Handheld Guides in Inspection Tasks: Augmented Reality vs. Picture

Jarkko Polvi, Takafumi Taketomi, *Member, IEEE,* Atsunori Moteki, Toshiyuki Yoshitake, Toshiyuki Fukuoka, Goshiro Yamamoto, *Member, IEEE,* Christian Sandor, *Member, IEEE,* and Hirokazu Kato, *Member, IEEE*

Abstract—Inspection tasks focus on observation of the environment and are required in many industrial domains. Inspectors usually execute these tasks by using a guide such as a paper manual, and directly observing the environment. The effort required to match the information in a guide with the information in an environment and the constant gaze shifts required between the two can severely lower the work efficiency of inspector in performing his/her tasks. Augmented reality (AR) allows the information in a guide to be overlaid directly on an environment. This can decrease the amount of effort required for information matching, thus increasing work efficiency. AR guides on head-mounted displays (HMDs) have been shown to increase efficiency. Handheld AR (HAR) is not as efficient as HMD-AR in terms of manipulability, but is more practical and features better information input and sharing capabilities. In this study, we compared two handheld guides: an AR interface that shows 3D registered annotations, that is, annotations having a fixed 3D position in the AR environment, and a non-AR picture interface that displays non-registered annotations on static images. We focused on inspection tasks that involve high information density and require the user to move, as well as to perform several viewpoint alignments. The results of our comparative evaluation showed that use of the AR interface resulted in lower task completion times, fewer errors, fewer gaze shifts, and a lower subjective workload. We are the first to present findings of a comparative study of an HAR and a picture interface when used in tasks that require the user to move and execute viewpoint alignments, focusing only on direct observation. Our findings can be useful for AR practitioners and psychology researchers.

Index Terms—Handheld devices, augmented reality, user evaluation, inspection task

1 Introduction

Inspection tasks are required in many industrial domains. The term refers to the inspection of targets in an environment via observation, for example, the inspection of engine parts in a motor vehicle. The inspection of different types of targets in large environments such as factory floors or server rooms is also frequently required. The inspector is frequently aided by guides of various forms such as paper manuals that he/she uses to gain information and locate targets in the environment. In Japan, paper checklists are frequently used in real-world inspection tasks to record results. Neumann and Majoros [1] defined two forms of inspection task activities: 1) informal and 2) task-environment. Informal activities consist mainly of information matching, which refers to comprehending information and transposing it from a guide to the environment. As a part of information matching, the user is required to align the viewing angle. Task-environment activities refer to the actual inspection of a physical target in the environment by direct observation. In a situation where the user is a newly hired inspector or does not have knowledge about what he/she should inspect, an inspection using conventional guides entails two main problems: first, the informal activities of the user decrease his/her work efficiency [2], and second, the need to shift his/her gaze according to the two forms of activities can severely increase the user's cognitive workload [3].

Augmented reality (AR) allows the information in a guide to be directly overlaid on the environment. To overlay information, 6 degrees of freedom of the device's movement are estimated using tracking technologies, and then, the geometric relationship between the real and virtual worlds is unified. In general, this process is called registration in AR. AR has been shown to increase work efficiency [4], [5] in various tasks, but the effectiveness of AR displayed on a handheld device for enabling efficient inspection has not been studied in depth. The main benefit of using AR in inspection in the case of newly-hired inspectors is that it allows them to conduct information matching more efficiently and reduces the number of necessary gaze shifts between the guide and the environment. The reduction in the number of these shifts can vary according to the AR display technology used.

Handheld AR (HAR) refers to AR displayed on handheld devices such as smartphones or tablet PCs. HAR provides a good means for the use of AR to become widespread because a vast number of suitable handheld devices already exist [6]. Handheld devices allow users to shift their gaze between the augmented representation of the environment (the screen of the device) and the real environment [7]. The devices are also highly mobile and can be considered to provide better information input and sharing capabilities than head-mounted displays (HMDs). However, many existing HAR guides are not considered effective in task support [8], [9]. HMDs have frequently been used in task support because they allow users to constantly observe an augmented task environment, eliminating the need for gaze shifts entirely and thus increasing the efficiency of task completion. However, HMDs can give rise to severe safety issues in many industrial domains because the wearer's vision of the actual environment is impeded [10]. In addition, while physical manipulation is usually not required in inspection tasks, workers are frequently required to view traditional 2D information, such as text and charts, or share information with other workers. These tasks are more easily accomplished using a handheld device rather than the HMD.

In this work, we evaluated the effectiveness of using two different handheld guides in inspection tasks: a simultaneous localization and mapping (SLAM)-based AR guide (Fig. 1(a)) and a non-AR picture guide (Fig. 1(b)). Our evaluation involved two computer hardware inspection tasks that represented the real-world inspection of complex environments. In our opinion, the results of our research are in general transferable to many work support applications that require the observation of an environment

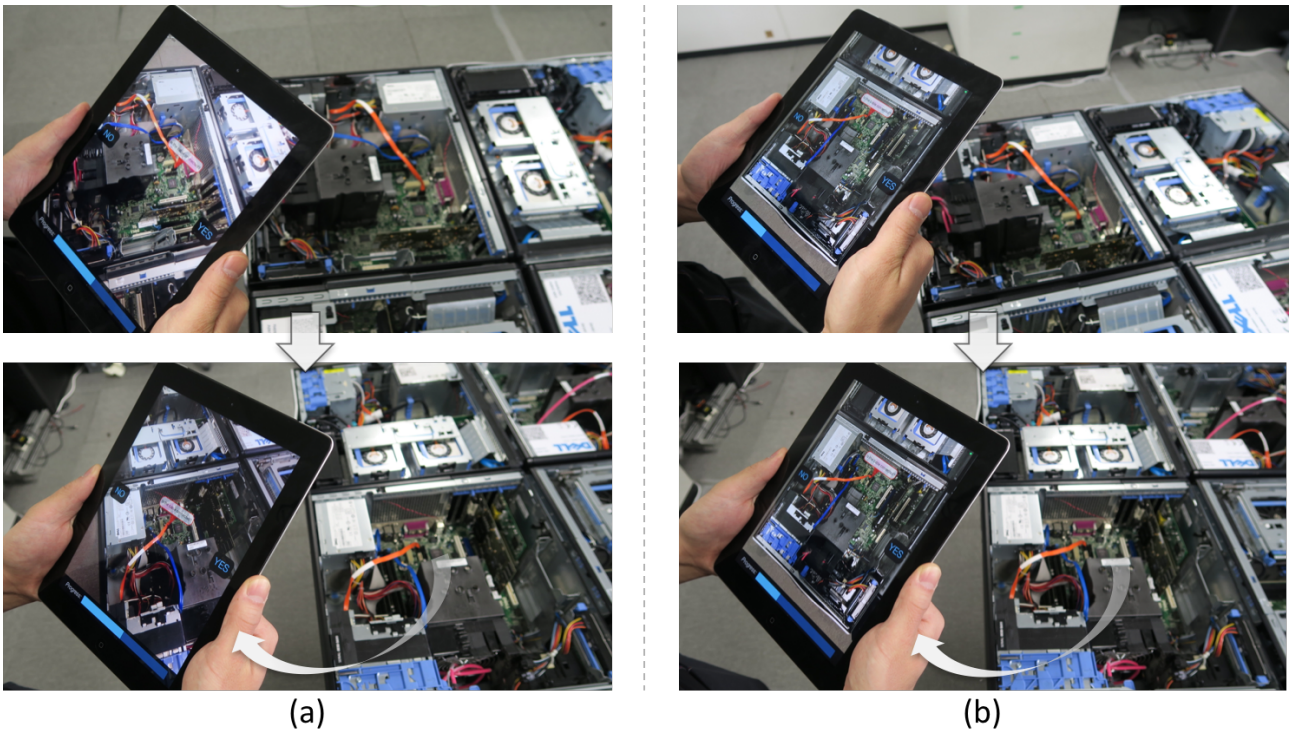


Fig. 1. We compared the effectiveness of a handheld AR (a) and a handheld picture (b) interface. The AR interface overlays 3D registered virtual annotations on the live representation of the task environment on the screen of the device. When the user changes the viewpoint to the left, the viewpoint on the screen changes correspondingly (a). The picture interface displays non-registered annotations on static images captured from a top-down viewpoint. When the user moves to the left, the viewpoint of the image does not change (b). Curved arrows illustrate the direction of movement.

