How Wrong Can You Be: Perception of Static Orientation Errors in Mixed Reality

Jacob B. Madsen* Rasmus Stenholt†

Dept. of Architecture, Design, and Media Techonology
Aalborg University, Denmark

ABSTRACT

Tracking technologies are becoming an affordable commodity due to the wide use in mobile devices today. However, all tracking technologies available in commodity hardware is error prone due to problems such as drift, latency and jitter. The current understanding of human perception of static tracking errors is limited. This information about human perception might be useful in designing tracking systems for the display of AR and VR scenarios on commodity hardware. In this paper we present the findings of a study on the human perception of static orientation errors in a tracking system, using two different setups leveraging a handheld viewfinder: a classical augmented scenario and an indirect augmented one. By categorizing static orientation errors by scenario and local orientation axis, new insights into the users' ability to register orientational errors in the system are found. Our results show that users are much more aware of errors in classical AR scenarios in comparison to indirect AR scenarios. For both scenarios, the users registered roll orientation errors differently from both pitch and yaw orientation errors, and pitch and yaw perception is highly dependent on the scenario. However, the users performance ranking for orientational errors in AR scenarios was unexpected.

Keywords: Augmented reality, perception, tracking errors.

Index Terms: H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Artificial, augmented, and virtual realities; H.5.2 [Information Interfaces and Presentation]: User Interfaces—Evaluation/Methodology

1 INTRODUCTION

As smartphones and tablets are becoming commodity hardware, equipped with numerous sensors, such devices have rapidly become an attractive platform for applications augmenting our everyday lives, ranging from GPS navigation to mixed reality (MR) gaming and location based browsers. One of the principal goals of most mixed reality applications is to deliver a convincing experience to the user in merging the real and virtual worlds. Errors in the tracking system are often one of the major factors in diminishing the overall perception or sense of presence of the experience for the user. These tracking performance errors can be split into two categories, dynamic errors such as measurement noise and jitter and static errors such as spatial distortion, calibration errors, and stability errors such as slowly accumulated drift. These are persistent problems in tracking systems, which can limit the usability of mixed reality applications. Therefore, researchers are working hard to overcome these problems [2, 3, 13, 23].

While there is no single best solution for motion tracking on mobile smart devices in unprepared environments, a possible design goal for any tracking system, according to Welch and Foxlin [20] is that: "Tracking artifacts remain below the detection threshold of a user looking for them." However, to the best of our knowledge, the field of human perception of tracking errors is still largely unexplored. Swan and Gabbard survey user-based experimentation within augmented reality (AR) [18], while mentioning nothing of studies or reports on user evaluation of current tracking or estimation accuracy. Also in a review by Zhou et al. [23], several studies with research in tracking techniques as well as hybrid tracking systems are surveyed. However, nothing on evaluating human perception of tracking accuracy is mentioned within these surveys of the field.

We present a study of human perception of static orientation tracking errors in video see-through augmented reality (classical AR) and indirect augmented reality (indirect AR), with an experiment designed to uncover the lower boundaries for human registration of tracking errors, specifically static orientation errors, in both a classical AR and an indirect AR setup. As any of the problems inherent to tracking systems affect this study as well, we present an experimental setup that to the best of our abilities attempts to overcome these problems in a laboratory setting.

In this paper, the term indirect AR is used to indicate the presentation of a purely virtual scene in a physical setting that matches the displayed virtual representation from some viewpoint, as defined by Wither et al in [21]. They present a system for displaying pre-captured panoramas to the user instead of the real camera feed. With this "indirect augmentation", it is only a convincing augmentation when viewed at its corresponding real location. In some cases, indirect AR is also known as a situated simulation [15].

Human perception of orientation tracking errors in a tracking system for commodity hardware may be useful in guiding the design of tracking systems for classical and indirect AR purposes. Knowing these boundaries might even relax the demands on the tracking system for some applications.

This paper is organized into the following sections: In Section 2, an overview of related works is given. In Section 3, the hypotheses that formed the basis of the experiments are presented and motivated. This is followed up by a detailed description of the main experiment and its follow-up in Sections 4 and 5. After this, the results of both experiments are presented together in Section 6 along with a discussion of the significance of these results. Finally, a conclusion is presented in Section 7.

2 RELATED WORKS

Surveys of MR indicate that, in general, tracking systems are a broad and well-researched field, and motion tracking is a hard problem with no fixed solution for all cases, as examined by Welch and Foxlin, who explain the problems of motion tracking in great detail in [20]. This is also adressed by Azuma [2, 3] and van Krevelen [13], both of whom state that this is a complex problem.

Multiple studies have evaluated and analyzed tracking performance from a technical perspective [4, 10, 11, 12]. E.g. Gilson et al. [10] performed a quantitative analysis of tracking systems and found that tracking performance deteriorates when the tracked ob-

*e-mail: jbm@create.aau.dk
†e-mail: rs@create.aau.dk

IEEE Symposium on 3D User Interfaces 2014
29 - 30 March, Minneapolis, Minnesota, USA
978-1-4799-3624-3/14/$31.00 ©2014 IEEE

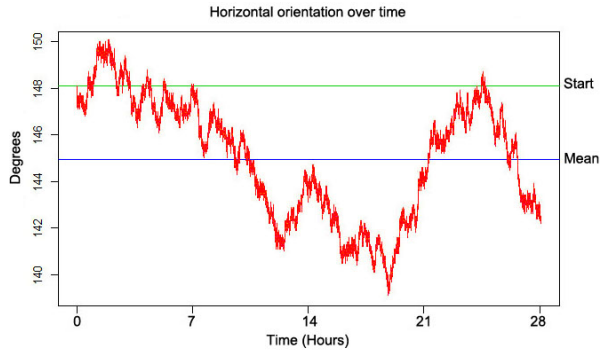

Figure 1: Illustration of yaw-drift from an iPad3's gyro, where the device is standing still over the course of 28 hours.

ject's speed increases. Others have attempted to estimate and fix these errors, such as Caarls et al. [6] who presents a framework for leveraging multiple sensors simultaneously to achieve a more precise and robust tracking system, providing 1 cm z-accuracy for distances up to 120 cm and small roll errors at distances under 70 cm.

