Subjective Evaluation of a Semi-Automatic Optical See-Through Head-Mounted Display Calibration Technique

Kenneth Moser, *Student Member, IEEE*, Yuta Itoh, *Student Member, IEEE*, Kohei Oshima, *Student Member, IEEE*, J. Edward Swan II, *Member, IEEE*, Gudrun Klinker, *Member, IEEE*, and Christian Sandor, *Member, IEEE*

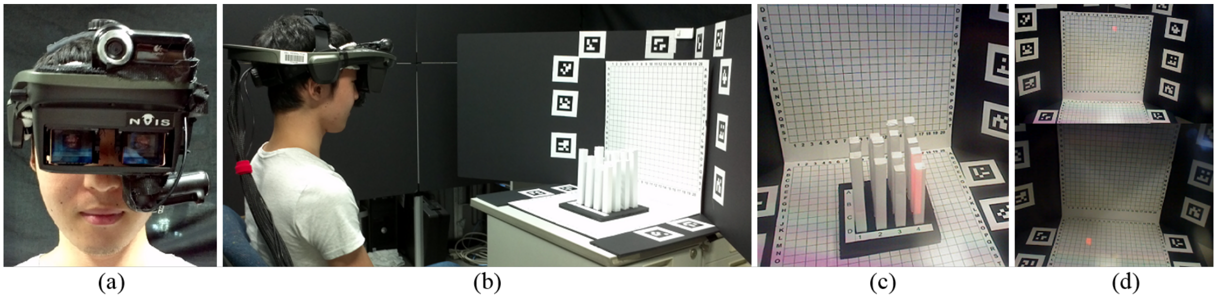
(a) (b) (c) (d)

Fig. 1. Experimental hardware and design. (a) Display and camera system. (b) Task layout. (c) Pillars task. (d) Cubes task.

Abstract— With the growing availability of optical see-through (OST) head-mounted displays (HMDs), there is a present need for robust, uncomplicated, and automatic calibration methods suited for non-expert users. This work presents the results of a user study which both objectively and subjectively examines registration accuracy produced by three OST HMD calibration methods: (1) SPAAM, (2) Degraded SPAAM, and (3) Recycled INDICA, a recently developed semi-automatic calibration method. Accuracy metrics used for evaluation include subject provided quality values and error between perceived and absolute registration coordinates. Our results show all three calibration methods produce very accurate registration in the horizontal direction but caused subjects to perceive the distance of virtual objects to be closer than intended. Surprisingly, the semi-automatic calibration method produced more accurate registration vertically and in perceived object distance overall. User assessed quality values were also the highest for Recycled INDICA, particularly when objects were shown at distance. The results of this study confirm that Recycled INDICA is capable of producing equal or superior on-screen registration compared to common OST HMD calibration methods. We also identify a potential hazard in using reprojection error as a quantitative analysis technique to predict registration accuracy. We conclude with discussing the further need for examining INDICA calibration in binocular HMD systems, and the present possibility for creation of a closed-loop continuous calibration method for OST Augmented Reality.

Index Terms—Calibration, user study, OST HMD, INDICA, SPAAM, eye tracking

1 Introduction

Optical see-through (OST) augmented reality (AR) systems allow the overlay of visual information onto a user's view of the real world. The primary benefit of these systems, specifically in conjunction with head mounted display (HMD) hardware, is the ability of the user to maintain a view of the real environment from the perspective of their own eyes. This is in contrast to video see-through (VST) AR systems, in which users view the environment from the perspective of a camera. Allowing a continuous hands-free view of the environment also lessens safety concerns from visibility loss due to hardware malfunction, as is the case in military AR usage [20], where constant sight is of the utmost importance. The utility of both OST and VST AR is further enhanced when the location of on-screen information is used to provide additional meaning and context.

In both VST and OST AR, information is fundamentally displayed in one of two reference frames: Screen Space and World Space. Information shown within screen space is statically positioned and does not appear displayed relative to any particular location or object within the world. This type of presentation is common in applications facilitating manufacturing or assembly style tasks where the AR visual components list instruction sets or other environment independent information [4]. On-screen items rendered within world space appear to have a 3D position within the environment and are registered, displayed relative, to fixed locations or objects. This display method requires sufficiently accurate registration between on-screen geometry and the visible environment in order to be effective [19]. Achieving an adequate level of world space registration accuracy requires the effectual employment of calibration mechanisms.

The goal of OST HMD calibration is to model the virtual camera projection matrix, used to render virtual geometry, such that it closely matches the real viewing frustum created by the user's eye and the display optics. Unlike VST AR, where computer vision techniques are used to correct registration [10, 16, 17], calibration of OST AR is not straightforward. VST techniques are difficult or impossible to use for correction in OST displays since the "camera" in these systems is the user's eye itself. Calibrating misalignment between the human eye and augmented geometry, in this case, is not a trivial task [22, 26]. Various calibration methods have been developed, which attempt to cor-

- *Kenneth Moser is with the Department of Computer Science and Engineering at Mississippi State University. E-mail: moserk@acm.org*
- *Yuta Itoh is with the Department of Informatics at Technical University of Munich. E-mail: itoh@in.tum.de.*
- *Kohei Oshima is with the Interactive Media Design Lab in the Department of Information Science at Nara Institute of Science and Technology. E-mail: oshima.kohei.of0@is.naist.jp.*
- *J. Edward Swan II is with the Department of Computer Science at Mississippi State University. E-mail: swan@acm.org.*
- *Gudrun Klinker is with the Department of Informatics at Technical University of Munich. E-mail: klinker@in.tum.de.*
- *Christian Sandor is with the Interactive Media Design Lab in the Department of Information Science at Nara Institute of Science and Technology. E-mail: sandor@is.naist.jp.*

rect misalignment by employing user feedback, in the form of visual alignments made between on-screen markers and real world locations. These techniques, however, are prone to alignment error and do not involve intuitive procedures easily understood by non-expert users.

