NAIST-IS-DD1261206

Doctoral Dissertation

Usefulness of Handheld Augmented Reality in Inspection Tasks

Jarkko Polvi

March 8, 2016

Department of Information Science
Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Jarkko Polvi

Thesis Committee:

Professor Hirokazu Kato	(Supervisor)
Professor Naokazu Yokoya	(Co-Supervisor)
Associate Professor Christian Sandor	(Co-Supervisor)
Assistant Professor Takafumi Taketomi	(Co-Supervisor)

Usefulness of Handheld Augmented Reality in Inspection Tasks*

Jarkko Polvi

Abstract

Augmented Reality (AR) refers to the combination of real world and virtual objects that are registered in 3D and can be interacted in real-time. Handheld AR (HAR) refers to AR on handheld devices such as smartphones or tablet computers. Handhelds can be considered practical, affordable and they provide easy means for information input and sharing. Thus, handheld devices have the potential to enable wide adaptation of AR. However, unlike AR in general, HAR is often not considered useful in goal-oriented tasks due to insufficient utility and usability.

In this thesis, we investigate the usefulness of HAR in inspection tasks. These tasks mean the inspection of targets in a workpiece via visual observation based on the information provided by a checklist or other type of guidance medium. In addition to observation, adding information to the checklist is often also necessary. While conducting an inspection, users have to divide their attention between the checklist and the workpiece. This action can decrease work efficiency. With AR, we can reduce divided attention in inspection tasks by directly overlaying information from a checklist to a workpiece. Furthermore, handhelds enable us to easily add information to a checklist and share it with other users.

Usefulness is a combination of utility and usability. In order for HAR to be consider useful in inspection, we have to confirm if HAR enables proper functionality for positioning virtual annotations accurately to a workpiece (utility). We also need to evaluate if HAR provides any benefits (usability) compare to

*Doctoral Dissertation, Department of Information Science, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD1261206, March 8, 2016.

non-AR interfaces in visual observation tasks. We have applied Simultaneous Localization and Mapping (SLAM) based HAR prototypes to tasks that focus on 3D positioning of virtual annotations and to visual observation of the task environment. We have conducted total of four user studies: two studies focus on 3D positioning and two to visual observation tasks.

This doctoral thesis has two main contributions: Firstly, we are the first to evaluate a ray-casting based AR 3D positioning method against a conventional device-centric method. We have confirmed the higher efficiency of a ray casting based method. Secondly, we are the first to apply HAR to complex visual observation tasks that require movement and viewpoint alignment. We have proven AR's higher efficiency over a non-AR picture interface. Even though we focus only on SLAM-based HAR, our findings can be applied to other types of AR display technologies and tracking techniques as well.

Keywords:

augmented reality, handheld devices, inspection tasks, usability

Contents

Contents

List of Figures

List of Figures

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1. Augmented Reality

Augmented Reality (AR) is a novel technology that can change the way people access, create and consume digital information in their everyday lives. In short, AR means the augmentation of the real physical world with computer-generated content. This superimposed content can be visual, auditory, or other sensory enhancements such as haptics.

1.1.1 Definition and Taxonomy

Milgram et al. [52] define AR as part a reality-virtuality continuum (Fig. 1.1) that consists of four areas: real environment, AR, Augmented Virtuality (AV), and virtual environment or Virtual Reality (VR). The different degrees of combinations of real and virtual are generally referred to as Mixed Reality (MR). AR is part of MR and in AR, the real environment is dominant and virtual augmentations only extends the reality without replacing it completely. In contrast, AV consists mostly virtual content and only a small portion of real enviroment. AR can be seen as a complimentary for immersive VR. One of the most well-known

Figure 1.1. The reality-virtuality continuum. AR and AV form the mixed reality that contains both real and virtual elements. [52].

definitions for AR is proposed by Azuma [2] and he defines that AR consist of three main characteristics:

1. Combines real and virtual
2. Interactive in real time
3. Registered in 3D

AR can also be seen as an effort to make conventional computer interfaces invisible and enhance user interaction with the real environment. Rekimoto and Nagao [72] have defined a taxonomy for different Human-Computer Interaction (HCI) styles (Fig. 1.2) in conventional and novel interfaces. They distinguishes between traditional Graphical User Interfaces (GUIs) and those (VR, ubiquitous computers, AR) that attempt to make the computer interface invisible. GUIs have a distinct separation between the on-screen digital domain an the real world. The separation in AR is less obvious and the main goal of AR is to enhance reality with digital content in a non-immersive way where the user still sees the real world.

1.1.2 Display Techniques

AR display techniques can be roughly divided into three categories based on the display position between the user's eye and the real world: head-mounted, handheld, and spatial displays [6]. Figure 1.3 illustrates the display categories and Figure 1.4 shows an example application from each category. Head-Mounted Displays (HMDs) and handhelds require a display surface, either video or optical

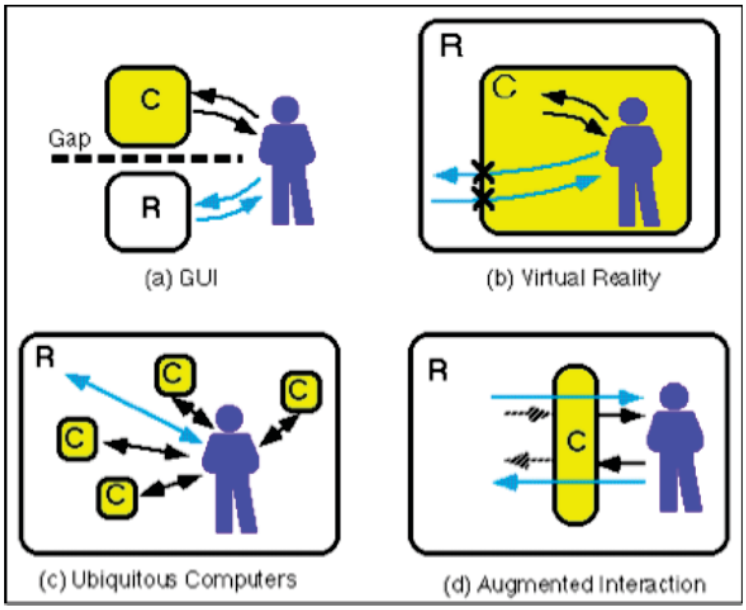

Figure 1.2. A taxonomy and comparison of different HCI styles (R = real world, C = computer)[72].

see-through. Spatial displays refer to large displays or projection of AR content directly to the real object. These systems can utilize optical see-through displays or no displays at all, only the projected surface [83]. Spatial displays are generally not movable. Head-mounted and handheld displays can be mobile and allow users to move in space. Movability is important in inspection tasks and because of this, we do not discuss about spacial displays more in this thesis.

1.1.2.1 Head-Mounted Displays

HMDs consist a wearable helmet or glasses that provides a display surface directly in front of user's eyes. An HMD can be used to show AR content either in an optical or in video see-through. Optical see-through HMDs use LCD screens and mirrors or light-field displays. This allows the user to see the real world with their unaided eyes while virtual images are overlaid on the view. A video see-through HMD display one or two video cameras are mounted on the HMD. The video stream of the real world captured by the video cameras is combined with the virtual content and the combination of real and virtual is then displayed to LCD

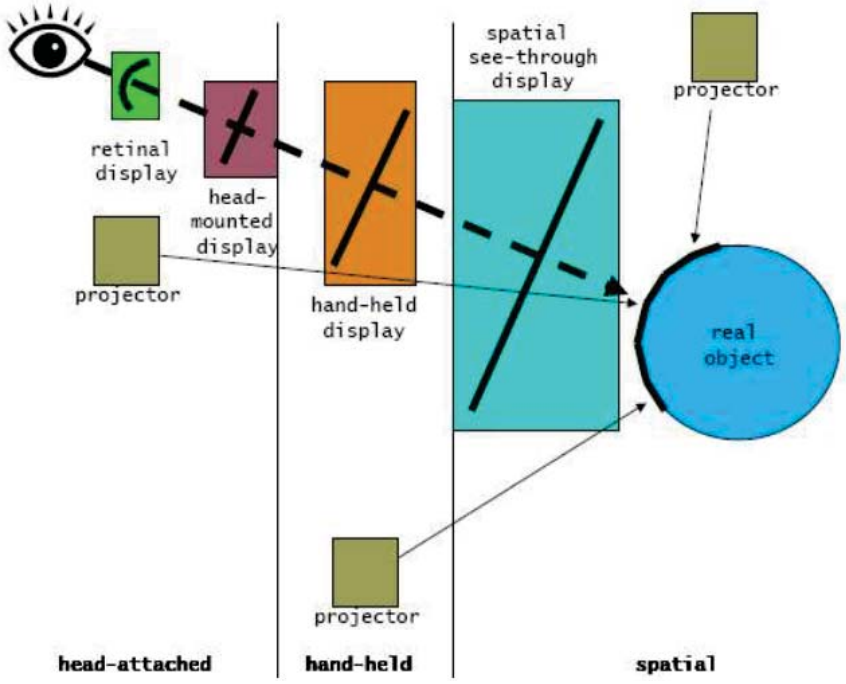

Figure 1.3. The three main categories of AR display types: Head-mounted, hand-held, and spatial displays [6].

displays in front of users eyes.

HMDs are the most common display technique that is used in AR research [6]. The main benefit of HMD is a visually immersive user experience as well as the possibility to use them handsfree. This leaves user's hands free for interaction with the real world or with various input devices. This makes observing the AR view also easier when it is not necessary to hold an external display in front of point of view. However, there are also still many technical limitations and HMDs often require an external computer. Even though HMDs are getting smaller, they can still be too cumbersome and current HMDs are still not cheap enough for commercial use. However, this might change when lower cost HMD are coming to market.

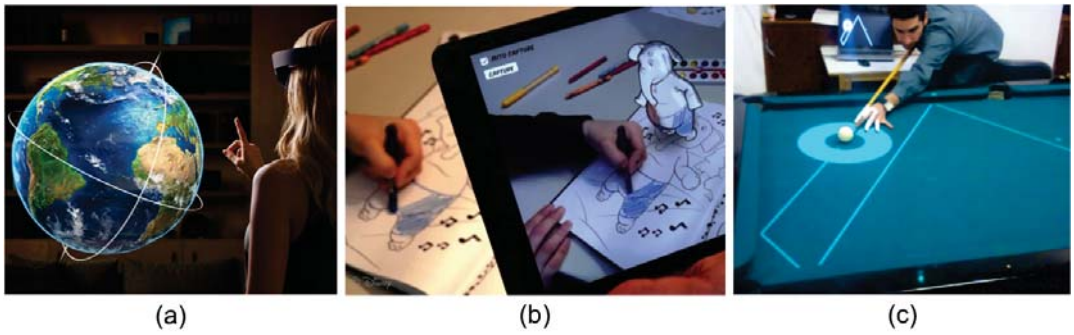

Figure 1.4. Example systems from the three AR display categories: (a) a HMD system (Microsoft HoloLens), (b) a handheld system (Disney Drawing), and (c) a spatial AR system (ARPool).

1.1.2.2 Handheld Displays

Handheld displays refer to AR on handheld devices such as modern smartphones or tablet PCs. Handhelds mainly utilize video see-through where live video stream of the real world is captured with device's camera that is augmented with virtual content before displaying it on device's display. Handhelds with optical see-through exist, but are still just experimental design concepts. Some micro projectors exist for handheld AR that mostly involve holding a projector in a similar manner to a flashlight.

HMDs have problem related to wearability, safety issues with indirect real world view, and social acceptance. Handhelds are considered to be mobile and personal, yet sharable with other users if necessary. They are also more socially acceptable compared to HMDs [5] due to wide adoption of suitable handheld devices. Furthermore, handhelds offer easy input metaphors via buttons or touchscreen displays. Current handhelds have powerful graphics processors, cameras, and various sensors that can run AR applications eliminating the need for external hardware.

The drawback of handhelds is their manipulability: in order to observe the augmented view, a user has to hold the device in a certain viewpoint that can be physically tiring. Fairly small screen size of handheld devices' can also be an issue and prevent users from observing the AR content clearly. However, the current trend in smartphone market has been towards bigger screen sizes and some

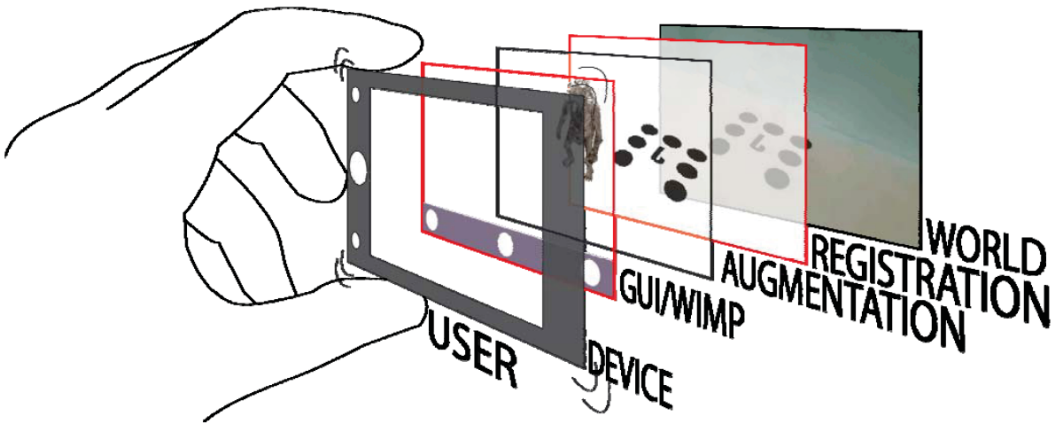

Figure 1.5. A visual representation of the HAR taxonomy and its' six layers. [21].

tablet PCs already utilize screens over ten inches. Even though handhelds can be considered as stand-alone devices for AR, they also have technical limitations and are not necessarily able to run the most sophisticated tracking technologies. AR on handheld devices is discussed more in the next Section.

1.2. Handheld Augmented Reality

We define HAR as AR on handheld devices such as smartphones, tablet PCs, ultra-mobile laptops and other small devices that can be operated while on a move. HAR can be seen as a subset of mobile AR: where mobile AR means AR on any type of mobile hardware, (e.g., HMDs or movable projectors), HAR refers to AR solely on handhelds. Handhelds are very appealing platform for AR and can introduce AR for consumer use due to widespread adoption of handheld devices[44]. Gjosater [21] has described a taxonomy for HAR applications. The first three layers (User, Device, and GUI) can be seen as traditional layers common to all handheld applications. The remaining three layers (augmentation, registration, and world) are more or less specific for HAR applications.

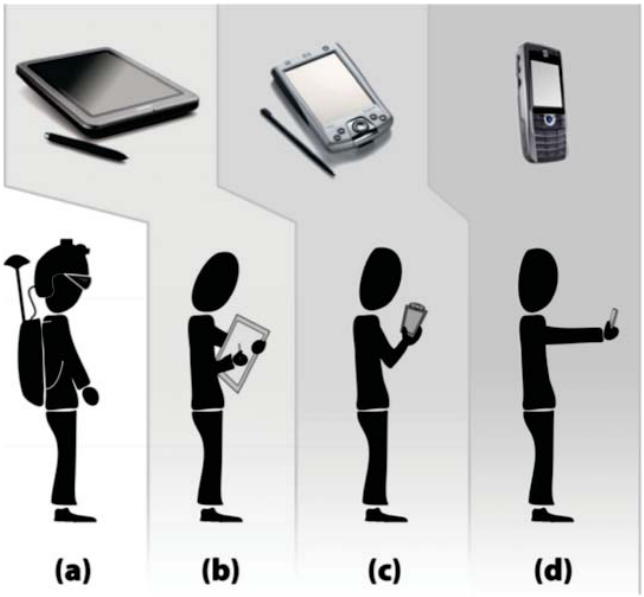

Figure 1.6. The evolution of HAR from mobile AR to it's current form factor: A backpack system with an HMD (a), ultramobile PCs (b), first tablet devices (c), and modern smartphones/tablets (d) [88].

1.2.1 History and Current State

HAR has evolved from cumbersome HMD-laptop systems all the way to modern handheld systems (Fig. 1.6) [88]. HAR was applied to tablets before current smartphones. However, tablet systems became popular for AR again approximately five years ago when Apple introduced it's first iPad[1] tablet. Modern smartphones and tablets are currently the most used platform for commercial HAR and for HAR research as well.

First HAR applications emerged at the beginning of 21st century when Personal Digital Assistants (PDAs) gained more popularity and their processing power was high enough to handle complex AR applications. Several researches used PDAs for various HAR. For example, Newman et al. [59] presented an indoor AR system called BatPortal and Vlahakis et al. [85] presented Archeoguide, a PDA-based AR system for outdoors. Later, Wagner and Schmalstieg [87] implemented ARToolkit [40] to an indoor AR guidance system running on a PDA

[1] https://support.apple.com/en-us/HT201471

device.

In 2004, the early smartphones were advanced to a state where they had enough processing power to run simple AR applications. Möhring et al. [53] were the first to demonstrate real-time tracking of colored 3D fiducial markers. Development of smartphones also led to the release of the first commercial HAR game, which was created by Siemens in 2004. In 2005, Henrysson et al. [32] presented the ARToolkit on an HAR system that demonstrated the complex virtual object manipulation possibilities of HAR. Later, Henrysson et al. [29] made a collaborative AR game using two smartphones. This, and many other early HAR systems utilized fiducial markers to track the environment.

The rise of smartphones was an important factor for AR because it enabled AR applications to be distributed amongst vast amount of potential users. The first wave of commercial HAR applications usually used sensor-based or marker based computer vision tracking. Several AR browsers, like Wikitude[2], Layar[3], or Junaio[4], were published in 2008 and later followed by similar applications. At that time, AR browsers were the most popular domain of consumer HAR applications and they were used for exploration, tourism, navigation and so on. AR browsers commonly used sensor based tracking and enabled the browsing of virtual information from different channels, such as blog posts or location information, overlaid to the camera image of the real environment.

The implementation of first markerless HAR systems happened in 2008 when Wagner et al. [86] developed a first AR natural feature tracking system for consumer smartphones. Later, Klein and Murray [42] implemented a real-time Parallel Tracking and Mapping (PTAM) to a consumer smartphone. First Simultaneous Localization And Mapping (SLAM) trackers for commercial handheld devices were not published until 2011. Tabletop Speed[5] and Minecraft Reality[6] were one of the first games to utilize markerless tracking. Wikitude and other companies later released their own SLAM trackers which allowed more complex AR applications to be developed.

[2]https://www.wikitude.com
[3]https://www.layar.com
[4]https://en.wikipedia.org/wiki/Junaio
[5]https://www.crunchbase.com/organization/dekko
[6]https://mojang.com/2012/11/announcing-minecraft-reality-for-ios/

More and more applications based on marker and markerless vision tracking techniques have emerged for marketing, education, and entertainment. An important aspect for the rise in popularity of HAR has been it's use in the advertisement domain for various types of advertisements. HAR provides a compelling experience that grabs user's attention in a way that makes it easy to draw attention to the marketing message[5].

The ongoing technical advancement of handheld devices has rapidly increased the amount of available HAR applications in general and the HAR development and research community is still growing strongly. The fast technical development has made more sophisticated tracking technologies possible. The development of visual tracking based HAR applications has became easier with free available tracking libraries such as Vuforia [7] and authoring toolkits like BuildAR [8].

The trend in state-of-the-art handheld systems, such as Google's Project Tango[9] or Intel's RealSense[10], has recently been to use two or more cameras for advanced tracking and visualization via better depth sensing of the environment. Furthermore, large mobile device manufactures like Apple have been rumored to include two back facing cameras for their flagship models in 2016. The spread of better depth tracking hardware on handhelds would inevitably increase the spread of markerless tracking, thus enabling more compelling HAR applications.

1.2.2 Tracking in HAR

In order to display virtual content augmented to the real world, it is necessary to know where the user is looking with the device's camera. Most commonly, HAR utilizes either sensor or computer vision based tracking (Fig. 1.7) technologies. Regardless of the used tracking technology, the purpose of tracking is to find the position and orientation of the handheld device in the real world so that the virtual content can be drawn correctly. The motions or changes made by the user need to result in the appropriate changes in the perceived virtual content

[7] http://www.vuforia.com

[8] http://www.buildar.co.nz

[9] https://www.google.com/atap/project-tango/

[10] http://www.intel.com/content/www/us/en/architecture-and-technology/realsense-overview.html

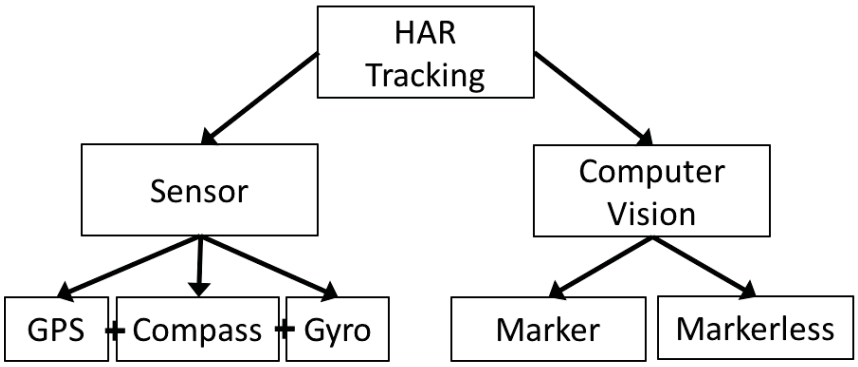

Figure 1.7. HAR tracking techniques divided into two branches: sensor and computer vision based tracking.

[1]. Figure 1.8 shows an example HAR system from three tracking categories: sensor based and two visual tracking based (markers and markerless) categories.

1.2.2.1 Sensor Based Tracking

Almost all modern handhelds contain several sensors, such as a GPS, a compass, an accelerometer, and a gyroscope. Sensor tracking uses information from these sensors to determine position (from GPS) and orientation (from the compass and by estimating gravity from the accelerometer). It does not recognize anything from the actual video image captured by the device's camera. Sensor based tracking can be considered really light weight and it is easy to implement, but it has several drawbacks. For example, GPS is useful for aligning the AR content over long-distances, but it does not work indoors and is only accurate up to an error of tens of meters, even under good conditions. Furthermore, a compass is accurate only to tens of degrees and is easily disturbed by metallic objects and electromagnetic interferences. Because accelerometers and gyros work only incrementally, they require constant recalibration from other sources [20]. These factors make sensor tracking too inaccurate for indoor use.

Using HAR indoors requires deploying tracking from the camera image. Tracking purely from sensors is not accurate enough in many situations, for example sensor based tracking might identify the street where the user is currently located, but it does not discriminate the items in a shop window, which would need be aug-

mented in case of marketing or advertising applications. Systems utilizing sensor based tracking do not fully match to the Azuma's [2] definition of AR, because the virtual content is not registered in 3D. However, from a user's point of view, the most important characteristic of AR is the combination of real and virtual. Users cannot necessarily distinguish how this combination is achieved technically. Thus, even if the virtual content is not registered in 3D, sensor tracking based systems can be considered as AR.

1.2.2.2 Computer Vision Tracking

Visual, or computer vision, tracking detects the real world based on the video image captured by the device's camera. Visual tracking can be either marker based or markerless. Marker based tracking requires physical fiducial markers, patterns, or figures placed into known positions in the real environment. The shape, material and texture of fiducial markers can vary considerably depending on the scenario and the system in question. The disadvantage of fiducial markers is that they require the modification of the real environment, which may not be desirable or even possible in some scenarios[5].

Here, with markerless tracking we refer to tracking of natural features from the real world. Natural feature tracking detects and tracks natural features such as corners, edges, or planar surfaces from the environment. The pose of the device's camera is determined relative to a known natural features in the environment in real-time using detected natural features. A markerless tracking system extracts feature points from the device's camera image and compares them to a database that stores known features together with their position in the environment. Given enough successful matches of surface textures or known shapes, the device can determine the camera's pose relative to the detected features [20].

Markerless systems can also dynamically reconstruct the physical environment and expand the natural feature map on-the-fly from the camera stream. These SLAM systems can comprehend changes in the real world, which enables them to be used in dynamic environments making SLAM suitable for tracking of unknown environments. The original motivation behind SLAM was for robot navigation in unknown environments but the technique was later adapted for AR [12]. Further optimizations of led to PTAM [41], where the tracking of the camera and mapping

Figure 1.8. Examples of HAR applications that utilize different types of tracking techniques: sensor (a), marker based (b), and markerless tracking (c).

of the environment components were separated. PTAM was specifically designed for AR, and improved both the accuracy of tracking as well as overall performance.

Even though SLAM tracking continuously scans and learns about the environment it requires an initialization from a certain viewpoint in order to start the SLAM tracking. Markerless tracking can be used without the need for adding fiducial markers or physical objects to the environment, but it commonly requires high processing power. Given the high computational demands, markerless tracking has been a large technical challenge for mobile devices for a long time and only the recent handheld devices, tablets especially, are capable for SLAM tracking.

Some technical difficulties have been overcome by combining visual and sensor tracking. A simple approach uses GPS information as a filter to narrow the search area for initialization of visual tracking [69]. Furthermore, most visual tracking techniques can be confused by fast rotational movements, which let observed feature points suddenly disappear from the image. This situation can be stabilized by fusing the information from sensors. E.g., using gyro informs the tracker about the expected rate of rotational movement so that it does not search for features that can no longer be observed.