that is unknown to the user. Observation is an essential part of inspection and other inspection activities are based on the results of an initial observation. Our results show that the AR interface allowed the user to work more quickly and produced a lower error rate, fewer gaze shifts, and a lower subjective workload score. A 3D virtual model-based guide can be considered an intermediate approach between the AR and the non-AR picture interface. However, the construction of a 3D model is not practical in an actual inspection scenario. For example, the arrangement of a server room is frequently changed as dictated by the conditions of the customer. In this case, the 3D model should be reconstructed to match the actual environment, a task that actual workers cannot easily accomplish. In contrast to the 3D virtual model-based interface, in SLAM-based AR [11] and non-AR picture interfaces the contents of the environment can easily be updated. For these reasons, we do not discuss the 3D virtual model-based interface in this paper.

The contribution of our work comprises the findings of a comparative user study. In contrast to a previous study [12], our study was focused only on observation and the use of handheld devices to demonstrate that improved inspection efficiency does not depend on the benefits of more immersive AR technologies and that HAR is suitable for widespread adoption. This type of comparison has previously been performed only with HMD-AR. In contrast to tasks using HAR that are conducted from a fixed position [13], our study tasks required users to move and to change their viewing angle. Our findings can facilitate HAR task support guide design as well as psychology research because they pertain to the effects of information matching in real-world tasks.

2 Related Work

In the following section, we summarize related work on information matching and AR task support systems that allow user mobility. We focus only on mobile AR guides because it is essential that the user be able to move during inspection tasks.

2.1 Information Matching

Most frequently, information matching requires a person to mentally or physically rotate the displayed information to align it correctly. Several studies have been conducted on low level mental rotation tasks involving 2D [14] and 3D objects [15]. These studies showed an increase in task time according to the complexity of the required rotation. The effect of mental rotation and information alignment has been evaluated in real-world tasks, such as outdoor navigation [16] or object assembly [17]. The results reported in both [16] and [17] indicated that allowing users to rotate the guide can improve task performance. As mentioned, AR can decrease the need for information matching, thus reducing the overall cognitive workload. Robertson et al. [18] evaluated the use of different AR graphical presentations in an object assembly task. Their results showed that fully aligned AR is more effective than non-aligned AR or non-AR presentations.

2.2 Head-Mounted Displays

HMDs are often utilized for AR tasks, leaving the hands of the user free to manipulate the physical environment [12], [19], [20]. Previous research on AR using HMDs demonstrated performance benefits for a variety of tasks [4], [21]. These benefits may be leveraged when using HAR when manipulation of the physical

environment is not required. However, the effectiveness of using HMDs purely for observation tasks has not been validated.

2.3 Handheld devices

Because handheld devices enable high mobility, HAR has frequently been evaluated in terms of outdoor navigation [22], [23] but has not been shown to offer significant benefits compared to conventional navigation guides. Jung et al. [24] presented an HAR guidance system that displays annotations related to various indoor locations but did not conduct a comparative evaluation. Rauhala et al. [25] developed an HAR system that visualizes network sensor data for walls, allowing the inspection of wireless networks. The authors conducted only a preliminary study, which did not include a comparison with other systems. Makita et al. [26] developed an HAR indoor navigation and machine inspection system. They confirmed its suitability for real-world tasks in a preliminary evaluation. All the studies mentioned above concentrated on observation and did not require the user to physically manipulate the environment.

Hakkarainen et al. [27] developed an HAR assembly guide for small objects that displays 3D augmentation on still images of the environment. The authors conducted only a quantitative pilot study and did not compare their HAR guide to more conventional guides. Karlsson et al. [28] developed an HAR system that detects objects in an environment and displays 3D models overlaid on them showing hidden targets inside the objects. The system allows users to inspect the internal components of various machines. They conducted pilot studies to confirm the functionality of their system, but they did not perform an extensive study with real users. Träskbäack and Haller [10] implemented a simple HAR tablet system for training oil refinery workers. They also created user requirements but did not evaluate their prototype. Liu et al. [13] evaluated HAR with and without real-time feedback in a simple device setup scenario where HAR was compared with conventional interfaces. They found that HAR did not offer any benefits, but the provision of real-time feedback allowed greater user efficiency than picture guides. Gauglitz et al. [29], [30] developed an HAR prototype for remote collaboration that provides local and remote users with their own separate views. They compared their system to conventional guides and found the subjective feedback to be significantly more useful when 3D-registered AR annotations were provided. However, their objective measurement results, such as task completion time, showed no benefits.

2.4 Study Motivation

Although many related studies on information positioning and mental rotation exist, only a few implications for practical real-world tasks have been offered. Understandably, HMD-AR is frequently applied to tasks that require physical manipulation of the environment because HMD-AR allows hands-free use. However, the results obtained for HAR in past comparative studies are not similar. Although many of the past HMD-AR and HAR guides have been evaluated in some form of user study, most of these studies did not include comprehensive comparisons. Furthermore, previous HAR studies did not take full advantage of the capabilities of AR in terms of information matching. Our comparative study is the first in which HAR was used in complex tasks that focused on observation only and required the user to execute viewpoint alignments. The main goal of this study was to provide empirical validation of the benefits of using HAR as opposed to images in procedural observation tasks, for which there currently exists only anecdotal evidence. This validation was the main motivation for our user study.

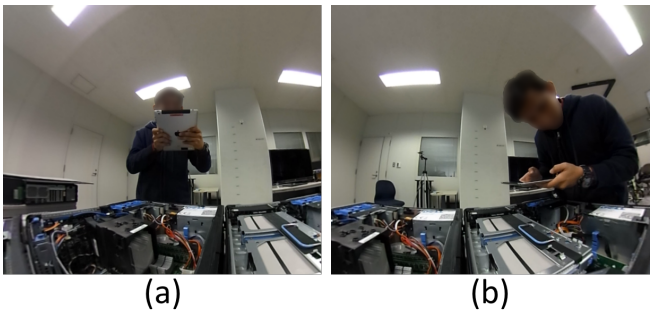


Fig. 2. Example of a test session. A test participant is performing the high mobility task using the AR (a) and the picture (b) interface. Images were captured by an omni-directional camera and used to calculate the number of gaze shifts.