In non-optical tracking systems for indirect AR or virtual reality (VR), inertial sensors are in many cases used for orientation estimation, where 3 degrees-of-freedom (DoF) tracking is sufficient, with examples being situated simulations for tablets/phones or VR environments and games for modern head-mounted displays (HMD) such as the Oculus Rift [1]. As depicted in Figure 1, the gyroscope in consumer products is prone to drift over time. Thus, many researchers have focused on minimizing this problem [5, 22]. One example is Won et al. [22] who presents a tilt angle correction method for handheld devices that detects if the system is stationary, and uses the gravity vector to stabilize and correct the yaw component, whereas the roll and pitch angle change in relation to the acceleration values. For dynamic movements, the tilt angles are corrected accurately, but the yaw angle shows no significant improvement with the proposed method.

In any tracking system, latency can be a problem, unless all tracked objects remain stationary. The size of the problem depends on the nature and size of the latency, as explained by [1, 8, 17, 19]. A small and constant latency might not be a problem, whereas any non-constant latency or a large latency poses a problem in any setup. Mania et al. [17] presents the results of an experiment of users experiences and perception of latency with varying degrees of scene complexity, and finds that the just noticeable difference (JND) is 15 ms or less. This is in line with earlier investigations, such as Adelstein et al. [1], who measured 8-17 ms and Ellis et al. [8], who found an average of 11.6 ms to be the JND.

Swan and Gabbard [18] and Kruijff et al. [14] review literature on user-based experimentation and perceptual issues respectively. In [18], the authors describe human perception and cognition research within AR, and note relevant examples of depth perception research to be taken into consideration when designing the experiment presented in this paper. In one example of depth perception and distance perception research by Ellis and Menges [9], the authors found that objects in the near field tend to suffer from perceptual localization errors in x-ray or monoscopic setups. Even though depth estimation is not a part of this study, it should be noted that there are no real or virtual objects in the field between the viewfinder and the objects of interest, i.e. the objects that participants are asked to align. Kruijff et al. [14] show that there is a current lack of research within the evaluation of human ability to notice tracking errors. This is in spite of the fact that human performance on occlusion handling, x-ray rendering, visual quality, depth perception, and accommodation are all areas of interest within the research community.

Livingston and Ai [16] present a user study on registration errors, i.e. latency, noise, and orientation errors. Their experiment focused on evaluating user performance in an AR environment using an optical see-through HMD. By adding high or low error variables to the system parameters (the latency, noise, and orientation), they found noise to have a limited impact on user performance, despite being displeasing in a subjective sense. Latency was shown to have a significant impact on localization performance, with users being slower under high latency in comparison to low latency. Orientation errors did not present a significant difference in localization accuracy. This provides a great step in the direction of expanding the knowledge of user performance in relation to registration errors. In this study, we look into the lower boundaries of orientation errors visible to the users. Setting a lower boundary might influence the level of artificially added offsets in future performance studies on user performance and registration errors.

3 Hypotheses

The experiments presented in this paper are based on a desire to test two main hypotheses:

1. It is more difficult to perceive static orientation errors in indirect AR than in classical AR.
2. There is a difference between the perception of static orientation errors w.r.t. yaw, pitch, and roll.
   (a) There is a difference between mean errors w.r.t. yaw, pitch, and roll.
   (b) There is a difference between error variance w.r.t. yaw, pitch, and roll.

Hypothesis 1 can be reasonably justified by two facts: 1) A person viewing an indirect AR scenario will *not* have the direct, pixel-to-pixel correspondence between the virtual and real worlds produced by having a live camera feed on the screen. This means that the viewer will have to resort to using his/her spatial abilities and a mental mapping between the virtual and the real world. 2) In the most common form of indirect AR, there is no tracking of the viewer's position relative to the screen. This means that the view of the virtual world seen on the screen will only be absolutely correct from a single viewpoint. This is furthermore complicated by the fact that the user does not know where this virtual sweet spot is. Instead, the only option is to rely on head motion to find the sweet spot, if the viewer does not want to rely purely on spatial imagination to mentally blend the virtual and real scene.

With respect to hypothesis 2, we believe that to be supported by the fact that the human sense of balance is governed by an external reference force, i.e. gravity. This allows people to gauge their own orientation as well as the orientation of other objects in relation to the local direction of gravity. This makes us well-equipped, even in the absence of technical equipment, to deal with tasks involving corrections on the roll and pitch axes. E.g. most people can tell if a picture hangs reasonably straight on a wall, without the use of a spirit level. Similarly, the sense of balance reliably tells people if they are falling forwards or backwards. This is contrary to the yaw axis, where there is no absolute, external reference that can be sensed by humans to use as guide. E.g. people in a windowless room will likely not be able to tell their absolute heading, without the use of a compass. For these reasons, we predict that the perceptible errors on the yaw, pitch, and roll axes will not be the same, neither in terms of accuracy (mean error), nor precision (error variance).

[1] http://www.oculusvr.com/

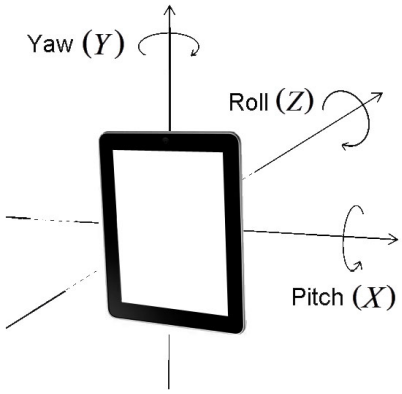

Figure 2: An illustration of the rotation axes relative to the tablet used in the experiment. Rotation around the local x-axis is named pitch, local y-axis rotation is named yaw, and local z-axis rotation is named roll.

4 Experiment 1: Classical and indirect AR

The first experiment is focusing on both classical AR and indirect AR. It investigates the lower boundaries for human perception of static orientation errors in both scenarios, and simultaneously investigates the effects of the error on the 3 local rotational axes. The purpose of the first experiment is to find the absolute minimum boundary for human error perception when calibrating an offset orientation in one axis. Thereby we also hope to find the lower threshold for user perception of registration errors.

