Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Investigating the relationships between user behaviors and tracking factors on task performance and trust in augmented reality

Matt Gottsacker[a,*], Zubin Datta Choudhary[a], Hiroshi Furuya[a], Austin Erickson[a], Ryan Schubert[a], Gerd Bruder[a], Michael P. Browne[b], Gregory F. Welch[a]

[a] *University of Central Florida, Orlando, FL, USA*
[b] *Vision Products, LLC, Campbell, CA, USA*

ARTICLE INFO

Article history:
Received 16 April 2024
Received in final form 2 August 2024
Accepted 2 August 2024
Available online 6 August 2024

Keywords: Augmented Reality, Tracking Factors, Trust

ABSTRACT

This research paper explores the impact of augmented reality (AR) tracking characteristics, specifically an AR head-worn display's tracking registration accuracy and precision, on users' spatial abilities and subjective perceptions of trust in and reliance on the technology. Our study aims to clarify the relationships between user performance and the different behaviors users may employ based on varying degrees of trust in and reliance on AR. Our controlled experimental setup used a 360° field-of-regard search-and-selection task and combines the immersive aspects of a CAVE-like environment with AR overlays viewed with a head-worn display.

We investigated three levels of simulated AR tracking errors in terms of both accuracy and precision (+0°, +1°, +2°). We controlled for four user task behaviors that correspond to different levels of trust in and reliance on an AR system: *AR-Only* (only relying on AR), *AR-First* (prioritizing AR over real world), *Real-Only* (only relying on real world), and *Real-First* (prioritizing real world over AR). By controlling for these behaviors, our results showed that even small amounts of AR tracking errors had noticeable effects on users' task performance, especially if they relied completely on the AR cues (AR-Only). Our results link AR tracking characteristics with user behavior, highlighting the importance of understanding these elements to improve AR technology and user satisfaction.

1. Introduction

Augmented reality (AR) technologies have seen major advances over the last decade in terms of their displays, sensors and tracking, and networking capabilities [1]. Especially head-worn displays (HWDs) are becoming increasingly attractive for a wide range of application domains, including architectural design, urban planning, simulation and training, and defense [2, 3, 4].

One of the key challenges in creating effective AR HWD systems in outdoor, dynamic, or generally uncontrolled environments and settings is maintaining accurate and precise head tracking so that virtual content can be properly placed in the real world [5, 6, 7]. Tracking the position of the physical objects to which the virtual content is registered poses additional challenges [8, 9, 10], but we focused on isolating the AR HWD head tracking errors in this study. Tracking technologies in AR have improved significantly over the last two decades by fusing sensor data from inertial measurement units [11], RGB and

[*] Corresponding author
e-mail: mattg@ucf.edu (M. Gottsacker)

depth cameras [12], skyline sensors [13], and related SLAM algorithms [14, 15, 16], cooperative methods [17], time warping [18], and other approaches. However, head tracking performance in AR is still often inadequate for registering virtual entities within the space around AR users, e.g., for training purposes, or fixing annotations to real objects. When relying only on head tracking for the placement of AR objects, even one degree of orientation error with respect to the user's actual head orientation may lead to an AR object appearing in the wrong location, or an AR annotation appearing over the wrong real object, especially for distant objects. This is particularly true for optical see-through (OST) HWDs [19], while video see-through HWDs often register AR overlays to the real world based on the video feeds of front-facing cameras which simplifies the registration problem [20].

The challenges include both the *accuracy* of the head orientation estimates and the *precision* of the orientation estimates. These two distinct types of head tracking errors have different effects on human perception, which may impact users' spatial task performance and/or affect their subjective sense of being able to rely on or trust the AR system [21, 22]. Further, while head tracking errors are comparatively easy to model and evaluate in virtual reality settings, their effects on users' perception with OST HWDs are more complicated [21]. For one, head tracking errors can introduce a perceptible discrepancy between the real scene and the AR content, which may affect the way users would leverage AR technologies for the completion of spatial tasks. Due to the complex relationships between the AR tracking accuracy and precision, and the way users would rely on and trust AR cues in a specific task context, we decided to evaluate these interrelations by evaluating and controlling for different behaviors when completing a visual search-and-selection task.

In this paper, we present two experiments where we investigated how tracking errors (accuracy and precision) affect OST HWD users' subjective assessment of the technology in terms of their trust in and reliance on the AR system as well as their objective performance (time and errors) during a 360° search-and-selection task. In these experiments, we also investigated users' performance with respect to different task behaviors. We describe both experiments in detail, including an analysis of the objective and subjective data, supporting the following findings:

- Effects of the AR head tracking accuracy and precision on participants' objective task performance: Each additional degree of accuracy/precision error led to an increase in error.

- Effects of the AR head tracking accuracy and precision on participants' subjective trust in and reliance on the AR system: Each additional degree of accuracy/precision error led to a decrease in trust.

- Interactions between participants' task behaviors and the AR tracking factors with respect to task performance: For +1° and +2° accuracy/precision angular offsets, relying completely on AR led to decreased performance compared to other task behaviors.

In particular, we found significant issues among objective task performance that were introduced by even small amounts of AR tracking errors. However, we also found that these issues were largely dependent on the task behaviors employed by the participants in our experiments, which implies that AR cues may be effective and helpful for users wearing OST HWDs even if the tracking accuracy and precision are not optimal.

This paper is structured as follows. Section 2 discusses background information in the scope of this paper. Section 3 describes the general experimental method we used for our two experiments. Sections 4 and 5 describe our two experiments, in which we evaluate the effects of AR head tracking accuracy and precision, respectively. Section 6 provides a general discussion of our findings. Section 7 concludes our paper.

2. Background

In this section, we provide background information on trust and related task behaviors, as well as AR tracking accuracy and precision.

2.1. Trust and Behavior with Registered AR Cues

OST HWDs are an attractive technology for various application domains due to their ability to present registered visual cues in a real environment. Such cues may be textual or symbolic annotations for objects or entities in the real environment, often providing orthogonal or redundant information to that which users can gain from looking at the real scene. In particular, there are situations in which OST HWDs can present information about real objects in a way that is easier to see and comprehend than the visual cues from the real world [23]. A basic example are AR object annotations, which can be designed to be visible, salient, and easy to understand, making it possible for AR users to find objects that otherwise would require extensive searching for small, hidden, or unclear information if users were to just rely on the real world, especially when under time pressure or in cluttered or unfamiliar environments [24, 25, 26]. Though technically redundant information is provided to users in these cases, meaning they could solve these tasks by relying only on the real world, AR tags afford another information channel that may have cognitive and task performance benefits. However, such AR task contexts are difficult to assess as both the subjective and objective results depend on the users' task behaviors, which in turn depend on their impression of the AR system. Specifically, AR users may choose to rely only on the AR cues, integrate both the AR and real-world cues (e.g., double-check the AR cues by confirming their information with the real world, or vice versa), or choose to ignore the AR cues entirely, depending on how much they trust this information channel. In this paper, based on these different approaches, we modeled and evaluated four behaviors for the completion of a 360° visual search-and-selection task (illustrated in Figure 4), each behavior differing in which source of information is prioritized and whether users fully trust it or not.

From related work we know that a user's trust in and reliance on an AR system depend on a multitude of factors related to the system, environment, and task context, which historically have been related to a perceived increase or decrease in performance [23, 27, 28]. In particular, a user may perceive the system to be unreliable, which can lead to a loss of trust, and eventual disuse of the AR system. Further, these situations may lead to users *undertrusting* the AR system, which may cause them to change their behavior by trying to confirm all AR information, potentially leading to worse performance than could be gained otherwise [29, 30]. Conversely, if users *overtrust* the AR system, it may lead to overuse and potentially increase errors than if they had a more well-calibrated trust in the system [31].

Recent work recognizes the importance of user behavior, reliance, and trust in relation to the performance of the AR system. Misfud et al. applied these concepts to joint terminal attack controller operations, which are extremely intolerant to error, revealing a potential for automation bias, a classic result of over-reliance on the AR system [32]. Other recent work demonstrated that even if AR cues occasionally highlight the wrong object, perceived effectiveness of the AR cue leads to over-reliance and elevated risk for error [25]. In our work, rather than assuming errors in the cue targeting itself, we investigate the effects of errors in registration, i.e., the cue points at the correct target but the AR system itself exhibits errors in properly aligning the cue with the physical world.

Evaluating these effects on reliance, trust, and overall system performance can be difficult, however, due to the complexities of human behavior. For example, users' reliance and trust in the system can change over time, leading to a subsequent change in task behavior and system performance. Further, external factors such as time pressure or distractions may induce behaviors that are influenced by trust and reliance but may deviate significantly from task behavior that would otherwise be expected. To better isolate the effects of registration errors on trust, reliance, and overall performance of AR systems, we chose to control for task behaviors by explicitly instructing participants to perform the task following procedures defined by the four aforementioned task behaviors. In doing so we introduce novel research questions related to how different behaviors, corresponding to different levels of system trust, affect AR cueing performance. This method also allowed us to quantify the performance and perceptual effects of different tracking errors at different levels.

2.2. AR Tracking Factors

The visual association of AR annotations with real world objects requires continuous information about the geometric relationship between both the AR HWD and the real world object. This geometric relationship is typically maintained by continuously estimating the position and orientation (pose) of the AR HWD relative to the origin of a three-dimensional coordinate system [6]. This process is often referred to as *head tracking*, and it requires access to external information about the position, orientation, and size of the real world object of interest, all in the same coordinate system with the same units.

In theory, access to all of this geometric information should allow one to accurately *register* the AR cues to their real world counterparts as seen in the HWD. Indeed, over the last decade tracking technologies have improved significantly, making AR systems in general useful in more circumstances, e.g., indoor work environments [33]. For reviews of existing tracking technologies, and discussion about the various types and sources of error, see [34, 35, 36, 37]. In practice, the registration of AR cues with real world objects presents a significant challenge [38], in particular with respect to head tracking for OST HWDs in outdoor, dynamic, or generally uncontrolled environments. The challenges include both the *accuracy* of the head pose estimates — the difference between the mean of the pose estimates and the true head pose (which is unknown) over a window of time, and the *precision* of the pose estimates — the variance of the pose estimates about the mean over the same window of time [39, 40, 41]. See Figure 1 for examples of both accuracy error and precision error on AR registration. The challenges persist even with systems employing multiple tracking modalities and sensor fusion paradigms, for example as afforded by the Kalman filter and related variants [42, 43, 44, 45, 46]. Livingston et al. [22] showed that precision errors in registration led to decreased performance when visually tracking distant objects, and accuracy errors led to more errors. Out of all the registration errors tested, participants reported precision errors were the most detrimental to their performance, even though this was not supported by their performance data [22]. Robertson et al. [47] found that accuracy and precision errors led to reduced performance on a block stacking task in terms of increased errors and task time, and that errors reduced users' confidence in their task performance. For certain AR surgical applications where AR users are viewing small 3D objects positioned within reach, positioning accuracy mean errors of less than 5 mm are considered acceptable [48, 49]. However, it is not known what is considered acceptable for AR annotations displayed at longer distances from the user.