3 User Study

3.1 Study Design

We conducted an evaluation using computer hardware inspection tasks where the participants assumed the role of newly-hired inspection personnel who were unfamiliar with the environment that they were required to inspect. Figure 2 shows an example of a test participant conducting the inspection. We used a within-group factorial study design with two independent variables (interface and task), both with two levels (2×2). We took three objective measurements: task completion time, error rate, and number of gaze shifts. We defined these as the main efficiency metrics in inspection. We also measured subjective feedback through a questionnaire and free-form comments. In addition, we observed the usage of the interfaces during the study sessions. A total of 24 graduate students (15 males and 9 females, mean age 28±5 years, age range 22-42 years) were recruited as test participants. All participants were graduate students from biological, material, and information science. The participants filled out a pre-test questionnaire in which they responded to questions addressing their familiarity with handheld devices, AR, and picture-based guides on a 7-point Likert scale, where 1=not very familiar and 7=very familiar. The mean (M) and standard deviation (SD) were M=5.71, SD=1.37 for handheld devices, M=3.50, SD=1,72 for AR, and M=5.00, SD=1.65 for picture-based guides. The condition order (4×3×2×1) was fully counterbalanced, meaning that each condition was conducted an equal number of times in each condition slot (slots: 1st, 2nd, 3rd, and 4th). The participants were given a total of three five-minute breaks between conditions to prevent the possible effects of fatigue from holding and moving with the iPad.

3.2 Interfaces

We developed two interfaces for a handheld tablet PC: an AR (Fig. 3(a)-(c)) interface and a non-AR picture interface (Fig. 3(d)-(f)). Both interfaces displayed textual annotations as white bubbles with red borders. A blue progress bar was shown to visualize how many targets remained to be inspected. Furthermore, both interfaces featured "YES" and "NO" buttons for answering

questions contained in the annotations (see Section 3.3). The test device was a 4th generation iPad tablet[1] with a 1.4 GHz dual core processor, 1 GB memory, and a 9.7-inch display with a resolution of 1536×2048 pixels. Its operating system was iOS version 7.2. Both interfaces could be used only in portrait orientation.

3.2.1 Augmented Reality Interface

The AR interface showed 3D-registered annotations overlaid on the environment in a live video feed (Fig. 3(a)-(c)). If an annotation was outside the field of view of the camera, a dynamic 2D indicator was displayed, pointing to the direction of the annotation (Fig. 3(b)), because the system has knowledge of other annotation' positions. We used an arrow to visualize direction because this has been shown in previous studies to be the most effective means of indicating off-screen content visualization [31]. That a direction indicator of the location of a component in a large workspace can be displayed is one of the advantages of using AR. We implemented SLAM and natural feature point tracking to track and map the environment. SLAM utilizes video images (resolution 480×640 pixels) of the camera of the handheld device and internal sensors to track the environment. SLAM constantly updates the tracked environment map. If the tracking fails, the interface displays instructions for initializing the tracking again (Fig. 3(a)). Tracking can be initialized only from one specific viewpoint. Our implementation does not visualize the feature points tracked by SLAM. The AR interface did not have a zoom function; zooming in on the view was possible only by moving the device closer to the object.

3.2.2 Picture Interface

The picture interface, considered a non-AR interface, displayed non-registered 3D annotations overlaid on images that were taken from a top-down viewpoint (Fig. 3(d)). The interface switched the image according to the section (see Section 3.2) of the task environment in which the target was located. The resolution of the images displayed by the interface was 1536×2048; they were captured with the same 4th generation iPad. An image was initially fully visible, and participants could zoom and pan (Fig. 3(e) and (f)) to obtain a better view of the target. When a user tapped either of the answer buttons, the initial level of the image zooming was restored. We chose a picture interface for comparison because it had been shown to be the most effective non-AR interface in previous studies [5], [13]. We used real pictures instead of sketches because our objective was to mimic a scenario where annotations could also have easily been created in situ using only a handheld device, as in various commercial handheld applications.

3.3 Tasks

The participants performed two test tasks (Fig. 4): a low mobility task and a high mobility task. In both tasks, the participants were instructed to answer 20 yes/no questions about the targets in the environment (for example, "Is the cable connected correctly?"). We used the same type of questions and wording with both AR and picture interfaces. An example question can be seen in Fig. 3. We did not use any specific wording pattern to form the interrogative sentences. The physical environment in both tasks was divided into sections. The environment of the low mobility task was divided into two sections and the high mobility task was divided into four sections or quadrants. The scale of the low mobility task was smaller than the scale of the high mobility task and the targets were distributed over a smaller area. Therefore, the low mobility task required less physical movement and fewer viewpoint alignments to see the targets clearly (for example, to read a serial code clearly, it was easier to view it from the correct direction). All sections consisted of one desktop PC placed on its side on top of a 70-cm high table. In the high mobility task, there were two section pairs that resembled each other with minor differences; the objective was to mimic a real-world environment such as a server facility that has many similarities.

The 20 targets in both tasks were distributed equally among the sections (yellow numbers in Fig. 4; five targets per section in the low mobility task and ten in the high mobility task). The targets were placed so that for each section of the environment there was a specific viewing angle from which to conduct the inspection (yellow arrows in Fig. 4). From these viewing angles, the targets could be clearly observed without occlusion. However, it was not, in fact, necessary to indicate the specific viewpoint because the participants needed only to stand on the correct side of the environment (two sides in the low mobility task and four in the high mobility task; see Fig. 4). The picture interface provided separate images for each section, two images for the low mobility task and four images for the high mobility task. We used separate images for each section, because if only one image had been used, some targets would have been occluded because of the 3D structure of the environment. Both tasks involved a total of eight viewpoint alignments between sections.

Only one annotation was displayed at a time. Annotations were binary yes/no questions that the participants were required to answer. For example, "Is the cable connected?" or "Have three screws been placed?" We consider the information density in both tasks high because the task environment contained several similar cables and other computer parts in close proximity to each other. Here, information density refers to the possible targets within an environment. Because of the within-group design of the study, we prepared two equally difficult versions of both tasks: two equally difficult versions of the low mobility task and two equally difficult versions of the high mobility task. The equal difficulty level similarity between the versions (e.g., between the two different versions of the low mobility task) were validated in a pilot study. Both versions included targets of the same type but in different locations. The distance between the targets was the same. Different task versions were randomized and counterbalanced within the test sessions.

3.4 Objective Measurements

We took three objective measurements: task time, error rate, and the number of gaze shifts. Task time refers to the total time taken to inspect all 20 targets in each condition. Participants started the timing manually, and it was automatically stopped after they had inspected all 20 targets. No time limit was set for the tasks.

Error rate measurements were based on the number of errors in the answers of the participants to the binary yes/no questions. For example, one question was "Is the blue cable connected?" After reading the question, a participant had to check the status of the blue cable. If the cable was connected and a participant answered "yes," it was counted as a correct answer. In addition to the cables, participants had to verify the serial numbers on parts, placement of memory and other card modules, number of screws,