1.3. HAR in Task Support

The use of HAR in various scenarios have been examined before and HAR has been found to provide highly positive hedonic experiences [63]. However, the

adoption of HAR for pragmatic goal-oriented tasks has not been as fast as expected. Despite the fact that AR on HDM system has proven to improve efficiency in goal-oriented tasks [27, 28], HAR has not managed to gain similar supporting research.

The purpose of task support is to aid users in goal-oriented tasks that are conducted in order to fulfill externally given or internally generated goals. Hassenzahl [26] defines systems to have either hedonic or pragmatic attributes. In order to be considered useful, a task support system need to have pragmatic attributes in a form of relevant functionality (e.g., utility) and easy ways to access this functionality (e.g., usability). Pragmatic attributes are the opposite to hedonic attributes found from games and other entertainment related systems. Usability and utility are the core attributes in task support systems, but many of the existing HAR system are not considered practical due to insufficient functionality and they do not fully answer to the needs of the users [64, 22].

The benefit of HAR in task support is that it allows users to easily input and share information among other local users. Furthermore, handheld devices are very practical and affordable. Non-AR handheld systems are already used in goal-oriented tasks. Possibly the most substantial drawback of handhelds is that they do not allow hands-free observation of the augmented real world. This makes two handed physical manipulation of the environment very difficult. Even one-handed manipulation can be difficult because device's camera has to be pointed to a certain direction in order to observe the augmented content. However, this drawback is only in tasks that require physical manipulation of the environment. Unlike maintenance or assembly tasks, inspection focuses on observation without the necessity for physical manipulation.

1.4. Inspection Tasks

Inspection refers to tasks where the condition or status of various targets in a workpiece are inspected by a user (Fig. 1.9 [11][12]). Here, a workpiece refers to the

[11] https://paysafeescrow.com/guide/buyer/inspection

[12] https://www.calvin.edu/admin/physicalplant/departments/ehs/policies/safety-inspections/safety-inspecs.html

Figure 1.9. Two examples of an inspection task: a worker is inspecting a car engine (a) and a worker is inspecting a factory machine (b). Both workers use a paper manual as a checklist.

apparatus to be inspected, such as a factory machinery, a motor vehicle, hospital equipment, or home appliances. For example, various engine parts (targets) inside a car engine (workpiece) require regular inspection. A checklist is commonly used as a guidance tool in unknown inspection environments. A checklist shows information of targets that are to be inspected. Some task support prototypes for inspection have been developed in the past [62, 76] but AR in inspection remains unstudied.

The precise task flow in inspection can vary depending on the specific inspection scenario in question. For example, inspection of a car engine can require different procedures compared to inspection of home appliances. In this thesis, we define inspection based on our collaboration with industrial workers and professionals. We have conducted user observations and interview with industry professionals to gain a better understanding of inspection tasks.

Figure 1.10 illustrates our definition of an inspection task. The left side illustrates the preparation for inspection which means the creation of a checklist by adding the appropriate amount of annotations to it. The right side of Figure 1.10 refers to conducting the inspection where user observes the workpiece based on the information on a checklist (Fig. 1.10, blue square on the right). In some situations, additional annotations might need to be added to the environment

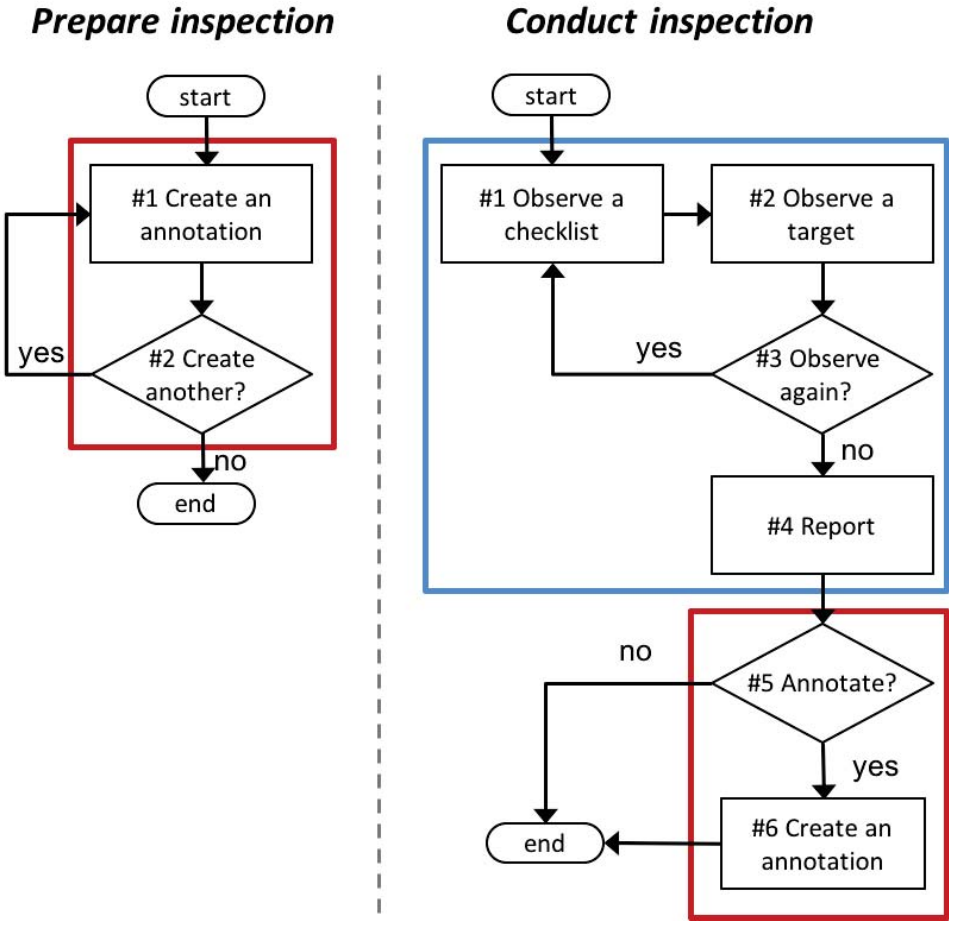

Figure 1.10. The inspection task flow. Left side represents the flow of preparing for inspection and right side visualizes the actual inspection task. The left side can be further divided into two parts: visual observation and report (the blue square) and 2) adding an annotation to a checklist (the red square). The colors highlight the two areas of AR in this thesis.

while user is conducting the inspection (Fig. 1.10, red square on the left). This might happen if a user notices a target not belonging to a checklist that should be inspected later. Reporting refer to confirmation of annotations already in the checklist, adding annotations means creating new annotations to the checklist.

The main and most often performed subtask in inspection is the visual observation & reporting (Fig. 1.10, blue square on the right). Preparing the inspection

by creating a checklist (Fig. 1.10, red square on the left) can be considered equally important because it enables users to perform the inspection in the first place. Furthermore, creating annotations are sometimes needed while conducting the inspection also (Fig. 1.10, red square on the right). Because of these reasons we consider creating annotations and observing them equally important. The procedures and decisions in Figure 1.10 are as follows.

- **Prepare for inspection**
    1. **Create an annotation:** A user creates a new annotation to a checklist.
    2. **Create another?:** A user adds more annotations if necessary.
- **Conduct inspection**
    1. **Observe a checklist:** A user observes a checklist in order to gain information of the target to be inspected.
    2. **Observe a workpiece:** A user observes and checks the status of a target in a workpiece based on the information in a checklist.
    3. **Observe again?:** If a user was unable to check the target, he needs to observe the checklist again.
    4. **Report:** A user reports target's status to a checklist based on the information gained during the mapping of the target. Reporting does not require users to create new annotations.
    5. **Annotate?:** If user notices something strange or abnormal, a new checklist annotation has to be added to a checklist.
    6. **Add an annotation:** A user adds a new annotation to a checklist.

In our definition, inspection focuses mainly on visual observation of a workpiece and physical manipulation of the environment is needed rarely or not at all. In some special cases, manipulation might be necessary, for example if the information required for the inspection is not directly visible [58]. AR research often focuses on various maintenance tasks. We separate these from inspection by defining maintenance as tasks that focus on physical manipulation. Sometimes

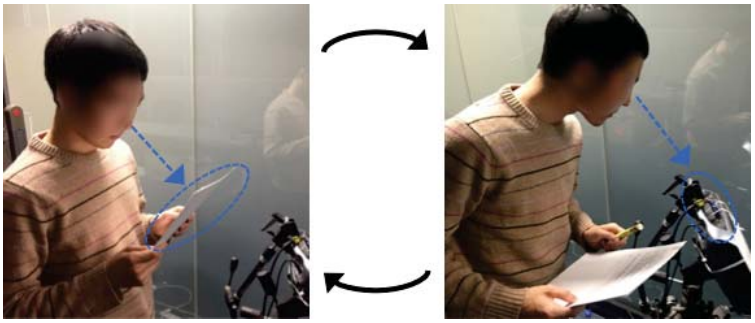

Figure 1.11. A user is inspecting a video projector. He has to repeatedly shift his gaze between the informal activities (left, observing the checklist) and workpiece related activities (right, observing targets in a workpiece).

inspection and maintenance tasks can overlap - inspection can included physical manipulation and maintenance can focus only on visual observation. Furthermore, our definition of inspection assumes that users are not familiar how to inspect the workpiece, making it mandatory for them to observe the checklist.

The largest issue in conducting inspection tasks is the added workload that comes from the divided attention between a workpiece related and informal activities [58, 65]. During the inspection, users have to repeatedly shift their gaze from a checklist to a workpiece because information in both needs to be observed. Figure 1.11 illustrates the problem caused by divided attention. Workpiece related activities refer to the observation and inspection of targets from a workpiece and informal activities refer to understanding, reading, or comprehending inspection checklists or other guidance mediums.

1.5. Goal and Approach

In this thesis, we investigate the usefulness of HAR in inspection tasks. In order to be considered useful, HAR needs to provide the relevant functionality (utility) and offer benefits compared to conventional non-AR inspection interfaces (usability). Based on these two requirements, we have formulated two research questions for this thesis:

1. **Add annotations.** Can we accurately place virtual annotations to targets

in a workpiece using HAR (Fig. 1.10, the red boxes)?

2. **Observe and report.** Does the use of HAR offer any benefits in visual observation tasks? (Fig. 1.10, the blue box)?

The use of AR for in goal-oriented tasks has been widely studied, and it has already been proven to have benefits compared to non-AR interfaces [27, 28, 78]. However the use of HAR in similar tasks has not been seen to provide similar benefits [48, 15]. The overall usefulness of HAR depends strongly on the task in question. HAR can be viable in inspection tasks because the information from a checklist can be directly overlaid to the workpiece. Furthermore, inspection focuses on visual observation so there is no requirement to physically manipulate the environment.

1.5.1 User Studies

In order to answer to the set research questions, we have conducted a total of four user studies. The connections between these studies and the three areas (Fig. 1.12, middle square) of inspection tasks are illustrated in Figure 1.12. First two studies (Fig. 1.12, the top most square) focused on HAR 3D positioning that is related to creating a checklist and adding annotations to a workpiece. Even though creating a checklist (preparing for inspection) is separate from conducting an inspection, it is essential. Thus, 3D positioning in HAR is related to not only conducting the inspection but also to preparing for inspection. The low level 3D positioning task is the same in both cases.

The later two user studies (Fig. 1.12, the lowest square) are related to using HAR in visual observation tasks. These studies are connected to the observe & report in conducting an inspection. Even though conducting inspection could include adding annotations, our studies focus solely on the observation because it is the most important part of an inspection task. The short summary of all four studies is as follows:

- **3D positioning in HAR**

    1. A qualitative evaluation of a HAR system in a checklist creation scenario for Audio/Video (A/V) equipment inspection. The purpose of

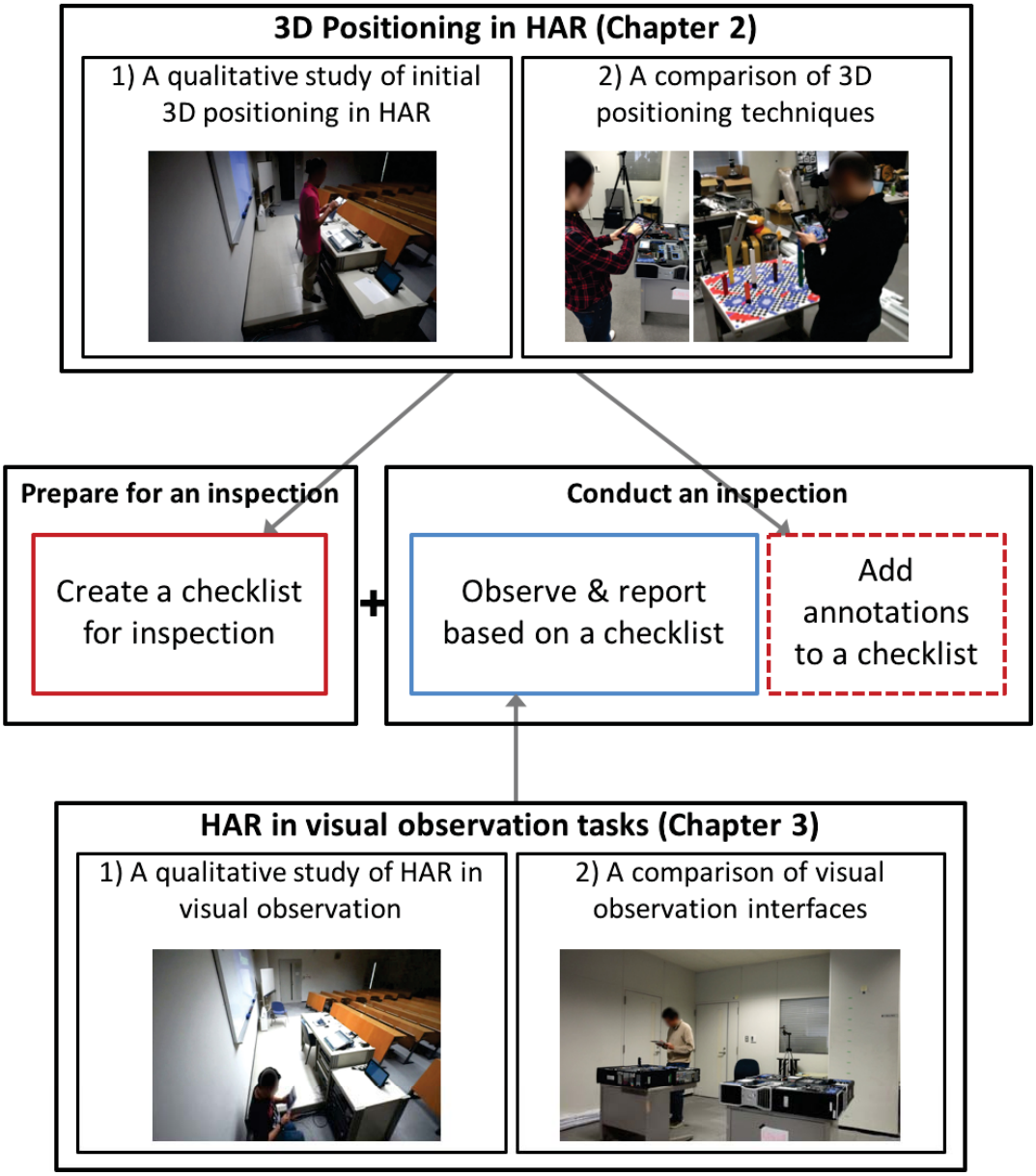

Figure 1.12. The conducted four user studies in two areas of HAR, 3D positioning (the upper square) and visual observation tasks (the lower square). Their connections to the three parts of an inspection task (the middle squares) are illustrated as grey arrows. The colors present the colors shown in Figure 1.10.

this study was to find issues from 3D positioning of annotations the environment while using HAR in a real world scenario.

2. A comparison between a ray-casting based 3D positioning method and a conventional device-centric method. This study focused on a specific 3D positioning problem found in the first positioning study.

- **HAR in visual observation tasks**

1. A qualitative evaluation of a HAR system in a real world A/V equipment inspection scenario. The purpose of this study was to find issues from HAR in a real world scenario that focused on visual observation.
2. A comparison between AR interface and a picture interface in a machine inspection. This study focused on comparing the benefits of AR against a conventional interface in a generic visual observation scenario.

1.5.2 Prototype Systems

In every study we used different versions of a SLAM-based HAR tablet prototype system 1.13. Our SLAM system consist of four main phases: 1) SLAM map is created with a side-to-side motion (Fig. 1.13a). 2) Initial positioning of an annotation is conducted by tapping to the screen and writing the annotations text (Fig. 1.13b). 3) If the tracking is lost, the system instructs a user to align the viewpoint according to the initial viewpoint (Fig. 1.13c). 4) Inspection is conducted by observing the target and answering either 'YES' or 'NO' (Fig. 1.13d). The prototypes were designed for small indoor workspaces where the target objects are the near-field distance [49] from the user. We did not design prototypes for any specific inspection scenario, but they are applicable to different scenarios. The interaction technique of our prototypes can be described as embodied [20], which means that the user focuses only on the device movements and its touchscreen display without manipulating the virtual annotations directly by hand.

The main reason for choosing handheld devices over HMDs was the easier information input in handhelds discussed already earlier in this Chapter. The reporting and adding annotations in inspection tasks requires users to input in-

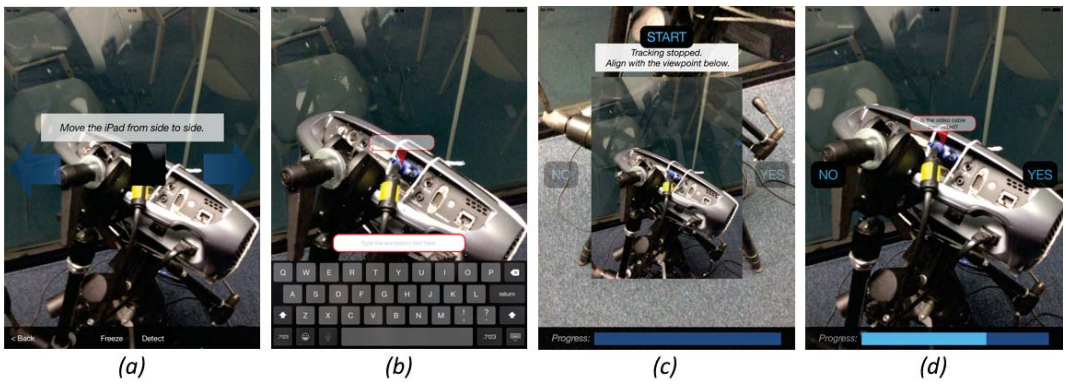

Figure 1.13. The four phases of our SLAM prototype: map creation (a), initial positioning (b), tracking initialization (c), and target observation (d).

formation. The complexity of information that needs to be inputted is scenario specific and can vary from simple confirmations to long textual inputs. Simple confirmations could be possible also with an HMD, but long textual inputs would be very difficult. Furthermore, handhelds enable easier information sharing and collaboration among local users. It might be necessary that two workers can observe the same handheld device simultaneously and co-operate during the inspection. This type of information sharing would not be possible with HMDs.

In inspection, a two-handed physical manipulation of the environment is not necessary and one-handed can be considered needed only in special situations. This allows users to comfordably manipulate the handheld devices with two hands instead of relying on one-handed manipulation of the device. Even though not essential from the research point of view, we use handhelds because we want to demonstrate that the use of AR does not depend on benefits of more immersive hardware technologies and is suitable for wide-spread adoption.

Hypothetically, we could have used smartphones instead of larger tablet devices. This would have enabled less demanding physical manipulability of the handheld device while conducting the tasks. However, the larger displays of tablet devices allow easier readability of static information, such as text, graphs, charts or images. Also, interaction and information input of AR and statistic information becomes less troublesome when using tablets with larger displays [74].

Even though our prototypes did not visualize static information, it is often necessary in real world inspection scenarios. In case of AR, handheld devices have also been shown to provide easier depth perception in some situations [13].

We used SLAM tracking because HAR inspection systems must be usable in any kind of indoor environment. From the tracking point of view, this means that only markerless tracking is feasible. There are often limitations where AR markers can be placed and all the virtual annotations would have to be in the vicinity of these markers. In theory, it could be possible to use pre-created 3D models of the environment. Nonetheless, these models would not react to changes made in the real environment, where as SLAM technology allows the real-time mapping and updating of the environment. Furthermore, creating predetermined 3D models is can be too time consuming and unpractical. This is important especially if annotations need to be created outside the tracked or modeled area during the conducting of an inspection.

1.6. Thesis Contribution

The contributions of this thesis are formalized from the research questions presented in previous Section. This thesis focuses on SLAM-based HAR in inspection tasks, but the contributions are applicable to other types of scenarios using different AR display techniques and tracking technologies. The contributions are related to the two areas of AR shown in Figure 1.12 (the upper and lower squares). The two main contributions of this thesis are:

1. **3D positioning in AR:** We are the first to evaluate 3D ray casting positioning method on AR and prove it's efficiency over a conventional device-centric AR 3D positioning technique. This type of 3D positioning based on ray-casting can be done with HMD-AR system as well and it does not require a handheld device. Also, the it can be done with other tracking techniques instead of SLAM. Thus, the findings from our user studies can be applied to design and development of HMD-AR based positioning or HAR positioning with other tracking tehniques.

2. **AR in visual observation tasks:** We are the first to use AR in visual

observation tasks and prove its higher usability over a conventional non-AR picture interface in tasks that have high information complexity and require viewpoint alignment. The results can be applied to HMD-AR to some extend, because we did not require complex information input. Furthermore, these results are useful in designing any type of HAR system that requires visual observation.

CHAPTER 2

3D positioning in HAR

In inspection tasks, adding annotations is fundamental in creating of a checklist. Furthermore, annotations sometimes need to be added during the conduction of inspection as well. Thus, it is important that we can accurately position the annotations to targets in a workpiece. 3D positioning is needed not only in inspection, but it is important for the wider acceptance of HAR in general: users must be able to create AR content by positioning virtual objects to the real environment in order for HAR to became widely popular [44, 45]. Furthermore, potential HAR users want to create AR contents in various indoor and outdoor environments [82]. The 3D positioning is part of the basic 3D virtual object manipulation tasks [7] that are fundamental in AR content creation.

We divide a HAR 3D positioning task into two phases: 1) initial positioning and 2) position adjustment (Fig. 2.1). In the first phase, the virtual annotation is created and it's initial position is decided. In the second phase, the 3D position is adjusted by translating the annotation from the initial position to the final desired position. In some cases, only initial position is required. However, sometimes it might be necessary to adjust the position after that.

In this chapter, we present two user studies: The first one focused on the usability of SLAM-based HAR 3D positioning using only a tap gesture for initial

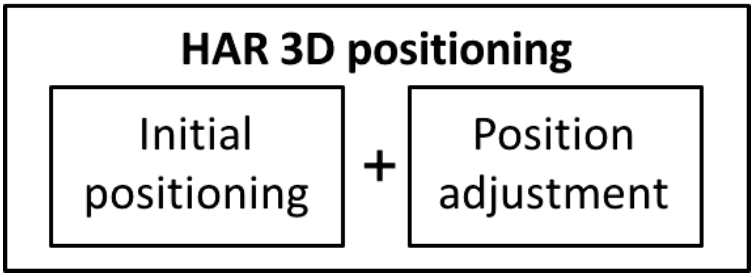

Figure 2.1. The two phases of 3D positioning in HAR: the initial positioning and the position adjustment. The first position of an object is determined in the initial positioning. If necessary, the position is adjusted after that.

positioning. In the second study, we evaluated a SLAM-based HAR 3D positioning method called SlidAR (Fig. 2.2) that uses 3D ray casting and epipolar geometry. This technique uses a tap gesture for initial positioning and a slide gesture for positioning adjustment. We compared this method against a conventional device-centric method first introduced by Henrysson et al. [30].

2.1. Related Work

Bowman et al. [7] have designated three basic virtual object manipulation tasks for VR and AR: selection, positioning and rotation. Authors define positioning as a task of changing the 3D position of a virtual object. In this section, we introduce HAR authoring tools, AR positioning methods specific for handheld devices, and methods that utilize ray-casting applied in hardware other than handheld devices.