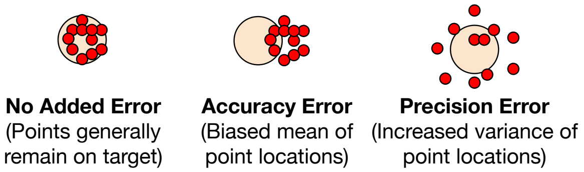

No Added Error
(Points generally remain on target)
Accuracy Error
(Biased mean of point locations)
Precision Error
(Increased variance of point locations)

Fig. 1. Examples illustrating the accuracy and precision tracking errors we investigated in our two experiments, respectively. In the scope of this paper, head tracking accuracy refers to a deviation in the mean among sampled head tracking orientations, while precision refers to the variance of these sampled head tracking orientations. In this illustration, the orange circles denote the ground truth and the red circles denote the sampled head tracking orientations.

In this paper we focus on the effects of head tracking *orientation* accuracy and precision under conditions when the user's head position is relatively static. Our work contributes to the body of tracking-related AR research by investigating the effects of tracking-related registration errors on users' task performance and trust in the AR system in a search-and-selection task. To complement this work from a human factors perspective, we also investigated how different task behaviors that users would adopt when experiencing different levels of trust in and reliance on the system affect their task performance.

3. General Experimental Method

We conducted two experiments to investigate the relationships between user behaviors and tracking factors on participants' task performance and trust. Each of these experiments utilized the same experimental configuration (participants, materials, procedure, etc.). The first experiment focused on tracking accuracy. The second experiment focused on tracking precision. In this section, we describe our general experimental method as well as how the experiments differ. Our experimental procedure and recruitment of participants in Experiments E1 and E2 were approved by the institutional review board of our university under protocol number ANONYMIZED.

3.1. Participants

We recruited 20 participants from our university community for these experiments, and we obtained participants' demographics using ACM's DIE Demographics Questionnaire[1]. Unfortunately, we had to remove 2 data sets from further analysis due to the data failing our internal sanity checks. Specifically, we observed a significant error rate among these two participants, bordering on chance level, pointing to the participants having misunderstood the task. The remaining 18 participants included 15 males, and 3 females, with ages between 18 and 41, $M = 23.6$, $SD = 6.7$. All of the participants had normal or corrected-to-normal vision and none of them reported any visual or vestibular disorders, such as color or night blindness, dyschromatopsia, or a displacement of balance. None of the participants reported any motor or cognitive disabilities. The participants were either students or non-student members of our university community who responded to open calls for participation and received monetary compensation for their participation. Both experiments together took participants on average 60 minutes to complete.

3.2. Material

Figure 2 shows the hybrid interaction space where we conducted our study. We used a CAVE-like 4 m×4 m square immersive projection environment to *simulate* the appearance of the real world, and an OST AR HWD to visually superimpose AR imagery over that "real world" imagery. Specifically, each of the four walls was covered from edge to edge with imagery from NEC U321H ultra-short throw projectors at a resolution of 1080p per wall, or 7680×1080 pixels total resolution. Further, participants wore a professional Vision Products SA-147/S OST HWD. The AR tags were presented via this AR HWD, while they were registered with and presented spatially above the heads of the simulated humans that were shown on the walls of the CAVE-like space. The Vision Products SA-147/S HWD includes four OLED microdisplays for a resolution of 3840×1200 per eye, with a 33° vertical FOV, a 143° horizontal FOV, and a 53° binocular overlap.

The 6 degrees of freedom (DoF) head pose was determined via a Vive Tracker 3.0 mounted on the HWD, along with two SteamVR Base Station 2.0 units mounted in opposite upper corners of the interaction space. We used a hand-held Vive Pro controller, which was also tracked in 6 DoF and provided point-and-click input from the participants. Rendering and experiment control were performed with the Unity engine (version 2021.3.2), which used SteamVR to receive the tracking data from the HWD and hand-held controller. The hybrid simulation environment was driven by a single BOXX APEXX X3 desktop workstation with two Nvidia Quadro RTX 6000 graphics cards.

Registration Calibration. Because our investigation is based on the participants' perceptions of angular accuracy and precision, we needed to ensure our baseline conditions exhibited no perceptible error. To do so we had each participant start with a self-calibration procedure under the supervision of the experimenter. Specifically they were presented with a grid of small spheres projected on a wall of our immersive projection environment, and an identical grid of spheres appearing in their AR HWD. They then used software controls to adjust the alignment of the two grids until any differences in alignment were imperceptible. At the completion of this self-calibration procedure we confirmed that there was no perceptible misregistration error. All of our subsequent simulated errors in registration accuracy and precision involved perturbations from this baseline.

We are confident that our baseline accuracy and precision persisted throughout the experimental tasks for each participant, based on our own pre-post tests, and because in a similar setup, Spitzley & Karduna found that a Vive tracker undergoing a series of rotations and translations maintained an angular variance of below one degree compared to a stable reference [50]. The rotations and translations of our participants' heads were very small compared to those of Spitzley & Karduna.

Hybrid Simulation Environment. The simulated environment that was presented to participants via the CAVE-like installation consisted of a desert landscape with ten clusters of three simulated humans sitting together in an evenly spaced-out circle, ten meters away from the center of the simulation space, with the center of the clusters at regular 36° intervals around the participant (see Figure 2). Each simulated human had a subtly colored arm band that was either red or blue to separate them into two groups (red team vs. blue team). In each cluster, one of the three humans had a red arm band while the other two had a blue arm band. The order of the colors was assigned randomly. The scene displayed on the projected walls was perspective-correct based on the user's tracked head position. The AR tags presented on the OST HWD were positioned to match the simulation. The corresponding accuracy and precision tracking offsets were applied to these and only to these AR tags, i.e., they were not applied to the head-tracked simulation of the CAVE-like environment around participants.

Tracking Error Simulation. To simulate *accuracy* errors as illustrated in Figure 1 (middle) we applied a condition-dependent horizontal offset to the apparent real world position of each AR tag. We chose to only perturb the horizontal positions because the corresponding humans rendered in the simulated real world

[1] https://community.acm.org/demographics/

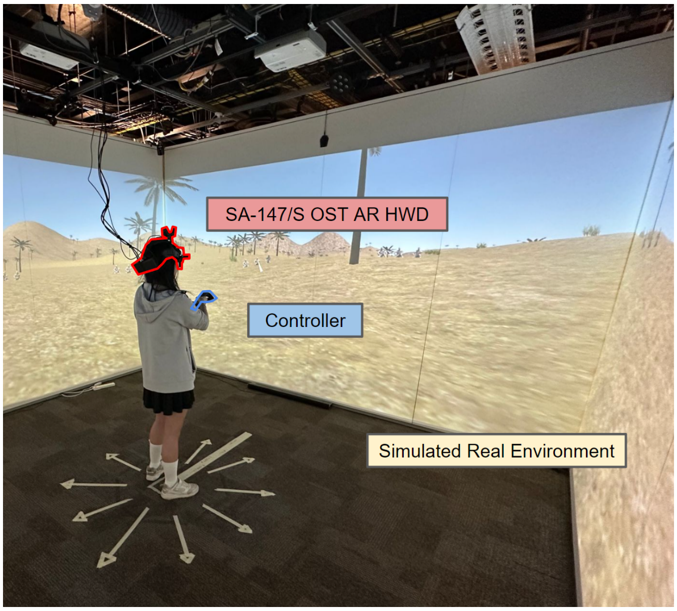


Fig. 2. Annotated photo showing our hybrid experimental space, in which we used a CAVE-like immersive projection environment to simulate a real space around our participants. As part of the simulated scenario, we placed ten clusters of three sitting humans in a circular pattern around the center of the simulation space (see Figure 3). Participants in our experiment further wore a professional Vision Products SA-147/S OST HWD, through which we presented the AR tags to participants. These and only these AR tags were affected by the simulated levels of head tracking accuracy and precision. These AR tags consisted of red diamonds that floated over the heads of some of the simulated humans. When the simulated tracking accuracy and precision were high, these AR tags were close to the corresponding simulated humans, while they were presented with increasing angular offsets in the other conditions.

environment were arranged horizontally. Specifically, we applied a condition-dependent fixed positive or negative offset of 0°, 1°, or 2° to the nominal horizontal angle of each AR tag as seen from the viewer's perspective in the HWD, in the world coordinate frame. The result was a condition-dependent horizontal shift of the perceived position of each AR tag relative to the associated human in the simulated real world imagery as illustrated in Figure 3 (a)–(c). All offsets were applied relative to each participant's individual baseline of imperceptible misregistration, determined during the initial registration calibration described above. All offsets were held constant throughout each trial, e.g., an angular offset of +2° meant an AR tag for a given selection target was always 2° to the right of its target during the trial as illustrated in Figure 3 (a).

To simulate *precision* errors, we followed a similar procedure to that described above for the accuracy errors, however rather than applying a constant integer offset for each trial, we applied an angular perturbation chosen randomly (uniform distribution) from a condition-dependent real number interval (e.g., {0}, [−1.0, +1.0], or [−2.0, +2.0]) at each rendering frame, throughout the entire trial. The effect was to produce a "jitter" of the perceived position of each AR tag relative to the associated human in the simulated real world imagery as illustrated in Figure 3 (d)–(f). Again, all angular perturbations were applied relative to each participant's individual baseline of imperceptible misregistration, determined during the initial registration calibration described above.