1. https://support.apple.com/kb/SP662

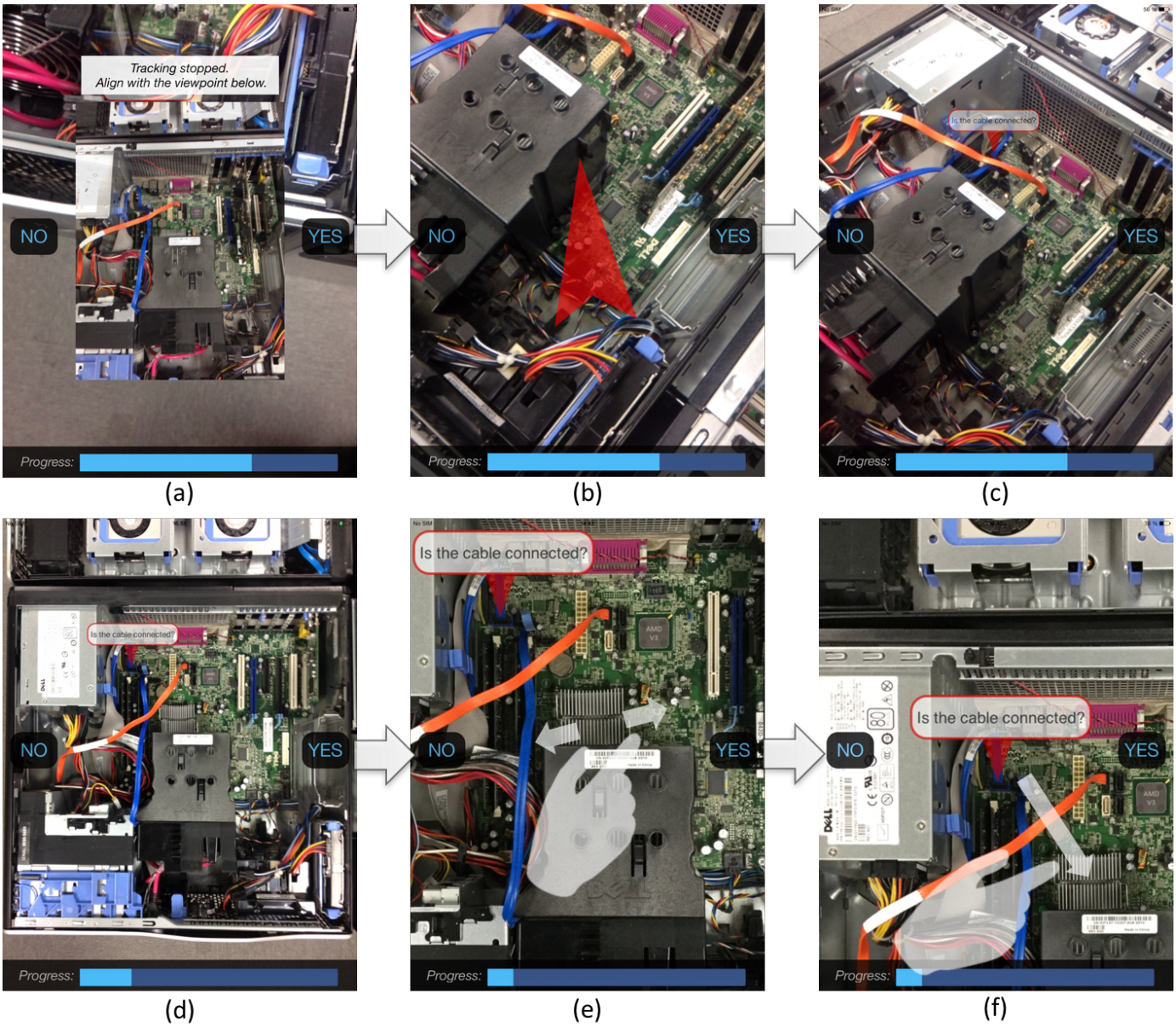


Fig. 3. (a)-(c): AR interface. If the system is not tracking, instructions and an image of the initial viewpoint are shown in the middle of the screen (a). When an annotation is off-screen, the interface displays an arrow pointing to the direction of the annotation (b). Finally, the interface shows a real-time augmented view with a 3D registered annotation (c). (d)-(f): the picture interface. A non-registered annotation is displayed in a static top-down image (d). A desired area of the image can be zoomed with a pinch-and-zoom gesture (e) and further panned if necessary (f). The transparent white hands illustrate touch gestures.

and other computer assembly-related issues. A correct answer was either "yes" or "no", both tasks included 10 questions to which the correct answer was "yes" and 10 where the correct answer was "no." The highest theoretical error count was thus 20. If a participant had answered only "yes" or only "no" the total number of errors would have been 10.

A gaze shift constitutes a sequence where a participant switches his/her eye focus from the display of the device to the environment followed by switching his/her focus back to the display of the device. This sequence was coded as one gaze shift event. All single gaze shift events during a trial were aggregated into one value for each participant.

3.5 Procedure

The user study procedure was as follows: a pre-questionnaire, instructions, tutorial tasks, all four conditions, and a post-questionnaire. For the AR interface, we explained the tracking operation to the participants and the actions that they should perform if tracking was lost. For the picture interface, we emphasized the zooming function and explained that the device could also be rotated if necessary to align the viewpoint more easily.

Participants could practice using both interfaces in tutorial tasks, during which they received feedback. The purpose of the tutorial tasks was also to allow participants to become familiar with the computer hardware. After each condition, the participants filled out a questionnaire about the condition and were given a five-minute break. During the break, a short introduction to the

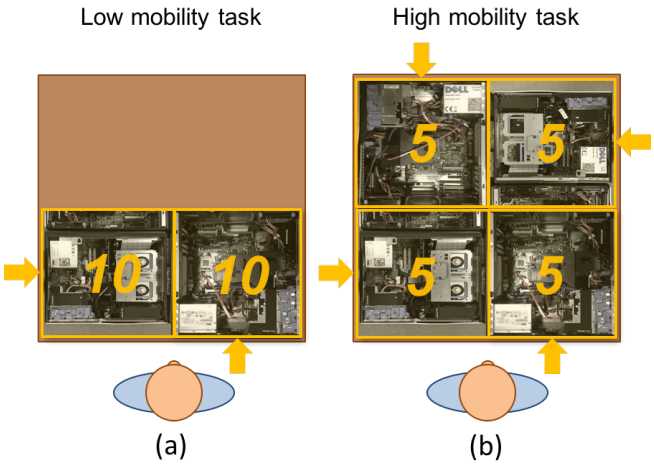


Fig. 4. Illustrations of the test tasks taken from a top-down viewpoint. The environment of the low mobility task (a) has two separate sections, both with 10 targets. The environment of the high mobility task (b) has four separate sections, all with 5 targets. The yellow arrows indicate the specific viewing angle for each section.

interface in the next condition was given. After all the conditions were completed, the participants were invited to give free-form comments. The study took approximately 50 min per test participant.

3.6 Hypotheses

Our hypotheses focused on factors that are important measures of efficiency. The time spent on a task and the number of errors incurred are fundamental factors for maximizing work efficiency. The number of gaze shifts is important from the point of view of cognitive workload, which can affect the overall task efficiency. We formulated the following three hypotheses for our user study.

H_1**: The task completion time when using the AR interface is less than the task completion time when using the picture interface.** Because of the high information density and the several viewpoint alignments required, we assumed that information matching using the picture interface would require significantly more time and that the complexity of the task should not have a significant effect on the task time in the case of the AR interface because of its 3D-registered annotations.

H_2**: There is no difference between the interfaces in the number of errors.** We considered that the overall number of errors should be very small because the targets (cables, screws, etc.) were clear and there was no ambiguity regarding their status (e.g., whether or not a cable was connected). Despite a greater amount of time being required for the information matching, we considered that the picture interface should produce a number of errors that was as small as the AR interface errors.

H_3**: The AR interface causes fewer gaze shifts than the picture interface.** Because the AR interface included 3D registered annotations, thus decreasing the amount of information matching, we considered that the AR interface should cause fewer gaze shifts between the display of the device and the environment, in contrast to the picture interface, which required the participants to constantly shift their gaze between the device and the environment.