2.1.1 HAR Authoring Tools

Before focusing on 3D positioning of virtual objects, we briefly discuss about systems in the related work that are designed for virtual content creation as a whole. We call these systems authoring tools. The AR content for HAR systems is usually authored with conventional desktop systems. However, some systems exist where the authoring can be done using a handheld device. Castle et al. [10]

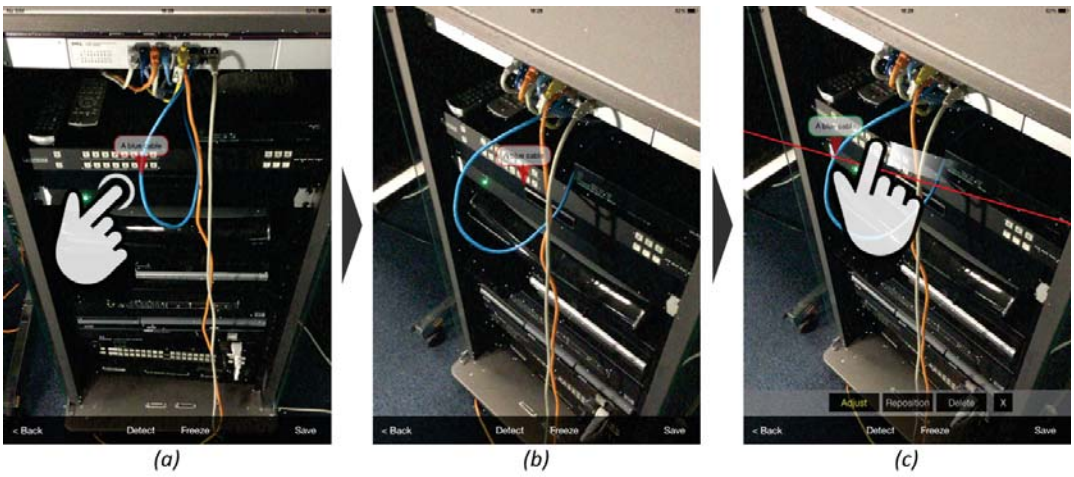

Figure 2.2. SlidAR: a virtual annotation is being positioned to a blue cable. The annotation is created by tapping the screen of the handheld device (a). The viewpoint is changed and the new viewpoint reveals that the annotation?s position is incorrect (b). From the new viewpoint, the position can be adjusted along the red epipolar line using a slide gesture (c).

have developed PTAMM, a SLAM-based system, which allows manipulation of AR content. The system uses a movable camera, a touchscreen display, and a laptop PC. With PTAMM, various predetermined 3D objects can be positioned and modified using the device's touchscreen. Langlotz et al. [46] have developed a HAR system for a smartphone, which enables simple content creating and sharing to other users. Their system enables the initial positioning of virtual object to the environment using vision-based tracking in indoors and sensor tracking in outdoors.

2.1.2 HAR 3D Positioning

In many systems, the AR content is placed relative to the tracked surface or a marker [47, 46]. No 3D positioning methods are used for position adjustment. In many scenarios, including inspection, simple initial positioning is not enough but more advanced techniques are required. Here, we focus on 3D positioning techniques that enable a user to adjust the 3D position.

Different 3D manipulation methods for HAR have been widely studied. In related work, a single method is commonly implemented and evaluated for more than one manipulation task. For example, a method that combines 3D positioning and rotation is often proposed. We present methods that have been designed solely for positioning or for more than one manipulation task including positioning. Here, the previous methods have been roughly divided into three groups: 1) buttons and touch-screen gestures, 2) mid-air gestures, and 3) device-centric movement.

2.1.2.1 Buttons and Touchscreen Gestures

Button-based positioning uses either the physical or the touchscreen buttons of a handheld device to position virtual objects. Henrysson et al. [30] have utilized smartphone's physical buttons for positioning where different buttons are mapped for different Degree Of Freedom (DOF). Castle et al. [10] have applied touchscreen buttons in tablet computer HAR system to position objects in three DOF. In the work of Bai et al. [15], the positioning in two DOF is conducted in a freezed AR view using a combination of buttons and gestures.

Touch gestures have become a standard for 2D manipulation on touchscreen handheld devices [25] and they have been used extensively in HAR 3D manipulation as well. Jung et al. [37] have developed a system where virtual objects can be positioned in 3D by controlling one DOF at a time with a single or multitouch drag gestures. The controlled DOF is based on the pose of the device relative to a ground plane. Marzo et al. [51] have used the DS3 technique [50] for 3D multitouch gesture positioning on a smartphone. Their method displayed a shadow on the ground plane below the virtual object as a depth cue.

Mossel et al. [54] have developed a method where the positioning is done with a slide gesture. The controlled DOF is based on the pose of the device relative to a ground plane. Kasahara et al. [39] have developed a tablet system where positioning is done only by tapping to the desired location on the device's display. The position of a virtual object is determined by the feature points detected from the live AR view, which is then compared to an image database. Touchscreen gestures have also been utilized in commercial HAR applications like Minecraft

Reality[1], Junaio[2] and the Ikea Catalog[3].

3D manipulation in VR has been widely studied and gesture-based positioning methods have also been applied for handheld VR systems. For example, Telkenaroglu et al. [79] and Tiefenbacher et al. [80] have experimented on 3D positioning in VR using touchsreen gestures. Interaction in handheld VR positioning shares similarities with HAR positioning, but there are also great differences related to scene navigation, etc. Thus, we will not discuss about handheld VR positioning methods more thoroughly.

2.1.2.2 Mid-Air Gestures

Mid-air gesture HAR positioning methods utilize the user's mid-air finger movement in front of the device's camera. Virtual objects can be positioned by moving the finger while the system tracks the finger movement. Henrysson et al. [31] have developed a 2D and a 3D mid-air gesture positioning methods using the front-facing camera of a smartphone. In the 2D method, the positioning of a virtual object is done in a freezed AR view. After freezing the AR view, the object's position is translated in two DOF by moving the finger in front of the camera. A small colored dot on the user's finger is tracked. In the 3D method, an AR marker is attached to the user's finger allowing a three DOF positioning in a live AR view. In the method presented by Hürst et al. [34], positioning is done with different finger gestures in front of the back-facing camera of a smartphone by tracking colored dots on user's fingers. Objects can be pushed with one finger or grabbed and moved with two fingers. Bai et al. [3] have also developed a finger gesture method where different axes of the objects position can be controlled by moving the finger in front of the back-facing camera.

2.1.2.3 Device-Centric Movement

Device-centric methods utilize the movability and small form-factor of a handheld device. Virtual objects are positioned by moving the device while the object's position is fixed relative to the device. Henrysson et al. [30] have developed a

[1]http://minecraftreality.com
[2]http://www.junaio.com
[3]http://www.ikea.com/gb/en/catalogue-2015/index.html

one-handed and bimanual device-centric methods to a smartphone system using AR markers. The object's position can be controlled by pressing a physical button from phone's keypad and moving the device. Mossel et al. [54] have implemented the same method for a modern touchscreen smartphone. In their method, virtual lines based on axes are used as depth cues. Marzo et al. [51] have also implemented a similar method for a touchscreen smartphone and they use a virtual shadow below the object as a depth cue. Hürst et al. [34] have implemented the device-centric method for a smartphone system that uses only sensor (a gyroscope, an accelerometer, and a compass) tracking. Güven et al. [23] have described three techniques for creating and moving AR annotations using a PDA and an external camera: The first technique enables freezing of the AR view for easier annotation creation. The second one pauses an AR object to the screen and allows it to be moved to different location by egocentric navigation. The third technique is used to link associated annotations from different paused frames together.

2.1.3 3D Ray-Casting

A 3D ray-casting for positioning is utilized widely in Head-Mounted Display (HMD) AR systems. Reitmayr and Schmalstieg [71] have presented a mobile AR system for outdoors that utilizes HMD and a handheld device. Positioning is done by using the handheld device for casting a 3D ray through a crosshair in the HMD. The system uses a predetermined 3D model of the buildings in the environment and the ray is intersected with the geometry of the buildings. Bunnus et al. [8] have developed an AR 3D modeling tool that uses 3D ray-casting and epipolar geometry to define the vertices of a plane. The system uses a handheld interface similar to a computer mouse with track wheels and buttons. A small camera is attached to this mouse-like interface, and the image is sent to a separate display. Wither et al. [90] have developed a mobile AR system with a mouse input interface. The ray is cast using the first person view of the HMD and the mouse interface. The target position is determined based on the intersection point of the ray and geometry of the buildings recognized from the aerial images. Later, Wither et al. [89] developed another positioning method using similar kind of hardware, but instead of aerial images, their method used a

single-point laser attached to an HMD. Lastly, Reitmayr et al. [70] have developed a SLAM-based method that allows pointing without pre-knowledge of the environment by detecting planar surfaces from the camera image.

2.1.4 Summary

The table 2.1 shows a summary of existing HAR and AR ray-casting based positioning methods. We did not include methods that have only initial positioning because it does not require any specific techniques. The main difference between SlidAR and other HAR positioning methods is that, unlike previous methods, it utilizes ray-casting and does not require virtual depth cues nor AR markers. Using markers is not always possible and there can be limitations depending on the kind of 3D structures and surfaces from the environment can be tracked. Previous 3D ray-casting based methods use HMD or other types of hardware and have not been used only on a handheld device. Some existing positioning methods do not use depth cues either, but their efficiency has not been confirmed in user studies. Few methods utilize slide gestures, but those are not based on epipolar geometry.

As we can see from the previous user studies and as stated by Bowman et al. [7], one manipulation method is not necessarily suitable for all basic manipulation tasks. On the other hand, the combination of different methods for two or more manipulation tasks can be beneficial [54]. Buttons and touch gestures methods have been proven to be very efficient for rotation and scaling tasks, they have difficulties in positioning tasks [30, 54, 51]. Mid-air finger gestures have been evaluated to be more suitable for entertainment purposes rather than practical use [31, 34]. The device-centric methods have been the most efficient for HAR 3D positioning [30, 31, 54, 51]. We chose to compare SlidAR method against the conventional device-centric method, because it has been the most efficient for 3D positioning in previous studies.

2.2. First Positioning Study

The purpose of the first study was to investigate the usability of SLAM-based HAR 3D positioning in a scenario where users create a checklist for Audio/Video equipment inspection. In this scenario, the virtual objects was created only by

Table 2.1. The summary of the HAR and ray-casting based positioning methods and their attributes from the related work. Several related publications present more than one method and those are marked with abbreviations: D = a device-centric method, B = a button or gesture based method, and M = a mid-air gesture based method.

	Utilizes ray-casting	Usable on a handheld device	Does **NOT** require AR markers	Does **NOT** require preknowledge	Does **NOT** require virtual depth cues	Evaluated
Henrysson et al. [30]: B & D		X			X	X
Reitmayr et al. [71]	X		X		X	
Wither et al. [90]	X		X			X
Henrysson et al. [31]		X				X
Reitmayr et al. [70]			X	X	X	X
Bunnus et al. [8]	X		X	X	X	
Castle et al. [10]		X	X	X	X	
Wither et al. [89]	X		X	X	X	
Hürst et al. [34]: M		X	X	X		X
Hürst et al. [34]: B		X	X		X	X
Bai et al. [3]: B		X		X		X
Bai et al. [3]: M		X		X		X
Jung et al. [37]		X		X		
Kasahara et al. [39]		X	X		X	
Mossel et al. [54]: B & D		X		X		X
Marzo et al. [51]: B & D		X		X		X
SlidAR	**X**	**X**	**X**	**X**	**X**	**X**

using initial positioning without any positioning adjustment methods. It is important to gather findings from the initial positioning in a real world scenario because it will teach us about the actual problems of 3D positioning task and what type of adjustment is needed.

2.2.1 Prototype System

The SLAM-based HAR iPad prototype used in this study utilizes PointCloud SDK[4] for feature point detection and tracking of the environment. The PointCloud SDK uses images and internal sensor information of the handheld device. A SLAM map can then be created for each of these areas with an initialization motion (moving the device from side to side) (Fig. 3.1a). Annotations can be created according to tracked feature points (white dots in Fig. 3.1b) by tapping the desired location.

[4]http://developer.pointcloud.io

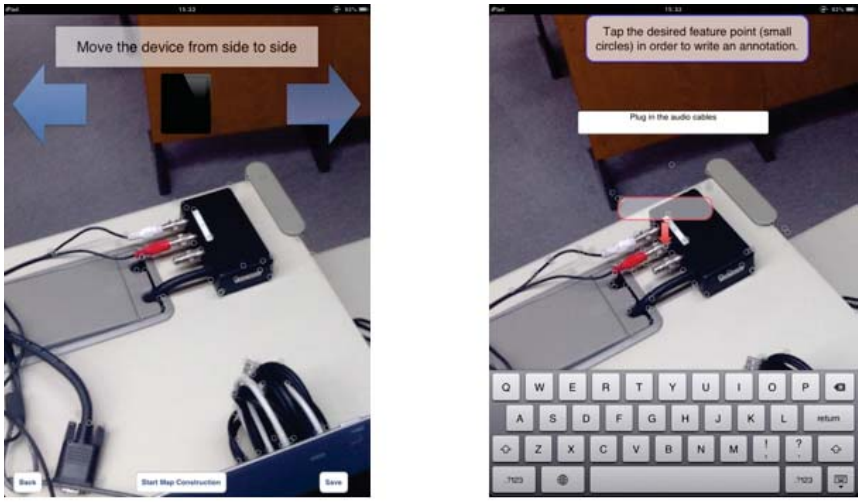

Figure 2.3. The interface of the prototype system used in the first study. SLAM tracking needs to be initialized with a side-to-side movement (a). After that, textual annotations can be created by tapping to a desired position on the handheld devices display (b).

2.2.2 Study Design

The first 3D positioning study had 10 participants (22-28 years old) who were asked to create a checklist for the inspection of A/V equipment in a large lecture room scenario (Fig. 3.2). The scenario represented real world scenario and we chose it because setting up these A/V devices can be difficult for those who do not use A/V equipment often or use them for the first time. A short introduction was given before the actual test and users were given feedback on how to use the system correctly during a tutorial task. The test included total of three tasks that required the creation of three annotation sets for a checklist:

- Create instructions on how to check the video and audio cables.
- Create instructions on how to check the projector and choose the correct video input.

Figure 2.4. A test participant conducting the first 3D positioning study.

- Create instructions on how to check that the main power is on and that the audio mixer is working correctly.

2.2.3 Results

We collected only qualitative results via video observation and subjective feedback. Video observation was based on videos captures from the environment and the device's display. Test participants were also encouraged to think out loud.

The SLAM map creation caused difficulties for some participants, especially in the second task where annotations had to be created on a reflective touchscreen display. Even if participants where able to create a SLAM map, the poor quality of the map forced them to add annotations to undesirable locations. Additionally, if the created SLAM map was not dense enough, the depth of the initial position of the annotations was often incorrect. It is important that the annotations can be placed to the exact desired location. In inspection and maintenance scenarios, the targets in the environment can be very small, such as cables. Thus, the 3D position needs to be exact. The chosen location of the annotation should not be

dependent on the quality of the SLAM map. In case of a poor SLAM map, users must be able to adjust the 3D position of annotations.

If the environment is unfavorable for a SLAM map creation, it is important that the system gives feedback on how users should proceed in order to be able to successfully create the SLAM map. We found out that participants tried to build the SLAM map from the same viewpoint several times without any improvements in the quality of the map. This was because there was no proper instructions or feedback how to improve the SLAM mapping.

We noticed that presenting 2D information as 2D augmentation can occasionally cause minor issues related to perceiving the direct real-virtual relationship when the viewpoint to the real object is changed. This problem is related to perceptual issues of AR [43] and the density of annotations in the augmented view. Our study confirms the importance of understanding the direct real-virtual relationship between an annotation and a real target object. This is especially important if several annotations are displayed in a small area. Also, 2D (e.g. textual) information should be presented in 2D also in the AR environment [17].

Several small issues where found related to layout of the GUI elements. With GUI elements we refer to static 2D elements illustrated in the taxonomy of HAR [21]. The placement of UI elements also affected the SLAM map creation, because participants did not pay enough attention to the initialization motion animation. Based on our studies we assume that a static instruction animation is enough, if it is placed correctly in the UI.

2.2.4 Summary

Many of the problem areas that we encountered have been studied and already solved in the related work. E.g., how to correctly instruct SLAM map initialization [56] or the connection between real and virtual objects [4]. However, the accurate 3D positioning while using HAR have not been comprehensively studied. Our study showed that in SLAM-based HAR, it is necessary to have a method to adjust the 3D position of annotations. This is because we cannot have a 100% accurate 3D reconstruction of environment if we use SLAM. It is fundamental for the efficiency in inspection tasks that the annotations can be placed accurately.

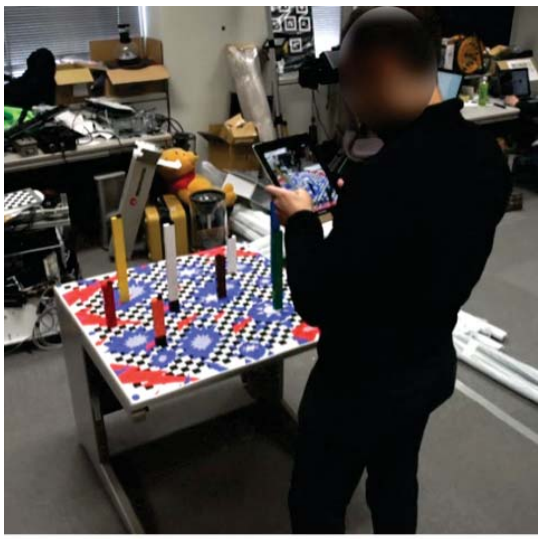

Figure 2.5. A user conducting the second 3D positioning study.

2.3. Second Positioning Study

In the second study we focused on the specific problem in a SLAM-based 3D positioning found in the first positioning study. We compared our implementations of a ray-casting based (SlidAR) and a device-centric based (HoldAR) HAR 3D positioning methods (Fig. 2.5). We evaluated these 3D positioning methods in SLAM-based HAR in order to learn more about the efficiency of each method. This study has been published in an academic journal [67].

2.3.1 Positioning Methods

We implemented two positioning methods, SlidAR and HoldAR [30], for a markerless SLAM-based HAR iPad system. The main difference between the two methods is that SlidAR relies on ray-casting and touchscreen gestures where HoldAR utilizes physical movement of the device. Also in this study, the SLAM system utilizes PointCloud SDK[5] for markerless feature point detection and tracking of the environment. The initial positioning in both SlidAR and HoldAR is

[5]http://developer.pointcloud.io

determined by tapping to the desired location on the representation of the real environment on the handheld device's display. The required level of accuracy in the initial positioning depends on the method used. The depth of the initial position is determined by the average depth $d_{average}$ of the surrounding feature points:

$$d_{average} = \frac{1}{|\boldsymbol{W}|} \sum_{i \in \boldsymbol{W}} d_i. \tag{2.1}$$

where $\boldsymbol{W}$ represents a set of natural feature points around the tapped area and d_i represents a depth value for each feature point. The position adjustment in both methods is explained separately in next Sections.

2.3.1.1 SlidAR

SlidAR utilizes 3D ray-casting and epipolar geometry for virtual object positioning (Fig. 2.6). After the initial positioning is conducted, a ray is cast from the handheld device's camera to the object's initial position. A ray can only be cast after initial positioning is done, because it requires camera pose and the ray direction information. This geometrical relationship between camera pose and a 3D point (in our case, the initial position) is known as epipolar geometry.

If the ray between the camera and the object's initial position intersects the target position (Fig. 2.6a), the object can be adjusted to the target position along the epipolar line using a slide gesture (Fig. 2.6b,c). The epipolar line is visualized as a red 2D line. We chose to use a slide gesture, because it gives smooth and precise control over the position of the virtual object. Furthermore, while conducting the slide gesture user's finger does not have to be directly on top of the epipolar line. This allows the adjustment to be done precisely without occlusion caused by the finger.

If the initial positioning was done incorrectly and the ray does not intersect with the target position, the ray must be recast by conducting the initial positioning again with a cut & paste function. The 3D position of the virtual object $\boldsymbol{p}_j$ is represented by the camera position $\boldsymbol{c}_i$, 3D ray direction $\boldsymbol{r}_j(|\boldsymbol{r}_j| = 1)$, and the distance from the camera position to the object's position l_j. The relationship between these parameters is as follows:

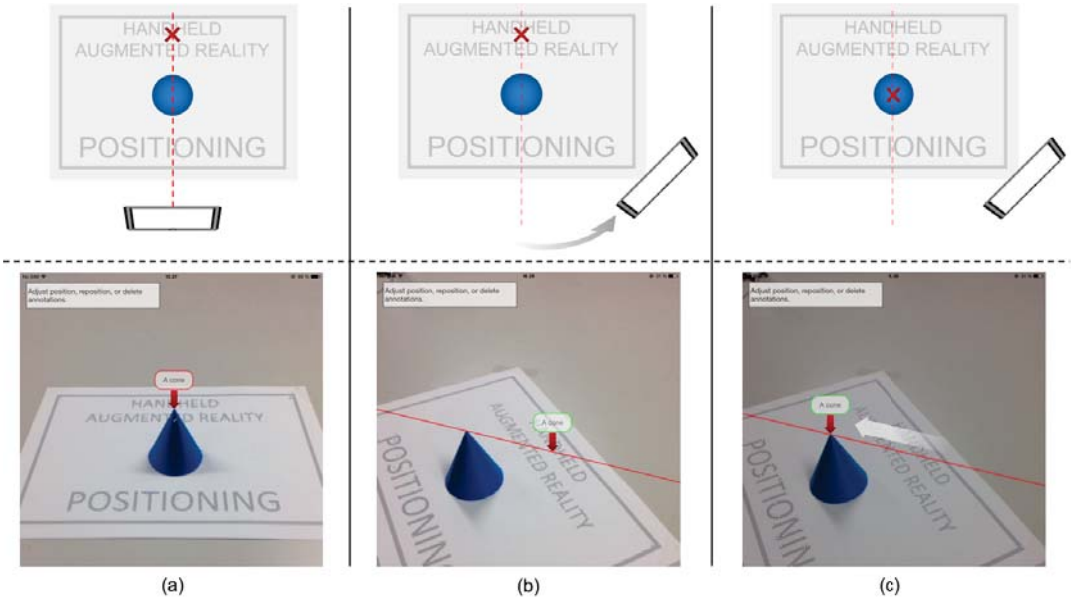

Figure 2.6. SlidAR: top-down (above the dotted line) and display (bellow the dotted line) views. A virtual object (a red bubble and an arrow) is being positioned to the tip of a blue cone (the target position). The object's position is perceived incorrectly from the first viewpoint (a). A new viewpoint exposes the correct position of the object (b). A ray from the device to the initial position intersects the target position and adjustment along the red epipolar line can be conducted with a slide gesture (c) (shown as a white arrow).

$$\boldsymbol{p}_j = l_j \boldsymbol{r}_j + \boldsymbol{c}_i \tag{2.2}$$

The l_j is changed during the adjustment phase where the epipolar line defined by $\boldsymbol{c}_i$ and $\boldsymbol{r}_j$ is first projected onto the current image using the current camera pose $\boldsymbol{M}_t$ and the intrinsic camera parameters $\boldsymbol{K}$. The current camera pose is estimated by the SLAM algorithm. In the position adjustment phase, a new 3D position of the annotation $\boldsymbol{p}'_i$ can be calculated based on the object's position on the epipolar line.

The main difference between SlidAR and other ray-casting based positioning methods is the used hardware. Ray-casting with HMD has different ergonomical

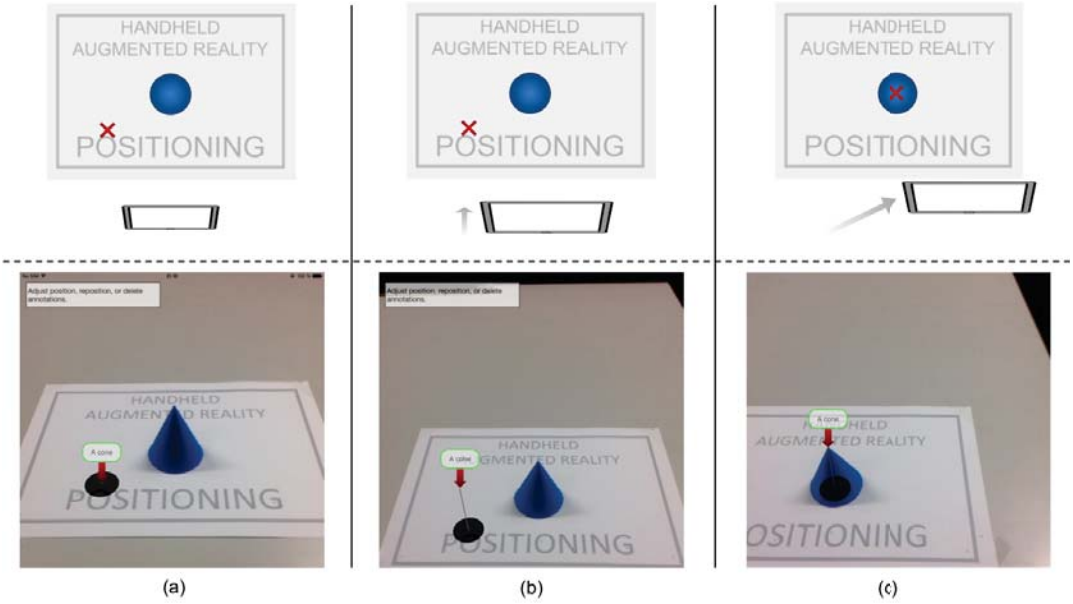

Figure 2.7. HoldAR: a top-down (above the dotted line) and display (bellow the dotted line) views. A virtual object (a red bubble and an arrow) is being positioned to the tip of a blue cone (the target position). The initial positioning is conducted near to the target position (a). A shadow is visualized below the object and a line between these two. While taping and holding the device's display, the device is moved up and the object also moves up (b). Again, the device is moved left and the object moves to same direction (c).

and perceptual issues because the ray is cast based on the head orientation instead of a handheld device's viewpoint. The actual adjustment in previous ray-casting based methods is done by using either special hardware or prior knowledge of the 3D structure of the environment. The positioning is easier this way, but our aim was to develop a method that is suitable for widespread adoption using low-cost hardware. Thus, we used only a handheld device without any prior knowledge of the environment, such as predetermined 3D model or aerial images. The system developed by Bunnus et al. [8] is closest to our work because it also uses ray-casting epipolar geometry. However, this system is not used for positioning virtual objects, but for making 3D models out of real world objects in AR. Furthermore,

the hardware they used is different and it requires an external display.