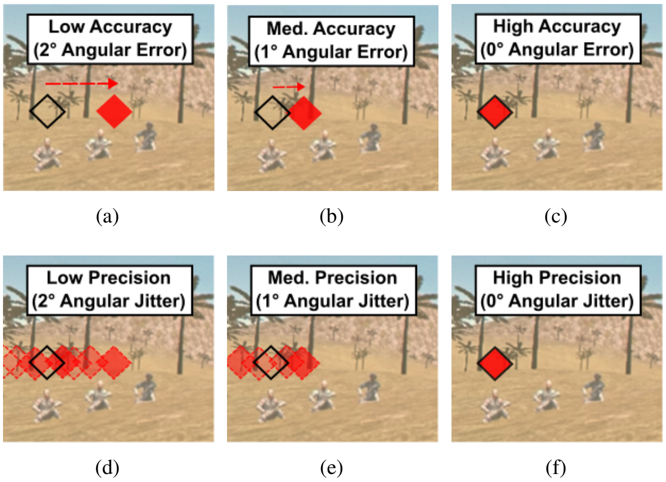
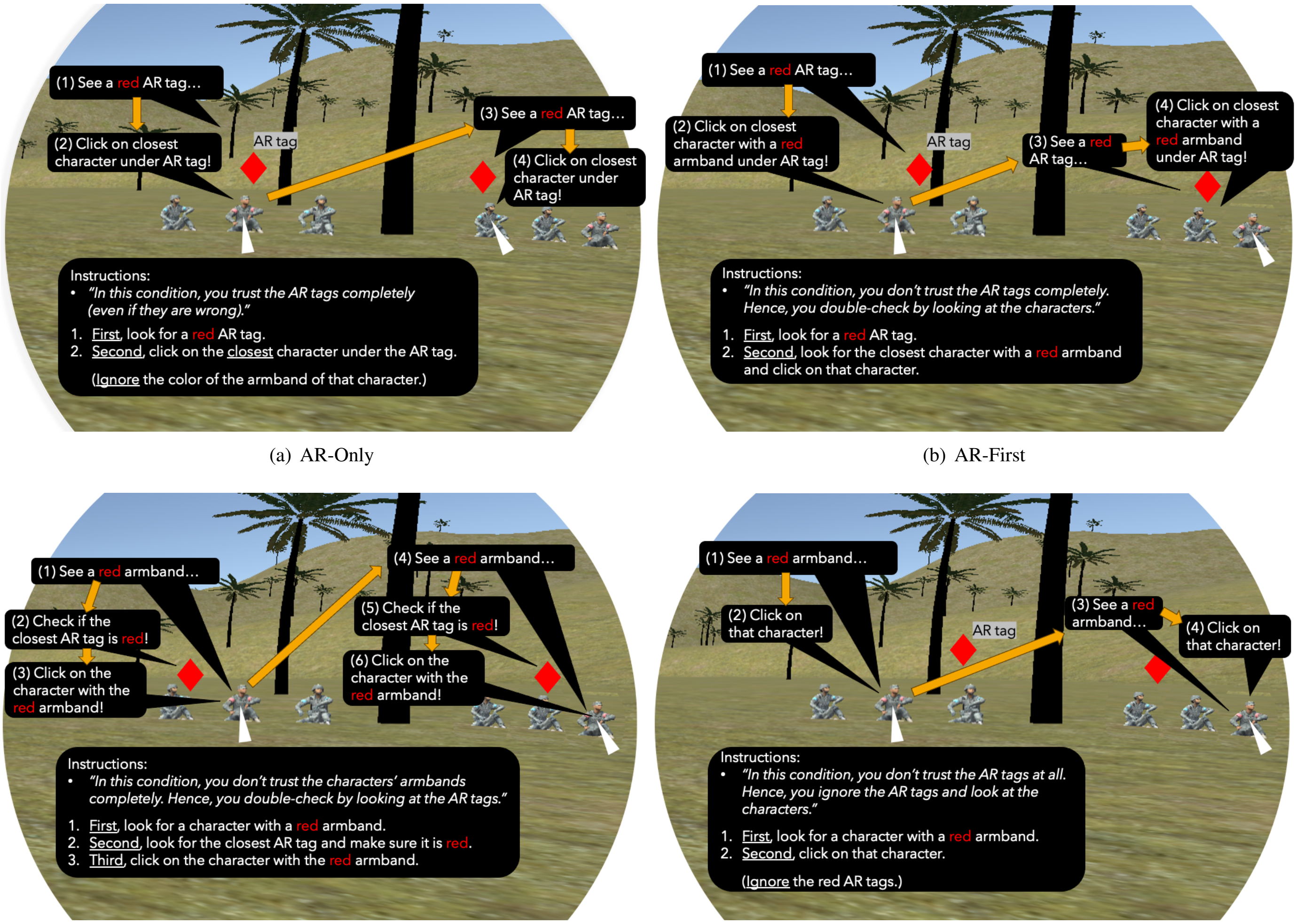

Low Accuracy (2° Angular Error)
Med. Accuracy (1° Angular Error)
High Accuracy (0° Angular Error)
(a)
(b)
(c)
Low Precision (2° Angular Jitter)
Med. Precision (1° Angular Jitter)
High Precision (0° Angular Jitter)
(d)
(e)
(f)

Fig. 3. (a-c) Experiment E1 (Accuracy): Illustration of the different accuracy levels participants experienced. The black outlined diamond represents where an AR tag would be positioned with zero error (perfect accuracy), while the solid diamond represents what a participant might have actually seen for each level of accuracy. This offset was consistent frame-by-frame. (d-f) Experiment E2 (Precision): Illustration of the different precision levels participants experienced. The black outlined diamond represents where an AR tag would be positioned frame-by-frame with zero error (maximum precision), while the dashed line diamonds with varying degrees of opacity represent the positions at which the tag was rendered across several frames (more opaque = more recent). The actual visual effect was that frame-by-frame, the tag appeared to jitter around its target location within the range of the precision condition's angular error.

3.3. Methods

We used a 3 × 4 within-subjects design for each of our experiments with the following tracking factors and task behaviors. Experiment E1 investigated the accuracy tracking factor — the difference between the mean of the orientation estimates and the true head orientation (which is unknown) over a window of time, and Experiment E2 investigated the precision tracking factor — the variance of the orientation estimates about the mean over the same window of time. Both experiments investigated the four different task behaviors.

- **Accuracy (3 levels):** As described in Section 3.2, we simulated three different tracking accuracy levels of +0°, +1°, and +2° of horizontal angular offset.
- **Precision (3 levels):** As described in Section 3.2, we simulated the three different precision range levels of 0°, ≤1°, and ≤2° of horizontal angular jitter.
- **Task Behaviors (4 levels):** Our study modeled four distinct behaviors for completing the visual search-and-selection task. These behaviors reflect varying degrees of trust in and reliance on AR for task assistance (see Figure 4 for examples):
  - *AR-Only*: Participants focused on the AR tags and disregarded the arm bands while completing the task.


(1) See a red AR tag…
(2) Click on closest character under AR tag!
AR tag
(3) See a red AR tag…
(4) Click on closest character under AR tag!

Instructions:
- "In this condition, you trust the AR tags completely (even if they are wrong)."

1. First, look for a red AR tag.
2. Second, click on the closest character under the AR tag.

(Ignore the color of the armband of that character.)

(a) AR-Only

(1) See a red AR tag…
(2) Click on closest character with a red armband under AR tag!
AR tag
(3) See a red AR tag…
(4) Click on closest character with a red armband under AR tag!

Instructions:
- "In this condition, you don't trust the AR tags completely. Hence, you double-check by looking at the characters."

1. First, look for a red AR tag.
2. Second, look for the closest character with a red armband and click on that character.

(b) AR-First

(1) See a red armband…
(2) Check if the closest AR tag is red!
(3) Click on the character with the red armband!
(4) See a red armband…
(5) Check if the closest AR tag is red!
(6) Click on the character with the red armband!

Instructions:
- "In this condition, you don't trust the characters' armbands completely. Hence, you double-check by looking at the AR tags."

1. First, look for a character with a red armband.
2. Second, look for the closest AR tag and make sure it is red.
3. Third, click on the character with the red armband.

(c) Real-First

Instructions:
- "In this condition, you don't trust the AR tags at all. Hence, you ignore the AR tags and look at the characters."

1. First, look for a character with a red armband.
2. Second, click on that character.

(Ignore the red AR tags.)

(d) Real-Only

Fig. 4. Illustrations of the four task behaviors we evaluated in this experiment. We used these illustrations to explain the behaviors to our participants: (a) *AR-Only*: participants were asked to scan the 360° environment for AR tags while completely disregarding the arm bands and selecting the character below the AR tag; (b) *AR-First*: participants were asked to scan the environment for AR tags but double-check the arm band to select the character with the red arm band near the AR tag; (c) *Real-First*: participants were asked to scan the environment for arm bands but double-check the nearest AR tag each time they identified a potential target based on a red arm band; (d) *Real-Only*: participants were asked to scan the environment for arm bands while completely disregarding the AR tags.

- *AR-First*: When making a selection, participants initially looked at the AR tag and then double-checked their impression by looking at the simulated humans' arm bands.
- *Real-First*: When making a selection, participants initially looked at the simulated humans' arm bands and then double-checked their impression by looking at the AR tag.
- *Real-Only*: Participants focused on the simulated humans' arm bands and disregarded the AR tags while completing the task.

Before and after each trial, we incorporated two **baseline** assessments where participants wore the AR HWD but it did not display any AR imagery. In these baselines, participants relied solely on the simulated real environment for decision-making. The primary purpose of these baselines was to serve as a control mechanism for evaluating potential learning effects throughout the study.

We tested all task behaviors in randomized order. For each task behavior, we further randomized the order of the tested tracking factor conditions. Each condition was tested twice, once while rotating clockwise and once in counterclockwise direction to avoid that our participants get entangled in the cables that were suspended from the ceiling.

3.4. 360° *Search-and-Selection Task and Behaviors*

Participants were tasked with swiftly scanning the entire 360° environment within the experimental area, identifying and selecting all simulated humans marked as red while refraining from selecting those marked as blue. Simulated humans were distinguished by colored arm bands and AR tags above their heads. Only the "red team" had AR tags appear above their heads. These AR tags, presented via the SA-147/S AR HWD, took the form of a red diamond.

As detailed in Section 2.1, the design of stimuli and tasks mirrored applications utilizing AR spatial cues integrated with real objects to aid users in spatial tasks. Despite AR cues technically offering redundant information compared to environmental cues, they benefit from their clear visibility, salience, and ease of comprehension.

