second component (guidance).

Assisting a blind person while taking pictures, whether for leisure [10], to document environmental features [18], or for remote assistance by sighted helpers [6,11,2], also requires some form of guidance to ensure that a good picture of a target from a close distance is produced. For example, EasySnap [10] is a mobile application that gives feedback to a blind photographer about the scene light, or about the presence and localization in the picture of an object or of a person. LocateIt [5] uses 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 get closer to it). The system developed by Vázquez and Steinfeld [18] uses a general-purpose saliency map to select a region of interest. Feedback to the user is provided through audio tones or synthetic speech; it was noted that visually impaired users slightly preferred speech feedback over audio tones, consistent with earlier findings from other research groups [9].

We note that similar guidance mechanisms, whereby the blind user receives feedback about the correct pointing of a hand-held device, were studied in the context of other navigational technology. For example, Talking Signs [8] uses beacons transmitting modulated infrared light. A user carrying a hand-held receiver in the proximity of the beacons hears audio (typically, informational speech) from the demodulated received light only when the receiver is aimed towards the beacon. Similar interface mechanisms (using audio and/or vibration) were studied in wayfinding systems using GPS [15] or digital compass [12].

ASSESSING VISUAL-BASED GUIDANCE

Overview and Rationale

Using a camera system for guidance towards a target can be challenging without sight. The characteristics of the image acquisition and processing and of the user interface both play a role in the user's experience with such a system. In this work, we concentrate on the characteristics of the vision system, assuming a simple user interface modality, described later in this section. Important system specifications include: the camera's resolution

(number of pixels); its field of view (FOV); the speed of the acquisition/processing system (in terms of frames processed per second, FPS); and the quality of the target detection system, which can be expressed, for example, in terms of false negative and false positive rates. It is clear that these characteristics are interrelated. For example, a system with higher resolution may take longer time to process, resulting in lower FPS. The FOV and resolution, combined with the characteristics of the detection algorithm, determine the distance at which a target of a certain size can be detected.

Two characteristics (FOV and FPS) are considered in this work. Specifically, we study how execution of a given guidance task is affected by the camera FOV as it is switched from narrow to wide, and by the FPS as it is switched between fast and slow. These two system specifications are both important for the design of a visual guidance system. The camera FOV can be modified (changing the lens, using a zoom lens, or using an add-on lens). Intuitively, a wide FOV ensures that the target is seen from a wider span of camera orientations, but also reduces the apparent size of the target, which affects the ability of the computer vision algorithm to recognize it. A wide FOV also affects the ability of a blind user to determine the target's precise location. As for the FPS, our goal is to understand whether algorithms characterized by low frame rate (due e.g. to heavy computational load or the need to access a remote server) are still usable for the purpose of blind guidance, or result in excessive user frustration.

We would like to emphasize that we have deliberately chosen to use an object recognition system with near-perfect performance in order to simplify our experimental study. Our system is almost entirely unaffected by the kinds of issues prevalent in real-world object recognition systems, such as false and positive negative detections and confusion with clutter. We felt that it would be difficult to model the occurrence of false positive/negative detections in a realistic way that would generalize to real-world tasks, since these detections are heavily influenced by the nature and quantity of occlusions and clutter in the scene. However, all object recognition systems (whether real or ideal) are subject to the FOV and FPS constraints, which we feel are the most fundamental constraints in the search process; by using our object recognition system we were able to design a tractable study that focuses on these two variables, thereby minimizing possible confounds with other search variables.

Participants

The experiments were conducted in two different locations. We recruited a total of 20 participants, 10 at each location. The oldest participant (age 86) was unable to hold the smartphone steady, even after training, and was therefore unable to perform the search task, so this person was excluded from the study. Another participant had to terminate the study early, so her data was also excluded from the study. As a result, our study includes a total of N=18 participants (six females and 12 males), with ages ranging from 18 to 71 years, with median value 48.5 years. Nine participants had no light perception and the rest had very limited light perception (insufficient to see the color markers in the experiment). Five participants have had their current degree of vision impairment since birth.

Apparatus

Detection Software—For our experiments we employed a system that uses specially designed fiducials ("markers") in the form of a color pie with four sectors (see Fig. 2). The detection software uses the algorithm described in [4], which returns the position (in the image) of four equi-spaced keypoints on the marker's circumference as well as on the marker's center. Given the known size of the marker (16 cm in diameter) and the optical/imaging characteristics of the camera, the camera's pose (position and orientation) can be estimated from these five keypoints. The detection algorithm assumes that the camera is kept approximately vertical, with a roll angle (around the optical axis) between -45° and 45°. If the user inadvertently rotates the phone by more than the allowed roll angle, a short warning vibration is produced. The algorithm is sensitive to the order of the colors in the sectors, which enables us to define a variety of different color markers by permuting the same four colors. Each color permutation is assigned an ID, and the system can be set to detect only markers with a specified ID.

The marker detection algorithm was implemented on an iPhone 4. At VGA resolution (640 by 480 pixels), the achievable frame rate varies from 9 frames per second (FPS) when no marker is visible, to 3 FPS when the marker is detected. The frame rate can also be artificially decreased to 0.5 FPS, to achieve the low frame rate modality used in our experiments. We chose this value of FPS as it can be reasonably expected that most standard computer vision algorithms for target detection would take no longer than 2 seconds per frame to execute on an iPhone.