Recently, calibration methods have been introduced, which seek to utilize eye imaging in order to ease the user performance burden and reduce user driven error sources in the system. A novel OST calibration approach, presented by Itoh and Klinker [14], incorporates such techniques and shows the possibility of reducing or even eliminating user error. Their Interaction Free Display Calibration (INDICA) method is a promising step toward closed-loop continuous on-line calibration of OST AR. Modern advances in miniature camera hardware and robust eye feature and pattern recognition algorithms allow the implementation of eye imaging methods on existing OST HMD hardware at consumer level cost. Quantitative analysis performed by Itoh and Klinker suggests their INDICA technique is capable of producing comparable results to standard interaction dependent methods. These results, though, are based on measures taken from a single expert user. A formal study, subjectively investigating the utility and performance of this new calibration method, when completed by inexperienced users, has yet to be conducted. Methods for adequately measuring calibration results in an OST AR system are highly user dependent and often rely on analysis of user performance in registration critical tasks.

The same factors that make calibration of OST HMDs difficult, primarily the inaccessible view from the user's eyes, also make it difficult to objectively measure the quality of registration produced by the calibration result. While a calibration technique can be quantitatively shown to produce appropriate results under ideal conditions, in practice, it must also be able to produce acceptable results across a varied user base and require minimum effort to effectively implement and perform. Swan and Gabbard [7] explain the importance of user-study driven design and evaluation for measuring utility and identifying user performance issues requiring further investigation.

In this work, we present the results of a user study which subjectively evaluates registration accuracy produced by Recycled INDICA, a semi-automatic OST HMD calibration method suited for use by non-expert users [14]. Our experimental design is the first to measure the performance of INDICA in two registration critical tasks, and we compare the results against a traditional user interaction based method, the Single Point Active Alignment Method (SPAAM), and Degraded SPAAM, an interaction free variant of SPAAM. We provide measures for both subjective and objective registration accuracy obtained from subjects in order to compare perceived registration quality in three dimensions, allowing for a more thorough accuracy evaluation over previous studies. We also present the results of a quantitative analysis, comparing the computed differences between the SPAAM and Recycled INDICA results. Our final discussion connects the subjective and objective results focusing on a surprising outcome of the Recycled INDICA calibration, and a possible issue with a quantitative analysis technique. We close with remarks addressing the calibration performance requirements for non-expert users, further investigative needs, and future extension of INDICA into closed-loop continuous calibration techniques for OST AR.

2 Background and Related Work

2.1 OST HMD Calibration

The goal of OST HMD calibration is to generate the 11 free parameters of a 3×4 projection matrix which most accurately describes the visual system created by the display screen and the user's eye. These parameters consist of intrinsic display specific and extrinsic user dependent values. The Single Point Active Alignment Method, introduced by Tuceryan and Navab [27], utilizes user feedback, in the form of screen to world alignments, in order to directly approximate all 11 parameters simultaneously. The benefit of this approach is the total independence from display hardware, ensuring applicability over a broad range of system types. An additional benefit of SPAAM is the freedom of movement allotted to users during the screen-world correspondence phase. Users are encouraged to move freely about the work space, in contrast to earlier alignment methods using bore sighting [3] or rigid fixation of the user's head upon a rest [6, 18]. The ease of implementation and relatively simple user requirements of SPAAM has made it the focus of numerous investigations into improving robustness to user and system errors, as well as expansion of its applicability to a broader range of OST AR systems. These investigations have produced a number of SPAAM variants and two stage calibration methods aimed at reducing the amount of necessary user interaction.

Genc et al. [8] devised a method for applying SPAAM to stereo OST HMDs, calibrating both left and right views simultaneously. Alignments for stereo SPAAM are performed between 3D world and virtual points by utilizing stereoscopic depth cues available to users wearing binocular displays. Varying the location at which screen-world alignments are taken is presented by Tang et al. [26]. Depth-SPAAM has shown increased robustness to user induced error but still requires that a significant number of alignments be made. Further development, given the denotation Easy SPAAM [9, 23], simplifies the calibration process by utilizing data from a previous SPAAM result. The new projection matrix is then optimized with accommodation to the eye location of the user through standard user driven 2D-3D alignments. This approach, however, significantly reduces the number of correspondence alignments the user is required to perform and isolates the impact of the alignments to only the extrinsic components of the final projection matrix. Similarly, Owen et al.'s Display Relative Calibration (DRC) [24] decouples the calibration process into two distinct phases, each producing different values of the projection matrix. The first phase, conducted offline prior to any user action, involves direct measurement of the intrinsic display parameters. The second phase, similar to Easy SPAAM, uses a small number of user driven screen-world correspondences to determine the eye pose. One option for DRC, provided by Owen et al., is completely interaction free when the second phase is ignored and systematic assumptions are made about the eye position within the system. The decoupling of intrinsic display and extrinsic user parameters is also utilized in a recently introduced calibration method incorporating eye imaging for eye position estimation.