Baseline trials conducted before and after each condition omitted AR tags, requiring participants to rely solely on colored arm bands, simulating a scenario without an AR display.

Participants were timed for each trial (comprising one 360° sweep) and instructed to complete the task swiftly and accurately. Each trial started with a click on a ”start” button and ended with a click on an ”end” button.

The characteristics of AR imagery varied across experimental conditions. Given the multitude of variables influencing AR users' behavior, participants were given explicit instructions regarding their expected behaviors as illustrated in Figure 4.

3.5. Objective Data

We assessed the duration it took in each trial for participants to conduct a complete 360° sweep of the environment. This measurement encompassed the time from initiating the trial by clicking the “start” button to indicating completion by clicking the “end” button.

Additionally, we recorded the instances of false-negative (Type II errors) where participants missed simulated humans marked red and false-positive (Type I errors) where they selected simulated humans marked blue.

3.6. Subjective Data

We gathered subjective data from our participants by having them complete questionnaires on a laptop immediately after finishing the search-and-selection trials:

TOAST Questionnaire [51]: In the Trust of Automated Systems Test (TOAST), participants read ten statements and indicate on a 7-point Likert scale the extent to which they disagree (1) or agree (7) with each. The TOAST measures “proximate causes of trust in an automated system” using a two-factor structure: system performance and system understanding [51].

- *System Understanding:* Higher understanding scores indicate that users have a high confidence that their trust in the system is well calibrated.
- *System Performance:* Higher performance scores indicate that users trust the system to help them perform their task.

For each of these two subscales, the average of the responses to items in that subscale is computed.

Single Item Questionnaires: We asked participants to rate their perception of the different experimental conditions with respect to the following 7-point, single-item Likert scales:

- *Trust*: On a scale from 1 (Not Trustworthy) to 7 (Very Trustworthy), how much would you trust the AR system?
- *Reliance*: On a scale from 1 (cannot rely) to 7 (very reliable), how much reliance on the AR system would you have?
- *Confidence*: How would you rate your confidence in your performance? (1=low, 7=high)
- *Difficulty*: How would you rate the difficulty of the task? (1=low, 7=high)
- *Advantageous*: On a scale from 1 (Disadvantageous) to 7 (Advantageous), how would you rate the AR system?

3.7. Task Behavior Sanity Check

As discussed in Section 2.1, we modeled four task behaviors (see Figure 4) according to different levels to which participants rely on the AR system when completing the experimental task. This allowed us to isolate the effects of errors in registration accuracy and precision for users experiencing different levels of reliance in the system. To be confident with the results of these analyses (reported in the following sections), we performed a sanity check to see whether the instructed task behaviors corresponded to participants' actual perception of how much they relied on the AR system. We analyzed participants' subjective assessments of their own reliance in the AR system in each of the four behavior conditions. Our results confirmed our assumption, showing that our participants rated the four behaviors in descending reliance order as AR-Only > AR-First > Real-First > Real-Only (see Table 1 for descriptive statistics). In other words, despite using a controlled experimental method, in which we instructed the participants to adopt a specific behavior, it matched participants' estimated reliance of the AR system, thus passing our sanity check.

Table 1. Results for reliance estimates by task behavior.

	Mean	SD
AR-Only	3.69	0.98
AR-First	3.60	1.29
Real-First	3.53	1.56
Real-Only	2.83	1.77

3.8. Procedure

Upon arrival, participants were provided with a physical copy of the informed consent document, which they read and signed along with the experimenter. Participants were escorted to the CAVE-like simulation environment, where verbal consent was obtained to place the SA-147/S AR HWD on their head.

An experimenter adjusted the AR HWD for the participants so that it fit properly and the AR imagery was displayed clearly. Then, participants were instructed to look straight ahead at a simulated horizon line, and a one-time tracker pitch rotation offset was applied (boresight calibration). Then, participants performed the horizontal angular registration calibration described in Section 3.2. Subsequently, the experimenter briefed participants on the search-and-selection task, behaviors, and experimental conditions. Participants were then allowed to practice the task until they felt comfortable performing it.

The experiment comprised baseline tasks conducted before and after a series of tasks that combined accuracy and precision tracking factors and behaviors as described in Section 3.3. Upon completing the selection tasks, the AR HWD was removed, and participants completed questionnaires for each experienced accuracy and precision level. Finally, participants completed a demographics questionnaire, received a debriefing, and were compensated monetarily.

4. Experiment E1: Tracking Accuracy

As described in Section 3.3, Experiment E1 investigated the relationships between tracking accuracy and participants' task behaviors, task performance, and trust. Experiment E1 used a 3 × 4 within-subjects design with the following factors:

- **Accuracy (3 levels):** As described in Section 3.2,we simulated three different tracking accuracy levels of +0°, +1°, and +2° of horizontal angular offset.
- **Task Behaviors (4 levels):** As described in Section 3.3, we modeled four distinct task behaviors corresponding to different degrees of trust in and reliance on the AR system: AR-Only, AR-First, Real-First, and Real-Only.

Experiment E1 involved the participants, materials, methods, task, data, and procedure described in Section 3. The rest of this section describes and discusses our experimental results.

4.1. Results

We analyzed the responses with repeated-measures analyses of variance (RM-ANOVAs) and Tukey multiple comparisons with Bonferroni correction at the 5% significance level. We confirmed normality with Shapiro-Wilk tests at the 5% level and QQ plots. Degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity when Mauchly's test indicated that the assumption of sphericity was violated.

We found no significant difference between the clockwise and counterclockwise trials as well as the pre and post baselines, so we pooled the responses with respect to our analysis.

Objective Data. The descriptive statistics for the elapsed trial times as well as the Type I and Type II errors are shown in Figures 6 and 7. The statistical test results are shown in Table 2.

Our results show significant main effects for the accuracy levels and task behaviors on elapsed time as well as Type I and Type II errors. We further found significant interaction effects for Type I and Type II errors. Specifically, we found that elapsed times were significantly higher (= worse) for the Real-First than the Real-Only condition. Additionally, we found that both Type I and Type II errors were significantly higher for the AR-Only condition than all other task behaviors. Moreover, in the AR-Only condition, the +2° accuracy offset resulted in significantly higher errors than the +0° and +1° accuracy offsets.

Lastly, it is important to note that "Baseline" in Figures 6 and 7, as discussed in Section 3.3, refers to trials conducted without any AR imagery shown.

Subjective Data. The descriptive statistics for the subjective responses are shown in Figure 5, and the statistical test results are shown in Table 2.

Our findings reveal significant main effects of accuracy levels on both subscales of the TOAST questionnaire and our single-item questionnaires for Trust, Reliance, Confidence, and Difficulty. Regarding the TOAST subscales, participants exhibited the least trust in the system for the +2° accuracy offset compared to the +1° and +0° accuracy offsets. Consistent responses were also observed across the single-item Trust, Reliance, and Confidence scores. Additionally, participants perceived the +2° accuracy offset as significantly more difficult than the +1° and +0° accuracy offsets, while no significant difference was observed between the +0° and +1° accuracy offsets.

4.2. Discussion

In this section, we discuss the results of Experiment E1.

Accuracy. The results of Experiment E1 indicate that the accuracy of the AR tracking system had a significant effect on participants' performance in our search-and-selection task. In particular, the accuracy of the system had a significant main effect on participants' task completion time. Further, each additional degree of accuracy error caused an increase in the Type I and Type II errors that participants made. In other words, participants made more false-positive (selecting a blue team member) and more false-negative (failing to select a red team member) mistakes as the system's tracking accuracy decreased. In general, as the accuracy of the system decreased, so did participants' task performance, even for +1° of accuracy error. This finding is not particularly surprising by itself, though it is important to keep in mind when developing AR HWD systems for deployment into real-world situations that have high time pressure and performance demands with cluttered target environments.

It is further important to note that participants' reported trust in the AR system decreased significantly with each decrease in system accuracy. Participants felt that decreased accuracy made the task significantly more difficult and overall reduced their confidence in their own performance with the AR system. Similarly, decreased accuracy significantly reduced participants' sense of wanting to rely on the AR system and how advantageous they rated the AR system. This finding shows that users have a low tolerance for accuracy errors before it starts impacting their sense of trust in and reliance on the AR system, which in turn may push them towards behaviors that correspond to lower trust levels. If the tracking accuracy of the AR system is not perfect, it could pose substantial barriers to users adopting the technology in real-world settings [52], or getting the best out of it by adopting sub-par behaviors.

Task Behaviors. The task behaviors participants used to make their decisions had a significant main effect on participants' task completion time, but the only significant pairwise difference we found was that the Real-First behavior slowed participants down compared to the Real-Only behavior. One could argue that this particular result is expected because participants had an additional step in their decision process in the Real-First condition: the Real-First condition required participants to look at the simulated humans' arm bands (as they do in the Real-Only condition) and then check the AR tag to confirm their red/blue team membership.

Moreover, there were significant differences in how the task behaviors affected the number of errors participants made. The AR-Only task behavior led to significantly more false-negatives and false-positives than the Real-Only, Real-First, and AR-First

Table 2. Results for Experiment E1: Statistical test results for the accuracy tracking factor.