4 Results

In this section, we describe the results of each objective and subjective measurement separately. Table 1 summarizes the results of the quantitative measurements. Basically, we employed a repeated-measure ANOVA to verify the interaction effect of interface × task.

4.1 Task Completion Time

Figure 5(a) shows the average task completion times. We noticed a significant difference between the interfaces in terms of the overall completion of time for both tasks. We checked the variance of the data by using Bartlett's test. According to the result of this test, we could not reject the assumption of equal variances ($p = 0.015$). Therefore, we decided to use a repeated-measure ANOVA, which showed that, when using the AR interface (both tasks: $M = 208$, $SD = 21.21$), the participants accomplished the task significantly faster than when using the picture interface (both tasks: $M = 270$, $SD = 56.6$): $F(1, 23) = 13.162$, $p < .001$. Similarly, we noticed a significant effect of the task difficulty on the completion time: $F(1, 23) = 14.81$, $p = .001$. The results support H_1. We did not notice any significant interaction effect of interface × task.

4.2 Number of Errors

We noticed a significant difference between the interfaces in terms of the overall error count for both tasks (Fig. 5(b)). We checked the variance of the data by using Bartlett's test. According to the results of this test, we could not reject the assumption of equal variances ($p = 0.020$). Therefore, we decided to use a repeated-measure ANOVA. To verify H_2, i.e., the AR interface produces a smaller number of errors, we used a repeated-measure ANOVA, which showed that the AR interface (both tasks: $M = 0.87$, $SD = 0.93$) caused significantly fewer errors than the picture interface (both tasks: $M = 1.48$, $SD = 1.35$): $F(1, 23) = 6.279$, $p < .05$. Similarly, and as expected, we noticed a significant effect of the task difficulty on the completion time: $F(1, 23) = 4.613$, $p < .05$. The results reject H_2. We did not find a significant difference in terms of the interaction effect of interface × task.

4.3 Gaze Shifts

Figure 5(c) shows the average number of gaze shifts per minute. Gaze shifts were calculated manually from video recordings that were captured using an omni-directional camera placed in the middle of the environment. In the low mobility task (see Figure 4) , the camera was in the upper corner between the two sections, and in the high mobility task, the camera was in the middle of the four sections. The calculation was performed three times by a single person. We checked the variance of the data by using Bartlett's test. According to the results, we could not reject the assumption of equal variances ($p = 0.025$). Therefore, we decided to use a repeated-measure ANOVA. We noticed a significant difference between the interfaces in terms of the overall gaze shift number for both tasks. A repeated-measure ANOVA showed that the AR interface (both tasks: $M = 5.89$, $SD = 0.47$) caused significantly fewer gaze shifts than the picture interface (both tasks: $M = 13.91$, $SD = 0.74$): $F(1, 23) = 244.162$, $p < .001$. The results support H_3. We did not notice a significant difference according to the task difficulty or any significant interaction effect of interface × task.

2. http://human-factors.arc.nasa.gov/groups/TLX/

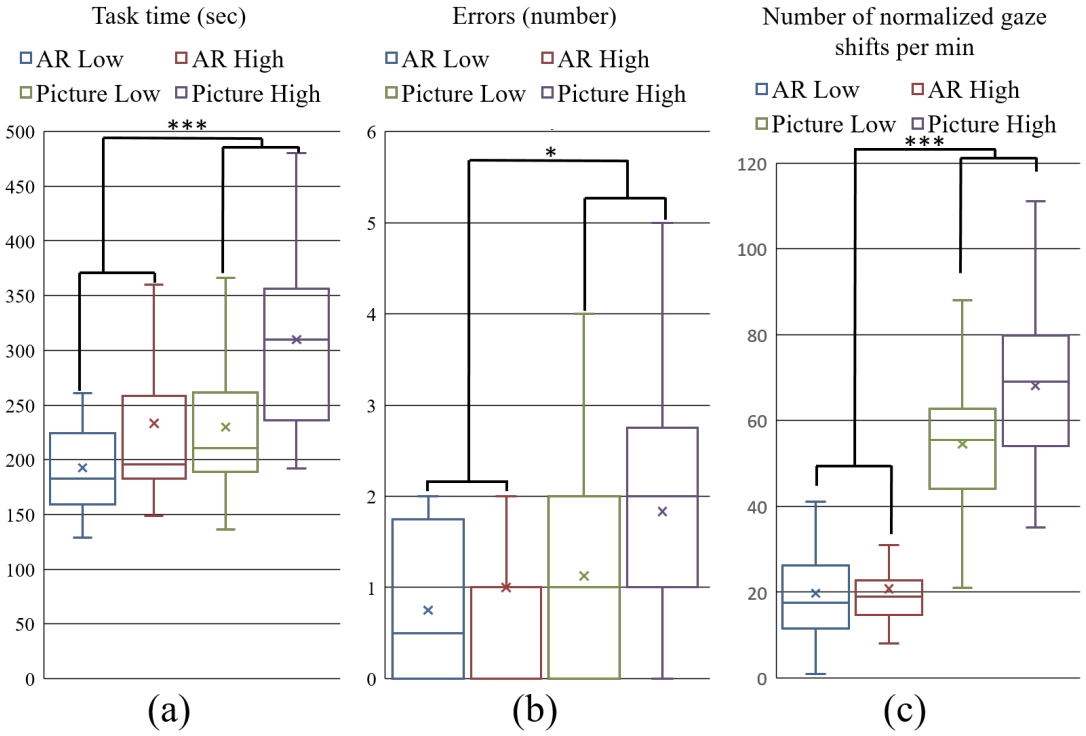


Fig. 5. Box plots of measurements (× represents average values): task completion times in seconds (a), number of errors (b), and number of gaze shifts per minute (c). Connected bars represent significant differences between means for the AR interface and picture interface (* = significant at < 0.05 level, *** = significant at < 0.001 level). N = 24.

TABLE 1
Results from the quantitative measurements (N = 24)

Interface	Task	Task Time (sec)		Errors (Number)		Normalized Gaze Shifts (Number per min)		NASA TLX (unweighted score)	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
AR	Low mobility	193	52	0.75	0.84	6.22	3.15	2.38	0.58
Picture	Low mobility	230	70	1.11	1.11	14.43	2.46	3.92	0.96
AR	High mobility	233	86	1.00	1.02	5.56	2.16	2.88	0.68
Picture	High mobility	310	93	1.83	1.49	13.39	2.16	4.76	0.96

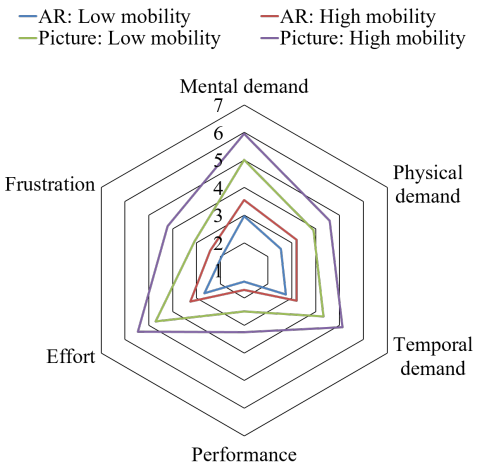


Fig. 6. Average scores for each NASA TLX questionnaire.

4.4 Subjective Feedback

We collected subjective feedback through a NASA TLX[2] questionnaire and written free-form comments. NASA TLX had a 7-point Likert scale. Table 1 shows the unweighted scores of the NASA TLX. A lower score indicates a better result. The average scores for each question are shown in Fig. 6. The figure shows that the participants felt that the AR interface was better than the non-AR picture interface.