4.1 Method

In evaluating user perception of static orientation errors, the attitude of a tablet device is used to describe the orientation, i.e. yaw, pitch, and roll angles, as depicted in Figure 2. After initial calibration, the attitude displayed is offset from the calibrated attitude to simulate a static orientation error. In order to simplify the experiment, and to make the task easier for the participants, only a single axis is offset in a single trial, since early pilot testing revealed that simultaneous calibration of multiple axes greatly increased the overall difficulty of the task.

The error term is defined as the difference between the final user input and the actual offset in the system from the calibrated ground truth. This implies that an error of 0° is a perfect solution to the task. It seems reasonable to assume that the these errors will have a mean value of 0° across all participants and conditions. Any departure from this assumption in the experimental data will indicate that results are somehow biased. However, it is not useful to analyse the raw errors to find the desired lower boundary for perceptible errors. In order to find the desired bound, we instead use the absolute values of the errors, since errors of e.g. -1.35° and 1.35° both are equally wrong in terms of magnitude, and are equally far from the calibrated ground truth. Both raw errors and absolute errors are logged, such that it is both possible to detect any bias and the desired perceptual bound.

We suggest that the task of asking users to correct an artificial, static tracking error can be expected to uncover where the lower threshold of error perception is. If the user can detect an error, then he/she is expected to continue to correct it, until no error is perceived anymore. When the user stops correcting the error, it is therefore resonable to assume that the current error is below what can be perceived by that person. Furthermore, in a realistic usage scenario, it is expected that an error of similar magnitude will also not be perceived by the same person. This is especially true, if the user has not been instructed to be alert of any errors in the realistic usage scenario.

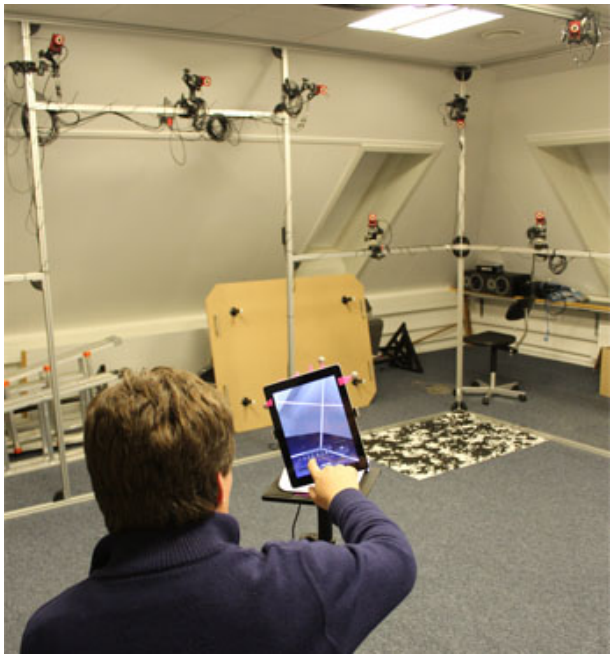

Figure 3: Example of the lab setup. The participant is currently working on an indirect AR task, hence all of the contents on the screen are virtual. The iPad stand is placed near the center of the room on a platform, facing one corner in which the trackable marker is placed, to ensure correct and stable tracking. The marker is printed on A0 format paper (841mm x 1189mm canvas paper). The user must then re-orient the offset rotation of the visualization towards the calibrated setting.

To reduce any initial estimation error in the optical tracking system, the system was manually calibrated by the experimenters prior to any participant interaction. This calibration step only took place once, after which the calibrated setup was left unmoved.

The procedure is that each participant is placed in a chair in front of a tablet device (see Figure 3). Then the participant is introduced to the purpose of the experiment, and the controls of the application. We embedded a control system for this experiment within the application, allowing the user complete control of the rotation on a single predefined axis according to the current task. I.e. the user is only able to adjust the axis currently offset from the calibration. The control system allows for adjustment of angular orientation in increments down to $\pm 0.0001^\circ$. The controls are shown and explained in Figure 4.

Prior to the actual experiment, the user goes through a training stage, in which random trials are presented in a configuration which does not occur in the actual experiment to mitigate any learning bias. Once the user is confident with the controls and the purpose of the experiment, the actual experiment commences. During the experiment, the participants were allowed to take breaks or ask questions, if needed.

All participants were told to solve the given tasks as well as possible, not paying any attention to the time spent. This was done in order to get the participants to emphasize quality over speed in their responses.

After completing all trials, the participants responded to a small questionnaire surveying their subjective evaluation of their performance along with explaining their strategies for solving the tasks for both the classical AR and indirect AR scenarios.

4.2 Materials

The experimental software is developed for tablet devices. The experiment is carried out using an iPad3[2] (2048x1536 resolution at 264 pixels/inch) using the back-facing camera (5 MP, 54.4° vertical field of view and 42.0° horizontal field of view) for tracking (indirect and classical AR) and background display (classical AR only). We generated the classical AR and indirect AR scenes using Unity3D[3] as the rendering engine, and Vuforia[4] for camera background (for the AR part) and as marker detection and tracking system. A trackable surface was created for the Vuforia tracking system, and printed as an A0 non-glossy poster (841mm x 1189mm canvas paper) placed at a 2 meter distance to the iPad stand (Figure 3).

The laboratory is specifically chosen for this experiment in order to provide a manageable setting with many options for comparing orientations displayed on the device with the setup of the interior of the laboratory. The main feature of the laboratory is a 4.5m × 4.5m × 2.5m aluminium cage used for mounting lab equipment. This cage serves as the main feature of the virtual scene in both scenarios. For the classical AR scenario the image captured by the built-in camera is rendered as background with the aluminium cage augmented on top, and in the indirect AR scenario, a simplistic model of the room is rendered in addition to the cage. To make this model feasible to produce, it only included the major features of the room, i.e. floor, ceiling, walls, windows, door, and large pieces of furniture, but not any smaller objects lying around the room. We have not investigated the consequences of this choice of 3-D model, but it may be an interesting subject of study in the future.