2.3.1.2 HoldAR

Device-centric positioning method for HAR was first introduced by Henrysson et al. [30] and we chose that for comparison, because it has been the most efficient for 3D positioning tasks in previous studies. We call our SLAM-based implementation of this method HoldAR. Despite the different tracking technoloqy, the interaction metaphor in HoldAR is similar to the marker-based device-centric methods introduced in the related work. With HoldAR, the position of virtual object is controlled by physically moving the device (Fig. 2.7). Unlike with SlidAR, the initial positioning can be done anywhere in the environment (Fig. 2.7(a)).

When a tap-and-hold gesture is performed on the handheld device's display, the position of the virtual object is fixed in the camera coordinate system and the object can be adjusted by moving the handheld device (Fig. 2.7(b) & 2.7(c)). When the tap-and-hold is released, the position of the object is set to the final adjusted position in the world coordinate system. HoldAR shows two virtual depth cues: 1) a shadow (D = 5cm, alpha value = 0.8) directly below the object on a ground plane and 2) a line between the object and the shadow. If the initial position is unclear or far away from the target position, the initial positioning can be done again with a cut & paste function. In order to visualize the depth cues correctly, a ground plane below the virtual object needs to be detected.

2.3.2 Pilot Study

Before the actual comparative study, we conducted a small-scale pilot study 2.8 in order to compare SlidAR and HoldAR in an real world 3D positioning scenario were annotations are created for the inspection of a large computer machine. A similar scenario is used later in Chapter 3. The pilot had a within-group (2×1) design with four test participants. In the pilot scenario, participants had to annotate four targets in a computer machine. Targets were various cables and other computer parts. None of the targets were mapped correctly so the test participants had to use SlidAR or HoldAR method to adjust the position of all

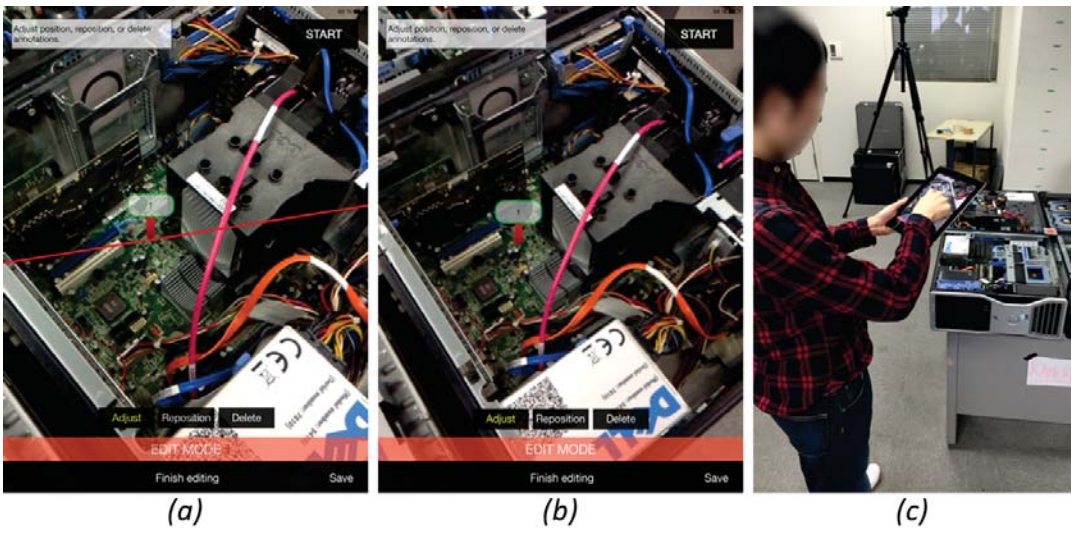

Figure 2.8. SlidAR (a), HoldAR (b), and a participant conducting the pilot study (c). SlidAR can be used regardless of the 3D structure of the environment. HoldAR shows the shadow incorrectly: it is in mid-air instead of at the surface of the ground plane (the green motherboard).

four annotations. We measured task completion time and subjective feedback in form of freeform comments.

In terms of average task completion time, SlidAR (M = 139, SD = 29) was faster than HoldAR (M = 179, SD = 27). Due to a small sample size, we did not perform statistical analysis. All participants preferred SlidAR over HoldAR. The largest drawback of HoldAR was the incorrect visualization of the depth cues. Even though the computer machine had well distinguishable horizontal planes, the depth cues were misleading and not shown correctly. If the environment has a complex 3D structure, correctly visualizing the shadow with a SLAM system is very difficult. The inaccurate shadow caused users to perceive the position incorrectly and adjust annotation positions too high from the target positions 2.9. If they noticed the mistakes, correcting the position caused them to additional adjustments. The difference in mean task completion time can be acceptable from the point of view of the creating a checklist for inspection. However, the inaccurate position is a severe problem because it could cause users to inspect

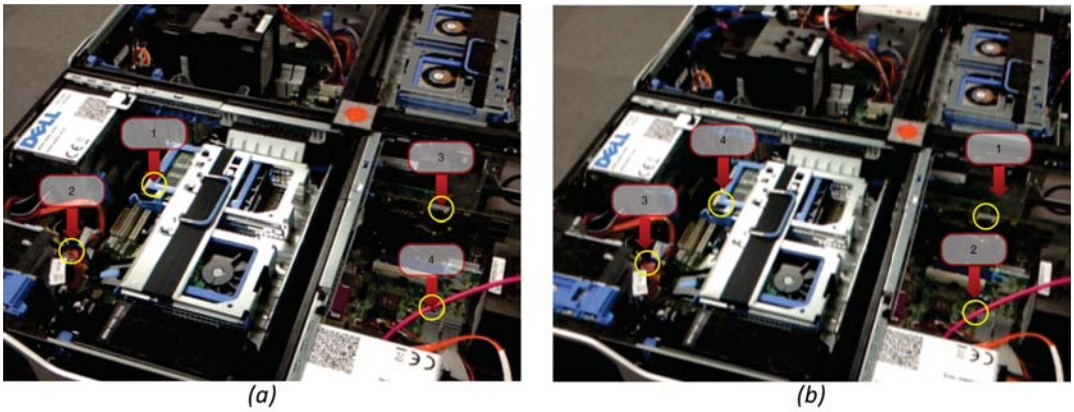

Figure 2.9. The positioning results from one participant: SlidAR (a) and HoldAR (b). The yellow circles illustrate the target positions. The final annotation positions while using the HoldAR where incorrect and too high from the desired target positions.

wrong targets.

Even though 3D positioning methods similar to HoldAR has been very efficient in past, they have been evaluated only in scenarios with an easily trackable and distinguishable ground plane [30, 54, 51] and without any surrounding real world objects. Visualizing the depth cues properly can be possible in a complex environment, but it is still very difficult and a technical challenge as itself. Thus, evaluating SlidAR against HoldAR in a scenario similar to the one in this pilot study would not be fair with the current depth cue visualization technique used in the device-centric methods.

2.3.3 Study Design

After the pilot study, we conducted the actual comparison. The purpose of the comparison was to evaluate the efficiency of use [60] and subjective feedback of SlidAR method against the HoldAR method. In our study, the efficiency consists of three objective quantifications: 1) the average time needed to complete a task, 2) the average magnitude of positioning errors (accuracy), and 3) the average amount of device movement needed to complete a task. In addition, we observed

the usage of the positioning methods during the study sessions. The pilot study showed us that showing the depth cues correctly in HoldAR is very important. Thus, we decided to choose a generic scenario for the actual comparative study in order to visualize the depth cues correctly for HoldAR. This allows us the focus on the low level interaction and efficiency of both methods.

Furthermore, HAR 3D positioning, and manipulation in general, has a large amount of possible scenarios where it can be needed, not only inspection. We did not choose a certain application domain, because we wanted to focus on a single problem in HAR 3D positioning that is similar to all domains. Furthermore, different domains can have specific use environment related issues that would affect to the generalization of results. Thus, we chose a laboratory scenario for easier generalization of the positioning task itself and for better controllability. In addition, we had to take into account the requirement of depth cues for HoldAR method. That is, in order for the comparison to be fair, we needed a test scenario that has a ground plane.

We used a within-group factorial design that included two independent variables (2×2): the positioning method (*SlidAR, HoldAR*) and the test task difficulty (*Easy, Hard*). The dependent variables were task completion time, positioning accuracy, device movement, and subjective feedback. Four conditions were evaluated and counterbalanced measures were taken (counterbalanced condition orders and breaks between conditions) to prevent possible learning effects. A total of 23 graduate school students (16 male and 7 female; mean age, 29±5 years; age range, 22 to 41; mean height, 167.5±12.8cm) were recruited as test participants. None of the test participants participated to the pilot study. All of them successfully completed the study. On a 7-point Likert scale (1= not familiar at all and 7 = very familiar), participants estimated their previous experience with touchscreen handheld devices (M = 6.4, SD = 0.9), AR (M = 4.2, SD = 1.4), HAR (M = 3.7, SD = 1.5), and 3D user interfaces (M = 4.6, SD = 1.4).

We used a 4th generation iPad[6] as a test device. The 4th gen. iPad has a 1.4 GHz dual-core processor and a 9.7 inch display with the native resolution of 1536 x 2048 pixels. In our HAR system, the resolution of the camera's video output was set to 480 x 640 pixels due to performance limitations of the iPad and

[6]https://support.apple.com/kb/SP662

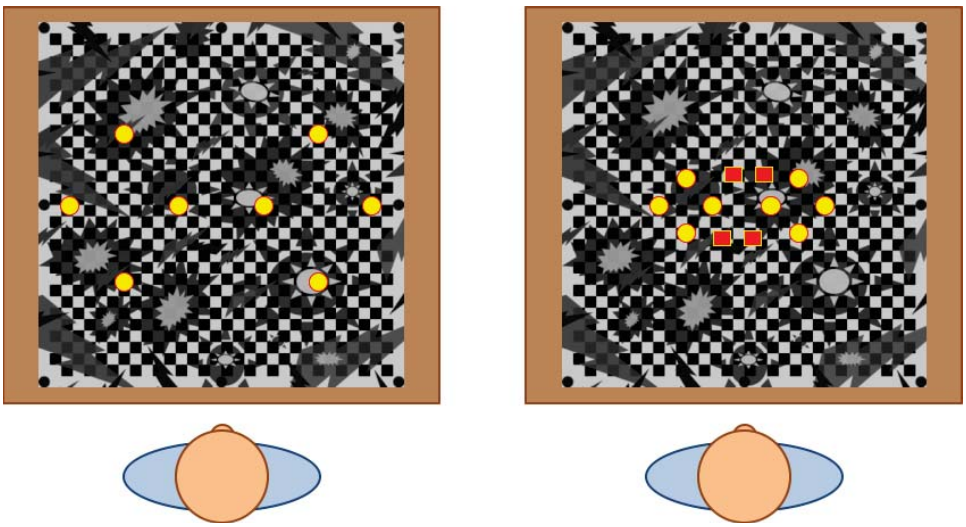

Figure 2.10. An illustrated top-down view of the study tasks. The easy task with eight target positions on top of eight Lego structures (left). The hard task with eight target position on top of eight Lego structures and four faux Lego structures (right). The yellow circles represent the target structures and red squares are the faux structures.

the PointCloud SDK. The system was usable only in a portrait orientation. The SLAM maps of the test environment were created in advance and the detection of additional feature points was disabled during the study. The detected feature points were not visible to the participants. Every participant used the same SLAM maps.

2.3.4 Study Tasks

In both tasks, participants had to position virtual objects relative to real world objects (Fig. 2.10). Here, with a virtual object we refer to a short virtual 2D textual annotation. Textual information is 2D by nature and there is no need to present it in 3D [17]. This withdraws rotation and other manipulation tasks from the scope of this study. The participants were asked to number the virtual objects using the device's touchscreen keyboard.

Tasks contained eight target positions at the top of eight Lego structures,

and featured a predetermined order for conducting the eight positioning tasks. Because each participant did both tasks twice, two equally difficult versions of the predetermined positioning order were prepared. The structures were placed on a small table (length = 80cm, width = 80cm, and height = 70cm). The participants were allowed to move around the table if they felt it necessary. The pattern on the surface of the table served as a ground plane for the depth cues were of the HoldAR method. The Lego structures on the table were not part of the SLAM maps, which means that participants had to always conduct the position adjustment.

In both tasks, the eight target positions were on top of four low (height = 16.32cm) and four high (height = 31.68cm) Lego structures. The hard task was more dense because the structures were placed in closer proximity to each other and four faux structures were added. The faux structures did not have target positions. The purpose of the hard task was to investigate the effect of higher object density. Our HAR system did not inform when the positioning was accurate enough and the level of accuracy was based on the participants' own perception. The target positions were located at the top most blocks in Lego structures and in order to avoid the ambiguity in accuracy measurement, they did not have a volume (Fig. 2.11). We used real world target positions instead of virtual ones in order to simulate a practical scenario.

2.3.5 Study Procedure

The user Study consisted of a pre-questionnaire followed by the all four conditions and a post-questionnaire. The whole study took approximately 80-90 minutes per participant. After the pre-questionnaire, instructions (a slide presentation and a video demonstration) to both methods were given. Finally, participants were able to practice the methods in a tutorial tasks sequentially. Feedback was given to participants during the tutorial tasks. In the tutorial for SlidAR method, we emphasized two main points: 1) The initial positioning should be done as accurately as possible; 2) To do the adjustment, the viewpoint needs to be changed from the initial viewpoint. For HoldAR, the following three main points were instructed: 1) the initial positioning can be anywhere in the environment. 2) The shadow is always directly below the virtual object on the ground plane;

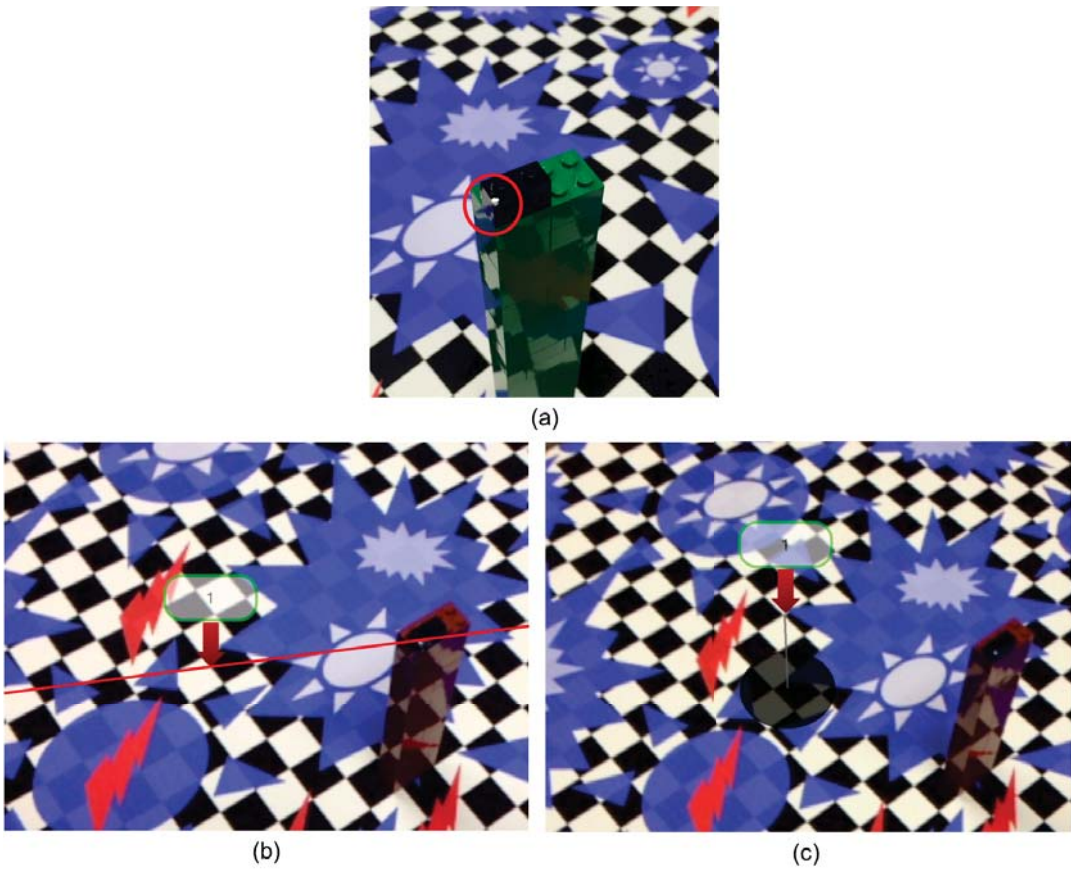

Figure 2.11. A target position on the top most block of the Lego structure (a). The positioning being conducted with SlidAR (b) and HoldAR (c).

and 3) the movement of the device also moves the virtual object similarly.

The participants were instructed to position the virtual objects as accurately as possible, and move on whenever they felt the positioning was accurate enough or that they could not conduct it more accurately. We also instructed how to use the cut & paste function in situation where initial positioning was not done correctly. This was important especially with SlidAR where the position could be adjusted only along the epipolar line. The participants were told that the Lego structures are not part of the SLAM maps. They were also encouraged to check the position of the objects from different viewpoints. After each condition,

there was a four minute break. During the break, the participants were reminded of the main points of the positioning method in the next condition, but did not receive any further feedback on their performance. In case of tracking failures, the system instructed participants to return to a marked starting point in front of the table and to initialize the tracking again.

Table 2.2. The results from the objective measurements. N = 23.

Method	Task	Task time (seconds)		Positioning error (mm.)		Device movement (m.)	
		Mean	SD	Mean	SD	Mean	SD
SlidAR	Easy	361.04	122.84	11.5	16.0	35.8	15.6
SlidAR	Hard	403.65	172.04	12.0	20.2	37.4	13.4
HoldAR	Easy	488.61	248.04	14.0	11.0	53.3	23.9
HoldAR	Hard	601.96	255.73	14.3	11.1	65.7	31.4

2.3.6 Hypotheses

We formulated the following four hypotheses for the positioning study. H1-H3 address the different quantifications of efficiency of use and H4 deals with the effect of the task difficulty to HoldAR method. Because the device movement required in SlidAR is more consistent and fewer DOFs are controlled at the same time, we hypothesize that it should performed significantly better against HoldAR (H1-H3). HoldAR relies heavily to virtual depth cues and because of this we assume that the environment has a higher effect to it's efficiency compared to SlidAR (H4).

- **H1**: SlidAR has a lower task completion time.
- **H2**: SlidAR has a lower error rate in positioning accuracy.
- **H3**: SlidAR requires less device movement.
- **H4**: HoldAR has a higher efficiency in the easy task than in the hard task.

2.3.7 Results

In this section, we describe the results of each objective and subjective measurement separately. Table 3.1 shows the summary of results for the objective

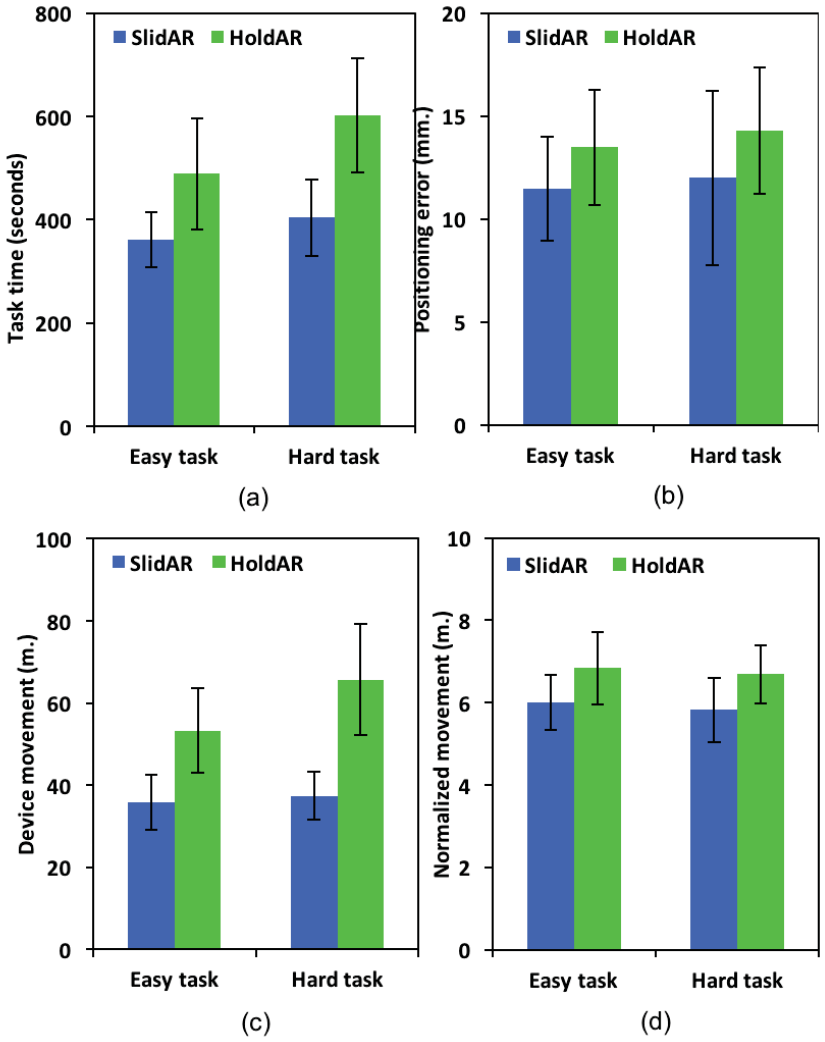

Figure 2.12. The results from the objective measurements. (a) The average task completion times in seconds. (b) The average positioning errors in millimeters. (c) The average amount of device movement in meters. (d) Normalized device movement per minute in meters. Connected bars represent significant differences between means (* = significant at 0.05 level ** = significant at 0.01 level, *** = significant at 0.001 level). N = 23 and error bars = +-95% CI.

measurements.

2.3.7.1 Task Completion Time

Figure 2.12(a) shows the average task completion times. The measurement included all eight target positions in each task. The participant started the timing and stopped it after the task was completed. We noticed a significant difference between the methods in terms of overall task time from both tasks. A repeated-measure ANOVA showed that SlidAR (M = 382, SD = 149) method was significantly faster than HoldAR (M = 545, SD = 256) method: F (1, 22) = 28.08, $p < .001$. Similarly and expectedly, we noticed a significant effect of the task difficulty on the completion time: F (1, 22) = 16.61, p = .001. We did not notice any significant interaction effect of Method × Test task. The results support the **H1**, but not the **H4**.

2.3.7.2 Positioning Accuracy

We calculated the average positioning errors in order to determine the overall accuracy (Fig.2.12(b)). We measured the positioning error by calculating the distance between the positioning done by the participants and the target positions (Fig. 2.11). The 3D coordinates of the SLAM maps and absolute coordinate system were registered by manually specified corresponding points. Although SlidAR (M = 14.8, SD = 7.98) caused less error than HoldAR (M = 18.3, SD = 6.71), we did not notice any significant difference between the two: F(1, 22) = 2.66, p = .117. Similarly, the hard task did not cause more errors than the easy task: F(1, 22) = 2.81, p = .113. There was no significant interaction effect either. The results do not support **H2** nor **H4**.

2.3.7.3 Device Movement

Figure 2.12(c) shows the average amount of movement during the task. We measured the overall trajectories of the device's movement based on the pose of the device's camera related to the tracked environment. The camera pose information was saved 30 times per second and the trajectories between each pose were added together. The movement was calculated only while the environment was tracked and the extra movement caused by the loss of tracking was not included in the overall trajectories.

We analyzed the movement data using a repeated measures ANOVA. The analysis revealed that overall, when using SlidAR (M = 36.60, SD = 14.42), participants had to move the display significantly less than HoldAR (M = 59.50, SD = 28.31) $F(1,22) = 31.47$, $p < .001$. Expectedly, during the easy task (M = 44.54, SD = 21.83) participants had moved the device significantly less than during the hard task (M = 51.56, SD = 27.85) $F(1,22) = 18.04$, $p < .001$. We have also noticed a significant interaction effect of Method × Test task $F(1,22) = 4.4$, $p < .05$. Both of the methods required less display movement for the easy task than the hard task, however, this decrease in movement was significantly more in the case of HoldAR than SlidAR.

Additionally, we analyzed device movement data normalized by time, i.e. device movement per minute 2.12(d). Our analysis revealed that SlidAR (M=5.91, SD=0.25) had significantly less device movement per minute than HoldAR (M = 6.76, SD = 0.27); $F(1,22) = 11.91$, $p = .002$. Interestingly, we did not notice a significant effect of task on normalized device movement. The device movement results support **H3** and **H4**.

2.3.7.4 Subjective Feedback

We collected subjective feedback with the Handheld Augmented Reality Usability Scale (HARUS) [75] and written freeform comments. We also asked participants which method they preferred.