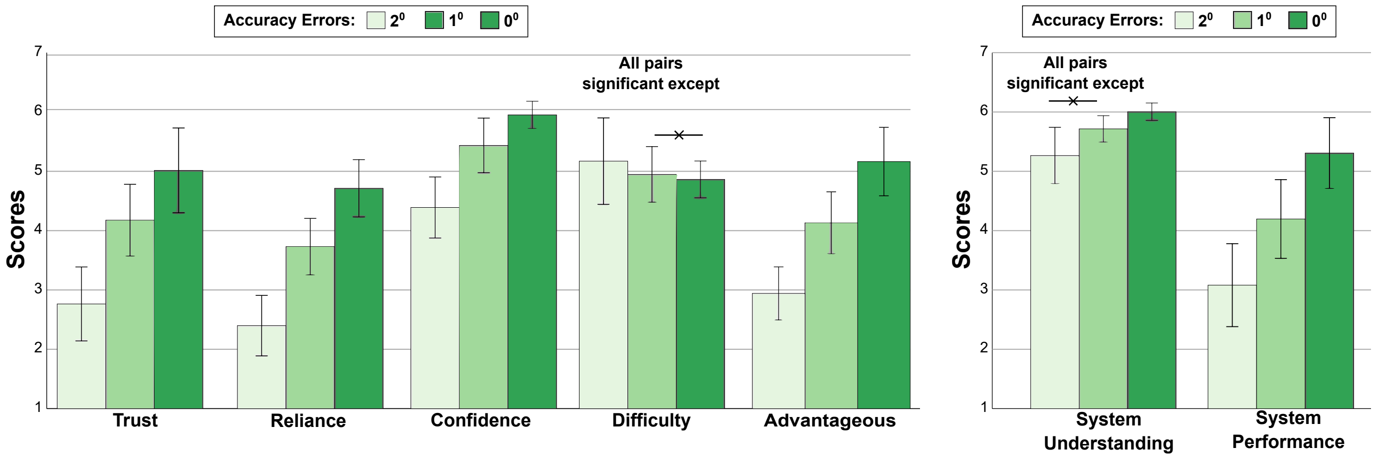
Measures	RM-ANOVA	Factors	df$_G$	df$_E$	F	p	η_p^2	Pairwise Comparisons
Elapsed Time	Two-way	Accuracy	2	34	3.76	**0.034**	0.18	None
		Task Behavior	3	51	3.57	**0.02**	0.17	$\boldsymbol{p}$**<0.05**: (Real-First > Real-Only)
		Accuracy * Task Behavior	3.71	63.13	0.99	0.44	0.06	N/A
Type I Errors	Two-way	Accuracy	2	34	12.91	< **0.001**	0.43	All $\boldsymbol{p}$**<0.05**: (+2° offset > +0° offset), (+2° offset > +1° offset)
		Task Behavior	1.01	17.24	34.87	< **0.001**	0.67	All $\boldsymbol{p}$**<0.05**: (AR-Only > Real-Only), (AR-Only > Real-First), (AR-Only > AR-First)
		Accuracy * Task Behavior	2.09	35.54	11.22	< **0.001**	0.40	All $\boldsymbol{p}$**<0.05**: For +0° offset: None For +1° offset: (AR-Only>Real-Only), (AR-Only>Real-First), (AR-Only>AR-First) For +2° offset: (AR-Only>Real-Only), (AR-Only>Real-First), (AR-Only>AR-First) For Real-Only: None For Real-First: None For AR-Only: (+2° offset > +0° offset), (+2° offset > +1° offset) For AR-First: None
Type II Errors	Two-way	Accuracy	2	34	14.22	< **0.001**	0.46	All $\boldsymbol{p}$**<0.05**: (+2° offset > +0° offset), (+2° offset > +1° offset)
		Task Behavior	1.19	20.28	26.49	< **0.001**	0.61	All $\boldsymbol{p}$**<0.05**: (AR-Only > Real-Only), (AR-Only > Real-First), (AR-Only > AR-First)
		Accuracy * Task Behavior	6	102	7.84	< **0.001**	0.32	All $\boldsymbol{p}$**<0.05**: For +0° offset: None For +1° offset: (AR-Only>Real-Only), (AR-Only>Real-First), (AR-Only>AR-First) For +2° offset: (AR-Only>Real-Only), (AR-Only>Real-First), (AR-Only>AR-First) For Real-Only: None For Real-First: None For AR-Only: (+2° offset > +0° offset), (+2° offset > +1° offset) For AR-First: None
TOAST System Understanding	One-way	Accuracy	1.44	24.61	18.38	< **0.001**	0.52	$\boldsymbol{p}$**= *0.033***: (+2° offset > +3° offset), All other pairs $\boldsymbol{p}$**<0.001**
TOAST System Performance	One-way	Accuracy	1.5	25.5	68.27	< **0.001**	0.8	All pairs $\boldsymbol{p}$**<0.001**
Trust	One-way	Accuracy	1.29	21.96	36.76	< **0.001**	0.68	All pairs $\boldsymbol{p}$**<0.001**
Reliance	One-way	Accuracy	1.13	19.21	47.92	< **0.001**	0.74	All pairs $\boldsymbol{p}$**<0.001**
Confidence	One-way	Accuracy	1.18	23.23	20.16	< **0.001**	0.66	All pairs $\boldsymbol{p}$**<0.001**
Difficulty	One-way	Accuracy	1.36	22.23	6.97	**0.009**	0.29	$\boldsymbol{p}$**= *0.04***: (+0° offset easier than +2° offset), $\boldsymbol{p}$**= *0.024***: (+1° offset easier than +2° offset)
Advantageous	One-way	Accuracy	1.23	20.96	53.36	< **0.001**	0.76	All pairs $\boldsymbol{p}$**<0.001**

task behaviors. Even in the 0° angular offset AR-Only condition participants made both Type I and Type II errors — around one error each among ten target selections in one full 360° sweep — due to the base accuracy error of our system. This finding is not unexpected: When AR is the only source of information to be relied upon, a less accurate AR system will produce more errors.

For real-world scenarios where tracking accuracy will not be better than in our laboratory settings, this means that this AR-Only task behavior is not a good choice in practice. In other words, users should not rely only on AR information to make their decisions in such task environments, even if they have high trust in the AR system. Instead, it appears preferable to rather rely on one of the other task behaviors, which were shown to effectively reduce Type I and Type II errors. If faced with the decision between one of these task behaviors, a decision which will likely be affected by participants' level of trust in and reliance on the system. Both the AR-First and Real-First task behaviors reduce errors, but the Real-First task behavior was the slowest (= worst) condition among the tested task behaviors, i.e., significantly slower than the Real-Only task behavior.

5. Experiment E2: Tracking Precision

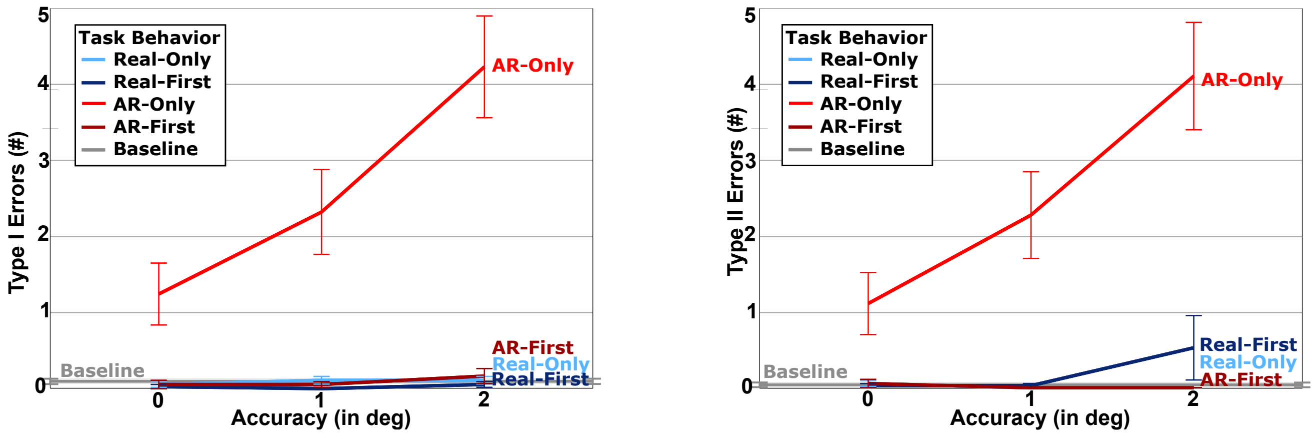
As described in Section 3.3, Experiment E2 investigated the relationships between tracking precision and participants' task



(a) Single-Item Questionnaire

(b) TOAST Questionnaire

Fig. 5. Experiment E1: Subjective data results for the three levels of accuracy with (a) Single-Item Questionnaires and (b) TOAST Questionnaire. The colored bars indicate the accuracy levels. Higher is better (except for Difficulty). The error bars show the standard error. The horizontal lines with crosses in the middle indicate those pairs that were *not* significant (p > 0.05).



(a) Type I Errors

(b) Type II Errors

Fig. 6. Experiment E1: Results for (a) Type I errors and (b) Type II errors for the four task behaviors and baseline condition. Lower is better. The error bars show the standard error.

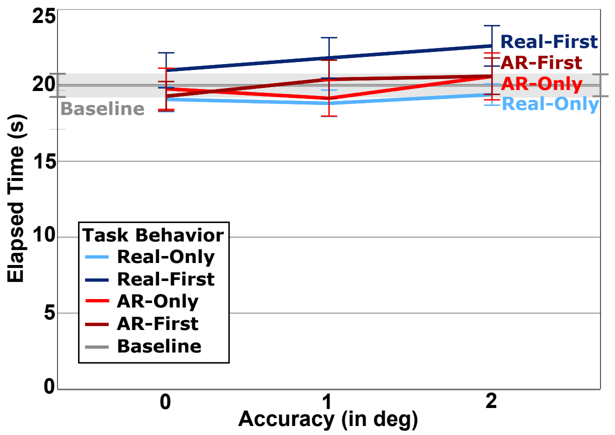


Fig. 7. Experiment E1: Results for elapsed time for the four task behaviors and baseline condition. Lower is better. The error bars show the standard error.

behaviors, task performance, and trust. Experiment E2 used a 3×4 within-subjects design with the following factors:

- **Precision (3 levels):** As described in Section 3.2, we simulated the three different precision range levels of 0°, ≤1°, and ≤2° of horizontal angular jitter.
- **Task Behaviors (4 levels):** As described in Section 3.3, we modeled four distinct task behaviors corresponding to different degrees of trust in and reliance on the AR system: AR-Only, AR-First, Real-First, and Real-Only.

Experiment E2 involved the participants, materials, methods, task, data, and procedure described in Section 3. The rest of this section describes and discusses our experimental results.

5.1. *Results*

We analyzed the responses with repeated-measures analyses of variance (RM-ANOVAs) and Tukey multiple comparisons with Bonferroni correction at the 5% significance level.

Table 3. Experiment E2: Statistical test results for the precision tracking factor.