An interaction free calibration method, utilizing an eye imaging camera, has been developed by Itoh and Klinker[14]. Images from the eye camera are processed using methods based on Swirski's iris detection [25] and Nitschke's algorithm [5], to estimate 3D eye position. Itoh and Klinker present two variants of INDICA: (1) Recycled INDICA, which extracts intrinsic parameters from a previously performed calibration, such as SPAAM and (2) Full INDICA, which similar to DRC, measures the display specific intrinsic parameters directly offline. Both variants require the extrinsic eye position values obtained from eye imaging in order to generate the final calibration result. Given that either INDICA variant is able to be performed independent of any user interaction, both methods could potentially be incorporated into an OST AR system with continuous calibration. Such a system would be ideal for users unskilled in performing calibration procedures and would exhibit increased robustness to HMD movement on the user's head during operation. Given the potential applicability of INDICA, thorough quantitative and subjective evaluation of the method is essential to not only verify correctness in an active setting but also gauge utility and ease of use by non-expert users.

2.2 Quantitative Evaluation of OST HMD Calibration

Itoh and Klinker [14] quantitatively compare the registration accuracy of INDICA with that of a standard SPAAM implementation. Reprojection of the screen-world alignment pairs from multiple SPAAM calibrations, using the INDICA generated projection matrices, produced pixel locations with low variance and only slight error from the original SPAAM "ground truth" projections. In a follow-up study, [15], sensitivity analysis of the separate parameters used in the derivation of both the SPAAM and INDICA projection matrices is performed. Their analysis shows that the Full INDICA setup is sensitive to errors in intrinsic parameters, particularly virtual screen distance and orientation, whereas both Recycled INDICA and SPAAM are more sensitive to eye position estimates. SPAAM results, in particular, are more sensitive to vertical and horizontal eye position error, relative to the screen.

These findings correspond to those reported by Axholt et al. [1], where SPAAM's insensitivity to depth error produces eye position estimates varying greatly in screen distance. Their analysis showed that SPAAM calibration performed with significant distance variation between screen world correspondences resulted in more consistent eye position estimates. Genc et al. [9] also analyzed errors in a number of SPAAM calibration models using reprojection error comparison. Implications from their findings suggest that simpler models, that is those comprised of fewer parameters, should result in more accurate calibration results. Quantitative analysis techniques are only capable of providing partial insight, however. Subjective analysis is still required to detect influences from the uncontrolled component in every AR system, the user themselves.

2.3 Subjective Measures in OST HMD Calibration

Axholt et al. [2] assessed the contribution of human error, in the form of involuntary postural sway during visual alignments, on calibration results. Their findings show that postural motion is highly dependent upon user stance and the type of head worn display used, and thus reducing user interaction in calibration methods is preferable. A study by Maier et al. [21], examining methods for reducing human error, focused on the effect that input methods, for recording a user's alignment response, have on contributing error to the calibration. They consider standard entry mechanisms, such as keyboard and mouse, but also vocal response and timed input. Their results indicate that the timed input method, having the user hold the alignment for a set interval, resulted in more accurate calibration results over traditional input methods.

Studies from Mcgarrity and Tang [22, 26] provide interaction methods for users to directly indicate the perceived registration of on-screen items using a stylus and tablet. Navab et al. [23] extend the functionality by allowing users to correct registration on-line. Grubert et al. [11] similarly conducted a user evaluation study of SPAAM and variants, in which subjects indicated the real world correspondence point of on-screen items using a laser pointer. Their discussion indicates that this method was time consuming for subjects to complete, however.

We developed our user evaluation method to be similar to the previously mentioned studies, by utilizing direct user feedback to determine registration accuracy. However, we limit user response to values along discrete grids or a limited number of real world objects. Our approach limits ambiguity from subject provided values and also allows independent evaluation of perceived registration in three dimensions, an aspect not considered in some of the aforementioned studies.

3 Experimental Design

Our experimental design consists of two separate tasks, through which we obtain both registration accuracy and perceived quality measures for each of three calibration methods: SPAAM, Degraded SPAAM, and Recycled INDICA. We utilize a within-subjects design, with each participant performing both tasks under each calibration type, for a total of six experimental conditions per subject.

3.1 Subjects

13 subjects, 6 male and 7 female, ranging in age from 22 to 26, participated in the user study. All subjects were students recruited through email solicitation to all departments at the university. Each participant was required to provide written consent before proceeding with the study and was provided monetary compensation for their time spent during the experiment. All subjects stated that they possessed normal or corrected-to-normal vision with no prior experience using HMDs. Subjects were provided a thorough explanation of the experimental hardware and procedure before beginning any calibration or task.

3.2 Hardware and Software

An NVIS ST50 OST HMD is used as the primary display worn by each subject. The ST50 possesses a display resolution of 1280 × 1024 with a 40° horizontal and 32° vertical field of view, and a manufacturer specified spatial resolution of 1.88 arcmn/pxl. The display utilizes pupil forming optics with a manufacturer's specified exit pupil size of 10mm. While the ST50 is a binocular HMD, subjects are allowed to view images through only the left eye piece. A piece of black opaque cloth adhered to the inside of the display, shown in Figure 2, prevents imagery on the right eye piece from being visible to the subjects.