We used HARUS (Table 2.3) in the questionnaire that measures participants' overall opinion about the manipulability (Table 2.3, S1-S8) and comprehensibility (Table 2.3, S9-S16) of HAR on a 7-point Likert scale. The manipulability and comprehensibility statements consider different ergonomic and perceptual issues common to HAR, respectively. To analyze HARUS data we used paired two-tailed t-tests for the HARUS scores. For manipulability, the SlidAR method (M = 70.83, SD = 10.69) was significantly easier to handle than HoldAR (M = 48.57, SD = 18.54); $t(22) = -4.82$, $p < .001$. For comprehensibility, the SlidAR method (M = 76.3, SD = 10.83) was significantly easier to understand than HoldAR (M = 66.96, SD = 15.71); $t(22) = -2.61$, $p = 0.02$. Overall, SlidAR (M = 73.57, SD = 6.54) was significantly more usable than HoldAR (M = 57.76, SD = 15.39); $t(22) = -4.54$, $p <$. 001. Figure 2.13 illustrates the results of individual statements. A

Table 2.3. The HARUS statements

	Manipulability:
1	I think that interacting with the positioning method requires a a lot of body muscle effort.
2	I felt that using the positioning method was comfortable for my arms and hands.
3	I found the device difficult to hold while operating the positioning method.
4	I found it easy to manipulate information through the positioning method.
5	I felt that my arm or hand became tired after using the positioning method.
6	I think the positioning method is easy to control.
7	I felt that I was losing grip and dropping the device at some point.
8	I think the operation of the positioning method is simple and uncomplicated.
	Comprehensibility:
9	I think that interacting with the positioning method requires a lot of mental effort.
10	I though the amount of information displayed on screen was appropriate.
11	I though that the information displayed on screen was difficult to read.
12	I felt that the information display was responding fast enough.
13	I though that the information displayed on screen was confusing.
14	I though the words and symbols on screen were easy to read.
15	I felt that the display was flickering too much.
16	I though that the information displayed on screen was consistent.

significant differences with $p < .001$ where found from S1, S4, S6, S8, and S9. A significant difference with $p < .05$ were found from S2, S3, S5, and S12.

In the freeform comments and rankings, overall, 14 participants preferred the SlidAR, seven preferred HoldAR, and two could not say. SlidAR was seen straightforward and fast. It did not require participants to move a lot, because in most cases the viewpoint had to be changed only once: from the initial viewpoint to a new viewpoint to conduct the adjustment. The drawback of SlidAR was the unclear visualization of the epipolar line. Moreover, the initial positioning was considered difficult because it had to be very precise. Even though it was not

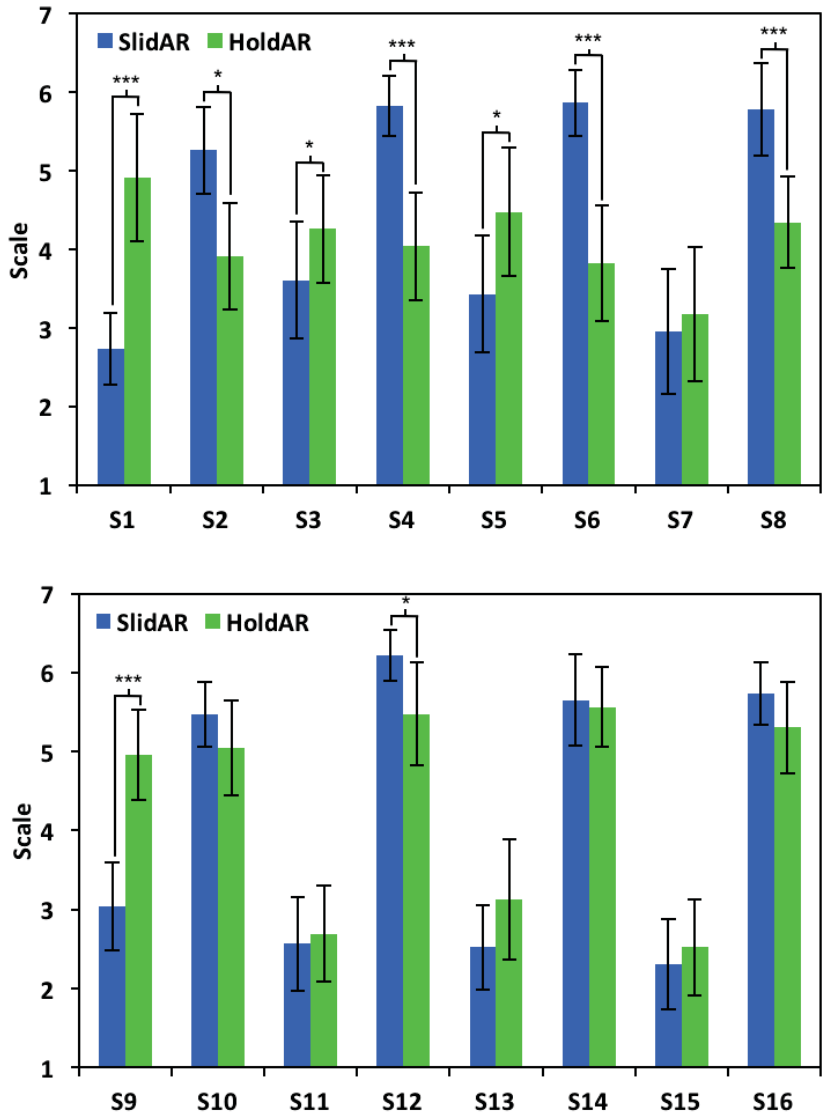

Figure 2.13. Subjective feedback results from the HARUS in a 7-point Likert scale: manipulability statements (S1-S8) and comprehensibility statements (S9-S16). S1-S16 represent statements from Table 2.3. Connected bars represent significant differences between means (* = significant at 0.05 level, *** = significant at 0.001 level). N = 23 and error bars = ±95% CI.

necessary to keep the fingers directly on top of annotation while conducting the slide gesture, some participants mentioned that their fingers sometimes block the view to the target position. Furthermore, holding the device with one hand while conducting the initial positioning and position adjustment can be tiring.

The initial positioning with HoldAR was reported as fast because it did not have to be accurate. The position adjustment, in general, was seen as intuitive, but a precise matching of the virtual object with its shadow to the target position was difficult. The simultaneous use of 3 DOF for adjustment was considered as unwanted because the method was sensitive to small movements and requires very fine adjustments and steady hands. The adjustment was seen more difficult in the hard task because the real objects were often occluded and it was difficult to perceive the position of the shadow correctly. Some participants felt that it was more intuitive to conduct the initial positioning precisely to the target position, similar to SlidAR, instead of positioning it freely to the close proximity of the target position.

2.3.7.5 Observations

The observations were conducted based on the video recordings of the device's display. In SlidAR, the 2D visualization of the epipolar line caused issues because participants were not always sure of the direction of the line. If the participants forgot the viewpoint of the initial positioning, they sometimes tried to conduct the position adjustment from the initial viewpoint. In HoldAR, the visibility of the shadow was sometimes an issue. If the ground plane had same coloring as the shadow, the shadow can get lost to the environment. This caused participants to perceive the depth incorrectly. The participants often adjusted the virtual objects to directions where they did not want objects to be adjusted because they were controlling all 3 DOF at once. With HoldAR, the positioning had to be confirmed from multiple viewpoints.

2.3.8 Summary

This study showed that SlidAR was beneficial and easy enough to use for positioning virtual content. Even though it would still need some improvements,

we can consider that it solves the problem of 3D positioning in SLAM-based HAR systems that are unable to create a completely accurate 3D model of the environment

2.4. Discussion

In this section we discuss about the findings from both user studies and how they could be applied to the real world scenarios.

2.4.1 First Study

In the first study, the test scenario represented a real world scenario making the findings applicable to other real world scenarios as well. Many of the issues found are were related to conventional GUI design issues that can be already considered solved. Other issues were related to SLAM tracking and it's initialization. This could also be improved based on the resent related work. The largest unsolved problem was the 3D positioning. Even though the study showed that positioning adjustment is often needed, in many situations only the initial position is still enough. This shows us at the SLAM-based HAR is viable option for creating annotations for inspection and other goal-oriented tasks.

2.4.2 Second Study

We did the second study was done in an abstract scenario because of the results of a pilot study. We discuss about the results in this abstract scenario and also how our findings could be applied to real world scenarios.

2.4.2.1 Test Scenarios

We assume that SlidAR was faster mainly due to very specific target positions. Even though the accurate initial positioning took some effort, the position adjustment was quick and accurate because only 1 DOF was controlled. There was no need to constantly change the viewpoint and the adjustment was not affected by the unintentional movement of the device. The initial positioning with HoldAR

was fast, but the position adjustment was time consuming because 3 DOF were controlled. This made the adjustment vulnerable to unintentional movement and perceptual errors.

A direct tap gesture is very intuitive as an initial position, but it has problems regarding the ambiguity caused by user's finger size and the shakiness of the handheld devices. This can be an issue SlidAR if the target positions in the real world are very small in which case initial positioning has to be very precise. Initial direct tap based positioning could be improved with view freezing [23] or with a combination of view freezing and Shift [84].

HoldAR does not require a precise initial position because target position does not need to be on the ray cast from the camera. According to participants' comments, however, more mental effort is required if they have to decide the initial position based on how effectively they can translate the virtual object from the initial position to the target position.

The perceptual issues [43, 14] can have a considerable effect on positioning accuracy when target positions are real instead of virtual. The combined average error rate in all conditions (M = 12.8mm, SD = 1.3mm) can be due to the issues in perception and the participants' judgment of the sufficient level of accuracy. A small positioning error can be very difficult to detect if the position is not checked from several viewpoints and at a close distance. Furthermore, the low resolution (480 x 640 pixels) of the video output in our implementation and the 2D representation of virtual objects can affect the accuracy in both methods. The large amount of variation (Fig. 2.12(b)) in the positioning errors of SlidAR can be explained with the threshold of adjusting the objects position away from the epipolar line. Because an arbitrary adjustment with SlidAR was impossible, the virtual objected had to be first repositioned and then adjusted again along the new epipolar line. Some participants may have settled with a certain level of accuracy due to the required effort in repositioning, even if they were aware that the position was not accurate enough.

The overall and normalized device movement needed was significantly higher because the position had to be adjusted and confirmed several times with HoldAR. The movement required while using SlidAR was more consistent. Furthermore, the adjustment was done with gestures without the need to move the handheld

device. The significant difference in movement between the easy and hard task with HoldAR can be associated with perceptual issues in understanding depth cues. The viewpoint had to be changed if the position of the object and it's shadow was unclear. We did not find significant differences between easy and difficult tasks. As such, based on our observations, the efficiency of SlidAR was not dependent on the environment's complexity.

The subjective results strongly correlate to the results from the objective measurements. Completing the tasks with SlidAR took less effort in terms of time and movement, which is reflected to overall manipulability scores. The comprehensibility scores were also significant, but this was mainly due to S9 and S12, which are related to the difficulties of controlling and perceiving the position accurately. The remaining comprehensibility statements were expectedly not significantly different, because both positioning methods were implemented to the same HAR system and their user interfaces were very similar.

Although the study results only support **H1** and **H3** but not **H2**, we argue that the SlidAR was more efficient in our test scenario. It can achieve the same level of accuracy with significantly less time and less effort compared to the HoldAR method. The **H4** was supported only partially, but it shows that the environment can affect those methods that require virtual depth cues to be displayed in the environment. The study gave important knowledge about the real world object annotation and HAR 3D positioning.

2.4.2.2 Real World Scenarios

In the test scenario of our comparative study, we considered two important aspects that are often missing from HAR positioning studies in the related work: 1) We used real target objects (Lego structures, Fig. 2.11) instead of virtual ones (e.g. target zones visualized with virtual rectangles) to simulate a practical scenario where virtual objects is very often spatially dependent on the environment [91]. 2) We did not have predetermined initial positions for the virtual objects.

The initial positioning is a fundamental part of the AR content creation in practical scenarios and it should not be separated from the position adjustment. Especially in case of SlidAR, where position adjustment is highly dependent on the accuracy of initial positioning. In practical scenarios, simply adjusting the

position between two points can be unrealistic if we are unable to justify why user would have chosen the specific initial position. In addition, we forced participants to move around while doing the tasks instead of just standing still or sitting. This is important, because HAR is used in mobile context, which can require users to move around.

There are still few matter that should be considered when applying our findings to practical scenarios, such as creating AR annotations to machines inside of factory or to medical equipment inside a hospital. In the test scenario, the participants were aware that the real objects are not mapped by the SLAM-system. This was because we wanted to focus on the specific HAR positioning problem that can occur often, but not every time. We designed the test scenario in a way that the positioning problem occurs every time. In practical scenarios, users might not always what in the environment is mapped and what is not. Thus, we would not know if the virtual object's initial position going to be correct or is position adjustment also needed. If we use SlidAR, it is not necessary to know is the real world object mapped or not because the initial positioning is conducted in similar manner in both situations. With HoldAR, however, the user might need to choose a initial position differently if it is too far away from the target position.

We did not limit the movement in any way and participants were allowed to freely move around the scene. Neither method did not require users to move 360 degrees around the target position, but in practical scenarios the environment might set limitations to the movement. This could possibly affect the performance of both methods: SlidAR requires the user to move to a new viewpoint and HoldAR relies to movement entirely.

Our test scenario was ideal for HoldAR, because we had an easily trackable ground plane in order to correctly show the depth cues. We chose this scenario because of the findings from the pilot study. The complexity and the structure of the environment can vary a lot depending on the scenario, which can make it more challenging to display depth cues correctly. This was proven in our pilot study where some participants made large errors in terms of position accuracy. The pilot study also showed us SlidAR does not require depth cues to be visualized on the environment, thus making it easier to use various practical scenarios, like the one

that we had in the pilot, with different level of environmental complexities.

The required level of accuracy can also depend highly on the scenario and the real objects that are being annotated. Small objects, such as buttons or cables, require precise positioning. Larger objects, like factory machinery, can allow more ambiguity. Positioning to a larger object is easier, regardless of the used method. Initial positioning would be easier with SlidAR and the position adjustment would be easier with the HoldAR.

We used a tablet device, but both methods could also be used on a smartphone. Tablets are beneficial because it provides more screen estate thereby easing perception and gesture-based interactions [13]. This can be beneficial for example in industrial or medical systems where in addition to AR it is necessary to view traditional 2D information also. The form-factor of the device and the amount of movement needed can affect the usability of HoldAR because it relies on the physical movement. With SlidAR, the form-factor affects the initial positioning because the device had to be kept as still as possible in order to perform the positioning correctly. This could be improved by adding view freezing discussed in previous section.

We chose a generic test scenario instead of a practical one, because the positioning problem can occur in any kind of practical scenario. Conducting the study in a practical scenario, such as inside a hospital or a factory can be risky, because the results could be affected by a scenario itself. This would steer the research focus away from the fundamental object positioning problem that is not specific for any type of scenario. A generic test scenario allowed us to focus more closely to the positioning problem and it gave us a solid implications regarding the efficiency of SlidAR. Practical scenarios might have some differences compared to our test scenario, but these are rather minimal. Furthermore, we believe that in practical scenarios SlidAR would provide even greater efficiency over HoldAR, because HoldAR requires more movement and virtual depth cues. Despite the possible differences between our test scenario and practical scenarios, we strongly argue that our results can be applied to various scenarios, because the fundamental 3D virtual object positioning task is required in any kind of practical scenario where we want to create AR content to the environment.

CHAPTER 3

HAR in Visual Observation Tasks

In the previous Chapter, we confirmed that virtual annotations can be positioned correctly using only initial positioning or a ray casting based position adjustment. This functionality works as a basis for the actual inspection task. In this Chapter, we focus on the evaluation of HAR in the visual observation part of an inspection task.

We have conducted two visual observation user studies: first study focused on finding usability issues from the use of HAR in a real world A/V equipment inspection scenario. In the second study we compared AR against a non-AR picture interface in a generic machine inspection scenario. Even though creating annotations can be sometimes necessary during the inspection task, in our studies we focused only to the visual observation without the annotation creation. This was because observation is the main and mandatory part of an inspection task and creating annotations can be seen as optional.

We also conducted a comparative study for HAR in a physical manipulation task. This study is explained in the appendices because it is outside the scope of visual observation tasks. However, the results from this manipulation study was used in the design process of the test tasks for the comparative study explained in Section 3.3.

3.1. Related Work

The use of AR in difference task support scenarios has been widely studied using various AR display techniques. Mobility is an essential part of conducting an inspection. Thus, we focus only on mobile AR task support systems. We divide the task support systems in related work based on two most commonly used AR display techniques: Head-Mounted Displays (HMDs) and handhelds.

3.1.1 HMD Systems in Task Support

AR on HMDs has been often utilized for tasks that require physical manipulation of the environment. Henderson and Feiner [27] have evaluated the benefits of an HMD system in a vehicle maintenance task showing a 50% increased task performance. Later, same authors [28] evaluated AR in a psychomotor phase of a procedural assembly task. Their results showed that AR offers higher efficiency compared to conventional task support mediums. Tang [78] et al. have compared the effectiveness of AR in an object assembly task where they compared AR against picture-based guides.

Ishii et al. [35] have conducted various studies to AR systems in nuclear power plant maintenance. Authors tested different kinds of display techniques in a large nuclear facility. The ARVIKA project [92] examined the use of AR in various industrial tasks, raging from production to service. Platonov et al. [66] developed and evaluated a mobile HMD AR system for maintenance and repair tasks. HMD AR systems have been also utilized in collaborative various decision making and planing tasks [61] where they are used to visualize important information.

3.1.2 Handheld Systems in Task Support

HAR has a high mobility and several studies have evaluated its efficiency in outdoor navigation [57, 15, 16]. All the systems mentioned above use sensor-based tracking of the environment and do not require user to physically manipulate the environment. However, using AR has not been shown to offer significant benefits compared to conventional navigation systems. Jung et al. [36] have presented a

HAR guidance system that displays related to various indoor locations. Rauhala et al. [68] developed a HAR system that visualizes network sensor data to walls. Their system allows users to inspect sensor data using AR. The system does not require physical manipulation of the environment. A common factor for all the studies mentioned above is that they do not require users to physically manipulate the environment.

Hakkarainen et al. [24] have developed a marker-based HAR maintenance assembly guidance system for small-scale objects. Their system displays complex 3D models. Due to technical limitations, their system does not work in real-time, but the instructions are shown on static images captured by the user. Karlsson et al. [38] have developed a markerless HAR system, which detects and tracks predetermined objects in the real world and displays corresponding 3D models overlaid to real objects. Their system can be used to examine hidden information via transparency. Träskbäck and Haller [81] have developed a simple HAR tablet system for oil refinery work training. Authors also created user AR requirements for the oil refinery systems. Liu et al. [48] have evaluated HAR in a simple device set-up scenario where AR was compared against conventional interfaces. Gauglitz et al.[18, 19] have developed a HAR prototypes for remote collaboration. Their systems allow the remote user to control their own separate view of the environment and create annotations to the local user.

3.1.3 Summary

AR systems using HMDs have been widely evaluated as a support system in various goal-oriented tasks. HMD-AR has been often shown to provide higher efficiency in tasks that require physical manipulation. However, tasks that focus on visual observation have not been evaluated. HAR has not been seen to offer large benefits in comparative studies tasks that require physical manipulation. However, tasks that focus on visual observation of near-field real world have not been evaluated using HAR interfaces.

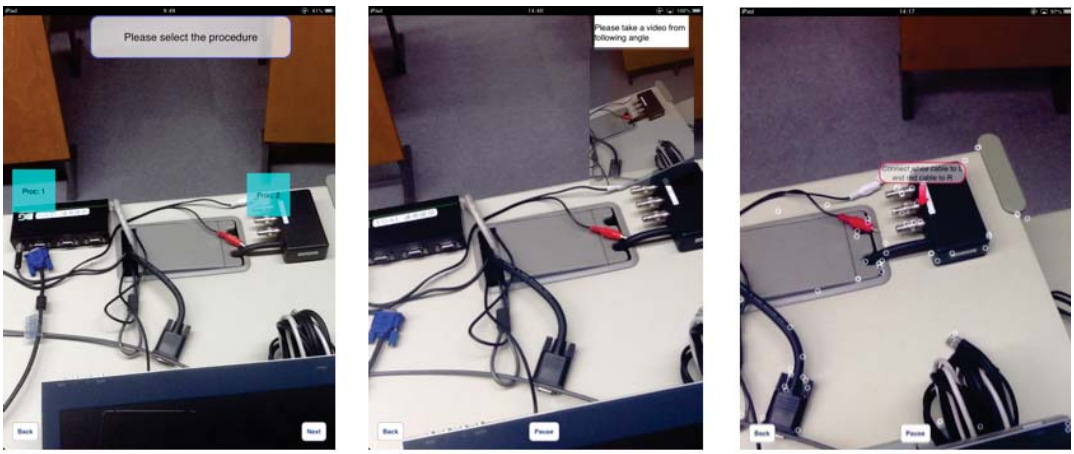

Figure 3.1. The prototype system used in the first observation study. Overview of all annotations (a), instructions for the correct viewpoint (b), and system displaying one annotation (c).

3.2. First Observation Study

The first study focused on finding usability issues from the use of HAR in a real world inspection scenario. The task focused on visual observation of A/V equipment. We wanted to find out what are the benefits and drawbacks HAR and it is fundamental to gain findings from a real world scenario. This will teach us about the actual problems of visual observation task. The results form the first study and the manipulability study have been published in a conference proceedings [67].

3.2.1 Prototype System

We used the same SLAM-based HAR prototype system from previous studies also in this study. However, the provided functionality of the prototype differed from previous studies because this time the focus was on visual observation instead of 3D positioning.

Figure 3.1 illustrates three main views of the prototype system. The desired

area of interest can be selected from the overview picture (Fig. 3.1, left). The system uses manual SLAM map selection and the correct SLAM map is selected based on the selected area. This is different to PTAMM, which selects the SLAM map automatically based on the viewpoint of the device's camera. The purpose of manual selection is to achieve reliable SLAM map selection. This is important for the usefulness of the system in practical use scenarios. After the area selection, the SLAM map needs to be detected by choosing the correct viewpoint (Fig. 3.1, middle) and then the annotations can be seen (Fig. 3.1, right).

3.2.2 Study Design

The study scenario was similar to the scenario in the first 3D positioning study, except that this time users had to inspect the A/V equipment instead of creating annotations to them. A lecture room was used as a test environment and it represented a real world scenario where people set up A/V devices prior to holding a lecture. We chose this scenario because setting up these A/V devices can be difficult for those who do not use the equipment often or for the first time. The study had 10 participants (22-28 years old) who were asked to inspect the A/V equipment using our HAR system. None of the test participants had previous experience with the A/V devices in question. A short introduction to the system was given to the participants before the test began and participants were able to practice in a tutorial task. The test was divided into three subtasks:

- Check the VGA-cable and two audio cables from the laptop PC.
- Check the video projector and check the correct inputs.
- Check the audio devices and the audio levels.

We gathered only qualitative results via video observation and subjective feedback. The test environment was recorded with a video camera and the tablet screen was captured with a screen recording software. Furthermore, test participants were encouraged to think out loud while they were conducting the test.

Figure 3.2. A user conducting the first observation study in a A/V equipment inspection scenario.

3.2.3 Results

The study showed that inspection can be conducted using HAR because all the participants were able to complete the task. HAR was seen as intuitive, because the annotations were visualized on top of the real world and there was no need to search for the targets from the environment. Participants said that the overview image is useful because it shows them the approximate location of annotations. All the issues were related to lack of instructions and feedback. For example, the desired viewpoint was not shown clear enough in the SLAM map initialization phase. Most of the participant had some level of problems with the map detection and tracking.

Some of the targets in the test scenario were very small and users had to go really close in order to conduct the inspection. Sometimes this caused the SLAM map tracking to fail and it had to be initialized again. Some participants had problems figuring out how many annotations a certain area of interest contained. The total amount of annotations was mentioned, but it was not clear enough. The largest issue was finding the annotations that the system did not visualize off-

screen annotations in any way. If all of the annotations were not in the camera's FOV, participants did not know to look for them. One test participant did not see any annotations because they where all outside camera's FOV and only feature points were displayed. Two participants tried to tap the annotation text fields, but no interaction was designed for this action.

3.2.4 Summary

The study showed us that and inspection task can be completed using HAR. Many of the problems found were related to the GUI design, not necessarily to the use of AR itself. Still, this study does not yet confirm us the overall usefulness of HAR, because we did not compare it against any conventional visual observation interfaces.

3.3. Second Observation Study

We improved the UI of the HAR prototype system based on the findings gained from the first visual observation study. The second study focused on finding benefits from HAR compared to a conventional picture interface in a generic machine inspection scenario.

3.3.1 Interfaces

In the study, we had two user interfaces: an AR and a Picture interface. Both interfaces displayed annotations as white text bubbles with a red border and a red arrow pointing to the part in question. In addition, the interfaces had a blue progress bar to visualize how many targets remain to be inspected. Both interfaces were design for a 4th generation iPad[1] tablet. This iPad model has a 1.4 GHz dual core processor, 1 gigabyte of memory, and a 9.7 inch display with the native resolution of 1536×2048 pixels. Both interfaces were usable only in the portrait orientation. Even though inspection is conventionally done with only pen and paper, we did not want to include this to the comparison, because

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

Figure 3.3. A test user conducting the machine inspection study. The machine is on top of a table and an omnidirectional camera is in the middle to capture participant's face for the gaze shift measurements.

it does not allow easy information sharing. With digital systems, information of the inspection can be hypothetically shared with other parties involved.