The free-form comments revealed that several participants felt that locating targets with the AR interface was easy because they needed only to follow the red indication arrow if the annotation was not immediately visible on the screen of the devise. Participants also mentioned that it was frequently possible to conduct the inspection via the screen of the devise without looking directly at the target. However, some participants said that this was not always possible because of the low resolution (480×640 pixels) of the video feed of the devise. The most serious issues in the AR interface were considered the occasional tracking problems, jittering of the 3D-registered annotations, and occlusion. A few participants stated that it was sometimes physically demanding to

manipulate the device because the viewpoint had to be maintained in a certain position to allow them to see the annotations.

The participants commented that the picture interface was intuitive and easy to learn because they had previous experience with similar guides. Furthermore, they could hold the device in any position and orientation, which demanded less physical effort. The most serious drawback of the picture interface was the difficulty in locating the information of the test environment based on the non-registered annotations. Finally, a few participants stated that it was easy to make mistakes when using the picture interface because of the high information density.

4.5 Observations

When using the AR interface, the participants occasionally (0-2 times per participant) experienced loss of tracking because they moved the device fast or tried to zoom in by moving the device too close to targets in the environment. If an annotation was off-screen, several participants panned and moved the device rather than first rotating the viewpoint. This sideways movement caused greater vulnerability to tracking issues. The tracking initialization was not always instant, and sometimes participants had to spend time making very subtle viewpoint adjustments to operate the tracking function. A few participants perceived the location of the target incorrectly because of occlusion, which caused them to answer the yes/no questions incorrectly. Sometimes, participants bumped into the computers in the environment while moving because they were focusing only on the augmented representation of the environment on the display of the devise, ignoring the real environment.

When they used the picture interface, the high information density caused participants to occasionally inspect incorrect targets within a section, for example, when several cables having a close resemblance were in near proximity to each other. A few times, participants inspected an entirely incorrect section, which occurred more frequently because of several similarities between the sections. Most participants rotated their body according to the angle in the images, but some rotated the handheld device in their hands to match the viewpoint in the picture with the environment.

5 Discussion

The discussion is divided into two sections. In the first, we interpret the results in terms of the hypotheses, and in the second, we concentrate on the various aspects of the interfaces and tasks used in the study.

5.1 Measurements

We speculate that the AR interface produced faster results for two main reasons. First, participants did not need to match the information in the guide with the environment. Second, they could conduct most of the inspections through the representation of the environment on the display of the tablet. Comprehending and transforming information from annotations was more demanding when using the picture interface in our test tasks. However, we assume that the difference in task completion times would decrease if the users were more familiar with the environment because locating the correct section and the targets within a section would become easier in the case of the picture interface. Increased familiarity would not affect the use of AR significantly because users would still simply follow the instructions on the screen. In the case of the AR interface, the most important factor affecting the task time may be the scale of the environment, because users must move more between targets, and movement can cause more tracking issues.

The overall average error rates were higher than expected. Based on the observations and subjective feedback, we speculate that the most frequent cause of errors was that the participants did not inspect the correct targets. In the case of the AR interface, errors were caused by the issues in the technical implementation of our HAR system itself because it did not include means of handling occlusion and did not provide the participants with instructions about the specific viewing angles. In the picture condition, errors were related to issues in information matching and the alignment of the images and the task environment.

These observations showed that the problems in the AR condition were technology-related, whereas in the picture condition, they were caused by more fundamental issues related to information matching. Furthermore, a longer task time and increased fatigue may be additional reasons for increased errors, since they may cause users to lose concentration. Although the task time for each test was quite short, the effort involved in holding a tablet device and moving around can affect the concentration of the user. In addition, the average scores in the NASA TLX questionnaire for the AR interface are lower (lower is better) than the average scores of the picture interface. Based on these observations, we can assume that the information density affected the number of errors in the case of both interfaces. The test participants were graduate students, and inspection experts might have made fewer errors during the inspection, as they would have been more familiar with complex machine environments.

The differences between the interfaces in the number of gaze shifts were very high, as expected. Because of the non-registered annotations, the picture interface requires users to constantly shift their eye gaze between the device and the environment. Furthermore, the fact that the gaze shifts were very frequently very quick may have substantially affected their overall number. Although a small number of gaze shifts is, in general, considered positive [32], focusing solely on the augmented representation of the environment on the screen of the devise can cause safety issues; we observed that sometimes participants bumped into the computers in the environment. The smaller average number of gaze shifts in the high mobility task could have been caused by the participants spending more time on moving from one target to another and not shifting their gaze while moving. However, we did not measure this.

Based on the objective results, we can state that AR was more effective in terms of efficiency. Our subjective measurements strongly correlate with the objective results. We did not objectively measure time spent on mental rotation and angle alignments; however, the subjective feedback suggests that this was an issue in the case of the picture interface, providing additional implications that the need for mental rotation can affect the work efficiency of users in real-world inspection tasks.

We consider that two factors can explain the differences in our results compared to the results of previous studies involving HAR and picture interfaces [11], [13]: our tasks did not require physical manipulation and the information density of our test environment was higher. However, a few interface improvements could render the use of HAR viable, even when physical manipulation is necessary.

We did not measure device movement, and, as mentioned, some of the participants noted the manipulability issues of HAR.

Even if the AR interface requires the user to move the device more frequently, we do not consider this was an issue in the case of our test tasks, because they were rather short, and the other measurements were in favor of AR. However, in inspection tasks that take a longer time, the amount of device movement required and the fatigue of the user would be a very important measurement.

5.2 Interfaces and Tasks

The interfaces we evaluated were very simple presentations of AR and picture guides. Since the minimum features and functionality involved in the study allowed us to focus better on the differences in the information presentation media, we were able to obtain more generalizable results. However, both interfaces could be improved. For example, the AR interface could provide instructions about specific viewpoints for avoiding occlusion. In the AR interface, the feature-based SLAM library was used for tracking the device movement. Feature-based SLAM can estimate a sparse 3D structure of a target environment. To handle occlusion, a dense SLAM algorithm, such as DTAM [33], is required to obtain surface information of the target environment. Furthermore, the addition of a freeze mode could facilitate manipulability. The picture interface could display an overview or an indication of the location of the section where the target is to be found. However, a separate overview might have rendered the use of the interface more complex.

One of the advantages of a picture interface is that users can easily obtain an overview of the environment and zoom in and out without moving the device or changing their position. Thus, if users are familiar with the environment, they can more quickly obtain an approximate location of the section and the target when using a picture interface. However, one of the benefits of AR is that no additional time is required for familiarization with the environment.

We did not objectively measure learnability, but we can speculate that the AR interface is more difficult to learn simply because users are not familiar with the technology, whereas most users of handheld devices are familiar with interacting with pictures via touch gestures, suggesting that even a simple AR interface may not be suitable for widespread adoption without appropriate user training or improved instructions within the interface.

We did not include a task that does not require viewpoint alignments because AR has not been shown to offer significant benefits over picture interfaces in this type of task [11], [13]. If the viewpoint is not fixed, while using the picture interface the user could place the handheld picture guide in a fixed position, which would allow easy manipulability. The complexity of inspection tasks can vary considerably according to the real-world domain, meaning that the effectiveness of an AR compared to a picture interface can also vary. The purpose of our test tasks was to provide a generic representation of complex inspection tasks and shed light on the advantages and disadvantages of both interfaces in these types of task. The inputs to our interfaces were only binary, but real world inspection may require more complex inputs and the observation of 2D content, such as text or charts. Thus, handheld devices are a more attractive option than HMDs in inspection tasks.