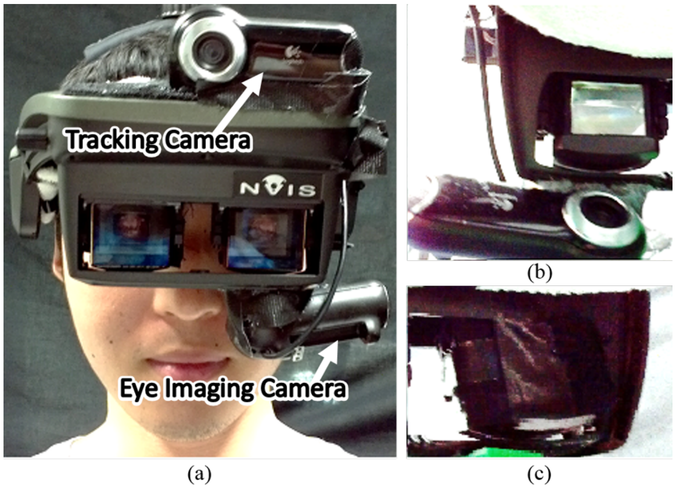

Fig. 2. (a) Front view of the HMD with mounted cameras labeled. The tracking camera is rigidly attached to the top of the HMD. Eye imaging is performed via the camera mounted beneath the left eye piece. (b) View of the eye imaging camera from within the head band of the HMD. (c) View of the right HMD eye piece covered by a patch of black opaque cloth.

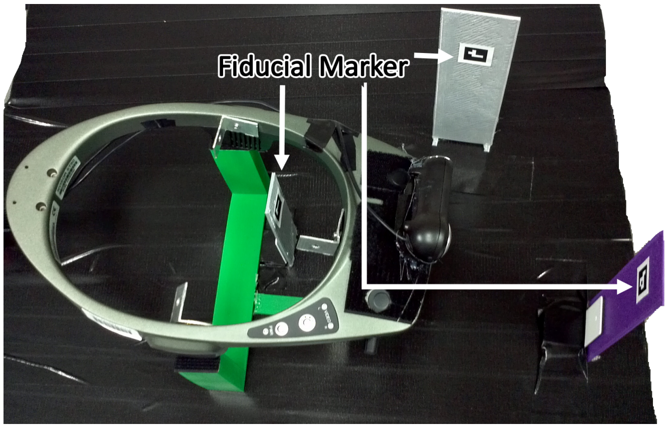

Fig. 3. Eye imaging to tracking camera calibration setup. The multi-marker configuration consists of 3 fiducial markers of known position relative to one another. The HMD assembly is positioned such that at least one marker is visible by each camera.

Logitech Quickcam Pro 9000 CMOS cameras are employed for both head tracking and eye image capturing. The location of the cameras relative to the HMD can be seen in Figure 2. The resolution of the head tracking camera is 640 × 360 with a framerate of 30 fps. The eye image camera is set to a resolution of 1280 × 720 with a framerate of 15 fps. Both cameras possess a diagonal field of view of 75°.

A multiple marker tracking configuration, shown in Figure 3, is used to determine the transformation from eye imaging to head tracking camera coordinate frames. Three fiducial markers are positionally arranged relative to one another in order to create a single coordinate frame. The HMD-camera system is then placed within the multi-marker system, such that at least one of the three markers is visible by each camera. The transformation from the eye imaging to head tracking camera coordinate frame is then calculated.

An Alienware m18 laptop, i7-4700MQ 2.4GHz processor with 16 GB RAM running Windows 7 x64, is used to drive the graphics for the display as well as execute the experiment control and eye imag-

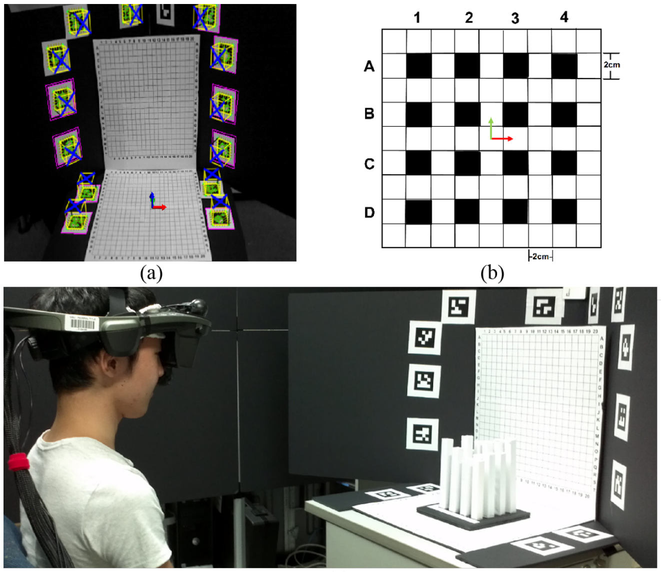

Fig. 4. Experimental task setup. (a) Multi-marker tracking coordinate frame with origin set to the approximate center of horizontal grid square J11. (b) Illustration of the Pillars configuration relative to the tracking coordinate frame. Each black square represents the location of a real pillar, with each row and column separated by 2cm. (c) Multi-marker tracking grid and task item placement relative to the subjects.