In some experimental settings, we increased the camera's FOV (normally approx. 48° by 61°) by means of a fisheye lens from Photojojo.com that snaps on and off the iPhone with a magnetic attachment. With this lens, the FOV increases to approximately 87° by 130°. This lens introduces very noticeable radial distortion which, however, does not affect marker detection, even from very large slant angles (see Fig. 2). The distance to the marker can also be computed accurately even with the fisheye lens on, except in situations with large horizontal and vertical off-axis angles. (The *off-axis angle* is the angle between the optical axis and the ray pointing from the camera center to the center of the marker, which is 0° when the marker center appears in the center of the image. It can be decomposed into horizontal and vertical components, which we refer to as the horizontal and vertical off-axis angles.) Marker detection in the conditions considered in our experiments, from a maximum distance of about 1.5m, under controlled illumination, and with solid white background, is extremely accurate and reliable, with virtually no false positives or missed detections. The application logs time-stamped data relative to acquired frame and detection results.

User Interface—The system has the following acoustic UI. When no marker is detected, the system is silent. When a marker is detected at a vertical or horizontal off-axis angle of more than 10°, a recorded sentence is uttered, giving directions to the user about how to rotate (pivot) the camera in order to reduce the off-axis angle. These sentences take the form of "Turn right", "Turn up", "Turn left and down", etc. When the marker is "well centered" (meaning that both vertical and horizontal off-axis angles are less than 10°), the system beeps periodically. Beeps are repeated at a rate of 2 beeps/sec. when the marker is at a

distance of more than 50 cm, and of about 5 beeps/sec. at lower distance. When the frame rate is artificially reduced to 0.5 FPS, one beep is emitted for each processed frame when the distance to the marker is 50 cm or more, and a sequence of three short beeps is emitted for each processed frame at shorter distances.

Procedure

The experimental set-up consisted of eight color markers (each with a distinct permutation of colors), affixed to a wall as shown in Fig. 2. The markers were arranged in three rows ranging in height (measured from the floor to the center of the target) from 112 cm to 162 cm, and the horizontal spacing between adjacent markers in each row was 45 cm (measured between the marker centers).

For each participant who volunteered for the experiment, the experimenters first obtained his or her consent to participate in accordance with an IRB protocol. The participant was then given a training and practice session to acquaint him/her with the purpose of the study and the operation of the iPhone app, including the proper way to hold and move the smartphone, and was asked to try out the system a few times to find and approach one or more markers.

The experiment consisted of four sessions of 12 trials each, for a total of 48 trials for each participant in the experiment. In each trial, the participant was asked to find and approach the target, starting from a point 150 cm from the wall (centered relative to the set of targets), and ending when the system announced the target was successfully localized in the camera's FOV from a distance of approximately 30 cm. Specifically, the following termination criteria were implemented: (1) the target is 30 cm or closer to the camera (*distance constraint*); (2) the magnitudes of the horizontal and vertical off-axis angles are both 10° or less (*angular constraint*); and (3) the entire target is contained in the image (*visibility constraint*). This last constraint is dictated by the fact that our system can compute the distance to the target only when the target is fully visible. In practice, this constrains the distance to the target to be larger than a certain amount (approximately 20 cm) for successful termination. We refer to these combined termination criteria as *Scenario 0;* in the next section, we will introduce four additional "Derived Scenarios," called Scenarios 1, 2, 3 and 4, based on modified criteria (defined ex post facto) that are less stringent than the criteria defining Scenario 0.

When the Scenario 0 termination criteria are met, the system declares success by uttering, "You have reached the target. Congratulations!" Note that, for a given lens setting (FOV), compliance with the termination criteria is determined solely by the *pose* (3-D location and orientation) of the camera. (Note that the camera orientation is specified by three angles, roll, yaw and pitch.) We will say that a pose is *compliant* if it satisfies the termination criteria. For a certain FOV, we define the *compliant pose set C(FOV)*, which contains all poses that are compliant for that specific FOV.

At the start of each trial, the participant was asked to face away from the wall at the starting position, and to turn to face the wall and begin searching for the target when the experimenter told him/her to begin. If success was not attained within 180 sec. from the time the experimenter told the participant to begin, or within 180 sec. from the first feedback

produced by the system, a “time-out” was declared for the trial. In each trial the target was chosen uniformly randomly from the set of eight targets on the wall.

Two factors, the FOV and frame rate (together these factor levels are jointly referred to as the experimental “settings”), were fixed for each entire session of 12 trials. The FOV had two possible levels: normal, using the standard iPhone camera lens, and wide, using the fisheye lens. The frame rate also had two possible levels: fast (several FPS), and slow (0.5 FPS). We will use the following notation for these factors: FOV = N or W denotes the narrow (normal) or wide-angle lens, and FPS = F or S denotes the fast or slow frame rate. The two factors imply a total of four possible settings: NF, NS, WF, and WS. Each of the four settings was applied to exactly one of the four sessions in the entire experiment. For each participant, the order in which the settings were assigned to the sessions was chosen at random in advance of the experiment. This randomization was done to minimize the confound between experimental settings and learning effects. The participant was informed of the FOV and FPS settings at the start of each session. The first two trials of each session were timed practice trials during which the experimenter was free to help out the participant, and the participant was free to ask for help; the purpose of these trials was to acquaint the participant with each setting, so these trials were not used in the data analysis. The next ten trials of the session were recorded and analyzed for a total of 40 trials recorded for each participant.

After the four sessions were completed, a brief questionnaire was administered to the participant, and the experimenter solicited feedback about the system and the experiment. Participants were also asked to report the perceived difficulty of completing the tasks for each one of the four settings on a scale between 0 and 5.

Derived Scenarios

In order to draw meaningful conclusions on how the experimental settings affect search performance, fair comparisons need to be drawn between the FOV = N vs. W levels. This need arises since the two FOV levels have different camera resolutions, which implies that a target can be satisfactorily resolved from a greater maximum distance with FOV = N than with FOV = W. Thus, it may be necessary to bring the camera closer to the target with FOV = W than with FOV = N for successful recognition. Moreover, even for Scenario 0 (the actual scenario used in the experiment, in which the termination distance of 30 cm was used for both N and W), different angular considerations apply for N vs. W: at 30 cm, in the narrow FOV case the marker had to be seen at an off-axis angle no larger than approximately 6° for the marker to be entirely contained in the image (visibility constraint). By contrast, this was not an issue for the wide FOV.