The cage was chosen as the main object, as the beams are linear objects connected at right angles, allowing for many opportunities in choosing which parts of the scene to use as calibration targets, such as horizontal/vertical beams, or the joints of beams at the corners of the cage. Modelling the cage is simple, since it is of known dimensions, cut to a precision of ±0.5mm. The room and its major features were modelled in Google SketchUp and Autodesk Maya to be imported into Unity3D. Figure 4 presents screenshots of the two scenarios tested in the experiments, i.e. classical and indirect AR.

A single iPad standing on a raised platform is used for all trials as shown in Figure 3. All user translation and orientation of the device is disallowed. As the setup is static and never moves, the estimation performance should be static, and latency and jitter, as well as other dynamic errors, are eliminated.

The iPad3 provides an average angular resolution of about 70 pixels per degree at a viewing distance of 38 cm, which is a realistic viewing distance in the experiments of this paper[5]. This number can be calculated using Equation 1. If the viewing distance is d and the height of the iPad is s, we get a total visual angle of $\alpha = 29.06°$. Using the iPad3's vertical resolution of 2048 pixels, we get $2048/29.06$ pixels$/° = 70.47$ pixels$/°$.

$$\alpha = 2 \cdot \arctan(\frac{s}{2d}) \frac{180°}{\pi} \; [°] \tag{1}$$

Given that the number of pixels per degree is higher than the human threshold according to [7], we assume that participants can, at least theoretically, detect a single pixel change from controlling the orientation in the application. The lower limit for in-app rotation needed for a single pixel change is more difficult to estimate, due to aliasing. Even the slightest rotation may contribute to one or more pixels changing value, if the involved pixels are currently located right on the verge of a change.

We therefore do not assume a minimum rotation threshold that changes a pixel value. However, for the virtual camera which has a FoV corresponding to the real camera (54.4° vertical field of view) and the iPad having an output resolution of 2048 pixels vertically, for a rotation of approximately $54.4°/2048 = 0.026°$, every virtual object will have shifted a minimum of 1 pixel on the screen. Allowing the user to control up to 4 decimals after the decimal point is considered reasonable to allow for the participants to accurately recalibrate the system.

[2]Featuring an Apple A5X chip (Dual-core 1 GHz Cortex-A9 processor with a PowerVR SGX543MP4 GPU)

[3]http://www.unity3d.com

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

[5]The angular resolution varies non-linearly across the field-of-view, because the screen is flat. However, this variation is quite small since the chosen field-of-view angle is also reasonably small.

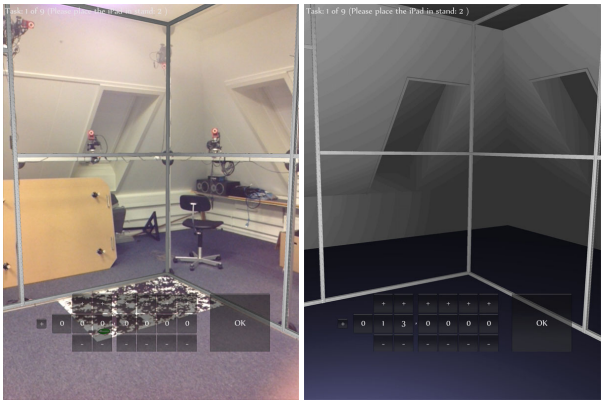

Figure 4: Illustration of the two scenarios, [Left] classic AR and [Right] indirect AR. Note that in the classical AR scenario, only the grey aluminium cage is virtual, whereas in the other scenario, all screen contents are virtual. The GUI controls used by particpants for re-calibrating the errors are also visible in both screenshots. For each digit of the user input, there is a plus and a minus button, allowing for precise control of every digit from ±10° down to ±0.0001°. Once the user is satisfied that the calibration error is gone, the large OK button is pressed.

4.3 Study design

Experiment 1 is a randomized, within-subjects experiment with 3 repetitions of 30 different trials for each participant, giving a total of 90 trials per participant. The factors of the experiment are scenario (2 levels; classical or indirect AR) and rotation axis (3 levels; yaw, pitch, and roll). Each combination of these factors are tested at 5 different, random initial offsets from the ground truth calibration. Inside each block of 30 trials, the order is completely randomized, meaning that classical and indirect AR trials are randomly distributed among each other. The offsets used are all in the range $\pm[10;30]°$, implying that the smallest initial offset is 10°, and the largest possible initial offset is 30°. This is done to ensure that 1) there clearly is an error to correct and 2) the participant does not get confused about what part of the virtual cage to match to the real one (the cage is symmetrical).

4.4 Participants

A total of 30 (2 female, 28 male) unpaid participants took part in the first experiment. As such the data collected from experiment 1 comprises 2700 trials. All participants were students and staff recruited at the local university campus, which implies that all have a background in media technology. The mean age of the participants in experiment 1 was 24.07 years. The average total completion time was approx. 22 minutes in this experiment.

4.5 Issues

Both during and after the first experiment, it became clear from the observations and data gathered that the experiment's indirect AR

scenario had an unfair advantage in comparison to a realistic indirect AR scenario. Even though the sequence of trials and conditions was completely randomized, the participants' memory of the final appearance of recent classical AR trials made them better at correcting the errors in any subsequent indirect AR trials. In the questionnaire, many participants mentioned that they learned the correct orientation after a while and as a consequence started to rotate the scene in indirect AR scenarios to match their memory of any previous classical AR trial solutions. This implies that the estimated perceptual threshold for indirect AR in experiment 1 is likely much too optimistic. However, the validity of the estimated thresholds for classical AR are unaffected by this problem. The classical AR scenario does not suffer from any such advantage, as the connection is happening directly in screen space, and the user simply has to match the virtual and the real representation on the screen. Furthermore, several participants mentioned that after the first few indirect AR scenarios, they stopped using the actual surroundings (i.e. the room) as a means of solving the task. Instead, they relied on their memory of the correct relation between the classical AR model and physical features on the iPad, or even the placement of GUI elements on the screen to complete the indirect AR trials.