Measures	RM-ANOVA	Factors	df$_G$	df$_E$	F	p	η_p^2	Pairwise Comparisons
Elapsed Time	Two-way	Precision	2	34	3.31	**0.049**	0.16	None
		Task Behavior	1.91	32.44	3.84	**0.034**	0.18	All **p<0.05**: (Real-First>Real-Only), (Real-First>AR-First)
		Precision * Task Behavior	2.91	49.48	0.77	0.60	0.04	N/A
Type I Errors	Two-way	Precision	2	34	4.88	**0.014**	0.22	None
		Task Behavior	1.01	17.21	27.34	< **0.001**	0.62	All **p<0.05**: (AR-Only>Real-Only), (AR-Only>Real-First), (AR-Only>AR-First)
		Precision * Task Behavior	1.89	31.99	5.43	**0.010**	0.24	All **p<0.05**: For 0° offset: None For 1° offset: (AR-Only>Real-Only), (AR-Only>Real-First), (AR-Only>AR-First) For 2° offset: (AR-Only>Real-Only), (AR-Only>Real-First), (AR-Only>AR-First) For Real-Only: None For Real-First: None For AR-Only: (1° offset > 0° offset) For AR-First: None
Type II Errors	Two-way	Precision	2	34	3.74	**0.034**	0.18	None
		Task Behavior	1.07	18.16	20.35	< **0.001**	0.55	All **p<0.05**: (AR-Only>Real-Only), (AR-Only>Real-First), (AR-Only>AR-First)
		Precision * Task Behavior	2.14	36.45	4.48	**0.016**	0.21	All **p<0.05**: For 0° offset: None For 1° offset: (AR-Only>Real-Only), (AR-Only>Real-First), (AR-Only>AR-First) For 2° offset: (AR-Only>Real-Only), (AR-Only>Real-First), (AR-Only>AR-First) For Real-Only: None For Real-First: None For AR-Only: None For AR-First: None
TOAST System Understanding	One-way	Precision	1.28	21.72	16.57	< **0.001**	0.49	All pairs **p<0.05**
TOAST System Performance	One-way	Precision	1.78	30.25	37.63	< **0.001**	0.69	All pairs **p<0.02**
Trust	One-way	Precision	1.21	20.61	39.72	< **0.001**	0.7	All pairs **p<0.001**
Reliance	One-way	Precision	1.16	22.29	19.8	< **0.001**	0.69	All pairs **p<0.001**
Confidence	One-way	Precision	1.07	20.08	18.28	< **0.001**	0.47	All pairs **p<0.005**
Difficulty	One-way	Precision	1.06	28.05	2.71	0.116	0.13	N/A
Advantageous	One-way	Precision	1.04	17.69	42.08	< **0.001**	0.71	All pairs **p<0.001**

We confirmed normality with Shapiro-Wilk tests at the 5% level and QQ plots. Degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity when Mauchly's test indicated that the assumption of sphericity was violated.

We found no significant difference between the clockwise and counterclockwise trials as well as the pre and post baselines, so we pooled the responses with respect to our analysis.

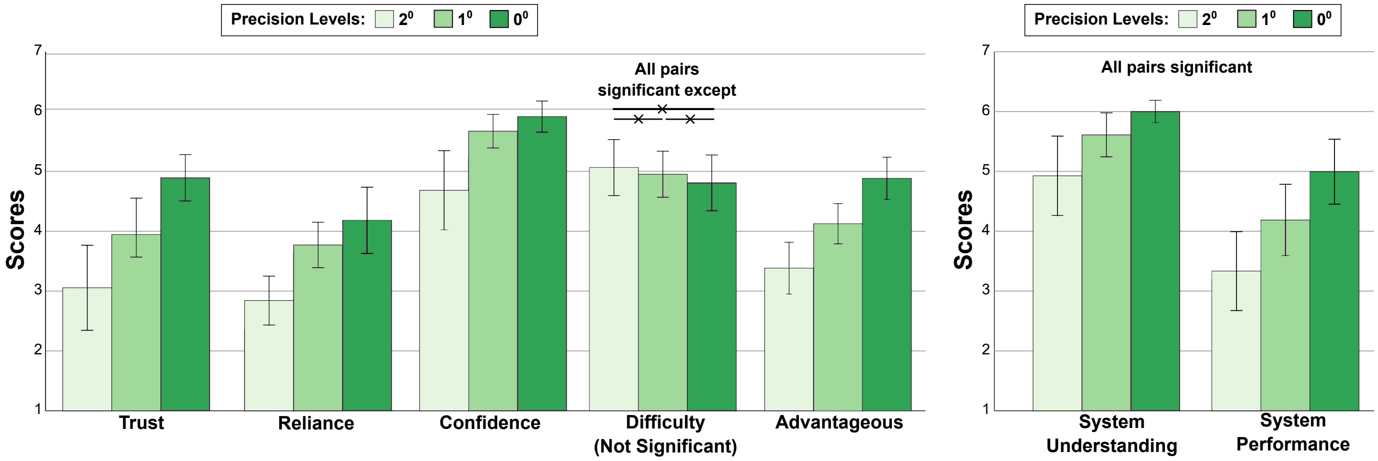
Objective Data. The descriptive statistics for the elapsed trial times as well as the Type I and Type II errors are shown in Figures 9 and 10. The statistical test results are shown in Table 3.

Our results show significant main effects and interaction effects for the precision levels and task behaviors on Type I and Type II errors but not for elapsed time. Specifically, we found that both Type I and Type II errors were significantly higher for the AR-Only condition than the other conditions. Moreover, in the AR-Only condition, the 1° angular jitter resulted in significantly higher Type I errors than the 0° angular jitter.

Lastly, it is important to note that "Baseline" in Figures 9 and 10, as discussed in Section 3.3, refers to trials conducted without any AR imagery shown.

Subjective Data. The descriptive statistics for the subjective responses are shown in Figure 8, and the statistical test results are shown in Table 3.

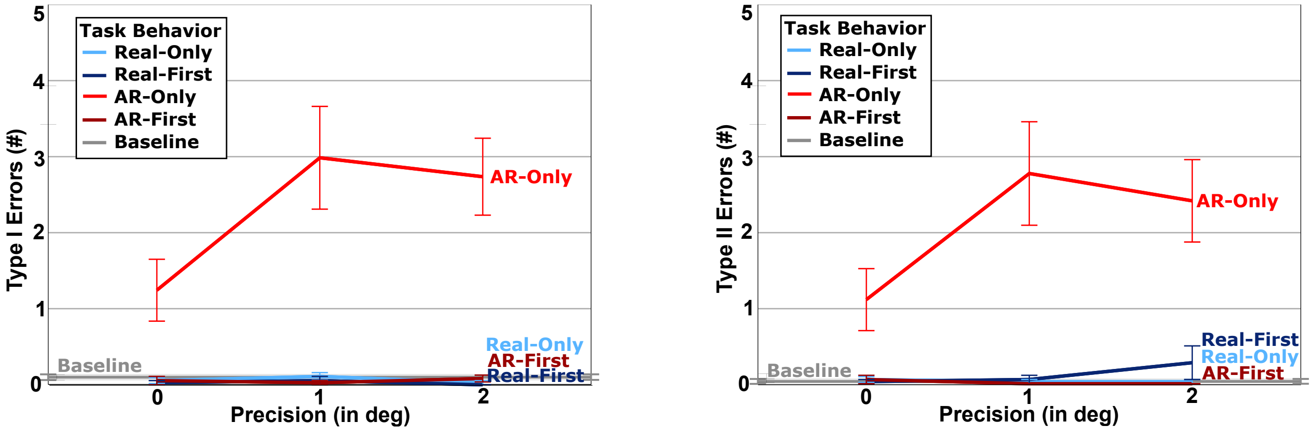
Our results show significant main effects of the precision levels on both subscales of the TOAST questionnaire and on our single-item questionnaire for Trust, Reliance, and, Confidence, but not for Difficulty. Regarding the TOAST subscales, participants exhibited significantly less trust in the system for 2° angular jitter compared to 1° and 0° angular jitter. Consistent responses were received for the single-item Trust, Reliance, Con-



(a) Single-Item Questionnaire

(b) TOAST Questionnaire

Fig. 8. Experiment E2: Subjective data results for the three levels of precision with (a) Single-Item Questionnaires and (b) TOAST Questionnaire. The colored bars indicate the precision levels. Higher is better (except for Difficulty). The error bars show the standard error. The horizontal lines with crosses in the middle indicate those pairs that were *not* significant (p > 0.05).



(a) Type I Errors

(b) Type II Errors

Fig. 9. Experiment E2: Results for (a) Type I errors and (b) Type II errors for the four task behaviors and baseline condition. Lower is better. The error bars show the standard error.

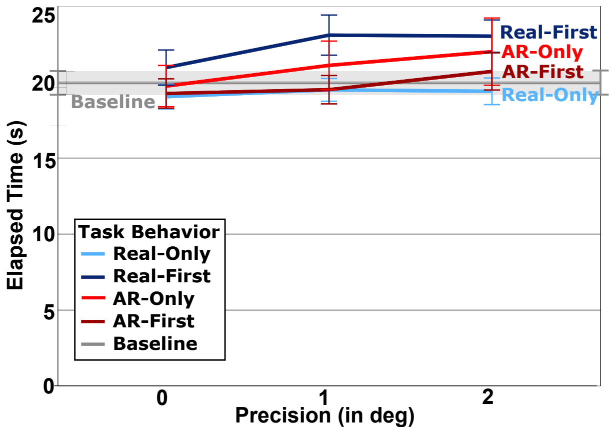


Fig. 10. Experiment E2: Results for elapsed time for the four task behaviors and baseline condition. Lower is better. The error bars show the standard error.

fidence, and Advantageous scores, but not for Difficulty scores, for which we did not find significant results.

5.2. *Discussion*

In this section, we discuss the results of Experiment E2.

Precision. Our results show that the precision of the AR tracking system had a significant effect on participants' performance in the search-and-selection task. In particular, the tracking system's precision had a significant main effect on participants' task completion time as well as Type I and Type II errors. However, compared with the results for accuracy in Experiment E1, we did not observe as strong and noticeable of increases in time and errors for increased precision errors.

Participants' subjectively reported trust in the AR system decreased significantly with each decrease in system precision. Participants felt that decreased precision made the task significantly more difficult and overall reduced their confidence in their own performance with the AR system. Further, decreased precision significantly reduced participants' sense of wanting to rely on the AR system and how advantageous they rated the AR system. While objectively the precision levels did not have a strong effect on participants' performance, their subjective ratings indicate that users have a low tolerance for precision errors before such errors impact their sense of trust on and reliance in the AR system, which could then push them towards behaviors that correspond to lower trust levels.