As a result, although all the tests were conducted under the same Scenario 0 termination criteria (see the “Procedure” section), for ex post facto data analysis we considered other, less restrictive *derived scenarios* to take into account various practical consequences of different FOV settings. Each derived scenario (see Fig. 3) corresponds to a specific set of search criteria and fulfills the following property: if a trial meets the Scenario 0 termination criteria, then for each Scenario 1, 2, 3 and 4 there must exist some contiguous subset of the time series for the trial (formed by omitting some data points at the beginning and/or at the

end of the original trial) that also fulfills all the criteria for that scenario. Thus, each "successful" trial satisfying Scenario 0 can be analyzed under Scenarios 1 through 4, and analysis under these derived scenarios permits meaningful conclusions on how to compare the search process across different settings. We note that using these derived scenarios, these conclusions can be drawn solely from Scenario 0 trials, *without* having the participants perform multiple versions of the experimental trials (up to four versions would be necessary, corresponding to Scenarios 1 through 4).

For example, one could look at the collected time series and artificially terminate it at the first occurrence of target detection at a distance of D_{stop}>30 cm, with both visibility and angular constraint satisfied. This is equivalent to changing the distance constraint to a higher value D_{stop} of distance. Conversely, one may find the first occurrence of a target seen at distance larger than or equal to a certain value D_{start}, and remove all data points in the time series before that. This would effectively modify the starting location of the participant. Or, one could artificially terminate the time series at the first occurrence of a target detected with both visibility and distance constraint satisfied, but regardless of whether the angular constraint is satisfied.

The four derived scenarios we considered are described below. Each scenario sets a value of termination distance, D_{stop}, and stipulates whether the angular constraint needs to be satisfied or not. In all cases, the starting distance D_{start} defined above is set to be equal to D_{stop} + 70 cm. This ensures fairness of comparison between settings that have different termination distance.

Scenario 1: Same distance, No angular constraint—In this case the termination distance D_{stop} is set to 57 cm for all settings, and the angular constraint is omitted. This scenario models a situation in which a blind person uses the guidance system to get to approximately arm's distance to the target, so that he/she can search for the target with his/her hand.

Scenario 2: Same distance, Angular constraint—The termination distance is the same as in the previous case, but the angular constraint is enforced. This means that, at termination, the user is pointing the phone fairly accurately towards the target. This may simplify subsequent tactile exploration, as one would need to search only along the direction pointed at by the phone.

Scenario 3: Same resolution, No angular constraint—This scenario models the case in which a picture should be taken of the target at a fixed resolution (number of pixels) in the image. The resolution at which an object is seen is an important parameter for image processing. For example, suppose that the target is a sign posted on the wall, which needs to be read by OCR. The camera needs to be moved close enough to the target to ensure sufficient resolution of the imaged text for OCR reading.

It is important to note that the resolution of a target seen from a certain distance by the same camera depends on the focal length of the lens. Hence, to ensure the same resolution using the N and W settings (i.e., with and without the fisheye lens add-on), the termination

distance should be different in the two cases. We verified experimentally that the image width of the target is the same when seen under the N setting at 57 cm and under the W setting at 30 cm. Hence, in this scenario, we set D_{stop}=57 cm for NF and NS, and D_{stop}=30 cm for WF and WS.

Scenario 4: Same resolution, Angular constraint—This scenario models situations in which centering the target in the image may be important (for example, to reduce radial distortion that may occur in the periphery of the image.)

RESULTS

Quantitative Analysis

We considered two observables to compare the different settings in each scenario. The first observable is the *time-to-target T*, defined as the time it takes from the starting to the termination distance. The second observable is the *out-of-FOV fraction V*, defined as follows. Consider the time series of measurements between D_{start} and D_{stop}. Let N_{iv} be the number of frames in which the target was detected (*in-FOV*), and N_{oov} the number of *misses*, that is, frames in which the target was not seen (*out-of-FOV*). Then $V=N_{oov}/(N_{iv}+N_{oov})$. The out-of-FOV fraction can take values from 0 to 1; small values indicate that the target was visible most of the time. Thus, while the time-to-target T is an objective measure of how quickly the user can reach a target compliant pose, the out-of-FOV fraction can be interpreted as a quantity that indirectly affects the user's experience, given that when the target is out of view, there is no feedback provided by the system.

The median of the time-to-target T and the median of the out-of-FOV fraction V for each participant over the 10 trials in each session were first computed. Computing the median for T has the benefit of removing the effect of censored data (time-outs) – since the number of time-outs was always less than 5 for all sessions and all participants, this implies that the median could be calculated even with the time-outs. The data was then represented as a 3-way vector T_{ijk} or V_{ijk}, where i is the participant index, j is the FOV level, and k is the FPS level.

We tested for equality of row and column mean treatment effects using standard two-factor, within-subject repeated measures ANOVA analysis at 0.05 significance level. (performed in the log domain for T_{ijk}, as this was shown to improve Gaussianity for this data, and directly on V_{ijk}). If the effects of both main factors were found to be significant and interaction was also found to be significant, simple effects were tested for significance using Bonferroni-corrected paired t-tests.

The results of this analysis are shown in Tab. 1 for T, and Tab. 2 for V. Each row in the table corresponds to one specific scenario, while columns indicate the factor that was found to be significant. More precisely: for a given scenario (row), if a certain factor (e.g., FOV) was found to be significant without interaction, the mean values of the observable (T or V) for the two levels of this factor (e.g., N=23.1, W=13.6) are reported in the corresponding cell. In the case of significant interaction, the cell reports the mean values of the observable for a fixed level of the other factor, if the simple effect was found to be significant (e.g., NF: 14.8,

WF: 20.7: in this case, FPS was fixed to F). Note that we included results for Scenario 0, even though, as discussed earlier, this scenario is excluded from the set considered in our analysis. This data is only meant as a reference for the comparative study of subjective user evaluation (see later in this section).