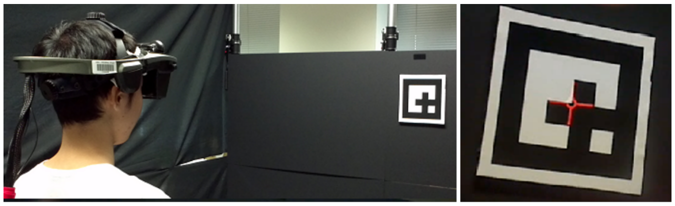

Fig. 5. (a) View of a sample participant conducting screen to world alignments. (b) View through the HMD of an on-screen cross-hair aligned with the center of the tracking marker.

ing software. Video from the cameras is captured through a USB 2.0 interface and graphics provided to the HMD via HDMI. The experiment control software is written in C++ using the Fast Light Toolkit (FLTK) for the interface components and Ubitrack [13] to manage and synchronize tracking, camera image capture, as well as the transformation and creation of on-screen objects displayed through the HMD. The eye image capturing software uses identical algorithms to that described by Itoh and Klinker [14]. Eye pose relative to the tracking camera coordinate frame is determined using the eye imaging software. SPAAM calibration is also facilitated using the same package as detailed by [14], with tracking and projection matrix creation managed by Ubitrack combined within Unity for graphics generation.

The entire experiment is conducted within a 3m×6m space enclosed by black cloth and foam core board in order to increase the contrast between the environment and the HMD visuals, as well as reduce the possibility of visual distraction by the subject during any of the experimental procedures.

3.3 Calibration Methods

Three OST HMD calibration methods are evaluated by each participant. The calibration methods are chosen for both their ease and speed to perform prior to the evaluation tasks, as well as their applicability to real world usage with current OST HMD hardware.

SPAAM: The Single Point Active Alignment Method as described in [27] is used as the control condition for the experiment. Its extensive usage and investigation in other works make it ideal for providing baseline registration results with which to compare other methods against. SPAAM calibration, for this study, is performed using 20 screen-to-world alignments. Subjects are instructed to stand and align the center of an on-screen cross-hair with the center of a fiducial marker rigidly mounted within the world. The fiducial marker itself is also used as the coordinate frame reference for the tracking camera and transformations required for the SPAAM DLT calculations. The 2D pixel coordinates of each on-screen crosshair are chosen randomly at run time, and subjects are given the option to skip cross-hairs whose locations on screen make them difficult to see. New 2D pixel locations are generated for skipped cross-hairs. In order to reduce error due to subject movement during the alignment steps, a hand clicker is provided to subjects which allows them to non-verbally indicate when an adequate screen to world alignment is achieved. Subjects activate the clicker using one or more fingers, at which point the experimenter counts backward from 3 to 0 and records the correspondence measurement. During the calibration procedure, subjects are instructed to take a number of steps forward or backward so that alignments are performed at varying distances between 1.5m to 3m from the fiducial marker. Subjects only perform the SPAAM calibration once and always at the beginning of the experiment before any tasks are started.

Degraded SPAAM: Degraded SPAAM (DSPAAM) refers to the reuse of a projection matrix obtained from a previously performed SPAAM calibration [14, 15]. This calibration method is chosen to replicate the real world condition where an HMD may shift or slip on a user's head, degrading the effects of calibration. At the start of this condition, the HMD is removed from the subject and then replaced with only minimal care to ensure the subject's left eye is within the exit pupil of the HMD and that on-screen visuals can be clearly seen. No further procedures are performed to correct any misalignment resulting from placement of the HMD. The projection matrix produced by that subject's initial SPAAM calibration is then reused to produce the on screen geometry for all tasks performed within the Degraded SPAAM condition.

Recycled INDICA: The Recycled INDICA setup, described in detail in [14, 15], comprises the third calibration method examined in this study. At the start of this condition, identically to DSPAAM, the HMD is removed from the subject and then replaced with only minimal care to ensure the subject's left eye is within the exit pupil of the HMD and that on-screen visuals can be clearly seen. Recycled INDICA utilizes intrinsic parameters obtained from decomposing an existing projection matrix, as well as an estimated distance from the user's eye to the perceived image plane of the display. A value of 1.855m, obtained using the focal distance of a camera placed within the display, is used as the distance from the subject's eye to the virtual image plane. The remaining intrinsic values are pulled from the results of that subject's initial SPAAM calibration using a standard decomposition technique. Eye imaging software is used to capture 10 images of the subject's left eye which are then processed, per the procedure outlined in [14], to estimate the location of the eye center relative to the tracking camera. The eye center is independently determined from within each of the 10 images by fitting ellipses to the detected iris and estimating the center of the spheroid produced from inverse projection of the ellipse. The resulting values are compared across the 10 images to derive a more accurate final estimate of the eye center. More detailed exposition of the algorithms can be found in Swirski [25] and Nitschke [5]. The estimated eye pose measures are then combined with the intrinsic values to produce the projection matrix used to display the virtual geometry for each task.