Another weakness of the approach is that a static setup might result in participants using techniques for solving the task that would not be possible in a realistic usage scenario, e.g. with more degrees of freedom in handling the device. Such techniques include remembering settings from one trial to the next. For instance, given a yaw correction trial, the user might remember the correct floor placement when doing the next pitch correction task. Given that the user cannot employ these tricks in realistic settings, we believe that the experimental results will estimate the absolute lower thresholds for human perception of static orientation errors.

It is also a relevant concern that the accuracy of the experiment is limited by the calibration accuracy attainable by the experimenters and the tracking software used. As we cannot guarantee the entire setup to be perfectly calibrated, the calibration can only be performed to the level where the experimenters cannot reliably tell the difference in the best case (classical AR). Through early pilot testing, this limit was found to be somewhere around $\pm 0.01^{\circ}$. For this reason, any results of the experiment that go below this limit should be treated with caution.

One final difficulty in the experiment is the spatial, mental mapping in the indirect AR scenarios. In comparison to the classical AR scenarios, where the connection of the real and virtual scenes happens on screen, this does not happen in the indirect AR scenario. Here, the user must mentally connect the real and virtual scenes. The indirect AR scenarios are made even harder by the fact that there is no tracking of the viewing position of the user in relation to the screen. We see this as a necessity, as this is how most indirect AR experiences are presented to users in current, realistic usage scenarios. However, in future experiments, the inclusion of head tracking is definitely a worthwhile direction to take.

5 Experiment 2: Indirect AR

In this experiment, we attempted to eliminate the possibility of relying on memory and learning when solving indirect AR tasks. Furthermore, we sought to eliminate solution tricks involving fitting the position of physical iPad features or GUI elements to virtual features on the screen. This was achieved by using three different physical platforms for the iPad during the experiment on which only one set of yaw, pitch, roll trials would be performed, and by removing the classical AR scenarios from the experiment. The new setup is illustrated in Figure 5.

5.1 Method

The experimental setup is very similar to the first experiment. The tasks of the participants are exactly the same as in experiment 1, and the tablet is still static during each trial, i.e. all user translation and rotation of the device itself is disallowed. However, the iPad will be moved from platform to platform by the experimenters as the experiment progresses. It is still assumed that other types of tracking errors are eliminated.

Figure 5: The lab setup for the second experiment. The iPad is placed on one of three platforms, all facing one corner of the room, in which the trackable marker is placed. In this example, the iPad is placed on platform #1. During the actual experiment it is moved between the platforms in random order.

The user is placed in front of a fourth iPad platform during the training stage to eliminate the possibility of any memory of trials actually used in the test. In the training stage, random tasks are presented one at a time in a configuration which does not occur in the actual experiment, until the participant feels confident about the tasks and controls.

During the experiment, the user is placed in front of one of the three platforms at a time, chosen at random. Furthermore, the sequence of yaw, pitch, and roll trials at each platform is randomized.

After completing all of the trials, the participants responded to a small questionnaire surveying their subjective evaluation of their performance, along with explaining their strategies for solving the tasks.

5.2 Materials

The materials used for the second experiment were identical to those used for the first experiment, with the exception that three different iPad platforms were used for the trials of experiment 2, along with a fourth plaform for the training session.

5.3 Study design

Experiment 2 is a randomized, within-subjects design with one factor of interest, axis (3 levels; yaw, pitch, and roll) and one blocking factor, platform (3 levels; named 1, 2, 3). The platforms are used in random order, and three trials (one for each axis) is conducted on one platform in random order before proceeding to the next. This produces a total of 9 trials per participant. This design should ensure that the benefit of learning from trial to trial should be minimized.

5.4 Participants

A total of 16 (all male) participants took part in the second experiment, all recruited from the same population of students and staff as that of experiment 1. However, no participant took part in both experiments. Since each participant performed 9 trials, the data from experiment 2 comprises 144 observations. The average age was 23.25 years in experiment 2. Each participant spent an average of approx. 7 minutes on the trials of the second experiment.

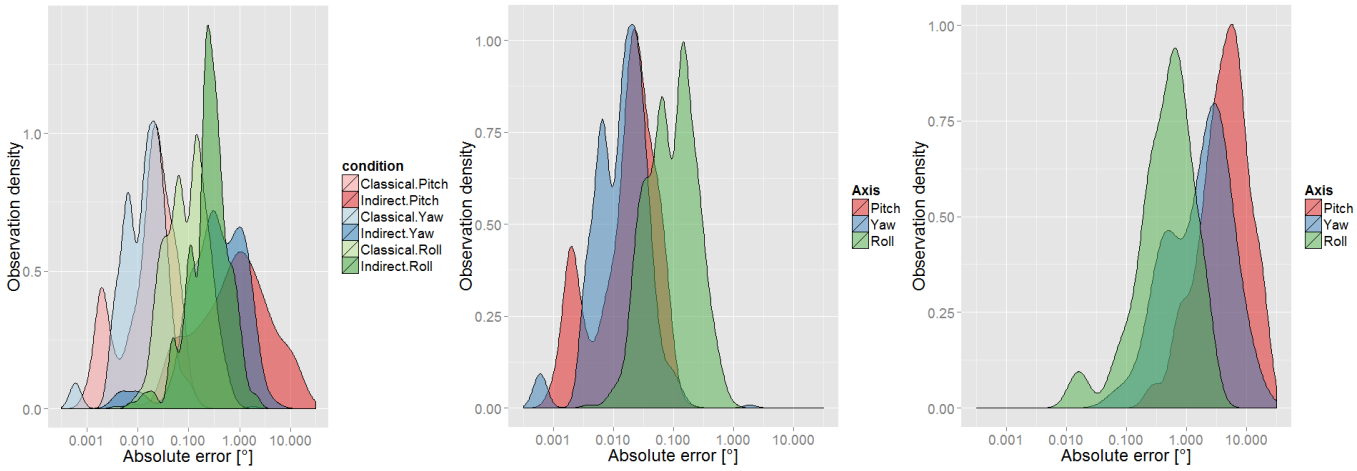

Figure 6: A density plot presentation of the angular error data from experiments 1 and 2. [Left] Density plot of the absolute angular errors in all conditions from *experiment 1 only*. The indirect conditions (darker regions) generally have larger absolute errors than the classical ones (brighter regions). [Middle] The classical AR condition absolute angular errors from *experiment 1 only*. The indirect AR results from experiment 1 have been removed from this plot, since these results are suspect. The roll errors (green) are larger than the yaw (blue) and pitch (red) errors. [Right] A density plot of the absolute angular errors from *experiment 2 only* (i.e. indirect AR). The roll errors (green) are smaller than yaw (blue) and pitch (red). Note that the horizontal axes are logarithmic in all the plots.