Task Behaviors. The results we found for our tested task behaviors with respect to the precision levels are very similar to those we found with respect to our accuracy levels in Experiment E1. The task behavior showed a significant main effect on participants' task completion time. Specifically, the Real-First task behavior slowed participants down compared to the Real-Only and AR-First task behaviors. Moreover, the AR-Only task behavior led to significantly more false-negative and false-positive errors than the Real-Only, Real-First, and AR-First task behavior. As in Experiment E1, the AR-Only behavior even resulted in higher errors than the other behaviors for the condition where we did not introduce any additional precision errors due to the base accuracy error of our system. Hence, practically, the AR-Only task behavior does not stand out as a good choice for such cluttered target scenarios, and users should not rely entirely on AR information to make their decisions, even if they have high trust in the AR system. Instead, the other task behaviors in which participants either ignored (Real-Only) or used and integrated information from both the AR system and the real environment (AR-First and Real-First) are preferable as they effectively reduced Type I and Type II errors. The AR-First task behavior is generally preferable to the Real-First task behavior as it not only reduces errors but also does not slow users down.

6. General Discussion

In this section, we summarize the main results from our two AR tracking factor experiments and provide a general discussion of our findings.

6.1. Higher task performance for higher accuracy and precision

We found significant main effects of both accuracy and precision on participants' task performance (see Figures 6, 7, 9 and 10). We found a significant impact of these factors on users' task completion time, and each additional degree of accuracy/precision error leading to an increase in errors. This decline in accuracy/precision corresponds to a decrease in overall task performance, even with just one degree of angular error.

6.2. Higher trust, reliance, confidence, and perceived advantages of the AR system for higher accuracy and precision

Our results showed that subjective ratings of trust in AR (measured via the TOAST questionnaire and single-item questions) significantly increased for both tested tracking factors, i.e., for higher accuracy and higher precision. In particular, our results show that trust ratings were significantly higher when no additional accuracy errors were introduced above our baseline system accuracy (+0° accuracy offset condition) compared to additional errors of one degree (+1° accuracy offset condition), which in turn resulted in significantly higher trust than an additional error of two degrees (+2° accuracy offset condition). We found similar effects for our three tested precision levels (0°, ≤1°, ≤2° precision error conditions). In other words, participants indicated that both tracking factors are important for trust in AR and their effects should not be ignored by researchers and practitioners when developing systems and applications that require users to trust AR. It is clear that trust is highly affected by the AR system being able to present AR information to users in a way that is free from ambiguities, i.e., the accuracy and precision of the AR tracking system must exceed the task demands.

We found these patterns in our results not only for participants' ratings of trust in AR but also with respect to their indicated reliance on AR, confidence in their task performance, and how much they perceived AR to be advantageous for the completion of the task.

6.3. Differences between subjective trust and objective performance differences

It is interesting to observe that participants' subjective responses for trust and reliance were more sensitive than their objective performance to differences in accuracy and precision error levels. In other words, differences in accuracy and precision errors have significant effects on user experience even when user task performance is not affected. While it is not generally surprising that lower accuracy/precision may reduce task performance, these findings together highlight the critical importance of maintaining a sufficiently high tracking accuracy/precision to mitigate substantial barriers to user adoption and ensure optimal performance in real-world settings, preventing the adoption of subpar task behaviors.

6.4. Highest trust in and reliance on AR does not necessarily result in highest task performance

In both Experiments E1 and E2, we observed that participants made significantly more errors (false-positives and false-negatives) with the AR-Only task behavior compared to the Real-First, Real-Only, and AR-First task behaviors. The AR-Only task behavior indicates the behaviors users adopt when they reach the highest trust in and reliance on the AR system. It is important to note that for this experimental task, low accuracy or precision would result in visual ambiguities between the registration of AR tags and simulated humans in the environment. Hence, if participants were asked to rely only on AR information (i.e., the AR-Only task behavior), a less accurate or precise AR system will result in more errors. This point is emphasized by the fact that in both experiments participants made errors in the 0° offset conditions using the AR-Only task behavior, indicating that there was registration error present in the AR cues even after completing the self-calibration procedure described in Section 3.2. Even so, the total levels of accuracy and precision error we tested in these experiments were relatively low compared to levels of error that can be expected in more realistic situations "in the wild," such as in outdoor environments. Current consumer AR tracking systems may be better suited for tasks less sensitive to accuracy and precision errors, such as cueing in clutter-free environments. In sum, for tasks that are sensitive to ambiguities caused by an AR system's accuracy and precision, users should not adopt the AR-Only task behavior as it will cause significant errors.

While the AR tracking errors affected not only the AR-Only task behavior but also the AR-First and Real-First task behaviors, it is noteworthy that the latter two task behaviors effectively reduced the Type I and Type II errors by using and integrating information from both the AR system and cross-checking this information against the real environment. Hence, for situations in which users perceive value in AR information but do not fully trust it to the level of wanting to rely only on it (i.e., adopt the AR-Only task behavior), the two intermediate task behaviors of AR-First and Real-First appear to be valuable options as they reduce errors and result in high task performance. Additionally, the AR-First task behavior resulted in lower elapsed time compared to the Real-First task behavior in our experiment testing different precision error levels. However, we did not find any significant task performance differences between the Real-Only task behaviors and the AR-First or Real-First task behaviors. Based on these findings, we recommend that AR systems may include a self-monitoring system that assesses the quality of the tracking data in real time to potentially turn off the AR imagery in cases where the accuracy or precision is too low to be helpful and not distract users.

In addition, the differences we found based on task behavior have important implications for the design of future user studies and evaluations of AR systems. We recommend that researchers control for or account for task behavior, especially in studies evaluating the effects of factors directly impacting user trust and reliance. As we have found, different strategies reflecting varying levels of trust and reliance on the AR system can have significant effects on end user performance. Of note is that higher trust and reliance on a given AR system may not translate to optimal user behavior. It may no longer be sufficient to administer a usability questionnaire or a trust scale to assess AR systems—their downstream effects on user behavior also has significant effects on the performance of the overall human-AR system.

6.5. Evidence for task behavior compliance

As discussed in Section 3.7, it was important to perform a sanity check to see if the task behaviors were correctly implemented by participants. In addition to the sanity check, the conspicuously high error rate in the 0° offset conditions for the AR-Only task behaviors we discussed above would also suggest that participants correctly implemented the AR-Only task behavior. Further, we found that when utilizing the AR-First task behavior, the errors seen in the AR-Only condition diminish significantly, which would be expected if participants were following the task behavior instructions. Additionally, we found that in both experiments participants utilizing the Real-First task behavior produced higher elapsed times than with the Real-Only task behavior, an expected result as Real-First requires participants to perform a similar procedure to Real-Only, but with the added burden of verifying with the AR cue. This would support the notion that participants correctly implemented the Real-First and Real-Only task behaviors. While these comparisons do not serve as manipulation checks, they do support the notion that participants correctly understood and implemented the task behaviors in this study.

6.6. Limitations and future work

In the experiments presented here, we observed task behaviors that are presumed to be indicative of users' trust in and reliance on AR systems (see Figure 4). However, it is important to note that in our experimental setup, participants were specifically directed to exhibit certain behaviors. This is in contrast to real-world scenarios where AR users are not typically given explicit instructions on how to engage with a task; instead, they navigate the task based on their personal perceptions and attitudes towards the AR system. As outlined in Section 2.1, there are multiple factors that may influence a user's decision to adopt a certain behavior or to alter their behavior during a task. The exploration of the varying behavioral patterns that lead to different interactions with AR systems presents a promising avenue for future research.

In our two experiments, we manipulated the presentation of AR cues in terms of their registration accuracy and precision and investigate their effects on performance, reliance, and trust. However, participants were instructed that the AR cues were always correct. Introducing ambiguity in the trustworthiness of the AR cues themselves (i.e., uncertainty whether a given cue correctly corresponds to a selection target) is an interesting direction for future work in this area.

Because our experiments had a large number of conditions, we did not include more than three levels of the two tracking factors accuracy and precision. We chose the angular offset levels because they are whole numbers and appeared qualitatively different in our pilot testing. We acknowledge that the tested levels do not include many extremes and leave out many values in between them that are worthy of inquiry. Future work could investigate other ranges of angular errors.

Lastly, our experiment's sample size (N=18) and skewed gender representation (15 male, 3 female), limits the generalizability of our work. Though we did not find a reason to investigate the degree to which gender could have an affect on the way that registration errors change user performance and trust, it is historically known that gender differences can appear in many different spatial cognition contexts [53]. Given the movement towards improved rigor and replicability in HCI communities related to our work [54], it would be valuable to verify our results across a larger demographic in future work.

7. Conclusion

In this paper, we investigated the relationships between AR tracking accuracy and precision and four task behaviors (AR-Only, AR-First, Real-Only, and Real-First) on users' subjective assessment of the AR system and objective performance when completing a 360° search-and-selection task. We conducted two within-subjects experiments, one for each AR tracking factor, while evaluating all four behaviors. By controlling for different task behaviors in AR, we were able to show significant issues in task performance that were introduced by even small amounts of AR tracking errors. However, our results also show that some of the evaluated task behaviors in AR were able to compensate for the tracking errors, suggesting that users' ability to integrate cues from AR and the real world may be effective and helpful even if the tracking accuracy or precision are less than optimal.

Acknowledgments

This material includes work supported in part by Vision Products LLC via US Air Force Research Laboratory (AFRL) Award Number FA864922P1038, the Office of Naval Research under Award Numbers N00014-21-1-2578 and N00014-21-1-2882 (Dr. Peter Squire, Code 34), and the AdventHealth Endowed Chair in Healthcare Simulation (Prof. Welch).

References

[1] Kim, K, Billinghurst, M, Bruder, G, Duh, HBL, Welch, GF. Revisiting Trends in Augmented Reality Research: A Review of the 2nd Decade of ISMAR (2008–2017). IEEE Transactions on Visualization and Computer Graphics 2018;24(11):2947–2962.

[2] Welch, G, Bruder, G, Squire, P, Schubert, R. Anticipating Widespread Augmented Reality: Insights from the 2018 AR Visioning Workshop. Tech. Rep.; University of Central Florida and Office of Naval Research; 2019.

[3] Norouzi, N, Bruder, G, Belna, B, Mutter, S, Turgut, D, Welch, G. A Systematic Review of the Convergence of Augmented Reality, Intelligent Virtual Agents, and the Internet of Things. In: Springer Transactions on Computational Science Computational Intelligence. 2019, p. 1–24.