3.4 Tasks

We evaluate registration quality for each of the three calibration methods through two tasks. Both tasks remove ambiguity in subject selection by limiting responses to discrete values. Our selection of task design also allows for independent evaluation of user perceived registration horizontally, vertically, and in depth relative to the display screen. During each task, the subjects are seated approximately 1.5m from a

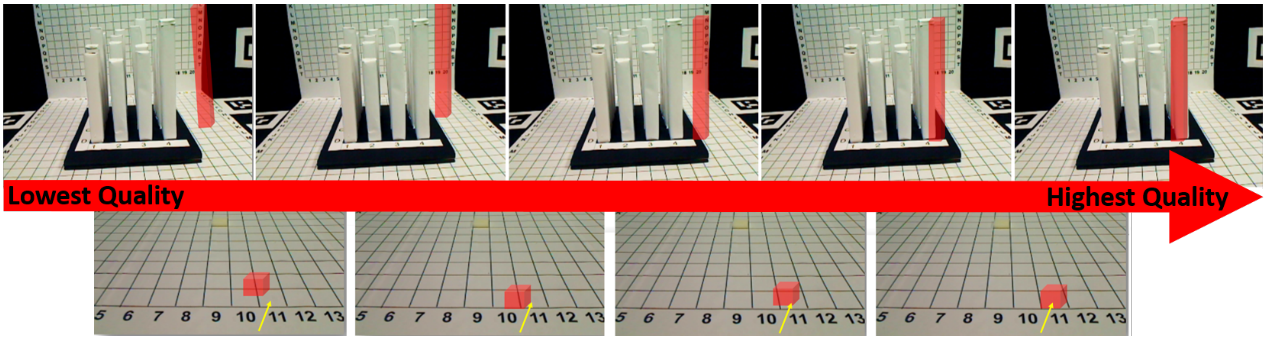

Fig. 6. Quality scale images provided to subjects prior to performing each task. Each view represents the approximate registration required for each level with quality increasing left to right along the scale. The images are only illustrations and not real views through the HMD. (Top) Quality scale for the Pillars task. The red virtual object is registered relative to the real pillar at location D4. From left to right: Quality 1, Quality 2, Quality 3, Quality 4, Quality 5. (Bottom) Quality scale for the Cubes task. The red virtual object is registered relative to grid location T11. From left to right: Quality 1, Quality 2, Quality 3, Quality 4.

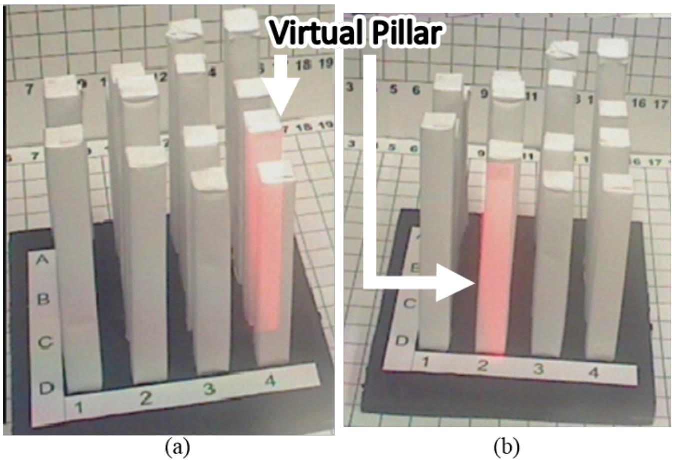

Fig. 7. The real pillar arrangement used in the Pillars task. (a) View through the HMD showing a virtual pillar aligned with the real pillar at location C4. (b) View through the HMD showing a virtual pillar aligned with the real pillar at location D2.

small table upon which the visible items for each task are placed. Subjects are allowed to move their heads and lean their bodies in any direction during the experiment but instructed to neither stand nor move the chair they are seated in during a task. Our independent measures for each task are subject perceived registration location, in terms of discrete grid coordinates, and a subjective registration quality rating based on provided quality scale images (Figure 6).

Pillars Participants are tasked with indicating which, of several, real world pillars an on-screen virtual pillar appears to be best registered with. The virtual pillar is rendered at each of the real pillar locations once, for a total of 16 trials per calibration method. Figure 7 shows the real pillar arrangement with on-screen virtual pillar, rendered in red, as it would appear during the task. During each measurement, the subject is able to freely choose any one of the sixteen real pillars they feel the virtual pillar is best aligned to. The ordering of virtual pillar locations is randomly permuted under the constraint that the next pillar location is chosen to be in both a different row and column as the previous. Heights for the real pillars cover the range 13.5 cm–19.5 cm varying by .25 cm increments. The pillars are arranged in a 4 × 4 grid such that the average height of the pillars in each row and column is between 16.25 cm–16.75 cm. The virtual pillar, displayed on-screen, is rendered such that it should appear to be a constant height of 15.5 cm. Once the virtual pillar is displayed at all sixteen real pillar locations, the task ends.

Subjects indicate their pillar selection by verbally stating the row and column designation of the desired real pillar. Visible labels along both the rows and columns of the pillar arrangement are provided, with the letters A through D denoting the rows and the numbers 1 through 4 denoting the columns, as shown in Figure 7. Subjects also verbally provide a quality rating for each trial of the task. A 1 to 5 subjective scale, with 1 denoting the worst registration and 5 denoting the best registration, are used for this metric. Before beginning the task, subjects are informed of the quality scale and provided printed images illustrating the expected visual quality that should be present at each quality level. The top row of Figure 6 shows the images provided to each user for the Pillars task.