In order to explain the observed results, we'll make extensive use of the concept of camera pose set C(FOV) introduced in the Procedure Section. Examples of compliant and non-compliant camera poses are shown in Fig. 3 in a simplified 2-D illustration.

Time-To-Target—The dependence of time-to-target T with camera settings is a function of the scenario considered, as discussed below. An example of the data distribution is shown by means of box plots in Fig. 4 for Scenario 4.

Scenario 1: Same distance, No angular constraint—It is easy to see that, in this scenario, if camera pose is compliant for FOV=N, then it is also compliant for FOV=W (while the opposite is not true). In other words, $C(N) \subset C(W)$, and thus we may expect that, using a wide FOV, one should be able to reach a compliant pose sooner. The experimental results confirm this conjecture: target is reached, on average, 10 sec. faster using the wider FOV.

Scenario 2: Same distance, Angular constraint—As shown in Fig. 3, the sets of compliant camera poses are identical for the two FOVs, thus similar performances could be expected. This is confirmed by the experimental results, which show no significant difference between the average time-to-target using narrow or wide FOV. The overall mean time-to-target is 20.5 s.

Scenario 3: Same resolution, No angular constraint—The two compliant camera pose sets overlap, but there are poses in each set that are not represented in the other set (see Fig. 3). Hence, it is difficult to speculate about the performance with different FOVs. Analysis of the experimental results shows that, at least when FPS=F, use of the narrow FOV results in shorter time-to-target (by almost 6 sec.).

Scenario 4: Same resolution, Angular constraint—For a given point in space, the set of camera orientations that satisfy the angular constraint is the same for both narrow and wide FOV. However, use of the wider FOV reduces the maximum allowable distance D_{stop} in this scenario, as explained earlier. Otherwise stated, $C(N) \supset C(W)$, which conforms to the experimental observation (average time to target is 6.8 seconds less, under fast FPS, for narrow FOV than for wide FOV).

Concerning the dependence on the frame rate, it is seen that the slower FPS increases the average time-to-target by a substantial amount (9.4 sec to 13.4 sec depending on the setting). In Scenario 3, the main effect of FPS was found to be significant with significant interaction with FOV; however, comparison of the different levels of FPS for any fixed level of FOV was not found to be significant.

Out-of-FOV Fraction—The observed out-of-FOV fraction (V) values confirm the intuition that wider field of view (FOV=W) should result in higher likelihood of the target being in view (fewer misses) during the guidance process. In the two same-distance scenarios, a somewhat surprising result emerges: lower frame rate slightly decreases the rate of misses. A possible explanation for this is that lower frame rate makes the system less responsive, which may prompt one to more carefully aim the camera. (This intuition was supported by the comment of at least one participant.)

Subjective Assessments of Difficulty—Participants were asked to assess, on a scale from 0 to 5 (0 being "very easy" and 5 "very difficult"), the perceived difficulty of the guidance process under each setting. The scores for each participant are shown in Fig. 5. Average scores were: NF=1.41; NS=3.76; WF=1.53; WS=2.59. Intra-class correlation analysis (ICC(1,1) [16]) reveals that the subjective scores are correlated across participants ($p<10^{-7}$). It is interesting to compare the subjective scores of each participant with the measured values of T and V for the same participant, to ascertain whether the participants' perceived difficulty correlates with the chosen measurements. Remember that participants only experienced the original Scenario 0, not the derived ones, and thus data from this scenario should be used (summarized in the last row of Tab. 1 and 2).

Kendall's τ rank correlation coefficients between each participant's ratings across the different settings and the measured values for the same participant were computed for T and for V. For all but one participant, a positive rank correlation with T was observed, whereas for three participants, the rank correlation with V was negative. This suggests that the execution speed may be a more important factor than the rate of misses in the participant's evaluation of task difficulty.

Qualitative Observations

Each participant was carefully observed by one of the authors at each session. We noted that the experiment was very challenging for a few participants, some of whom were clearly tired towards the end. However, many participants seemed to approach the experiment as a game that was mostly enjoyable. We observed large variability in the participants' search strategies and search performance, which we believe arose at least in part to factors that are separate from the system's frame rate and FOV: namely, the participants' abilities to orient themselves to their surroundings, to hold the camera properly and move it slowly and steadily, and to use their proprioception to maintain awareness of how they were holding the smartphone in relation to their bodies.

Orientation in the Environment and With Respect to Target—Some participants had great difficulty orienting themselves to the room, and in some cases wound up aiming the smartphone camera towards the wrong wall (adjacent and perpendicular to the wall containing the markers). Some participants also had a consistent directional bias, e.g., usually pointing the camera axis well to the left/right of the direction their torso was pointing. One participant tended to consistently point the camera towards the ground, which caused problems in finding higher markers.

Many participants did not get close enough to the marker. They seemed reluctant to advance forward, as if they were waiting for something to happen and did not proceed when they were short of 30 cm. Also, some participants seemed to focus too hard on "centering" the target and forgot to advance closer. To advance closer, some participants extended their arm but not enough. In fact, it seems that, when close to the wall and the system was beeping, some were not sure exactly where to move.

Some participants monitored their distance to the wall with their foot or hand while others did not. Overall, many participants often failed to realize when they were very close to the wall. In fact, sometimes participants came so close to the wall that the system would detect the marker but could not announce success (marker never in full view), or they lost the marker, resulting in a long search time or time-out.