[4] Gottsacker, M, Norouzi, N, Schubert, R, Guido-Sanz, F, Bruder, G, Welch, G. Effects of environmental noise levels on patient handoff communication in a mixed reality simulation. In: Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology (VRST). 2022, p. 1–10.

[5] Azuma, RT. The challenge of making augmented reality work outdoors. Mixed reality: Merging real and virtual worlds 1999;1:379–390.

[6] Azuma, R, Bishop, G. Improving static and dynamic registration in an optical see-through HMD. In: Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques. 1994, p. 197–204.

[7] Billinghurst, M. Grand challenges for augmented reality. Frontiers in Virtual Reality 2021;2.

[8] Hinterstoisser, S, Lepetit, V, Ilic, S, Holzer, S, Bradski, G, Konolige, K, et al. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. 2013, p. 548–562.

[9] Soleimanitaleb, Z, Keyvanrad, MA, Jafari, A. Object tracking methods: A review. In: Proceedings of the International Conference on Computer and Knowledge Engineering (ICCKE). 2019, p. 282–288.

[10] Wenhan Luo, , Xing, J, Milan, A, Zhang, X, Liu, W, Kim, TK. Multiple object tracking: A literature review. Artificial Intelligence 2021;293:103448.

[11] Azuma, R, Hoff, B, Neely, H, Sarfaty, R. A motion-stabilized outdoor augmented reality system. In: Proceedings of IEEE Virtual Reality. 1999, p. 252–259.

[12] Endres, F, Hess, J, Engelhard, N, Sturm, J, Cremers, D, Burgard, W. An evaluation of the RGB-D SLAM system. In: Proceedings of the IEEE International Conference on Robotics and Automation. 2012, p. 1691–1696.

[13] Ayadi, M, Scuturici, M, Ben Amar, C, Miguet, S. A skyline-based approach for mobile augmented reality. The Visual Computer 2021;37(4):789–804.

[14] Park, J, You, S, Neumann, U. Natural feature tracking for extendible robust augmented realities. In: Proceedings of the International Workshop on Augmented reality: Placing Artificial Objects in Real Scenes. 1999, p. 209–217.

[15] Davison, A, Mayol, W, Murray, D. Real-time localization and mapping with wearable active vision. In: Proceedings of the IEEE and ACM International Symposium on Mixed and Augmented Reality. 2003, p. 18–27.

[16] Bleser, G, Wuest, H, Stricker, D. Online camera pose estimation in partially known and dynamic scenes. In: IEEE and ACM International Symposium on Mixed and Augmented Reality. 2006, p. 56–65.

[17] Welch, G, Wang, T, Bishop, G, Bruder, G. A Novel Approach for Cooperative Motion Capture (COMOCAP). In: Proceedings of the International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments. 2018, p. 73–80.

[18] Kim, P, Orlosky, J, Kiyokawa, K. AR timewarping: A temporal synchronization framework for real-time sensor fusion in head-mounted displays. In: Augmented Human International Conference. 2018, p. 1–8.

[19] Erickson, A, Kim, K, Bruder, G, Welch, G. A review of visual perception research in optical see-through augmented reality. In: Proceedings of the International Conference on Artificial Reality and Telexistence & Eurographics Symposium on Virtual Environments. 2020, p. 1–8.

[20] Rolland, JP, Fuchs, H. Optical versus video see-through head-mounted displays in medical visualization. Presence 2000;9(3):287–309.

[21] Kruijff, E, Swan, JE, Feiner, S. Perceptual issues in augmented reality revisited. In: Proceedings of the IEEE International Symposium on Mixed and Augmented Reality. 2010, p. 3–12.

[22] Livingston, MA, Ai, Z. The effect of registration error on tracking distant augmented objects. In: Proceedings of the IEEE and ACM International Symposium on Mixed and Augmented Reality. 2008, p. 77–86.

[23] Yeh, M, Wickens, CD. Display signaling in augmented reality: Effects of cue reliability and image realism on attention allocation and trust calibration. Human Factors 2001;43(3):355–365.

[24] Feiner, S, MacIntyre, B, Höllerer, T, Webster, A. A touring machine: Prototyping 3D mobile augmented reality systems for exploring the urban environment. Personal Technologies 1997;1:208–217.

[25] Warden, AC, Wickens, CD, Mifsud, D, Ourada, S, Clegg, BA, Ortega, FR. Visual search in augmented reality: Effect of target cue type and location. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting; vol. 66. 2022, p. 373–377.

[26] Rusch, ML, Schall Jr, MC, Gavin, P, Lee, JD, Dawson, JD, Vecera, S, et al. Directing driver attention with augmented reality cues. Transportation Research Part F: Traffic Psychology and Behaviour 2013;16:127–137.

[27] Merlo, JL. Effect of reliability on cue effectiveness and display signaling. University of Illinois at Urbana-Champaign; 1999.

[28] Yeh, M. Attention and trust biases in the design of augmented reality displays. University of Illinois at Urbana-Champaign; 2000.

[29] Parasuraman, R, Riley, V. Humans and automation: Use, misuse, disuse, abuse. Human factors 1997;39(2):230–253.

[30] Sorkin, RD. Why are people turning off our alarms? The Journal of the Acoustical Society of America 1988;84(3):1107–1108.

[31] Mosier, KL, Skitka, LJ, Heers, S, Burdick, M. Automation bias: Decision making and performance in high-tech cockpits. The International Journal of Aviation Psychology 1998;8(1):47–63.

[32] Mifsud, D, Wickens, C, Maulbeck, M, Crane, P, Ortega, FR. The Effectiveness of Gaze Guidance Lines in supporting JTAC's Attention Allocation. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 2022;66(1):2198–2201.

[33] Carter, D. Immersive employee experiences in the metaverse: virtual work environments, augmented analytics tools, and sensory and tracking technologies. Psychosociological Issues in Human Resource Management 2022;10(1):35–49.

[34] Syed, TA, Siddiqui, MS, Abdullah, HB, Jan, S, Namoun, A, Alzahrani, A, et al. In-depth review of augmented reality: Tracking technologies, development tools, AR displays, collaborative AR, and security concerns. Sensors 2022;23(1):146.

[35] Welch, G, Davis, L. Tracking for Training in Virtual Environments: Estimating the Pose of People and Devices for Simulation and Assessment. In: The PSI Handbook of Virtual Environments for Training and Education: Developments for the Military and Beyond; chap. 30. Praeger Security International; 2008, p. 1–54.

[36] Welch, G, Foxlin, E. Motion Tracking: No Silver Bullet, but

a Respectable Arsenal. IEEE Computer Graphics and Applications 2002;22(6):24–38.

[37] Allen, BD, Bishop, G, Welch, G. Tracking: Beyond 15 minutes of thought. In: ACM Annual Conference on Computer Graphics & Interactive Techniques (SIGGRAPH) Course Pack. 2001, p. 1–193.

[38] Ferrari, V, Cattari, N, Fontana, U, Cutolo, F. Parallax free registration for augmented reality optical see-through displays in the peripersonal space. IEEE Transactions on Visualization and Computer Graphics 2020;28(3):1608–1618.

[39] Luckett, E, Key, T, Newsome, N, Jones, JA. Metrics for the evaluation of tracking systems for virtual environments. In: IEEE Conference on Virtual Reality and 3D User Interfaces (VR). 2019, p. 1711–1716.

[40] Niehorster, DC, Li, L, Lappe, M. The accuracy and precision of position and orientation tracking in the HTC Vive virtual reality system for scientific research. i-Perception 2017;8(3).

[41] Erickson, A, Norouzi, N, Kim, K, Schubert, R, Jules, J, LaViola Jr, JJ, et al. Sharing gaze rays for visual target identification tasks in collaborative augmented reality. Journal on Multimodal User Interfaces 2020;14(4):353–371.

[42] Weber, M, Hartl, R, Zäh, MF, Lee, J. Dynamic Pose Tracking Accuracy Improvement via Fusing HTC Vive Trackers and Inertia Measurement Units. International Journal of Precision Engineering and Manufacturing 2023;24(9):1661–1674.

[43] Julier, S, Uhlmann, J. Unscented filtering and nonlinear estimation. Proceedings of the IEEE 2004;92(3):401–422.

[44] Julier, S, Uhlmann, J. A general method for approximating nonlinear transformations of probability distributions. Tech. Rep.; 1996.

[45] Welch, G, Bishop, G. An introduction to the Kalman filter. Tech. Rep. TR95-041; University of North Carolina at Chapel Hill, Department of Computer Science; 1995.

[46] Kalman, RE. A new approach to linear filtering and prediction problems. Journal of Basic Engineering 1960;82(1):35–45.

[47] Robertson, CM, MacIntyre, B, Walker, BN. An evaluation of graphical context as a means for ameliorating the effects of registration error. IEEE Transactions on Visualization and Computer Graphics 2008;15(2):179–192.

[48] Jiang, T, Yu, D, Wang, Y, Zan, T, Wang, S, Li, Q. HoloLens-based vascular localization system: precision evaluation study with a three-dimensional printed model. Journal of Medical Internet Research 2020;22(4):e16852.

[49] Pratt, P, Ives, M, Lawton, G, Simmons, J, Radev, N, Spyropoulou, L, et al. Through the HoloLens™ looking glass: augmented reality for extremity reconstruction surgery using 3D vascular models with perforating vessels. European Radiology Experimental 2018;2:1–7.

[50] Spitzley, KA, Karduna, AR. Feasibility of using a fully immersive virtual reality system for kinematic data collection. Journal of Biomechanics 2019;87:172–176.

[51] Wojton, HM, Porter, D, T. Lane, S, Bieber, C, Madhavan, P. Initial validation of the trust of automated systems test (TOAST). The Journal of Social Psychology 2020;160(6):735–750.

[52] Wu, K, Zhao, Y, Zhu, Q, Tan, X, Zheng, H. A meta-analysis of the impact of trust on technology acceptance model: Investigation of moderating influence of subject and context type. International Journal of Information Management 2011;31(6):572–581.

[53] Coluccia, E, Louse, G. Gender differences in spatial orientation: A review. Journal of Environmental Psychology 2004;24(3):329–340. doi:10.1016/j.jenvp.2004.08.006.

[54] Echtler, F, Häußler, M. Open Source, Open Science, and the Replication Crisis in HCI. In: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems. CHI EA '18; New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-5621-3; 2018, p. 1–8. doi:10.1145/3170427.3188395.