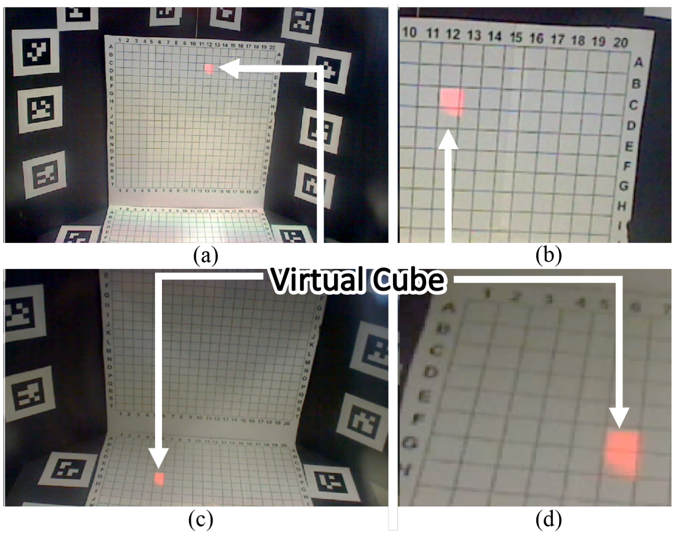

Fig. 8. Grid arrangements used for the Cubes task. (a) and (b) Views through the HMD showing a virtual cube displayed on the vertical grid at location C12. (c) and (d) Views through the HMD showing a virtual cube displayed on the horizontal grid at location G6.

Cubes Participants are tasked with indicating which, of many, grid locations a virtual cube appears to be best registered with. Two separate grids are used for this task, each comprised of 2 cm×2 cm squares in a 20 × 20 arrangement. Rows for each grid are labeled with letters from A-T and columns labeled with numbers from 1-20.

The first grid is positioned flat on the task table in front of the user and is referred to as the horizontal cubes grid. The second grid is placed perpendicular to the horizontal cubes grid so that it faces the user. This perpendicular grid is referred to as the vertical cubes grid. An array of fiducial tracking markers are placed around both grids with the origin of the tracking coordinate system aligned with the center of horizontal cubes grid square J11. The complete arrangement used for the task can be seen in Figure 8.

The virtual cube, shown on the HMD, is modeled such that its per-

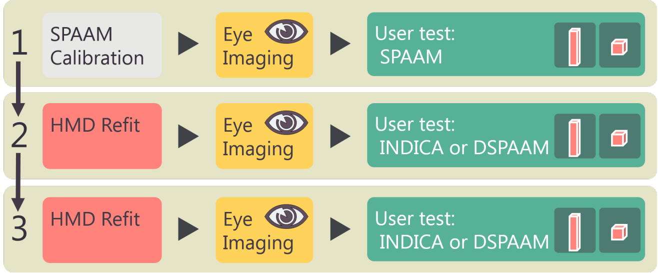

Fig. 9. Stages of the experimental procedure. Every subject performs an initial SPAAM calibration followed by the recording of eye images and performance of both tasks using the SPAAM results. The HMD is removed and refit to the subject, eye images recorded once again, and both tasks for one of the remaining conditions performed. The procedure is repeated a final time for the remaining calibration condition.

ceived size should be 2 cm×2 cm×2 cm and rendered red for increased contrast with the real environment. The virtual cube is presented at 10 grid locations on both the horizontal and vertical grid for a total of 20 trials per calibration condition. The positions of the virtual cube, on either grid, are randomly selected such that no location is repeated. The display order is chosen such that no consecutive virtual cubes will appear in the same row or column. Ordering of trials between the horizontal and vertical cubes grid locations are selected randomly, and subjects are verbally informed at the start of each trial which grid the virtual cube should appear upon. Once the virtual cube is displayed at 10 locations on both grids, the task ends.

For each of the 20 trials, subjects indicate their selection by stating the row letter followed by the column number of the grid location to which they feel the virtual cube is best aligned. Registration accuracy in the vertical direction, Y relative to the tracking coordinate system, is measured using the rows of the vertical cubes grid. Registration accuracy in depth, Z relative to the tracking coordinate system, is measured by the rows of the horizontal cubes grid. Registration accuracy in the horizontal direction, X relative to the tracking coordinate system, is measured by the columns of both grids. Subjects also verbally provide a quality value for each trial of the task. A 1 to 4 subjective scale, with 1 denoting the worst registration and 4 denoting the best registration, are used for this metric. Before beginning the task, subjects are informed of the quality scale and provided images illustrating the expected visual quality that should be present at each quality level. The bottom row of Figure 6 shows the images provided to each user.

3.5 Procedure

Both tasks are performed sequentially, though not always in the same order, for each of the three calibration methods. In order to balance against first-order residual effects, the sequence in which tasks are presented is arranged such that no subject performs the tasks, across calibration methods, in the same order. However, because the Degraded SPAAM, as well as the Recycled INDICA, calibration requires an existing projection matrix to already be accessible, the SPAAM calibration is always performed at the start of the experiment for each subject. Since the SPAAM calibration is performed first, the tasks for the SPAAM condition are also always performed before any other calibration condition, since any movement of the HMD on the subject's head would require SPAAM to be reperformed. Even though both tasks for SPAAM are conducted first, the ordering of all tasks over all three calibration conditions is never repeated between subjects. In addition, to counter any effect from the eye imaging phase required by Recycled INDICA, 10 eye images are recorded, though not used, before the start of both SPAAM and Degraded SPAAM task sets. Figure 9 illustrates the general task ordering and experiment flow for each subject.