6 Conclusions and Future Work

In this work, we compared the effectiveness of a handheld AR and a non-AR picture interface in terms of efficiency in inspection tasks with high information density that require the user to move and perform several viewpoint alignments. We showed that the AR interface was more effective in terms of the task completion time, error rate, and number of gaze shifts of the users, mainly because the 3D-registered annotations in the AR interface made matching information from the guide to the environment effortless.

Our research goal was to empirically validate the advantages of an HAR over a picture interface in procedural observation tasks, which are otherwise known only from anecdotal evidence. The results of our study provide valuable insights into the practicality of HAR. A quantitative evaluation can strongly support prior expectations, and our results can be used as evidence of the usefulness of HAR products. It has been proven that HAR provides easy-to-use techniques for correctly positioning the annotations in 3D [34], [35] without using any external hardware. The existence of these techniques and the results of our study allow us to state that AR by means of off-the-shelf handheld devices is already a practical option in real-world task support.

In future work, we plan to make several improvements to both interfaces. Better occlusion handling as well as a freeze mode and zooming capability could be added to the AR interface to prevent users from bumping into objects. In addition, the AR interface should provide real-time feedback about the tracking quality and initialization. The picture interface could provide better indications as to the section of the environment from which the images were captured. Finally, we will consider a combination of AR and a picture interface that allows users to seamlessly zoom in and out and switch between the two guide types.

We also need to evaluate the AR interface in additional real-world inspection tasks to gain a broader understanding of the benefits and issues related to this interface. We are also planning to compare the two interfaces in tasks that require physical manipulation in addition to direct observation. Furthermore, we need to evaluate interfaces in tasks where the targets are distributed in a larger 3D space, for example, on several horizontal and vertical surfaces.

References

[1] U. Neumann and A. Majoros, "Cognitive, performance, and systems issues for augmented reality applications in manufacturing and maintenance," in *Virtual Reality Annual International Symposium, 1998. Proceedings., IEEE 1998*, 1998, pp. 4–11.

[2] J. Ott, "Maintenance executives seek greater efficiency," *Aviation Week and Space Technology*, vol. 142, no. 20, p. 43, 1995.

[3] S. Kim and A. K. Dey, "Simulated augmented reality windshield display as a cognitive mapping aid for elder driver navigation," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '09. New York, NY, USA: ACM, 2009, pp. 133–142. [Online]. Available: http://doi.acm.org/10.1145/1518701.1518724

[4] S. Henderson and S. Feiner, "Augmented reality in the psychomotor phase of a procedural task," in *Proceedings of International Symposium on Mixed and Augmented Reality*, ser. ISMAR '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 191–200. [Online]. Available: http://dx.doi.org/10.1109/ISMAR.2011.6092386

[5] A. Tang, C. Owen, F. Biocca, and W. Mou, "Comparative effectiveness of augmented reality in object assembly," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '03. New York, NY, USA: ACM, 2003, pp. 73–80. [Online]. Available: http://doi.acm.org/10.1145/642611.642626

[6] D. W. F. R. van Krevelen and R. Poelman, "A survey of augmented reality technologies, applications and limitations," *International Journal of Virtual Reality*, vol. 9, no. 2, pp. 1–20, Jun. 2010.

[7] T. Vincent, L. Nigay, and T. Kurata, "Classifying handheld augmented reality: Three categories linked by spatial mappings," 2012.

[8] T. Olsson and M. Salo, "Online user survey on current mobile augmented reality applications," in *Proceedings of International Symposium on Mixed and Augmented Reality*, 2011, pp. 75–84.

[9] S. Kurkovsky, R. Koshy, V. Novak, and P. Szul, "Current issues in handheld augmented reality," in *Communications and Information Technology (ICCIT), 2012 International Conference on*, June 2012, pp. 68–72.

[10] M. Träskbäack and M. Haller, "Mixed reality training application for an oil refinery: User requirements," in *Proceedings of the 2004 ACM SIGGRAPH International Conference on Virtual Reality Continuum and Its Applications in Industry*, ser. VRCAI '04. New York, NY, USA: ACM, 2004, pp. 324–327. [Online]. Available: http://doi.acm.org/10.1145/1044588.1044658

[11] J. Polvi, T. Taketomi, G. Yamamoto, M. Billinghurst, C. Sandor, and H. Kato, "Evaluating a slam-based handheld augmented reality guidance system," in *Proceedings of the 2Nd ACM Symposium on Spatial User Interaction*, ser. SUI '14. New York, NY, USA: ACM, 2014, pp. 147–147. [Online]. Available: http://doi.acm.org/10.1145/2659766.2661212

[12] S. Henderson and S. Feiner, "Evaluating the benefits of augmented reality for task localization in maintenance of an armored personnel carrier turret," in *Proceedings of International Symposium on Mixed and Augmented Reality*, 2009, pp. 135–144.

[13] C. Liu, S. Huot, J. Diehl, W. Mackay, and M. Beaudouin-Lafon, "Evaluating the benefits of real-time feedback in mobile augmented reality with hand-held devices," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '12. New York, NY, USA: ACM, 2012, pp. 2973–2976.

[14] L. A. Cooper, "Mental rotation of random two-dimensional shapes," *Cognitive Psychology*, vol. 7, no. 1, pp. 20 – 43, 1975. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0010028575900031

[15] R. N. Shepard and J. Metzler, "Mental rotation of Three-Dimensional objects," *Science*, vol. 171, no. 3972, pp. 701–703, Feb. 1971. [Online]. Available: http://dx.doi.org/10.1126/science.171.3972.701

[16] E. Laurier and B. Brown, "Rotating maps and users: praxiological aspects of alignment and orientation," *Transactions of the Institute of British Geographers*, vol. 33, no. 2, pp. 201–216, 2008.

[17] G. W. Zimmerman, D. Klopfer, G. M. Poor, J. Barnes, L. Leventhal, and S. D. Jaffee, ""how do i line up?": Reducing mental transformations to improve performance," in *Proceedings of the 14th International Conference on Human-computer Interaction: Design and Development Approaches - Volume Part I*, ser. HCII'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 432–440. [Online]. Available: http://dl.acm.org/citation.cfm?id=2022384.2022435

[18] C. M. Robertson, B. MacIntyre, and B. N. Walker, "An evaluation of graphical context when the graphics are outside of the task area," in *Proceedings of International Symposium on Mixed and Augmented Reality*, ser. ISMAR '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 73–76. [Online]. Available: http://dx.doi.org/10.1109/ISMAR.2008.4637328

[19] K. M. Baird and W. Barfield, "Evaluating the effectiveness of augmented reality displays for a manual assembly task," *Virtual Reality*, vol. 4, no. 4, pp. 250–259. [Online]. Available: http://dx.doi.org/10.1007/BF01421808

[20] H. Ishii, Z. Bian, H. Fujino, T. Sekiyama, T. Nakai, A. Okamoto, and H. Shimoda, "Augmented reality applications for nuclear power plant maintenance work," in *ï£ijï£ijï£ijProceedings of International Symposium on Symbiotic Nuclear Power Systems*, 2007, pp. 262–268.

[21] J. Platonov, H. Heibel, P. Meier, and B. Grollmann, "A mobile markerless ar system for maintenance and repair," in *Proceedings of International Symposium on Mixed and Augmented Reality*, Oct 2006, pp. 105–108.