Holding and Moving the Camera—Most participants held the smartphone with one hand, but a few used two hands. Some pivoted the smartphone around the wrist, and others around the elbow. Most participants held the smartphone at approximately the same height off the ground, while a few lifted it up and down more in line with the heights of the targets. Some participants had trouble with markers that were very low or very high compared with their height.

Several participants tended to approach the target at a very large slant angle (i.e., the camera axis was far from perpendicular to the plane of the wall), largely because they tended to rotate the smartphone instead of translating it, and because the system provided no explicit feedback to differentiate between these two kinds of movements. For one participant, each time the system told him to turn left (or right), he tended to translate to the left (right) as directed but also inadvertently rotated to the right (left); thus over time he became increasingly slanted relative to the wall, while maintaining lock on the target.

Some participants moved the smartphone too fast, even after the experimenters trained them to move the smartphone slowly. Besides the risk of motion blur, moving the phone too fast may lead to a situation such that, by the time a speech feedback utterance is completed, the directions given in the utterance are no longer appropriate. This is because the smartphone moved too far between the time the image generating the feedback was acquired and the time the feedback utterance is completed. The problem is exacerbated at slow frame rate, where it may lead to "limit cycles," with the user rotating the phone periodically left to right and right to left, unable to find the precise direction to the marker.

When the target was lost, some participants were able to scan the scene methodically to rediscover the target, while others searched at random.

Hand-Body Coordination—Most participants move the smartphone "forward" with respect to their body, not to the optical axis. This is the main reason why, when the smartphone is very tilted, it is easy to lose the marker. The correct motion, when the phone is tilted but well centered (off-axis angle is small), would be along the optical axis.

When moving towards the target, participants have to decide when to take steps. Some took very small steps (which seems to be an effective strategy), while some move their arm and

then take a longer step. Clearly some were better than others at coordination. Some wait until the marker is well centered before taking a step (which also seems to be an effective strategy).

Feedback from Participants

This section summarizes the qualitative feedback offered by the participants at the end of the experiment.

General Feedback—Several participants expressed the desire for more feedback from the system, including feedback when the camera is too close to the target. One participant wanted continuous feedback, and said: “When I don’t hear anything, I feel I am lost”. A few participants noted that vibration feedback or the use of earphones/bone conduction headphones might be useful in a noisy setting, e.g., outdoors in an urban area. Some also suggested that a Google Glass-type interface might be more natural than the smartphone interface; another tried a body scanning strategy, in which the smartphone was held against the chest, which made the camera pose more stable but resulted in unwanted camera tilt. Finally, a few participants pointed out that the word “turn” in the “turn left/right/etc.” utterances is superfluous and should be omitted.

Feedback on Settings—Regarding the FOV, some participants said the fisheye lens provides more information and allows more rapid detection, but others said they preferred the normal lens, in one case because the fisheye “requires more filtering”; one person preferred the narrow lens when closer to the target because it provided more specific localization. Opinion was divided on the frame rate, with most participants preferring the higher rate of information provided by the normal frame rate (and disliking the wait imposed by the slow frame rate), yet a few preferring the slow frame rate either because the pace felt more comfortable or because there are “too many beeps” in the fast mode, causing worry when the system does not beep.

Other Feedback—One participant “Felt like the target was pulling me” while another lamented the fact that “When I walk, I feel that I lose it because I move.” Someone noted that some blind people are taught not to “experiment” with touching and are encouraged to keep their hands to themselves, while others explore freely by touch. Finally, other comments emphasized the need to listen carefully to the training instructions, to stay calm during the trials, and to pay close attention to their surroundings – underscoring how challenging the search tasks in these experiments are for many blind persons. (One person, however, went so far as to say that he wanted his own version of the system to use!)

DISCUSSION

Our experiments have reinforced the inherent difficulty in guiding a blind person to a target using acoustic feedback from a hand-held camera phone system. In the easiest scenario considered, it took our participants an average of more than 13 seconds to complete the task, which involved moving the phone by just 70 cm. The difficulty of the “last meter” guidance can be explained in simple geometric terms, as shown in Fig. 3. Simply put, *maintaining in-FOV visibility* (that, is, orienting and moving the camera so that the target remains in the

FOV) *becomes more challenging at closer distances*. Increasing the camera's FOV may appear at first to be a simple solution to maximize visibility and thus facilitate guidance. However, wider FOV comes at the cost of reduced angular resolution and thus reduced image resolution (size) of the target. In practice, this means that, with a wider FOV, one needs to get closer to the target before the system can detect and recognize it.

This work is, to the best of our knowledge, the first investigation into the non-trivial trade-off between spatial resolution and FOV. Our simple geometric analysis of this trade-off in different scenarios of usage is able to justify in good part the experimental results with our blind participants. Our results show that, in terms of execution time, *wide FOV is preferable to narrow FOV only when neither spatial resolution nor precise pointing of the camera towards the target is important*. If, however, a certain minimum spatial resolution is required, narrow FOV results in shorter average execution time, even though the out-of-FOV fraction is (as expected) consistently higher with narrow FOV. It is important to observe that, when the target is out-of-FOV, the system does not produce any feedback. This led to sporadic situations with a participant "losing" a target for some time, and having to explore the scene moving the camera around for a certain amount of time before the target is re-acquired. In a few cases, the participant was not able to re-acquire the target before time-out.

Concerning the FPS setting, the experiments showed that in most scenarios a slower FPS increases execution time. This was obviously expected; perhaps the biggest surprise was that, even with such a slow frame rate (one frame every other second), participants were by and large able to complete the task in the allotted time. Interestingly, the out-of-FOV fraction seems largely independent of the FPS, and in some cases a lower FPS makes for a lower miss rate.