6 Results and Discussion

All analyses have been carried out using the statistical software package R, using a significance level of $\alpha = 0.05$. The main analysis methods employed are type III ANOVA and the Friedman test. Post-hoc tests for pairwise comparisons following ANOVA has been performed using Tukey's honest significant difference (HSD) method. All reported confidence intervals have been computed using bootstrapping with bootstrap samples of size $n = 100000$.

None of the statistical analyses have used pooled data from experiments 1 and 2. Each experiment is separately analysed, and no cross-inferences about the results between the two are made. However, we do note that the results in the indirect AR condition in both experiments show the same general tendencies, which supports the validity of both experiments in spite of the flaws of experiment 1.

In both experiments, the collected absolute angular errors did not meet the standard assumptions of ANOVA analysis. Particularly, the normality of the residuals, and the homogeneity of the variance across the experimental conditions were found to be problematic. However, a logarithmic transformation of the angular errors completely solved this problem for the data from experiment 2, and greatly improved the situation for experiment 1. For this reason, the preferred scale for statistical analysis of the absolute angular errors is a logarithmic one. To further ensure that the conclusions were well-supported, a non-parametric Friedman test was run alongside the ANOVA to verify that no conflicting conclusions were found.

The data collected from experiment 1 and 2 is summarized in the density plots shown in Figure 6. The difference between errors, absolute errors, and logarithmic absolute errors is illustrated in Figure 7.

6.1 Hypothesis 1

The first hypothesis stated that people would be worse at detecting and rectifying static orientation errors in indirect AR than in classical AR, due to the missing information of the live camera feed. Running ANOVA on the data from experiment 1, this hypothesis is supported by the fact that scenario type is a significant main effect ($F_{1,29} = 677.1, p < 2.2 \cdot 10^{-16}$). This means that people are significantly worse at detecting and correcting errors in indirect AR than in a classical AR setup, in spite of the fact that the indirect AR errors from experiment 1 are likely too small to be realistic.

In experiment 1, the estimated mean absolute angular error in classical AR scenarios is 0.0610° with a 95% confidence interval of $[0.0559; 0.0670]^\circ$, whereas the same figures for indirect AR in experiment 1 are 1.06° and $[0.95; 1.19]^\circ$. In other words, the mean error is approx. 17.4 times larger with indirect AR than classical AR. As was previously explained, the conditions for indirect AR viewing in experiment 1 were in all likelihood unrealistically good, mainly due to the effect of learning from classical AR trials. This implies that the mean indirect AR figures above should be taken as an extreme best-case scenario for detecting errors. From experiment 2, the more realistic estimate of these figures in indirect AR scenarios are 3.11° and $[2.51; 3.78]^\circ$, implying that the realistic mean angular error is approx. 51 times larger in indirect AR than in classical AR.

Another way of checking this is to see the position of the least significant digit that participants chose to adjust in the two scenarios. This data is discrete by nature, so the two scenarios are compared using a Friedman test on the data from experiment 1 instead of ANOVA. The conclusion in this case is the same: People use more digits to adjust their estimate in classical AR than in indirect AR ($\chi^2_1 = 30, p = 4.3 \cdot 10^{-8}$). In the light of the analysis of the angular errors in the two types of scenarios, this means that not only are participants adjusting their responses more finely in classical AR, they are also reaching higher levels of accuracy by doing so.

6.2 Hypothesis 2

The second hypothesis stated that there would be a difference between the errors made on the three tested rotation axes, yaw, pitch, and roll. This difference was not only hypothesized to be a difference in mean error, but also in the error variance, as it seemed likely that, especially when adjusting the roll axis, participants might get extra help from their sense of the direction of gravity.

Running an ANOVA analysis on the data from experiment 1 reveals that the axis is both significant as a main effect ($F_{2,58} = 38.56, p = 2.2 \cdot 10^{-11}$) and as an interaction with the type of scenario, i.e. indirect AR or classical AR, ($F_{2,58} = 81.1, p < 2.2 \cdot 10^{-16}$). The interpretation of this result must therefore be that the mean errors for the axes are different, and that these differences are