3.3.1.1 AR Interface

The AR UI (Fig. 3.4) was an improved version of the same prototype used in the first study. It used SLAM and tracked natural feature points from the environment. If the tracking failed, the UI showed a user a message and instructions to return to starting position in order to initialize the tracking again. The starting position and the required angle were also instructed to the users before the test. The tracking could not be initialized from any other position other than the starting position. The virtual annotations were positioned according to guidelines presented by Müller [55]. The AR interface did not have an occlusion handling and this was also instructed to the participants.

If an annotation was outside the camera's field of view, a 2D red arrow was displayed pointing to the direction of the annotation (Fig. 3.4b). We used an arrow because arrows have been the most efficient for off-screen content visualization [9, 33]. Tracked feature points were not visible to the participants. The

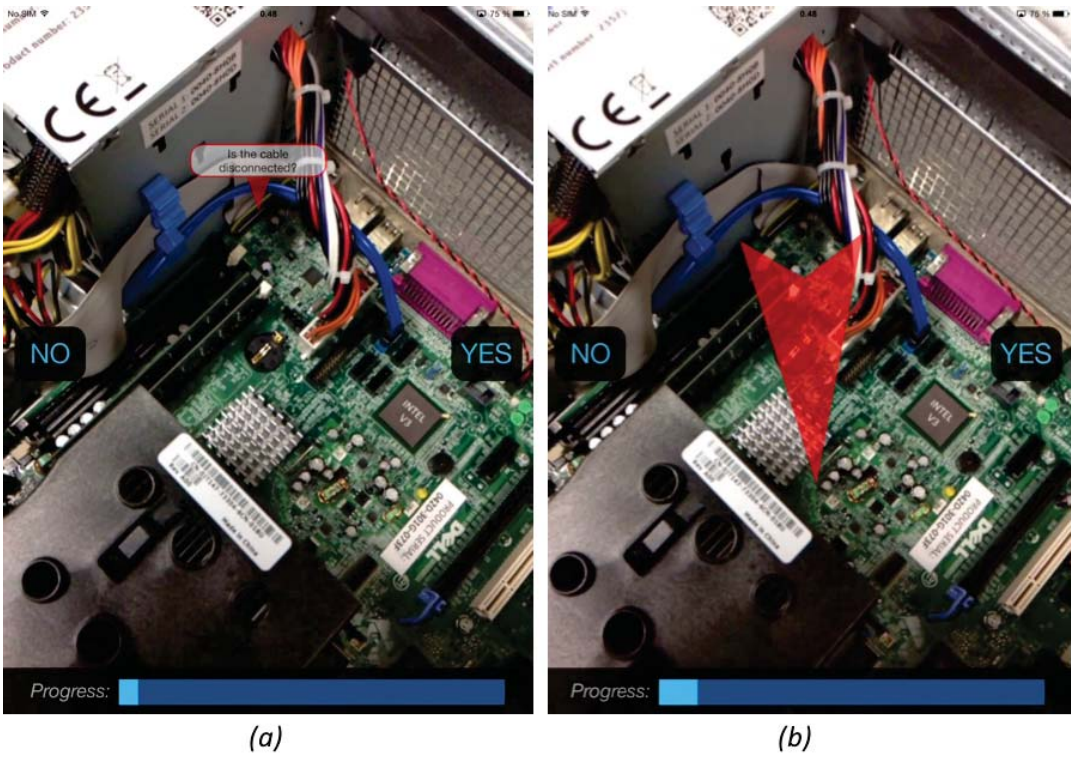

Figure 3.4. The AR interface: annotations are displayed on a live AR view (a) and if an annotation is off-screen, a red arrow is displayed pointing to the direction of the annotations (b).

SLAM map for the study was created in advance and the map expansion was disabled during the experiment, meaning that no additional feature points were tracked. In contrast to the first study, the AR interface did not show an overview of the environment but the annotations were placed within a single SLAM map. The resolution of the video output was 480×640 pixels.

3.3.1.2 Picture Interface

The picture interfaces uses screen-fixed images that depict a top-down image of the that portion of the environment where an annotation is (Fig. 3.5a). Other interface elements are similar to the AR interface. The images in the picture

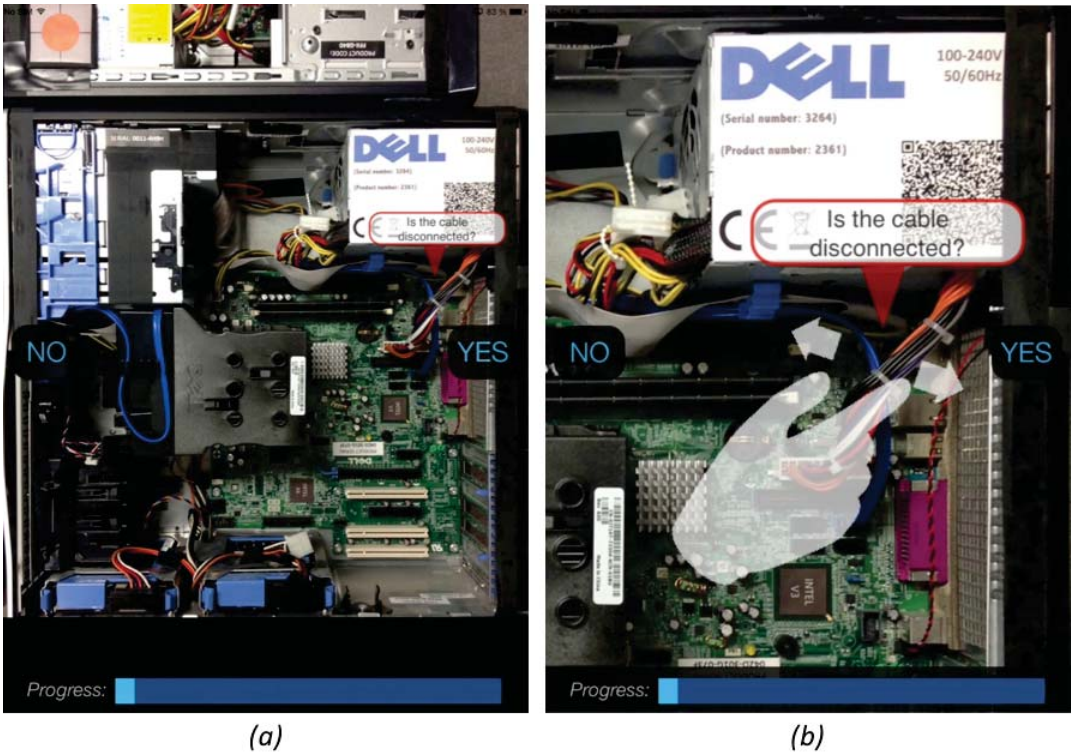

Figure 3.5. The picture interface: A static top-down view of the workpiece and an annotation overlaid to it (a). The view can be zoomed with a pinch-and-zoom gesture (b).

interface are presented in 1536×2048 resolution and taken with the same 4th generation iPad. Images are initially fully visible and users can zoom and pan (Fig. 3.5b) to get a better view of the target area.When a user taps either either of the answer buttons ('YES' or 'NO'), the image was restored to it's initial size. Images were taken from a top-down viewpoint in order to clearly display all annotations in a single image. This was because the targets were placed in 3D space.

3.3.2 Study Design

We used a within-group factorial design that included two independent variables (2×2): the interface and the task. In an inspection system, the efficiency [60] can be considered as the main usability attribute. Here, we define the efficiency to consist of three object dependent variables that we measured: task completion time, inspection errors, and gaze shifts. We also measured subjective feedback with a questionnaire and freeform comments. In addition, we observed the usage of the interfaces during the study sessions. In the test environment, we consider target objects that are within arm's reach (near-field) [11]. Four conditions were evaluated and counterbalanced measures were taken (fully counterbalanced condition orders and breaks between conditions) to prevent possible learning effects. A total of 24 graduate school students (15 male and 9 female; mean age, 28±5 years; age range, 22 to 42) were recruited as test participants.

We chose a to conduct the study in a laboratory, because it made collecting all the necessary measurements easier. The study scenario can be described as a generic visual inspection scenario, which does not directly represent any particular real world scenario. However, our test scenario represented a real world scenario. Often in real world inspection scenarios, such as in factories or other facilities, it is necessary to inspect various small targets, such as cables, meters or other parts.

We used a 4th generation iPad[2] as a test device. This iPad model has a 1.4 GHz dual core processor, 1 gigabyte of memory, and a 9.7 inch display with the native resolution of 1536×2048 pixels. The brightness of the iPad's screen was adjusted to maximum. Both interfaces were usable only in the portrait orientation. However, users were told that they can rotate the device if they want to do so. For the AR interface, the SLAM maps of the test environment were created in advance and the map expansion was enabled during the study. The detected feature points were not visible to the participants.

[2]https://support.apple.com/kb/SP662

3.3.3 Study Procedure

The user study procedure was as follows: a pre-questionnaire, instructions, tutorials, all four conditions, and a post-questionnaire. For the AR interface, we instructed how tracking works and what should be done if tracking is lost. For the picture interface, we emphasized the zooming function and instructed that the device can be rotated if for easier alignment if necessary. Participants were able to practice both interfaces in tutorial tasks and they received feedback during these tutorials. A purpose of the tutorial tasks was also to led participants familiarize themselves with different computer parts they were going to inspect in the actual test tasks. After each condition, participants answered a questionnaire about the previous conditions and had a five minute break. During the break, the a short introduction to the next interface was given. After all conditions were done, participants answered a final questionnaire that measured their overall opinion of the interfaces. The study took approximately 50 minutes per test participant.

3.3.4 Study Tasks

In the related work [48] and in our past studies [67] the results have shown that the use of HAR has not beneficial over picture interface in tasks that can be conducted from one viewpoint without the need for movement or mapping information. For this reason, we did not include this type of simple task in our study. Instead we focused on tasks that have high information density and require information alignment. Thus, we named the tasks 'medium' and 'high.' Furthermore, in order to do full counterbalance for our within-group design, we did not want to include more than four conditions.

In the test scenario, the participants assumed the role of a newly-hired inspection staff where they did not have previous knowledge of the specific workpieces to be inspected. The tasks were divided based on the scale and amount of movement required (Fig. 3.6): task with medium angle alignment (medium angles) and a task with high angle alignment (high angles). In the medium task, viewpoint had to be changed between two areas and between four areas in the high task. Both tasks had the same amount of viewpoint alignments. The information density in

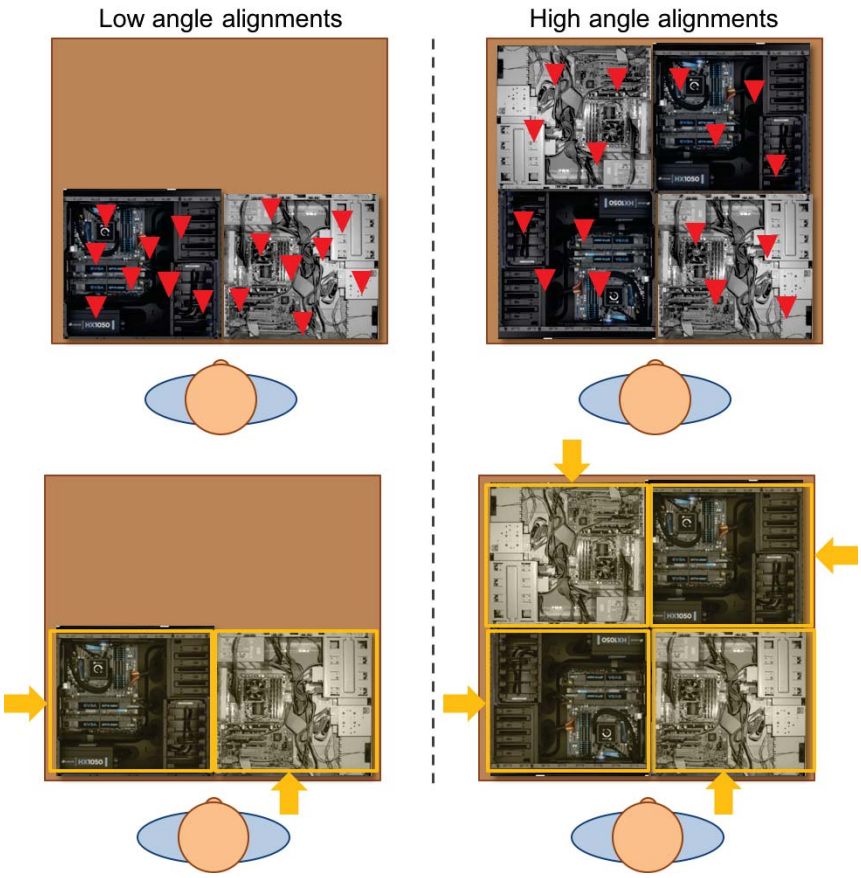

Figure 3.6. Top-down illustrations of the test task design: medium angles task (left) and high angles task (right). Upper part illustrates how targets and annotations were distributed in each task. Lower part illustrates the viewpoint required and the viewpoints used in the picture interface.

both tasks was higher compared to same style of studies in the related work. The targets were placed in a way that each section of the workpiece required a certain angle for inspection. This forced users perform angle alignments.

Same way, in the medium task the picture interface had two separate pictures of two areas and four pictures of four areas in the high task. Because the target were placed in 3D, not on a single plane, it would not have been possible to take only one picture of the whole workpiece. In this case, all of the targets would not

have been visible at the same time.

The medium angles task consisted two different desktop computers placed on their side on top of a table (height = 70 cm). The High angles task had two pairs, total of four computers. The computers in each pair resembled each other with minor differences. The resemblance was done in order to mimic a scenario where the environment looks similar and differentiating the correct area could potentially be difficult. Both tasks included 20 annotations for 20 targets. Only one annotations was displayed at the time. The reporting of an inspection was done by answering either 'YES' or 'NO'.

3.3.5 Hypotheses

We formulated the following three hypotheses for the study. We assume that viewpoint alignments and larger task scenario make the mapping of the information more difficult with non-AR interfaces (H1). We assume there to be very low amount of errors overall and because of this and no difference, and because of this there should be no difference in error rates. The targets could be observed with both interfaces, but using the picture interface would take more time (H2). AR interface should cause less gaze shifts, because participants can see the real world through the handheld devices display and there is a far smaller need to divide attention (H3).

- **H1**: AR interface has a smaller task completion time.
- **H2**: There is no difference in the amount of errors.
- **H3**: AR interface causes less gaze shifts.

3.3.6 Results

In this section, we describe the results of each objective and subjective measurement separately. Table 3.1 shows the summary of results for the objective measurements.

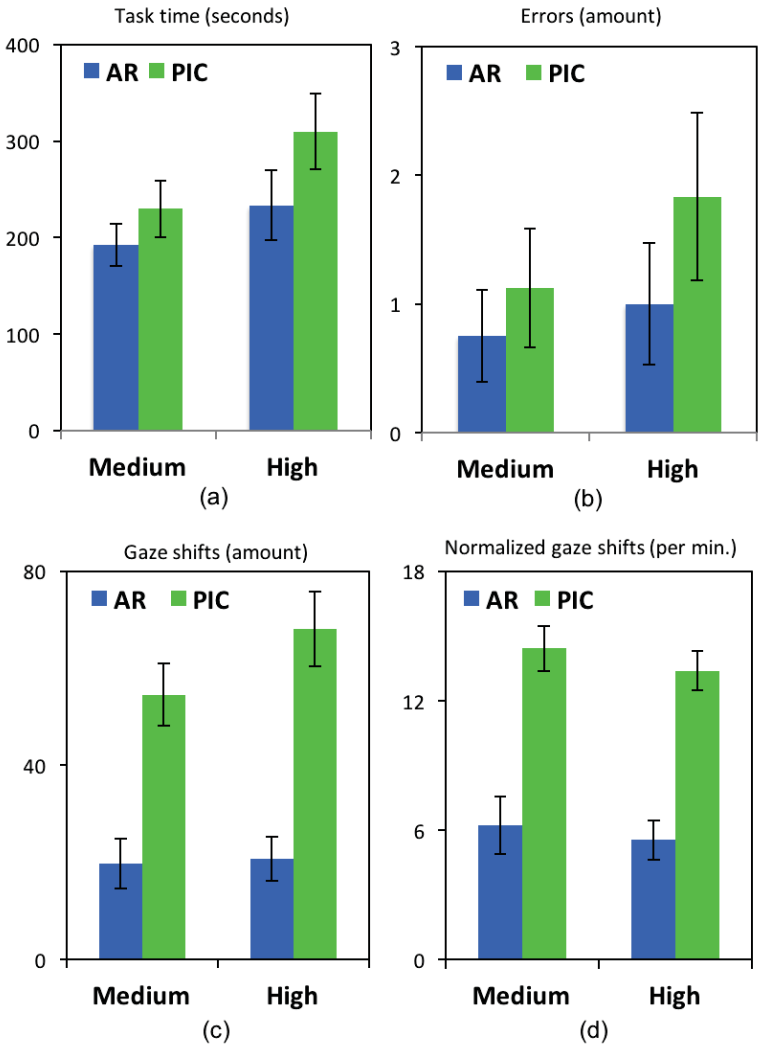

Figure 3.7. The results from the objective measurements: The average task completion times in seconds (a). The amount of errors (b). The amount of gaze shifts (c). The normalized amount of gaze shifts per minute (d). N = 24 and error bars = ±95% CI.

3.3.6.1 Task Completion Time

Figure 3.8(a) shows the average task completion times. The total times included all 20 targets. Participants started the timing manually and stopped it after they

Table 3.1. The results from the objective measurements. N = 24.

Inter-face	Task	Task time (seconds)		Errors (amount)		Gaze shifts (amount per min.)	
		Mean	SD	Mean	SD	Mean	SD
AR	Med.	193	52	0.75	0.84	6.22	3.15
AR	High	233	86	1	1.02	5.56	2.16
Pic	Med.	230	70	1.125	1.12	14.43	2.46
Pic	High	310	93	1.83	1.149	13.39	2.16

had completed the task. Tasks did not have a time limit. We noticed a significant difference between the interfaces in terms of overall task time from both tasks. A repeated-measure ANOVA showed that AR (Both tasks: M = 208, SD = 21.21) interface was significantly faster than the picture (Both tasks: M = 270, SD = 56.6) interface: $F(1, 23) = 13.162$, $p < .001$. Similarly and expectedly, we noticed a significant effect of the task difficulty on the completion time: $F(1, 23) = 14.81$, $p = .001$. We did not notice any significant interaction effect of interface × task. The results support the **H1**.

3.3.6.2 Amount of Errors

We calculated the amount of errors based on the participants' answers ('YES' or 'NO') to the questions in the annotations (Fig. 3.8b). The highest theoretical error amount would have been 20 if all answers would have been incorrect. We noticed a significant difference between the interfaces in terms of overall error amount from both tasks. A repeated-measure ANOVA showed that AR (Both tasks: M = 0.88, SD = 0.18) interface caused significantly less errors than the picture (Both tasks: M = 1.48, SD = 0.50) interface: $F(1, 24) = 6.279$, $p < .05$. Similarly and expectedly, we noticed a significant effect of the task difficulty on the completion time: $F(1, 23) = 4.613$, $p < .05$. We did not notice any significant interaction effect of interface × test task. The results do not support the null hypothesis **H2**, because it stated that the null hypothesis (no difference in errors) should be correct.

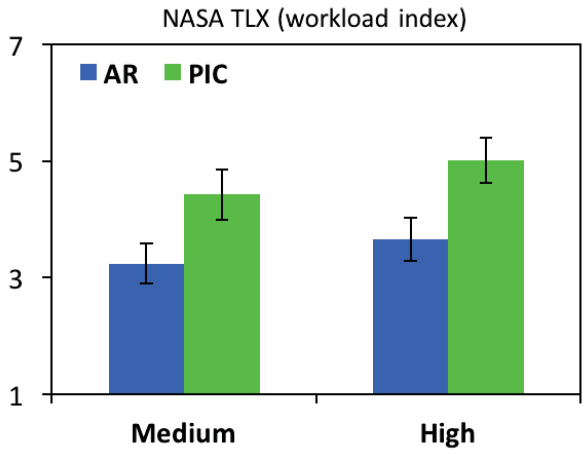

Figure 3.8. The results from the NASA TLX workload questionnaire. Results show a combination of the workload index. Lower is better. (* = significant at 0.05 level ** = significant at 0.01 level, *** = significant at 0.001 level). N = 24 and error bars = ±95% CI.

3.3.6.3 Gaze Shifts

Figure 3.8(c) shows the average amount gaze shifts per minute. A sequence where a participant switch eye focus from the device's display to a workpiece followed by eyes switch back to device's display was coded as one gaze shift. Similar counting manner has been used in the related work [73]. All single gaze shift events during a trial were aggregated into one value for each participant. Gaze shifts were calculated manually based on video recordings from an omni-directional camera that was placed in the middle of the task environment. We noticed a significant difference between the interfaces in terms of overall gaze shift amount from both tasks. A repeated-measure ANOVA showed that AR (M = 5.89, SD = 0.47) interface was significantly faster than picture (M = 13.91, SD = 0.74) interface: $F(1, 23) = 244.162$, $p < .001$. However, we did not find significant difference on the task difficulty nor on the interaction effect of interface × test task. The results support the **H3**.

Table 3.2. The NASA TLX questions

	Questions:
1	Mental demand: How much mental and perceptual activity was required?
2	Physical demand: How physically demanding was the task?
3	Temporal demand: How time-consuming was the task?
4	Performance: How successful were you in accomplishing the task?
5	Effort: How hard did you have to work (mentally and physically) to accomplish your level of performance?
6	Frustration: How insecure, discouraged, irritated, stressed or annoyed were you?

3.3.6.4 Subjective Feedback

We collected subjective feedback with the NASA TLX [3] questionnaire (Table 3.2 and written freeform comments. NASA TLX is a subjective workload assessment tool and it allows users to perform workload assessments on task support systems. Lower workload index is better. We also asked participants about their overall preference base on the usability attributes described by Nielsen [60]. We aggressed the 7-point Likert scale answers into one value for each participant. We analyzed the workload data using a repeated measures ANOVA. We noticed a significant difference between the interfaces in terms of the subjective workload index. A repeated-measure ANOVA showed that AR ($M = 3.45$, $SD = 0.29$) interface was significantly faster than picture ($M = 4.72$, $SD = 0.42$) method: $F(1, 23) = 46.928$, $p < .001$.

Expectedly, several participants saw that the greatest benefit the AR interface was the easy locating of targets by just following the red indication arrow. Participants also mentioned that it was often possible to conduct the inspection by looking through the screen. However, some participants said that sometimes inspection could not be done through the device's display due to low display resolution. The largest issues in the AR interface was seen to be the occasional lose of tracking, jittering of 3D registered annotations, and occlusion. The manipu-

[3] http://human-factors.arc.nasa.gov/groups/TLX/

lability was sometimes seen physically demanding because the device had to be held in a certain viewpoint in order to observe the augmented view. Participants commented that the picture interface was intuitive and easy to learn. Also, it caused less physical demand. It's largest drawback was the difficulty in locating the information from the test environment based on the non-registered annotations. Furthermore, few participants thought that it was easy to make mistakes with the picture interface because of high information density.

3.3.6.5 Observations

In the AR interface, the tracking was occasionally lost because participants moved the tablet too fast or when participants tried to zoom in by moving the device too close to targets in the task environment. If an annotation was off-screen, several participants tend to pan and move the device rather than rotating the viewpoint. This interaction was more liable for tracking losses. The tracking initialization was not always instant and sometimes participants had to do subtle viewpoint adjustments in order to get the tracking working. Few times participants perceived the location of the target incorrectly due to occlusion which caused them to answer incorrectly. Sometimes participants bumped to the task environment while moving because they were focusing only on the augmented view on device's display.

In the picture interface, high information density caused participants to occasionally inspect wrong targets within a section. For example, if several similar looking cables were in the close proximity from each other. Furthermore, few times participants inspected a wrong section. This happened more often in the high angles tasks some users inspected a wrong area of the workpiece due several similarities. Most of the participants rotated their own body according to the angle in the picture, but some participants rotated the handheld device in their hands in order to match the viewpoint in the picture with the task environment.

3.3.7 Summary

The user study showed that HAR can offer benefits over conventional non-AR interfaces in complex tasks. If the real world scenario requires align viewpoint

several times and map information from a dense environment, the use of AR could be justified.

3.4. Discussion

In this section we discuss about the findings and their implications from both visual observation studies separately.

3.4.1 First Study

The first study demonstrated that inspection tasks can be completed with HAR, although it still has some issues. The first study also included few physical manipulation subtasks in addition to visual observation, unlike the final comparative study that focused only on visual observation. Even though in this sense the studies are different, there are still many similarities and we assume that we can draw findings from both. Furthermore, the studies that included manipulation the required manipulation was simple and could always be conducted with one hand. We can also assume that manipulation affects non-AR interfaces too.