CONCLUSION

We have investigated the effects of two important constraints in object recognition technology, frame rate (FPS) and camera field of view (FOV), on the ability of blind users to search for visual targets and acquire well-framed images of them. The results of the analysis show that, while increasing the FPS generally improves search performance, in many cases increasing the FOV does not help, and may even hurt, performance. We hypothesize that an increased FOV confers a mixture of benefits and drawbacks: while it may help the user find the target from a distance, up close it provides less localization information than a narrow FOV and may therefore hinder the user's attempts to acquire a well-framed (and approximately head-on) image.

Although our experimental results are specific to the chosen apparatus, these trade-offs may have important implications for the design of any object recognition system for blind or visually impaired users. In general, a wide FOV is only preferable to a narrow FOV when neither spatial resolution nor precise camera pointing towards the target is important. A faster FPS generally improves search performance, but in many practical situations FPS is increased by reducing the image resolution, and this must be weighed against the cost of having to approach the target closer to resolve it.

While the current study assumed a fixed UI configuration, in the future we will also study how the UI may be jointly optimized with system parameters such as FPS and FOV. In particular, we will investigate how the UI can be optimized for a wide FOV, to provide more specific feedback about the target's location in the image and thereby speed the process of centering the target.

We will also investigate the role of other variables in search and target framing performance, such as how often (and for how long) a user loses "sight" of the target by the system, and what the user can do to recover from this setback as gracefully as possible. Finally, there is growing interest in new form factors such as a wireless camera mounted on the eyeglasses (as in Google Glass), which could facilitate entirely new visual search strategies and behaviors, and we plan to study these form factors in the future.

Acknowledgments

The project described was supported by Grant Number 1R21EY021643-01 from NEI/NIH. David Vásquez and Corinne Olafson are acknowledged for extensive assistance with participant recruitment and experimental procedures.

References

1. KNFB reading technology. http://www.knfbreader.com
2. Sight On Call. Blindsight, Inc; http://blindsight.com/sight-on-call-tm
3. Text Detective. http://blindsight.com/textdetective
4. Bagherinia H, Manduchi R. Robust real-time detection of multi-color markers on a cell phone. Journal of Real-Time Image Processing. 2011; 6
5. Bigham, J., et al. Vizwiz::LocateIt — Enabling blind people to locate objects in their environment. Proc CVAVI'10;
6. Bigham, JP., et al. VizWiz: Nearly real-time answers to visual questions. Proc ACM UIST ' 10;
7. Coughlan J, Manduchi R. Functional assessment of a camera phone-based wayfinding system operated by blind and visually impaired users. International Journal on Artificial Intelligence Tools. 2009; 18(3):379–397. [PubMed: 19960101]
8. Crandall, W.; Marston, J. Development, evaluation, and lessons learned: A case study of Talking Signs remote infrared audible signage. In: Manduchi, R.; Kurniawan, S., editors. Assistive Technology for Blindness and Low Vision. CRC Press; 2013.
9. Golledge R, Marston J, Loomis J, Klatzky R. Stated preferences for components of a personal guidance system for nonvisual navigation. Journal of Visual Impairment and Blindness, pages. 1998:135–47.
10. Jayant, C.; Ji, H.; White, S.; Bigham, JP. Supporting blind photography. Proc ACM ASSETS ' 11;
11. Kutiyanawala, A.; Kulyukin, V.; Nicholson, J. Teleassistance in accessible shopping for the blind. Proc ICOMP' 11; 2011.
12. Magnusson, C.; Rassmus-Grohn, K.; Szymczak, D. The influence of angle size in navigation applications using pointing gestures. Proc HAID' 10;
13. Manduchi, R. Mobile vision as assistive technology for the blind: An experimental study. Proc ICCHP' 12;
14. Manduchi, R.; Kurniawan, S., editors. Assistive technology for blindness and low vision. CRC Press; 2013.
15. Marston JR, et al. Evaluation of spatial displays for navigation without sight. ACM Trans on Applied Perception. 2006; 3(2):110–124.
16. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychological Methods. 1996; 1(1):30–46.

17. Tekin, E.; Coughlan, J. A mobile phone application enabling visually impaired users to find and read product barcodes. Proc ICCHP' 10;
18. Vázquez, M.; Steinfeld, A. Helping visually impaired users properly aim a camera. Proc ACM ASSETS 12;