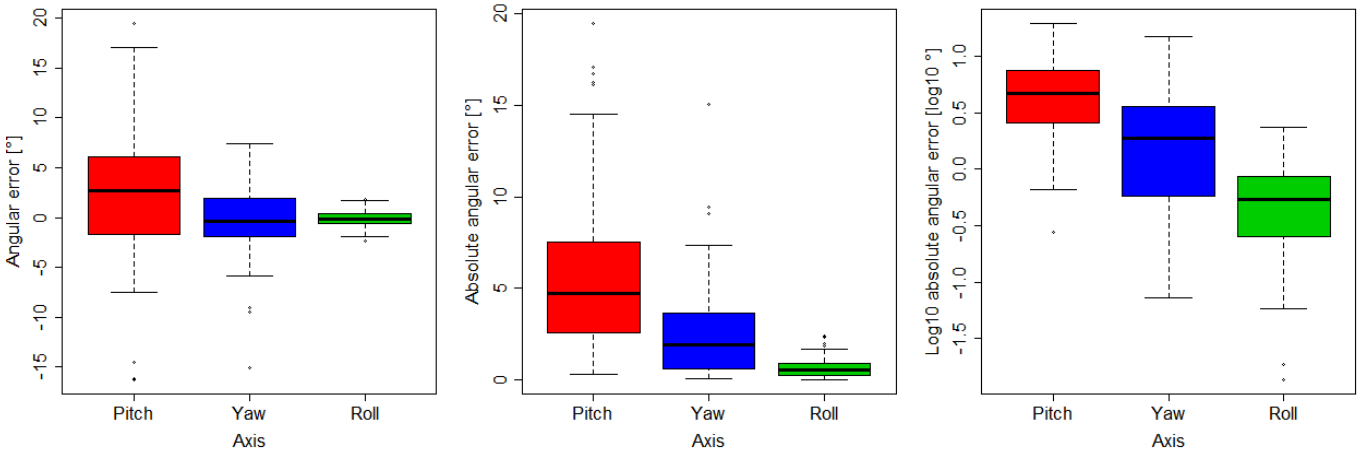

Figure 7: A boxplot presentation of the angular errors made in experiment 2. [Left] A boxplot of the untransformed angular errors, measured in degrees. This plot clearly shows that there is a difference in variance for the three different axes. Furthermore, the pitch axis is revealed to have an unexpected bias of about 2° that should not be there, if participants were equally likely to over- and underestimate the errors. [Middle] A boxplot of the errors transformed by taking their absolute values. This is a more useful representation, when the main concern is the magnitude of errors, rather than the direction of the errors. [Right] The absolute errors transformed by a $\log_{10}$ transformation that both has the effect of making the variance in all three axes equal as well as making the distributions on each axis closer to normal.

significantly affected by the type of scenario. Following this result up by a Tukey HSD test on just the classical AR data from experiment 1, it is revealed that yaw and pitch are both significantly different from roll, but not from each other (all significant $p < 1.0 \cdot 10^{-4}$). In the classical AR scenario, the errors on the axes are significantly smaller for yaw and pitch than for roll.

With the data from experiment 2, the ANOVA results also show that there is a significant main effect of the rotation axis ($F_{2,30} = 65.81, p = 1.1 \cdot 10^{-11}$). The follow-up Tukey HSD test in this case shows that all three axes are significantly different from each other (all $p < 1.0 \cdot 10^{-5}$). In the indirect AR scenario, roll has smaller errors than yaw, which in turn has smaller errors than pitch. Thus, hypothesis 2 is strongly supported by the data from both experiments.

With respect to variance, the hypothesis that the variance would be different for different axes is also strongly supported by the data. This has been tested using Levene's test of equal variance on the absolute angular errors. The results all come out with $p \ll 0.001$. This implies that people are not equally consistent about their error estimates in all axes.

The estimated mean error responses for the three axes are presented in Table 1, along with their associated 95% confidence intervals. The numbers in the table clearly indicate that the roll errors seem to be consistent across the two scenarios, whereas the errors on the other two axes are highly dependent on the scenario.

6.3 Other findings

There are several other interesting findings in the experiment that are not directly related to the hypotheses. First of all, when systematically asking all participants about their own perceptions of the experiment, many of them reported that they think that they somehow used a trick for some of the trials. When asked to elaborate, they told that they relied on help from aliased lines in the image (i.e. looking for the setting where a line no longer appears jagged because of the pixel grid), memory of the locations of specific linear features in the image relative to the frame of the tablet, by assuming that certain linear features on the screen had to be completely vertical or horizontal, or by expecting specific features to be centered on the screen. Most, if not all, of these tricks would not really be possible in a situation outside the lab, where the device is no longer stationary relative to the scene. For this reason, we believe that all estimated error tolerances are lower bounds, and that the error tolerance in a more realistic, dynamic situation might be somewhat larger. Furthermore, our results are unambiguous enough that these tricks have probably in most cases mainly helped in letting participants make more consistent responses, rather than more correct responses.

Another observation made in the data is that the mean (non-absolute) angular error on the pitch axis in the indirect AR condition in both experiments is approx. 2°. This is contrary to the expectation that the mean value should be around 0°, if it was equally likely to over- and underestimate the error. In other words, there is an unexpected bias of about 2° for indirect AR pitch tasks. We believe that this bias is caused by the combination of two conditions: 1) The tablet was pointing slightly downwards during both experiments, and 2) several people reported using a specific trick involving the expectation of seeing an equal amount of virtual floor and ceiling in the correct setting. However, since the tablet pointed slightly downwards, the correct setting shows more floor than ceiling. Therefore, these two facts will explain why there is a general tendency to overshoot on the pitch estimates in the experiments. In-

Table 1: A table of the estimated mean absolute angular error tolerance (with 3 significant digits), dependent on the two independent variables of axis and scenario. The estimated 95% confidence intervals for the mean values are given in brackets. All classical AR results are estimated from data gathered in experiment 1 only, and all indirect AR results are estimated using experiment 2 data only.

Axis	Scenario	
	Classical AR (exp. 1)	Indirect AR (exp. 2)
Yaw	$0.0235^\circ[0.0185; 0.0329]^\circ$	$2.74^\circ[2.01; 3.63]^\circ$
Pitch	$0.0254^\circ[0.0234; 0.0275]^\circ$	$5.88^\circ[4.59; 7.32]^\circ$
Roll	$0.134^\circ[0.123; 0.146]^\circ$	$0.680^\circ[0.525; 0.857]^\circ$
Overall	$0.0611^\circ[0.0559; 0.0670]^\circ$	$3.11^\circ[2.51; 3.78]^\circ$

terestingly, many participants have also reported that subjectively, they found the pitch trials to be much more difficult than yaw or roll. Conversely, several participants stated that the roll tasks were the easiest, which is also supported by the data.

7 Conclusion

There are several conclusions to be made, based on the performed experiments. The two main lessons learned must be:

1. People are much more perceptive of static orientation errors in classical AR than in indirect AR scenarios.

2. The ability to detect static orientation errors is *highly* dependent on the rotation axis affected by the error.

3. Roll error perception seems consistent in indirect and classical AR, but yaw and pitch errors are perceived differently in indirect and classical AR.