[22] A. Mulloni, H. Seichter, and D. Schmalstieg, "User experiences with augmented reality aided navigation on phones," in *Proceedings of International Symposium on Mixed and Augmented Reality*, Oct 2011, pp. 229–230.

[23] A. Dünser, M. Billinghurst, J. Wen, V. Lehtinen, and A. Nurminen, "Exploring the use of handheld ar for outdoor navigation," *Comput. Graph.*, vol. 36, no. 8, pp. 1084–1095, Dec. 2012. [Online]. Available: http://dx.doi.org/10.1016/j.cag.2012.10.001

[24] E. Jung, S. Oh, and Y. Nam, "Handheld ar indoor guidance system using vision technique," in *Proceedings of the 2007 ACM Symposium on Virtual Reality Software and Technology*, ser. VRST '07. New York, NY, USA: ACM, 2007, pp. 47–50. [Online]. Available: http://doi.acm.org/10.1145/1315184.1315190

[25] M. Rauhala, A.-S. Gunnarsson, and A. Henrysson, "A novel interface to sensor networks using handheld augmented reality," in *Proceedings of the 8th Conference on Human-computer Interaction with Mobile Devices and Services*, ser. MobileHCI '06. New York, NY, USA: ACM, 2006, pp. 145–148. [Online]. Available: http://doi.acm.org/10.1145/1152215.1152245

[26] K. Makita, T. Vincent, S. Ebisuno, M. Kourogi, T. Ishikawa, T. Okuma, M. Yoshida, L. Nigay, and T. Kurata, "Mixed reality navigation on a tablet computer for supporting machine maintenance in wide-area indoor environment," in *Conference Proceedings of ICServ2014, the 2nd International Conference on Serviceology*. Yokohama, Japan: Springer, 2014, Conférences internationales de large diffusion avec comité de lecture sur texte complet, pp. 41–47.

[27] M. Hakkarainen, C. Woodward, and M. Billinghurst, "Augmented assembly using a mobile phone," in *Proceedings of International Symposium on Mixed and Augmented Reality*, Sept 2008, pp. 167–168.

[28] N. Karlsson, G. Li, Y. Genc, A. Huenerfauth, and E. Bononno, "iar: An exploratory augmented reality system for mobile devices," in *Proceedings of the 18th ACM Symposium on Virtual Reality Software and Technology*, ser. VRST '12. New York, NY, USA: ACM, 2012, pp. 33–40. [Online]. Available: http://doi.acm.org/10.1145/2407336.2407343

[29] S. Gauglitz, C. Lee, M. Turk, and T. Höllerer, "Integrating the physical environment into mobile remote collaboration," in *Proceedings of the 14th International Conference on Human-computer Interaction with Mobile Devices and Services*, ser. MobileHCI '12. New York, NY, USA: ACM, 2012, pp. 241–250. [Online]. Available: http://doi.acm.org/10.1145/2371574.2371610

[30] S. Gauglitz, B. Nuernberger, M. Turk, and T. Höllerer, "World-stabilized annotations and virtual scene navigation for remote collaboration," in *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '14. New York, NY, USA: ACM, 2014, pp. 449–459. [Online]. Available: http://doi.acm.org/10.1145/2642918.2647372

[31] S. Burigat, L. Chittaro, and S. Gabrielli, "Visualizing locations of off-screen objects on mobile devices: A comparative evaluation of three approaches," in *Proceedings of the 8th Conference on Human-computer Interaction with Mobile Devices and Services*, ser. MobileHCI '06. New York, NY, USA: ACM, 2006, pp. 239–246. [Online]. Available: http://doi.acm.org/10.1145/1152215.1152266

[32] M. Rohs, R. Schleicher, J. Schöning, G. Essl, A. Naumann, and A. Krüger, "Impact of item density on the utility of visual context in magic lens interactions," *Personal and Ubiquitous Computing*, vol. 13, no. 8, pp. 633–646, 2009. [Online]. Available: http://dx.doi.org/10.1007/s00779-009-0247-2

[33] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "DTAM: Dense tracking and mapping in real-time," in *2011 International Conference on Computer Vision*, Nov 2011, pp. 2320–2327.

[34] A. Henrysson, M. Billinghurst, and M. Ollila, "Virtual object manipulation using a mobile phone," in *Proceedings of the 2005 International Conference on Augmented Tele-existence*, ser. ICAT '05. New York, NY, USA: ACM, 2005, pp. 164–171.

[35] J. Polvi, T. Taketomi, G. Yamamoto, A. Dey, C. Sandor, and H. Kato, "Slidar: A 3d positioning method for SLAM-based handheld augmented reality," *Computers and Graphics*, vol. 55, pp. 33 – 43, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0097849315001806

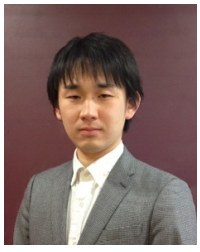
Jarkko Polvi received the B.S. and M.S. degrees in information processing science from the University of Oulu in 2011 and 2012, respectively. He received the Ph.D. degree in information science from Nara Institute of Science and Technology in 2016. His research interests include augmented reality and UI/UX research of mobile devices.

Takafumi Taketomi received the B.E. degree from the National Institute for Academic Degrees and University Evaluation in 2006. He received the M.E. and Ph.D. degrees in information science from the Nara Institute of Science and Technology in 2008 and 2011, respectively. He has been an assistant professor at the Nara Institute of Science and Technology since 2011.

Atsunori Moteki received the B.S. and M.S. degrees in information science and technology from the University of Tokyo in 2009 and 2011, respectively. He is a researcher in the Media Processing Laboratory at Fujitsu Laboratories Ltd. His main interests include augmented reality and SLAM, and their application to human-centric user interfaces.

Toshiyuki Yoshitake received the B.S. and M.S. degrees in electrical engineering from Nagoya University in 1996 and 1998, respectively. He is the Research Manager of the Media Processing Laboratory at Fujitsu Laboratories Ltd. His main interests include augmented reality, computer vision, and 3D computer graphics.

Toshiyuki Fukuoka received the B.E. and M.E. degrees in engineering science from Osaka University, Japan in 1989 and 1991, respectively. He is a director of the Media Processing Laboratory at Fujitsu Laboratories Ltd. His main interests include artificial intelligence, human machine interface, and affective media processing.



Goshiro Yamamoto received the B.E., M.E., and Ph.D. degrees in Engineering from Osaka University, Japan in 2004, 2006, and 2009, respectively. Since 2016, he has been a senior lecturer at Kyoto University Hospital. His major interests are human-computer interaction, projection-based augmented reality technologies, and wearable computing systems.

Christian Sandor is an Associate Professor at one of Japan's most prestigious research universities, the Nara Institute of Science and Technology (NAIST), where he is co-directing the Interactive Media Design Laboratory In 2005, he obtained a doctorate in Computer Science from the Munich University of Technology, Germany under the supervision of Prof. Gudrun Klinker and Prof. Steven Feiner. Before joining NAIST, he directed the Magic Vision Laboratory.

Hirokazu Kato is a professor at the Nara Institute of Science and Technology (NAIST), Japan. He received his B.E., M.E., and Dr. Eng. degrees from Osaka University, Japan in 1986, 1988, and 1996, respectively. Since 2007, he has been with the Graduate School of Information Science of NAIST. He received the Virtual Reality Technical Achievement Award from IEEE VGTC in 2009, and the Lasting Impact Award at the 11th IEEE International Symposium on Mixed and Augmented Reality in 2012.