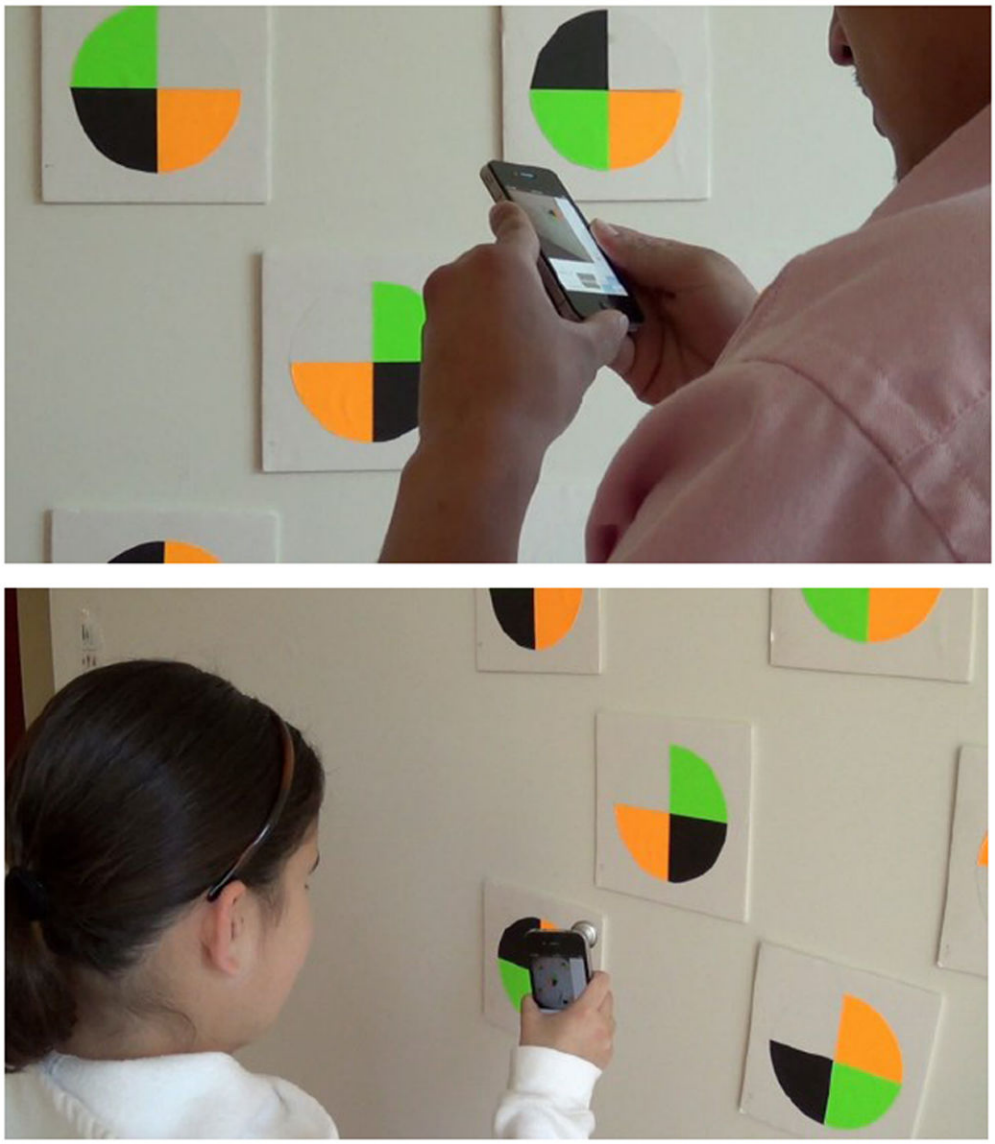

Fig. 1.
Experimental set-up showing two visually impaired participants using an object recognition smartphone app to locate and approach a specific target. Note that the participant shown in the lower picture is using the fisheye lens attachment.

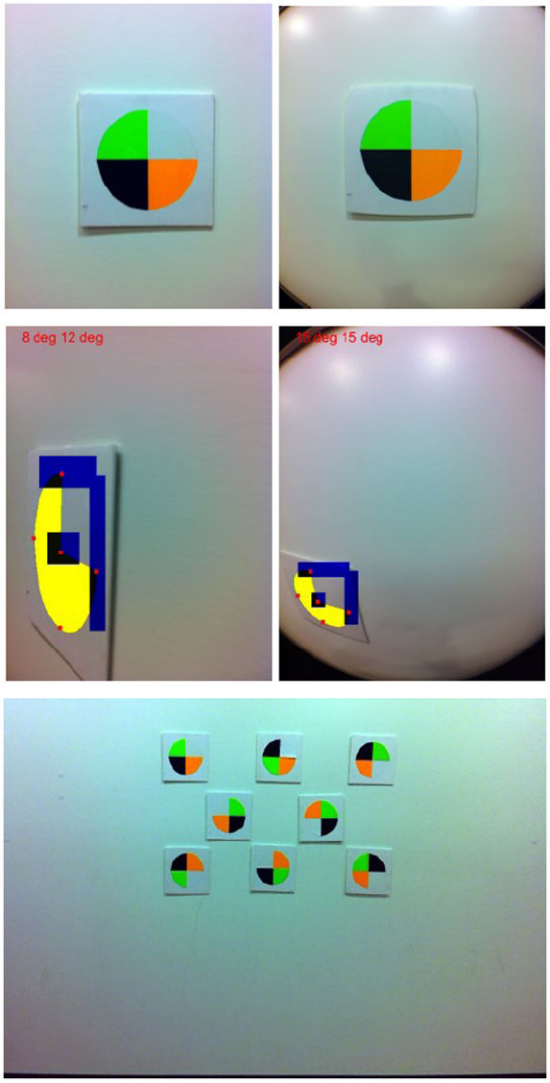

Fig. 2.

Top row: the color marker used in our experiments, seen at a distance of 57 cm with narrow FOV (left) and at 30 cm with wide FOV. Second row: the marker seen at a large slant angle (left) and off-axis angle (right) is still detected by our system. The red dots are the detected keypoints and the yellow pixels indicate the detected color sectors in the marker. Bottom row: the placement of markers on the wall for our experiments.

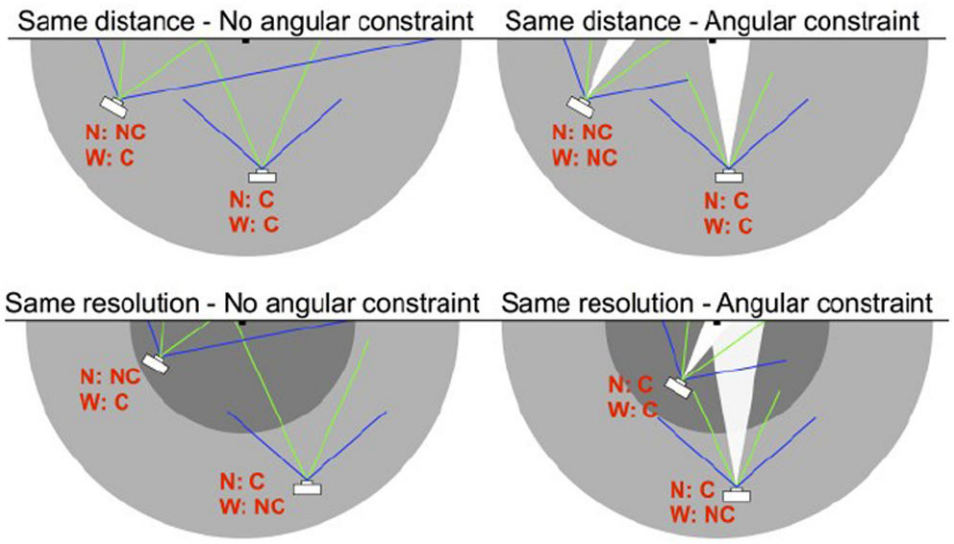

Fig. 3.