3.4.2 Second study

The discussion of the second study is divided into two sections. First sections interprets the findings based on the measurements of the user study. The later section focuses on the various aspects of the interfaces and tasks used in the study.

3.4.2.1 Measurements

We can identify two main reasons why AR interface was faster: Firstly, participants did not need to transport the information from the guidance medium (AR interface) to the task environment. Furthermore, the amount of divided attention is smaller because the targets could be inspected through the handheld representation of the real world. Comprehending and transforming was more demanding with the picture interface, especially because of high information density. However, we assume that the difference in time would get smaller when users are more

familiar with the task environment because they would have to spend less time on information transforming with the picture interface.

We expected there to be very total amount of errors in both interfaces because, because the targets to be inspected were very simple. Targets and questions were designed in a way that there should have been no possibility for misinterpretation regarding the target's status. Due to required mental mapping of the information, we expected the task to be more difficult with the Picture interface but the required mapping and the 3D mental rotation [77] should only affect task completion time.

Even though the amount of errors overall was very low (overall average: one per participant), the amount was higher than expected. The higher information density can be seen as the main cause for errors in both interfaces: in AR the lack of occlusion handling and information density caused participants to misinterpret to which targets the virtual annotations were indicating. In picture interface, even though participants were inspecting the correct section of the workpiece, they sometimes inspected a wrong target if it is surrounded by other similar types of targets. For example, if the target is a cable next to similar looking cables. We believe that these kind of errors are caused by the combination high information density and required angle alignments. Furthermore, if the workpiece section in the picture has enough similarities to other sections (the high angles tasks), some participants inspected the wrong section.

It is possible that in a task with no angle alignments, there would have been less errors with the picture interface because participants would not have had to do constant mental rotation or mapping. It must be noted that participants were not able to undo their answers so it also be possible that some of the answers were unintended in both interfaces. However, based on video observations we believe this amount the be very minimal.

Difference between the amount of gaze shifts within the two interfaces was expectedly very high. The main explanation for this is that AR interface allowed participants to observe the augmented representation of the task environment, making it possible to conduct inspections without looking at the real environment. Due to the complexity of the tasks, participants had to recheck information several times from the picture interface before finding the target in the work environment.

We did not objectively measure the time or effort used for mental rotation with the picture interface, but from the subjective data we can get implications that it caused difficulties for the participants.

Furthermore, often the gaze shifts with the picture interface were very quick, for example several times within few seconds. These quick shifts increased the total gaze amount with a fair amount. The average normalized amount of shifts was actually smaller in the high angles tasks with both interfaces. We assume that this is because users had to spend more time on moving from one angle to another. Even though the lower amount of gaze shifts can be seen as good because it affects to the mental workload [73, 58, 65], conducting the task without gaze shifts is not necessarily a desired outcome. Even though some targets can be observed trough the device's display, it is not always recommended or desired. For example, completely ignoring the real world might cause dangerous situation in factory inspection scenarios. These type of safety issues are one of the drawbacks of HMDs and they exist also when handhelds are used. The fact that participants sometimes bumped to the test tables supports this assumption.

Overall, the objective measurements clearly state the higher efficiency of an AR interface in our test scenario. AR also had a lower subjective workload index in both tasks, which means that the subjective results correlate to the objective measurements. One of the reasons why AR performs well in visual observation tasks is that users do not need to physically manipulate the environment. Manipulation while using a handheld device is generally troublesome, especially if it is necessary to keep a certain viewpoint with the device's camera. Thus, it is clear that physical manipulation only tasks are more suitable for HMD-based AR systems. However, we believe that HAR could be useful even in tasks that include physical manipulation if the interface would be improved.

3.4.2.2 Interfaces and Tasks

The interfaces we evaluated were very simple presentations of the two mediums. Even though the AR interface had an indication arrow and the picture did not, this was an essential feature. Without an indication arrow, the tasks would have been too difficult for AR. Both interfaces could be improved: For example, the AR interface could have visualized the optimal viewpoint to reduce the occlusion

related misunderstanding. Furthermore, it could have visualized the 3D distance to the next target and enable users to freeze the view if necessary. Also, SLAM systems suffer from one main drawback: They need to be initialized from a specific viewpoint. This was an issue in our system as well and it is a problem if a user has to move while conducting the inspection.

The picture interface could have shown an overview or an indication which section of the workpiece the picture is taken from. However, having an separate overview could have made the interface more complex to use. Nonetheless, we believe that having the bare minimum in terms of interface features and functionality, gave us the most generalizable results of the advantages and disadvantages of the two information representation mediums.

We did not measure learnability in this study, but based on observations during the tutorial tasks and comments gained from the participants, it takes time to get familiar with the tracking in the AR interface. Even though this was not a problem during the actual study because we had an intensive tutorial, it does tell us that a simple interface would not necessarily be ready for wider use without appropriate training or several interface improvements. Learnability is a crucial factor for the wider acceptance of HAR. The advantages of picture based interfaces is that they are really intuitive, because users area already familiar of viewing and manipulating pictures with handheld devices.

Our test task design tried to mimic inspection scenarios where several small near field targets [11] are inspected. In the medium angles task, the two sectors (Fig. 3.6, left) of the workpiece were distinguishable from one another, but the information density was high. In the picture interface, we assume that finding the correct sector was fairly easy and the most time consuming part was to find the correct target within the workpiece sector. The AR interface was faster in the medium angle than in the high angles task because of smaller amount of movement required. Furthermore, most of the time participants were standing approximately at the starting position so initializing the lost tracking was quick.

In the high angles task (Fig. 3.6, right), the four sectors were divided into pairs that had many similarities. In addition to higher requirement for movement, this can be seen as the reason for increased task time and errors in the picture interface. In the high angles task while using the picture interface, some of the

inspection errors were caused by participants inadvertently observing a wrong sector of the workpiece. The AR interface did not have this problem, because the information was overlaid directly to the target location. The increased task time for AR was because participants had to move more if tracking was failed.

The high angles task can be seen as a very special case in inspection, but it shows some of the benefits of an AR interface. However, it must be noted that if the environment looks exactly similar from more than one angle, also the SLAM tracking has difficulties in tracking it correctly. It must also be noted that the results might get more equal when users get more familiar with the environment. Now, the test participants were not given any time to observe at the environment before starting the test tasks. This can also be considered a benefit of AR, there is no extra time required for learning the task environment.

The user study was designed based on results from previous studies and it did not include tasks with no viewpoint alignment. Thus, we can say that even though AR interface was significantly better in both tasks in all measurements, it does not mean that picture interface is poor or badly designed. It means that in complex inspection tasks, the use of AR can be beneficial. Picture interface can still be useful, or even better than AR, in large variety of inspection tasks that were viewpoint alignment is not required or content density is low. As we stated earlier this study did not represent any real world scenario. Its purpose was to examine the possible benefits of HAR in more generic inspection scenario.

CHAPTER 4

Conclusion and Future Work

4.1. Review of the Thesis

In this thesis, we conducted four users studies related to inspection tasks. The studies focused on two areas in HAR: 3D positioning and visual observation. The 3D positioning evaluations confirmed the higher efficiency of a ray casting based SlidAR method against a conventional device-centric HoldAR method. The visual observation studies showed the benefits of an AR interface against a conventional picture interface. In both areas, we first conducted a small scale qualitative evaluations in a real world scenarios to learn more about the overall utility and usability. Then, we focused on specific areas in larger comparative studies.

4.2. Design Findings

In every study, we used a SLAM-based HAR system prototype. However, our findings can be applied to other types of AR systems and scenarios as well. We group the findings to 3D positioning in AR and AR in visual observation tasks.

4.2.1 3D positioning in HAR

When we talk about 3D positioning of virtual annotations, we can also consider this a task under a broader umbrella term: 3D pointing. 3D pointing is one of the basic manipulation tasks and it is necessary not only in AR, but also in VR. We believe that our findings can be applied to AR systems using other than SLAM tracking and to systems utilizing other types of AR display techniques.

- **Initial positioning.** In an optimal situation, there should be no need to adjust annotation's position. However, several errors can occur with the current hardware and tracking technologies because it is very difficult to track the environment completely accurately. Furthermore, if it is necessary to point into mid-air, we have to adjust the position even with accurate environment mapping.

- **Depth cues.** We found out that a ray casting based method allows the positioning to be conducted effectively without any additional depth cues visualized to the real environment. This is important, because we might have complex environments that make the accurate visualization of the depth cues impossible. The limitations of the used hardware might also affect on the quality of the depth cues.

- **Accuracy and effort.** The most substantial benefit of ray casting in 3D positioning is that it requires significantly less objective and subjective physical effort. It allow the 3D positioning to be conducted accurately with only a small amount of movement. Instead of a slide gesture, some other interaction metaphor could be used, depending on the used hardware and an interface. For example, a HMD system could allow the adjustment along the epipolar line with head movements.

4.2.2 HAR in Visual Observation Tasks

AR in task support is an umbrella term for HAR in visual observation tasks. Handheld devices have some handheld specific features, but some of our findings from the user studies could be expanded to other types of AR display techniques

as well because we did require complex information input or the observation of conventional static information.

- **Manipulation and observation.** The most substantial drawback of handheld devices is that the user does not have both hands free which can be a problem if physical manipulation of the environment is required. Nonetheless, if the task requires only visual observation, the use of HAR can be beneficial. The reasons to use handhelds instead of HMDs are the easier information input and shareablity. Real world observation scenarios often require users to conduct subtasks that cannot be done using only AR, manipulation and observation of conventional static information is also required.

- **Information mapping and task difficulty.** The main benefit of AR in any type of task support scenario is that it eases the workload required for mapping the virtual information to the real world. When the information is visualized directly to an object, less mental and physical effort is required to map virtual information to the real world. But this happens only in complex scenarios, the use of AR does not always decrease the workload. As we and the related work has shown, in simple scenarios AR can only add unnecessary complexity. Thus, use of non-AR picture or even text interfaces can be more suitable compared to AR in simple scenarios.

- **Manipulability.** The used hardware has a higher effect to the usability in AR interfaces compared to non-AR interfaces. For example, in scenarios that require physical manipulation of the real world, the use of HMDs can be more efficient instead of handhelds. Thus, we always need to consider the required amount of manipulability when we consider the usefulness of AR in a specific task. The required physical effort in task completion can increase greatly depending on the used hardware and system type (non-AR or AR).

- **Different tracking techniques.** In theory, other tracking methods could have been used in our test scenarios. like AR markers. These technologies would not be practical but still possible. We used SLAM and the findings

related to initialization and tracking can be considered SLAM specific to some extend. Like many other SLAM trackers, the SLAM tracking we used requires initialization from a specific viewpoint. This is a drawback of SLAM in scenarios where user is required move around and take different viewing angles, because he or she always have to return to the initial potion if tracking is lost.

4.3. Future Work

Several possible future work directions exist for our research. there are still many necessary interface improvements and user studies that need to be conducted before we could truly use off-the-shelf HAR in visual observation tasks. For the future work:

- General
    1. **Tracking.** Technical research was not in the scope of this research, but the tracking and tracking initialization have a considerable effect to the usability of HAR. Further development of SLAM tracking can substantially improve the usability of HAR in inspection and other goal-oriented tasks.
    2. **From the laboratory to the field.** We have conducted the comparative studies in a laboratory setting. After improving our interfaces, we should evaluate the system as a whole in real world scenarios in order to gain more insights.
- 3D positioning
    1. **Information visualization.** In SlidAR, the ray was visualized only as a 2D red line. We should investigate visualization techniques that are more easier to understand. For example, we could use 3D spheres instead of a 2D line. In a real world scenario, the visualization of the line should be easier to understand in order to decrease the learning time.

2. **Instructions and feedback.** We need to investigate better methods to instruct users how to change the viewpoint in order to perform the 3D position adjustment. Users should not rely on instructions given by a professional beforehand, instead the system should give them realtime feedback.

- Visual observation tasks

1. **Interface improvements**. Both AR and picture interfaces in the comparison were fairly simple and several improvements could be done to both. Better visualization of off-screen annotations, freeze mode and zooming could be added to the AR interface. Furthermore, AR interface should have realtime feedback about the tracking quality and how to initialize the tracking.
2. **Comprehensive studies**. We need to evaluate our system in comprehensive scenarios to get a broader understanding of the benefits and issues of HAR. We need include annotation positioning, more complex information input and possibly physical manipulation to the test scenarios.

Publication List

Journal Articles

1. **Jarkko Polvi**, Takafumi Taketomi, Goshiro Yamamoto, Arindam Dey, Christian Sandor and Hirokazu Kato. SlidAR: A 3D Positioning Method for SLAM-based Handheld Augmented Reality, *International Journal of Computers and Graphics*, 55(5): 33–43, 2015.

2. Marc Ericson C. Santos, **Jarkko Polvi**, Takafumi Taketomi, Goshiro Yamamoto, Christian Sandor and Hirokazu Kato. Towards Standard Usability Questionnaires for Handheld Augmented Reality, *IEEE Computer Graphics & Applications*, 35(5): 50–59, 2015.

Peer-Reviewed Conference Publications

1. Marc Ericson C. Santos, **Jarkko Polvi**, Takafumi Taketomi, Goshiro Yamamoto, Christian Sandor and Hirokazu Kato. A Usability Scale for Handheld Augmented Reality, In *Proceedings of ACM Symposium on Virtual Reality Software and Technology*, pp. 167–176, Edinburgh, United Kingdom, November, 2014.

2. **Jarkko Polvi**, Takafumi Taketomi, Goshiro Yamamoto, Mark Billinghurst,

Christian Sandor and Hirokazu Kato. Evaluating a SLAM-based Handheld Augmented Reality Guidance System, In *Proceedings of the 2nd ACM Symposium on Spatial User Interaction*, pp. 147–147, New York, NY, USA, October, 2014.

Other Conference and Workshop Publications or Presentations

1. Marc Ericson C. Santos, **Jarkko Polvi**, Takafumi Taketomi, Goshiro Yamamoto, Christian Sandor and Hirokazu Kato. On Usability Analytics and Beyond with Human-Centered Data Science, In *Workshop on Computer-Supported Cooperative Work and Social Computing*, San Fransico, USA, February, 2016.

2. **Jarkko Polvi**, Takafumi Taketomi, Goshiro Yamamoto, Christian Sandor and Hirokazu Kato. SlidAR: A 3D Positioning Technique for Handheld Augmented Reality, In *the 14th IEEE International Symposium on Mixed and Augmented Reality*, Fukuoka, Japan, October, 2015.

3. **Jarkko Polvi**, Takafumi Taketomi, Goshiro Yamamoto, Mark Billinghurst, Christian Sandor and Hirokazu Kato. Evaluating the Use of Handheld Augmented Reality Authoring and Guidance in Unknown Environments, In *USB Proceedings of Korea-Japan Workshop on Mixed Reality*, Seoul, South-Korea, April 2014.

4. **Jarkko Polvi**, Juhyun Kim, Takafumi Taketomi, Goshiro Yamamoto, Mark Billinghurst, Christian Sandor and Hirokazu Kato. User Interface Design of a SLAM-based Handheld Augmented Reality Work Support System, In *Virtual Reality Society of Japan Research Report*, Japan, September, 2013.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervising Professor Hirokazu Kato for accepting me as a doctoral student and patiently guiding me through the 3.5 year period of my doctoral studies. He made it possible for me to realize my dream of coming to Japan and working on something new and very existing - even if in the beginning I did not really have a concrete idea what I will do. Working with him allowed me to learn a lot about augmented reality and how to conduct research in general.

I want to thank all my other committee members. I am very grateful for the many things that I learned from our Associate Professor Christian Sandor. He gave me valuable knowledge, especially related to presentation skills and conducting user studies. I think that these skills are extremely valuable for me wherever I decide to go next. Special thanks go to my committee member and co-supervisor Assistant Professor Taketomi Takafumi. He helped me tremendously in every part of my research process, especially related to many technical problems that I encountered. The needed technical development for the systems in the thesis could not have been completed without his help. I am particularly grateful to Assistant Professor Goshiro Yamamoto who helped me beyond his assigned duties. I also would like to thank my thesis committee member Professor Naokazu Yokoya for giving me good feedback considering my thesis.

I want to extend my appreciation to all the current and past Interactive Media

Acknowledgments

Design laboratory members. My time at NAIST was made extremely enjoyable due to all the amazing and colorful individual that I met during my stay. The help and humor from my lab mates kept me going during these years. I would also like to thank our secretary, Makiko Ueno. She helped with many daily issues and made working in the much more lab easier.

My last and the most important gratitude goes to my parents and family back in Finland. They have supported me in everything that I do. Thank you so much. Kiitos!

Appendix

A. HAR Manipulation Study

An important factor in the overall usefulness of the system is to understand what are the benefits and disadvantages of AR compared to conventional guidance tools. The purpose of this manipulability study was to compare AR against two conventional interfaces on a handheld device.

A.1 Study Design

An assembly scenario with Lego blocks (Fig. 4.2) was used as a test scenario. We had a within-group 3×1 design with three interfaces as main independent variables. In the tasks, participants were asked to modify small block structures by following the instruction given by each guide interface. The use of Lego blocks minimizes the bias towards participants with experience related to a certain assembly tasks. Lego blocks also represent a general assembly task rather than an assembly in a specific scenario [78]. The task had 10 subtasks that consisted of assembly and disassembly of Lego blocks. All subtask were designed so that the tasks can be completed using only one hand. Because we had a within-group design, we constructed three equally difficult versions of the same task. The test task had low information density and did not require angle alignments.

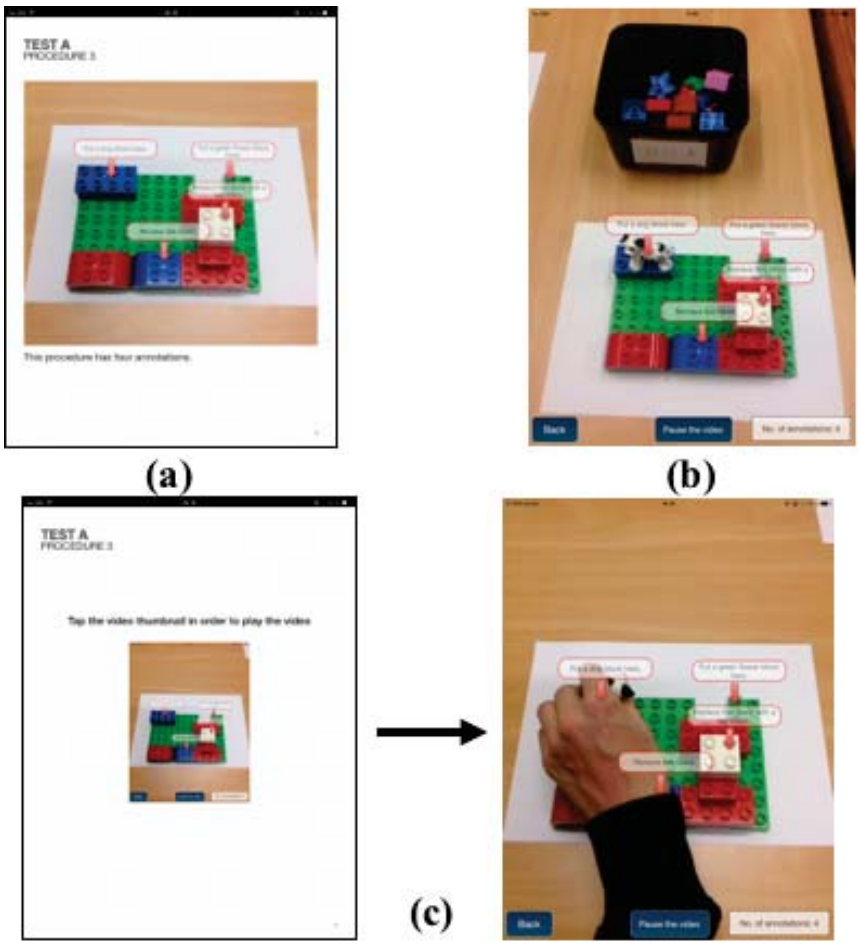

Figure 4.1. The interfaces used in the study: the picture interface (a), the AR interface (b), and two screenshots from the video interface. (c). After the thumbnail is tapped (c-left), an instruction video will be played (c-right).

The study had three interfaces: a picture guide, and AR guide, and a video guide (Fig. 4.2). The picture guide showed a 2D picture of each structure with annotations. The images in the picture guide were screenshots from the AR guide. The video guide had a video of each block structure showing a video how a subtask should be conducted. The videos captured for the video guide where from the AR guide as well. The picture and video guides were based on the AR guide in order to make the exact position and look of the annotations as similar

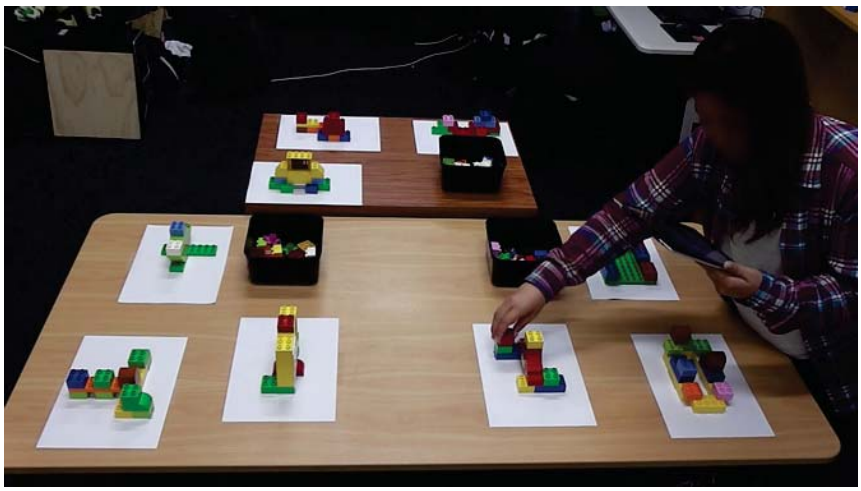

Figure 4.2. A user conducting the lego assembly study

as possible so that the differences would not affect the results.

The study had 27 test participants (17 males and 10 females) between 19 and 44 years old. The participants were asked about their experience with handheld devices in general (M = 4.6, SD = 2.1) and with HAR (M = 3.6, SD = 2.0) on a 7-point Likert scale. We used task performance (task completion time and error rate) as the objective measure and a questionnaire (ranking and freeform comments) as the subjective measurement. Ranking questions concerned learnability (Q1), easy of use (Q2), effectiveness (Q3), confidence (Q4), best indication to the real object (Q5), and overall preference (Q6). Before participants started the evaluation, they were given a change to practice in a tutorial scenario until they felt comfortable using the interface in question. The order of use for the interfaces and were mixed between test participants.

A.2 Results

We conducted a repeated measures ANOVA on all measurements. Average task completion times are shown in Figure 4.3. The picture interface was significantly faster than the video ($F(1, 26) = -4.60$, $p < .001$) and the AR interface ($F(1, 26) = -6.28$, $p < .001$). The total amount of errors between interfaces was very

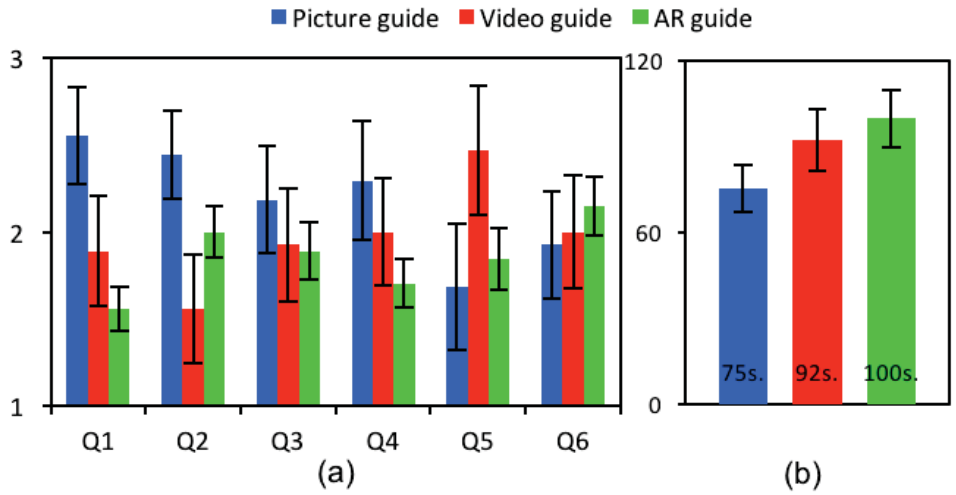

Figure 4.3. The Horizontal axis (Q1-Q6) represents the results of the ranking questions (a). The average task completion time in seconds (b). Error bars 95% CI.

low and there were no significant differences. We found significant differences from ranking (Fig. 5) in Q1, Q2, Q4, and Q5. Q1: the picture interface was significantly easier to learn than the video ($F(1, 26) = 2.550$, $p < .018$) and the AR interface ($F(1, 26) = 2.550$, $p < .001$). Q2: the picture interface was significantly easier to use than the video interface ($F(1, 26) = 4.253$, $p < .001$). Q4: test participants felt significantly more confident with the picture interface than with the AR interface ($F(1, 26) = 2.209$, $p < .037$). Q5: The video interface showed significantly better indications to the real object compared to the picture interface ($F(1, 18) = -2.616$, $p < .018$).