These points can be elaborated further. Quantitatively, the size of static orientation errors that can be detected in indirect AR is somewhere between 1 and 2 orders of magnitude larger than those detectable in classical AR. This implies that designers of indirect AR systems need not worry as much about static orientation errors, e.g. errors similar to those produced by a slow drift caused by quantization and integration errors. The fact that the axes are differently perceived means that precise orientation tracking is not equally important for all axes. In classical AR, people are more critical of yaw and pitch errors than roll errors. Conversely, in indirect AR where there is no help from a camera feed, the roll axis is the one where it is easiest to detect errors, whereas yaw and pitch are less important. It seems reasonable to speculate that, in the absence of help from a camera feed, the roll errors are much easier to detect because the roll axis is also tied to the human sense of balance which works irrespective of the imagery on the screen.

Another important contribution of this paper is the first estimation of detection thresholds for static orientation errors on each of the three tested axes, and in both classical and indirect AR. These thresholds were presented in Table 1. The difference in thresholds for the same axis depending on the scenario is quite large in some of the cases. For instance, in the case of the yaw, the threshold in the case of indirect AR is more than 100 times larger than that in classical AR. In the case of pitch, the difference is even more pronounced, the threshold being more than 200 times larger in indirect AR than classical AR. For the roll axis, the thresholds are much more consistent across the scenarios.

This study also leaves many open questions to be answered in the future, such as the effect of allowing the tablet and the participant to move around in a realistic manner and the effect of taking the study out of a controlled lab context and into a more realistic setting. Other interesting areas of future study might be incorporation of head tracking, a higher level of realism in the virtual model, and similar studies of static position errors.

Acknowledgements

The authors wish to thank Peter Skotte, Claus B. Madsen, and the anonymous participants.

References

[1] B. D. Adelstein, T. G. Lee, and S. R. Ellis. Head Tracking Latency in Virtual Environments: Psychophysics and a Model. In *Human Factors and Ergonomics Society Annual Meeting Proceedings (2003)*, pages 2083–2087, 2003.

[2] R. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier, and B. MacIntyre. Recent advances in augmented reality. *IEEE Computer Graphics and Applications*, 21(6):34–47, 2001.

[3] R. T. Azuma. A survey of augmented reality. *Presence*, 4(August):355–385, 1997.

[4] M. Bajura and U. Neumann. Dynamic registration correction in video-based augmented reality systems. *IEEE Computer Graphics and Applications*, 15(5):52–60, 1995.

[5] J. Borenstein and L. Ojeda. Heuristic Reduction of Gyro Drift in Gyro-based Vehicle Tracking. In E. M. Carapezza, editor, *Proc. SPIE Vol. 7305*, volume 7305, page 11, May 2009.

[6] J. Caarls, P. Jonker, and S. Persa. Sensor Fusion for Augmented Reality. In E. Aarts, R. Collier, E. Loenen, and B. Ruyter, editors, *Ambient Intelligence*, number 1, pages 160–176. Springer Berlin Heidelberg, July 2003.

[7] R. Clark. Notes on the resolution and other details of the human eye, Nov. 2013.

[8] S. R. Ellis, K. Mania, B. D. Adelstein, and M. I. Hill. Generalizeability of Latency Detection in a Variety of Virtual Environments. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 48(23):2632–2636, Sept. 2004.

[9] S. R. Ellis and B. M. Menges. Localization of virtual objects in the near visual field. *Human Factors*, 40(3):415–431, 1998.

[10] S. J. Gilson, A. W. Fitzgibbon, and A. Glennerster. Quantitative analysis of accuracy of an inertial/acoustic 6DOF tracking system in motion. *Journal of neuroscience methods*, 154(1-2):175–82, June 2006.

[11] W. Hoff and T. Vincent. Analysis of head pose accuracy in augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 6(4):319–334, 2000.

[12] R. Holloway. Registration error analysis for augmented reality. *Presence: Teleoperators and Virtual Environments*, pages 1–25, 1997.

[13] D. W. F. V. Krevelen and R. Poelman. A Survey of Augmented Reality Technologies, Applications and Limitations. *The International Journal of Virtual Reality*, 9(2):1–20, 2010.

[14] E. Kruijff, J. E. Swan II, and S. Feiner. Perceptual issues in augmented reality revisited. In *2010 IEEE International Symposium on Mixed and Augmented Reality*, pages 3–12. IEEE, Oct. 2010.

[15] G. Liestø l. Situated Simulations Between Virtual Reality andMobile Augmented Reality: Designing a Narrative Space. In B. Furht, editor, *Handbook of Augmented Reality*, chapter 14, pages 309–319. Springer New York, New York, NY, 2011.

[16] M. A. Livingston and Z. Ai. The effect of registration error on tracking distant augmented objects. *2008 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, pages 77–86, Sept. 2008.

[17] K. Mania, B. D. Adelstein, S. R. Ellis, and M. I. Hill. Perceptual sensitivity to head tracking latency in virtual environments with varying degrees of scene complexity. In *Proceedings of the 1st Symposium on Applied perception in graphics and visualization - APGV '04*, number study 1, pages 39–47, New York, New York, USA, 2004. ACM Press.

[18] J. E. Swan II and J. L. Gabbard. Survey of User - Based Experimentation in Augmented Reality. In *Proceedings of 1st International Conference on Virtual Reality*, pages 1–9, 2005.

[19] R. J. Teather, A. Pavlovych, W. Stuerzlinger, and I. S. MacKenzie. Effects of tracking technology, latency, and spatial jitter on object movement. *2009 IEEE Symposium on 3D User Interfaces*, pages 43–50, 2009.

[20] G. Welch and E. Foxlin. Motion tracking: no silver bullet, but a respectable arsenal. *IEEE Computer Graphics and Applications*, 22(6):24–38, Nov. 2002.

[21] J. Wither, Y.-t. Tsai, and R. Azuma. Indirect augmented reality. *Computers and Graphics*, 35(4):810–822, 2011.

[22] S.-h. Won, N. Parnian, F. Golnaraghi, and W. Melek. A quaternion-based tilt angle correction method for a hand-held device using an inertial measurement unit. *2008 34th Annual Conference of IEEE Industrial Electronics*, (1):2971–2975, Nov. 2008.

[23] F. Zhou, H. B.-L. Duh, and M. Billinghurst. Trends in augmented reality tracking, interaction and display: A review of ten years of ISMAR. In *2008 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, pages 193–202. IEEE, Sept. 2008.