Examples of compliant and non-compliant camera poses under all derived scenarios. The edges of the narrow (wide) FOV are shown by green (blue) lines. The compliant angular section is shown by a white angular triangle. The target is represented by a small black rectangle. Regions of space for which D ≤ 57 cm and D ≤ 30 cm (i.e., locations where the camera location is compliant) are represented by the light gray and dark gray semicircles respectively. At each camera position, the first letter in each line of red text represents the FOV (N: narrow; W: wide) while the subsequent letters indicate compliance (C: compliant; NC: non-compliant).

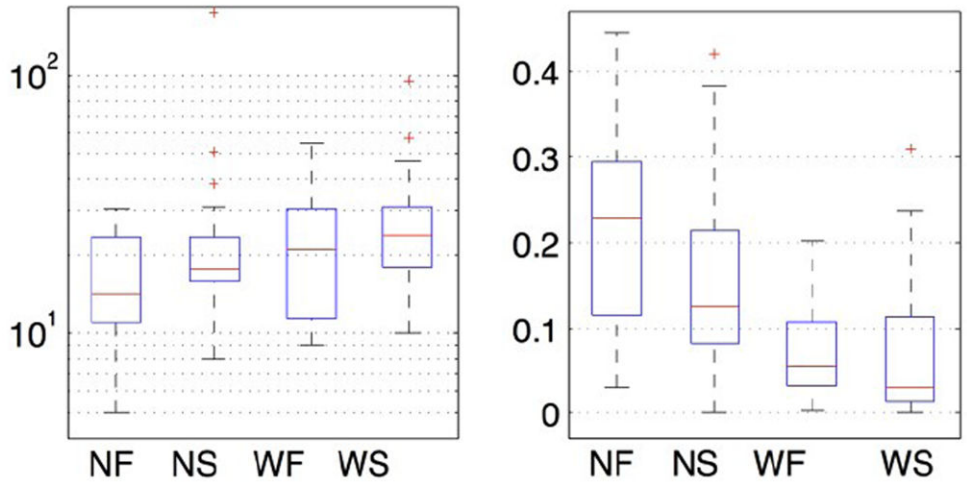

Fig. 4.

Box plots representing the distribution of the time-to-target T, in sec. (left), and out-of-FOV fraction V (right) across participants for the different settings for Scenario 4.

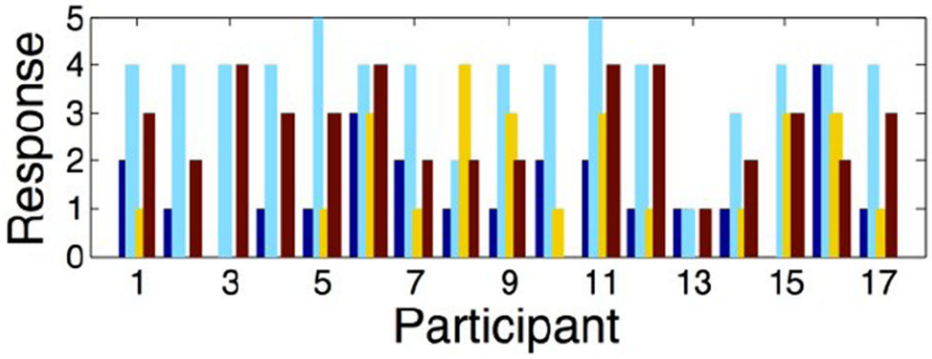

Fig. 5.

Subjective evaluations for the different settings across 17 participants (blue: NF; light blue: NS; yellow: WF; brown: WS).

Table 1

Average time-to-target T values across participants for all scenarios (in units of sec.). If a main effect was found to be significant, the average T for the two levels of the corresponding factor (N,W and/or F,S) is reported (with the level corresponding to the better of the two results in bold). If interaction is also significant, the average T value for a factor level is computed keeping the level of the other factor fixed (e.g., NF and WF) for all factors with a significant simple effect.

	FOV	**FPS**
Scenario 0	NS: 64.1, **WS: 30.1** (p<.00001)	**NF: 26.9**, NS: 64.1 (p<.001)
Scenario 1	N: 23.1,**W: 13.6** (p<.01)	**F: 13.6**, S: 23.1 (p<.01)
Scenario 2		**F: 15.7**, S: 25.2 (p<.001)
Scenario 3	**NF: 14.8**, WF: 20.7 (p<.01)	
Scenario 4	**NF: 16.3**, WF: 23.1 (p<.01)	**NF: 16.3**, NS: 29.7 (p<.01)

Table 2

Average out-of-FOV fraction V values across participants for significant main factors and interactions under different settings considered. (See caption of Tab. 1.)

	FOV	**FPS**
Scenario 0	N: 0.24, **W: 0.08** (p<.001)	
Scenario 1	N: 0.16, **W: 0.08** (p<.001)	F: 0.14, **S: 0.10** (p<.05)
Scenario 2	N: 0.18, **W: 0.08** (p<.01)	F: 0.15, **S: 0.12** (p<.05)
Scenario 3	N: 0.16, **W: 0.07** (p<.00001)	
Scenario 4	N: 0.18, **W: 0.08** (p<.00001)	