Several participants noted that the picture interface felt the most familiar and was very easy to use, because there was no interference from the onscreen movement. Some participants complained that it was difficult to see where annotations actually indicated due to the picture having only one viewpoint. The main benefit of the video interface was that it showed direct manipulation information and also made it possible to see the correct end condition (ranking Q5). Few participants mentioned that it was very difficult to follow the video and perform

the assembly at the same time (ranking Q2).

Participants said that the advantage of the AR interface was that the instructions could be viewed from different viewpoints and that it was easy to check the precise location by changing the viewpoint. It was also mentioned that the AR interface was easy to use because there was no need to split attention between the interface and the real world. The UI of the AR interface was said to be too complex and slow compared to the other two interfaces, and that too many steps were required to move from one task to another. Even though the amount of tracking error was almost non-existent, viewpoint alignment still took time within a task and between tasks. Several participants said that the AR interface should give feedback on a user's actions and show visual annotations instead of just plain text. View pausing was seen useful when both hands wanted to be used for the assembly.

A.3 Summary

We found that HAR can enable an intuitive inspection of the environment, but we were not able to find any benefits in the efficiency. We derived guidelines that will aid the designing of future AR guidance systems. In our evaluations, the tasks were fairly simple and the use of AR does not offer clear benefits if it is used solely as a substitute to conventional multimedia. However, the use of HAR in generic guidance can be justified if other benefits of HAR (information input, collaboration, etc.) are also made use of. Future work will see additional improvements made to our system based on the results of the comparative evaluation. We will then conduct another evaluation utilizing a more complex inspection scenario. Our qualitative findings from this study can be explained in the form of five design guidelines.

Location of the AR content: The system should instruct the user where AR content is located, particularly in situations when the environment itself does not provide enough physical cues (e.g. easily recognizable AR-markers) to make the location of AR content apparent. The system should inform about the approximate location of the AR content in the environment as well as with more precise information for correct viewpoint.

The Off-screen AR content: Even after AR content has been discovered in the

environment, it is still possible that the user might accidentally skip annotations not in the immediate FOV of the display. The system should indicate the amount of AR content contained in the environment or visually indicate the existence of content that is not in the immediate FOV.

View pausing: Pausing the current view of the AR environment is important in several situations: When a real object must be examined from a distance too close to make viewing it through the AR display practical. It is especially important when physical manipulation of the environment is required, with one hand or with both hands.

Navigational shortcuts: The significant differences in learnability (Q1), confidence (Q4), and in task time are related to the complexity of the AR interface. Switching from one task to another was time consuming, since it was necessary to return to the overview image, choose another area, and align the viewpoint again. If the user wishes to move from one area of AR content to another, and already knows the approximate location of the new area, the system should supply navigational shortcuts to quickly move between these areas without leaving the AR view.

Feedback: In our prototype, the AR annotations were static and did not react to changes in the real environment. The annotations were displayed even after the task was completed, which was considered confusing by the users. Even though the task contained several areas which were in procedural order, inside an area all the annotations were displayed at the same time. The system should provide a form of interactive feedback related to users' actions in the real world, for example, by removing the 'completed' annotations.

Bibliography

[1] Ronald Azuma. Tracking requirements for augmented reality. *Commun. ACM*, 36(7):50–51, July 1993.

[2] Ronald Azuma. A survey of augmented reality, 1997.

[3] Huidong Bai, Gun A. Lee, and Mark Billinghurst. Freeze view touch and finger gesture based interaction methods for handheld augmented reality interfaces. In *Proceedings of the 27th Conference on Image and Vision Computing New Zealand*, IVCNZ '12, pages 126–131, New York, NY, USA, 2012. ACM.

[4] Evan Barba, Blair MacIntyre, Rebecca Rouse, and Jay Bolter. Thinking inside the box: Making meaning in a handheld ar experience. In *Mixed and Augmented Reality - Arts, Media, and Humanities (ISMAR-AMH), 2010 IEEE International Symposium On*, pages 19–26, Oct 2010.

[5] Mark Billinghurst, Adrian Clark, and Gun Lee. A survey of augmented reality. *Found. Trends Hum.-Comput. Interact.*, 8(2-3):73–272, March 2015.

[6] Oliver Bimber and Ramesh Raskar. Modern approaches to augmented reality. In *ACM SIGGRAPH 2006 Courses*, SIGGRAPH '06, New York, NY, USA, 2006. ACM.

Bibliography

[7] Doug A. Bowman, Ernst Kruijff, Joseph J. LaViola, and Ivan Poupyrev. *3D User Interfaces: Theory and Practice*. Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA, 2004.

[8] Pished Bunnun and Walterio W. Mayol-Cuevas. Outlinar: An assisted interactive model building system with reduced computational effort. In *Proceedings of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, ISMAR '08, pages 61–64, Washington, DC, USA, 2008. IEEE Computer Society.

[9] Stefano Burigat, Luca Chittaro, and Silvia 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*, MobileHCI '06, pages 239–246, New York, NY, USA, 2006. ACM.

[10] Robert Castle, Georg Klein, and David Murray. Video-rate localization in multiple maps for wearable augmented reality. In *Wearable Computers, 2008. ISWC 2008. 12th IEEE International Symposium on*, pages 15–22, Sept 2008.

[11] James E. Cutting. Reconceiving perceptual space. *Looking into pictures: An interdisciplinary approach to pictorial space*, 11:215–238, 2003.

[12] Andew J. Davison, Ian D. Reid, Nicholas D. Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):1052–1067, June 2007.

[13] Arindam Dey, Graeme Jarvis, Christian Sandor, and Gerhard Reitmayr. Tablet versus phone: Depth perception in handheld augmented reality. In *Mixed and Augmented Reality (ISMAR), 2012 IEEE International Symposium on*, pages 187–196, Nov 2012.

[14] Arindam Dey and Christian Sandor. Lessons learned: Evaluating visualizations for occluded objects in handheld augmented reality. *Int. J. Hum.-Comput. Stud.*, 72(10-11):704–716, October 2014.

[15] Andreas Dünser, Mark Billinghurst, James Wen, Ville Lehtinen, and Antti Nurminen. Exploring the use of handheld ar for outdoor navigation. *Comput. Graph.*, 36(8):1084–1095, December 2012.

[16] Hitoshi Furata, Kyosuke Takahashi, Koichiro Nakatsu, Ken Ishibashi, and Mami Aira. A mobile application system for sightseeing guidance using augmented reality. In *Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), 2012 Joint 6th International Conference on*, pages 1903–1906, Nov 2012.

[17] Joseph. L Gabbard and Deborah Hix. Researching usability design and evaluation guidelines for augmented reality (ar) systems, 2001. Last checked: 2015-03-10.

[18] Steffen Gauglitz, Cha Lee, Matthew Turk, and Tobias 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*, MobileHCI '12, pages 241–250, New York, NY, USA, 2012. ACM.

[19] Steffen Gauglitz, Benjamin Nuernberger, Matthew Turk, and Tobias 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*, UIST '14, pages 449–459, New York, NY, USA, 2014. ACM.

[20] Michael Gervautz and Dieter Schmalstieg. Anywhere interfaces using handheld augmented reality. *Computer*, 45(7):26–31, July 2012.

[21] Tor Gjosater. A taxonomy of handheld augmented reality applications. In *Intelligent Human Computer Interaction (IHCI), 2012 4th International Conference on*, pages 1–6, Dec 2012.

[22] Jens Grubert, Tobias Langlotz, and Raphaël Grasset. Augmented reality browser survey. Technical report, University of Technology Graz, 2011.

[23] Sinem Guven, Steven Feiner, and Ohan Oda. Mobile augmented reality interaction techniques for authoring situated media on-site. In *Proceedings of the 5th IEEE and ACM International Symposium on Mixed and Augmented Reality*, ISMAR '06, pages 235–236, Washington, DC, USA, 2006. IEEE Computer Society.

[24] Mika Hakkarainen, Charles Woodward, and Mark Billinghurst. Augmented assembly using a mobile phone. In *Mixed and Augmented Reality, 2008. ISMAR 2008. 7th IEEE/ACM International Symposium on*, pages 167–168, Sept 2008.

[25] Mark Hancock, Sheelagh Carpendale, and Andy Cockburn. Shallow-depth 3d interaction: Design and evaluation of one-, two- and three-touch techniques. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 1147–1156, New York, NY, USA, 2007. ACM.

[26] Marc Hassenzahl. Funology. chapter The Thing and I: Understanding the Relationship Between User and Product, pages 31–42. Kluwer Academic Publishers, Norwell, MA, USA, 2004.

[27] Steven Henderson and Steven Feiner. Evaluating the benefits of augmented reality for task localization in maintenance of an armored personnel carrier turret. In *Mixed and Augmented Reality, 2009. ISMAR 2009. 8th IEEE International Symposium on*, pages 135–144, Oct 2009.

[28] Steven Henderson and Steven Feiner. Augmented reality in the psychomotor phase of a procedural task. In *Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality*, ISMAR '11, pages 191–200, Washington, DC, USA, 2011. IEEE Computer Society.

[29] Anders Henrysson, Mark Billinghurst, and Mark Ollila. Face to face collaborative ar on mobile phones. In *Mixed and Augmented Reality, 2005. Proceedings. Fourth IEEE and ACM International Symposium on*, pages 80–89, Oct 2005.

[30] Anders Henrysson, Mark Billinghurst, and Mark Ollila. Virtual object manipulation using a mobile phone. In *Proceedings of the 2005 International*

Conference on Augmented Tele-existence, ICAT '05, pages 164–171, New York, NY, USA, 2005. ACM.

[31] Anders Henrysson, Joe Marshall, and Mark Billinghurst. Experiments in 3d interaction for mobile phone ar. In *Proceedings of the 5th International Conference on Computer Graphics and Interactive Techniques in Australia and Southeast Asia*, GRAPHITE '07, pages 187–194, New York, NY, USA, 2007. ACM.

[32] Anders Henrysson, Mark Ollila, and Mark Billinghurst. Mobile phone based ar scene assembly. In *Proceedings of the 4th International Conference on Mobile and Ubiquitous Multimedia*, MUM '05, pages 95–102, New York, NY, USA, 2005. ACM.

[33] Niels Henze and Susanne Boll. Push the study to the app store: evaluating off-screen visualizations for maps in the android market. In *Proceedings of the 12th Conference on Human-Computer Interaction with Mobile Devices and Services, Mobile HCI 2010, Lisbon, Portugal, September 7-10, 2010*, pages 373–374, 2010.

[34] Wolfgang Hürst and Casper van Wezel. Gesture-based interaction via finger tracking for mobile augmented reality. *Multimedia Tools and Applications*, 62(1):233–258, January 2013.

[35] Hirotake Ishii, Zhiqiang Bian, Hidenori Fujino, Tomoki Sekiyama, Toshinori Nakai, Akihisa Okamoto, and Hiroshi Shimoda. Augmented reality applications for nuclear power plant maintenance work. In *ï£ijï£ijï£ijn Proceedings of International Symposium on Symbiotic Nuclear Power Systems*, pages 262–268, 2007.

[36] Eunsoo Jung, Sujin Oh, and Yanghee Nam. Handheld ar indoor guidance system using vision technique. In *Proceedings of the 2007 ACM Symposium on Virtual Reality Software and Technology*, VRST '07, pages 47–50, New York, NY, USA, 2007. ACM.

[37] Jinki Jung, Jihye Hong, Sungheon Park, and Hyun S. Yang. Smartphone as an augmented reality authoring tool via multi-touch based 3d interaction

method. In *Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry*, VRCAI '12, pages 17–20, New York, NY, USA, 2012. ACM.

[38] Nils Karlsson, Gang Li, Yakup Genc, Angela Huenerfauth, and Elizabeth Bononno. iar: An exploratory augmented reality system for mobile devices. In *Proceedings of the 18th ACM Symposium on Virtual Reality Software and Technology*, VRST '12, pages 33–40, New York, NY, USA, 2012. ACM.

[39] Shunichi Kasahara, Valentin Heun, Austin S. Lee, and Hiroshi Ishii. Second surface: Multi-user spatial collaboration system based on augmented reality. In *SIGGRAPH Asia 2012 Emerging Technologies*, SA '12, pages 20:1–20:4, New York, NY, USA, 2012. ACM.

[40] Hirokazu Kato and Mark Billinghurst. Marker tracking and hmd calibration for a video-based augmented reality conferencing system. In *Augmented Reality, 1999.(IWAR'99) Proceedings. 2nd IEEE and ACM International Workshop on*, pages 85–94. IEEE, 1999.

[41] Georg Klein and David Murray. Parallel tracking and mapping for small ar workspaces. In *Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, ISMAR '07, pages 1–10, Washington, DC, USA, 2007. IEEE Computer Society.

[42] Georg Klein and David Murray. Parallel tracking and mapping on a camera phone. In *Mixed and Augmented Reality, 2009. ISMAR 2009. 8th IEEE International Symposium on*, pages 83–86, Oct 2009.

[43] Ernst. Kruijff, J.Edward Swan, and Steven Feiner. Perceptual issues in augmented reality revisited. In *Mixed and Augmented Reality (ISMAR), 2010 9th IEEE International Symposium on*, pages 3–12, Oct 2010.

[44] Stan Kurkovsky, Ranjana Koshy, Vivian Novak, and Peter Szul. Current issues in handheld augmented reality. In *Communications and Information Technology (ICCIT), 2012 International Conference on*, pages 68–72, June 2012.

[45] Tobias Langlotz. *AR 2.0: Social Media in Mobile Augmented Reality*. PhD thesis, University of Technology Graz, 2013.

[46] Tobias Langlotz, Stefan Mooslechner, Stefanie Zollmann, Claus Degendorfer, Gerhard Reitmayr, and Dieter Schmalstieg. Sketching up the world: In situ authoring for mobile augmented reality. *Personal Ubiquitous Comput.*, 16(6):623–630, August 2012.

[47] Gun A. Lee, Ungyeon Yang, Yongwan Kim, Dongsik Jo, Ki-Hong Kim, Jae Ha Kim, and Jin Sung Choi. Freeze-set-go interaction method for hand-held mobile augmented reality environments. In *Proceedings of the 16th ACM Symposium on Virtual Reality Software and Technology*, VRST '09, pages 143–146, New York, NY, USA, 2009. ACM.

[48] Can Liu, Stephane Huot, Jonathan Diehl, Wendy Mackay, and Michel 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*, CHI '12, pages 2973–2976, New York, NY, USA, 2012. ACM.

[49] Mark Livingston, Arindam Dey, Christian Sandor, and BruceH. Thomas. Pursuit of "x-ray vision" for augmented reality. In Weidong Huang, Leila Alem, and Mark A. Livingston, editors, *Human Factors in Augmented Reality Environments*, pages 67–107. Springer New York, 2013.

[50] Anthony Martinet, Gery Casiez, and Laurent Grisoni. Integrality and separability of multitouch interaction techniques in 3d manipulation tasks. *IEEE Transactions on Visualization and Computer Graphics*, 18(3):369–380, March 2012.

[51] Asier Marzo, Benoît Bossavit, and Martin Hachet. Combining multi-touch input and device movement for 3d manipulations in mobile augmented reality environments. In *Proceedings of the 2Nd ACM Symposium on Spatial User Interaction*, SUI '14, pages 13–16, New York, NY, USA, 2014. ACM.

[52] Paul Milgram, Haruo Takemura, Akira Utsumi, and Fumio Kishino. Augmented reality: A class of displays on the reality-virtuality continuum. pages 282–292, 1994.

[53] Matias Mohring, Christian Lessig, and Oliver Bimber. Video see-through ar on consumer cell-phones. In *Mixed and Augmented Reality, 2004. ISMAR 2004. Third IEEE and ACM International Symposium on*, pages 252–253, Nov 2004.

[54] Annette Mossel, Benjamin Venditti, and Hannes Kaufmann. 3dtouch and homer-s: Intuitive manipulation techniques for one-handed handheld augmented reality. In *Proceedings of the Virtual Reality International Conference: Laval Virtual*, VRIC '13, pages 12:1–12:10, New York, NY, USA, 2013. ACM.

[55] Tobias Muller. *Augmented and Virtual Reality: Second International Conference, AVR 2015, Lecce, Italy, August 31 - September 3, 2015, Proceedings*, chapter Towards a Framework for Information Presentation in Augmented Reality for the Support of Procedural Tasks, pages 490–497. Springer International Publishing, Cham, 2015.

[56] Alessandro Mulloni, Mahesh Ramachandran, Gerhard Reitmayr, Daniel Wagner, Raphaël Grasset, and Serafin Diaz. User friendly slam initialization. In *Mixed and Augmented Reality (ISMAR), 2013 IEEE International Symposium on*, pages 153–162, Oct 2013.

[57] Alessandro Mulloni, Hartmut Seichter, and D. Schmalstieg. User experiences with augmented reality aided navigation on phones. In *Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on*, pages 229–230, Oct 2011.

[58] Ulrich Neumann and Anthony Majoros. Cognitive, performance, and systems issues for augmented reality applications in manufacturing and maintenance. In *Virtual Reality Annual International Symposium, 1998. Proceedings., IEEE 1998*, pages 4–11, 1998.

[59] Joseph Newman, David Ingram, and Andy Hopper. Augmented reality in a wide area sentient environment. In *Augmented Reality, 2001. Proceedings. IEEE and ACM International Symposium on*, pages 77–86, 2001.

[60] Jakob Nielsen. *Usability Engineering*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[61] Susanna Nilsson, Björn J. E. Johansson, and Arne Jönsson. A co-located collaborative augmented reality application. In *Proceedings of the 8th International Conference on Virtual Reality Continuum and Its Applications in Industry*, VRCAI '09, pages 179–184, New York, NY, USA, 2009. ACM.

[62] Jennifer Ockerman and Amy Pritchett. Preliminary investigation of wearable computers for task guidance in aircraft inspection. In *Proceedings of the 2Nd IEEE International Symposium on Wearable Computers*, ISWC '98, pages 33–, Washington, DC, USA, 1998. IEEE Computer Society.

[63] Thomas Olsson. *User Expectations and Experiences of Mobile Augmented Reality Services*. Tampereen teknillinen yliopisto. Julkaisu - Tampere University of Technology. Publication;1085. 11 2012.

[64] Thomas Olsson and Markus Salo. Online user survey on current mobile augmented reality applications. In *Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on*, pages 75–84, Oct 2011.

[65] James Ott. Maintenance executives seek greater efficiency. *Aviation Week and Space Technology*, 142(20):43, 1995.

[66] Juri Platonov, Hauke Heibel, Peter Meier, and Bert Grollmann. A mobile markerless ar system for maintenance and repair. In *Mixed and Augmented Reality, 2006. ISMAR 2006. IEEE/ACM International Symposium on*, pages 105–108, Oct 2006.

[67] Jarkko Polvi, Takafumi Taketomi, Goshiro Yamamoto, Mark Billinghurst, Christian Sandor, and Hirokazu Kato. Evaluating a slam-based handheld augmented reality guidance system. In *Proceedings of the 2Nd ACM Symposium on Spatial User Interaction*, SUI '14, pages 147–147, New York, NY, USA, 2014. ACM.

[68] Malinda Rauhala, Ann-Sofie Gunnarsson, and Anders 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*, MobileHCI '06, pages 145–148, New York, NY, USA, 2006. ACM.

[69] Gerhard Reitmayr and Tom W. Drummond. Initialisation for visual tracking in urban environments. In *Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on*, pages 161–172, Nov 2007.

[70] Gerhard Reitmayr, Ethan Eade, and Tom Drummond. Semi-automatic annotations in unknown environments. In *Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, ISMAR '07, pages 1–4, Washington, DC, USA, 2007. IEEE Computer Society.

[71] Gerhard Reitmayr and Dieter Schmalstieg. Collaborative augmented reality for outdoor navigation and information browsing. In *Proceedings of the Second Symposium on Location Based Services and TeleCartography*, pages 53–62. TU Wien, 2004.

[72] Jun Rekimoto and Katashi Nagao. The world through the computer: Computer augmented interaction with real world environments. In *Proceedings of the 8th Annual ACM Symposium on User Interface and Software Technology*, UIST '95, pages 29–36, New York, NY, USA, 1995. ACM.

[73] Michael Rohs, Robert Schleicher, Johannes Schöning, Georg Essl, Anja Naumann, and Antonio Krüger. Impact of item density on the utility of visual context in magic lens interactions. *Personal and Ubiquitous Computing*, 13(8):633–646, 2009.

[74] Lawrence Sambrooks and Brett Wilkinson. Handheld augmented reality: Does size matter? In S. Marks and R. Blagojevic, editors, *16th Australasian User Interface Conference (AUIC 2015)*, volume 162 of *CRPIT*, pages 11–20, Sydney, Australia, 2015. ACS.

[75] Marc Ericson C. Santos, Takafumi Taketomi, Christian Sandor, Jarkko Polvi, Goshiro Yamamoto, and Hirokazu Kato. A usability scale for handheld augmented reality. In *Proceedings of the 20th ACM Symposium on Virtual Reality Software and Technology*, VRST '14, pages 167–176, New York, NY, USA, 2014. ACM.

[76] Christoph Schlieder. Winspect: A case study for wearable computing- supported inspection tasks, 2001.

[77] Roger N. Shepard and Jacqueline Metzler. Mental rotation of Three-Dimensional objects. *Science*, 171(3972):701–703, February 1971.

[78] Arthur Tang, Charles Owen, Frank Biocca, and Weimin Mou. Comparative effectiveness of augmented reality in object assembly. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pages 73–80, New York, NY, USA, 2003. ACM.

[79] Can Telkenaroglu and Tolga Capin. Dual-finger 3d interaction techniques for mobile devices. *Personal Ubiquitous Comput.*, 17(7):1551–1572, October 2013.

[80] Philipp Tiefenbacher, Andreas Pflaum, and Gerhard Rigoll. Touch gestures for improved 3d object manipulation in mobile augmented reality. In *Mixed and Augmented Reality (ISMAR), 2014 IEEE International Symposium on*, pages 315–316, Sept 2014.

[81] Marjaana Träskbäack and Michael 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*, VRCAI '04, pages 324–327, New York, NY, USA, 2004. ACM.

[82] Tuomas Vaittinen, Tuula Karkkainen, and Thomas Olsson. A diary study on annotating locations with mixed reality information. In *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*, MUM '10, pages 21:1–21:10, New York, NY, USA, 2010. ACM.

[83] D. W. F. (Rick) van Krevelen and Ronald Poelman. A survey of augmented reality technologies, applications and limitations. *International Journal of Virtual Reality*, 9(2):1–20, June 2010.

[84] Thomas Vincent, Laurence Nigay, and Takeshi Kurata. Precise pointing techniques for handheld augmented reality. In Paula Kotzé, Gary Marsden, Gitte Lindgaard, Janet Wesson, and Marco Winckler, editors, *INTERACT (1)*, volume 8117 of *Lecture Notes in Computer Science*, pages 122–139. Springer, 2013.

[85] Vassilios Vlahakis, John Karigiannis, Manolis Tsotros, Michael Gounaris, Luis Almeida, Didier Stricker, Tim Gleue, Ioannis T. Christou, Renzo Carlucci, and Nikos Ioannidis. Archeoguide: First results of an augmented reality, mobile computing system in cultural heritage sites. In *Proceedings of the 2001 Conference on Virtual Reality, Archeology, and Cultural Heritage*, VAST '01, pages 131–140, New York, NY, USA, 2001. ACM.

[86] Daniel Wagner, Gerhard Reitmayr, Alessandro Mulloni, Tom Drummond, and D. Schmalstieg. Pose tracking from natural features on mobile phones. In *Mixed and Augmented Reality, 2008. ISMAR 2008. 7th IEEE/ACM International Symposium on*, pages 125–134, Sept 2008.

[87] Daniel Wagner and Dieter Schmalstieg. First steps towards handheld augmented reality. In *Wearable Computers, 2003. Proceedings. Seventh IEEE International Symposium on*, pages 127–135, Oct 2003.

[88] Daniel Wagner and Dieter Schmalstieg. History and future of tracking for mobile phone augmented reality. In *Ubiquitous Virtual Reality, 2009. ISUVR '09. International Symposium on*, pages 7–10, July 2009.

[89] Jason Wither, Chris Coffin, Jonathan Ventura, and Tobias Hollerer. Fast annotation and modeling with a single-point laser range finder. In *Mixed and Augmented Reality, 2008. ISMAR 2008. 7th IEEE/ACM International Symposium on*, pages 65–68, Sept 2008.

[90] Jason Wither, Stephen Diverdi, and Tobias Höllerer. Using aerial photographs for improved mobile ar annotation. In *Proceedings of the 5th IEEE*

and ACM International Symposium on Mixed and Augmented Reality, ISMAR '06, pages 159–162, Washington, DC, USA, 2006. IEEE Computer Society.

[91] Jason Wither, Stephen DiVerdi, and Tobias Höllerer. Technical section: Annotation in outdoor augmented reality. *Computers and Graphics*, 33(6):679–689, December 2009.

[92] Wolfgang Wohlgemuth and Gunthard Triebfürst. Arvika: Augmented reality for development, production and service. In *Proceedings of DARE 2000 on Designing Augmented Reality Environments*, DARE '00, pages 151–152, New York, NY, USA, 2000. ACM.