At the start of the experiment, the subject is given a thorough explanation of the hardware and why calibration is required. The SPAAM calibration process is then described to the user with emphasis placed on the need for stable accurate screen-world alignments. The HMD is then placed onto the subject's head and adjusted so that their left eye is visually centered behind the left eye piece. The SPAAM calibration procedure as described in section 3.3 is then performed.

Once SPAAM calibration is completed, the subject is seated in front of the task table. Using the eye imaging camera, 10 images of the user's left eye are taken. The subjective quality scale for each task is reviewed and each subject is given the images shown in Figure 6. The subject then performs both experimental tasks, as described in section 3.4, using their SPAAM results. After completion of both tasks, the HMD is removed from the subject's head, and a 5 minute break allotted to give the subject ample rest before the next set of tasks.

After the 5 minute break, the HMD is replaced onto the subject's head with care only taken to ensure their left eye is positioned properly behind the eye piece and graphics are clearly seen on the display. Eye imaging is performed once again, and 10 new images of the subject's left eye taken. Both experimental tasks are repeated again for either the Degraded SPAAM or Recycled INDICA condition. The order in which subjects perform the Degraded SPAAM or Recycled INDICA calibration condition is arranged to match the previously mentioned criteria, that no subject would perform tasks in the some order across conditions. Following the completion of both tasks, the HMD is removed once again and a 5 minute respite given to the subject. Afterwards, the HMD is refit a final time and 10 more images taken of the subject's left eye. Tasks for the remaining condition, either Degraded SPAAM or Recycled INDICA, are then performed.

4 Experimental Results

We obtain our experimental results by taking the difference between the subject reported row/column positions and the actual locations where the virtual object should have appeared. The difference along a row indicates registration error in the horizontal, X, direction relative to the tracking coordinate frame, with negative error indicating a user value that is to the left of the actual. We take the difference along a column to represent error in the vertical, Y, direction for measures taken during a trial on the vertical cubes grid, with negative error indicating a user value that is below the actual. Difference along a column in both the Pillars and horizontal cubes grid trials is interpreted as error in distance, Z, relative to the tracking coordinate frame, with negative error indicating a response that is closer to the user then the actual.

We also convert the error measures from the difference in grid squares to distance measures. The size of grid squares for both grids in the Cubes task is 2cm×2cm. Thus, we equate an error of 1 to an error of 2cm in the respective direction. Similarly, the spacing of pillars in the Pillars task is 4cm, since each 2cm×2cm pillar is separated by a 2cm row or column. Therefore, we equate an error of 1 pillar to an error of 4cm in the respective direction.

The subject-provided quality values are also normalized for our analysis. Measures for both tasks are normalized to values from 1 to 4. Converting both tasks to an identical scale allows for direct and fair comparisons between tasks.

4.1 Subjective Measures

We used repeated-measures analysis of variance (ANOVA) to test the effect of the different algorithms in each experimental condition. For each test, if Mauchly's test indicated non-sphericity, we adjusted the p-value according to the Huynh-Feldt ε value; in such cases we report ε along with the ANOVA F-test. In addition, we used the Ryan REGWQ post-hoc homogenous subset test to determine how the three algorithms differed from each other, as described by Howell [12].

Figure 10 provides mean normalized quality values across subjects for each task under each calibration method. Quality values obtained for the Cubes task are shown separated by grid type, Cubes-V representing measures for the vertical cubes grid and Cubes-H representing measures for the horizontal cubes grid. ANOVA, performed on the values within each subplot of Figure 10, shows a significant main effect of calibration method in both the Pillars task $(F(2,24) = 5.03, p = 0.015)$ and the horizontal cubes grid $(F(2,24) = 6.65, p = 0.013, \varepsilon = 0.71)$. The vertical cubes grid condition shows no significant difference between calibration methods $(F < 1)$. The normalized quality values also show that subjects report Recycled INDICA registrations,

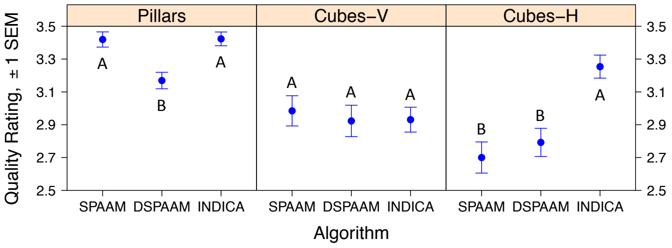

Fig. 10. Mean subjective quality values for each calibration method during each task, normalized to a 1–4 scale with 1 denoting the lowest quality and 4 the highest. The values shown are across subjects with individual plots for the Pillars task as well as each grid of the Cubes task. Cubes-V shows normalized quality for the vertical cubes grid. Cubes-H shows normalized quality for the horizontal cubes grid. Means with the same letter, within each plot, are not significantly different at $p \leq 0.05$ (Ryan REGWQ post-hoc homogeneous subset test).

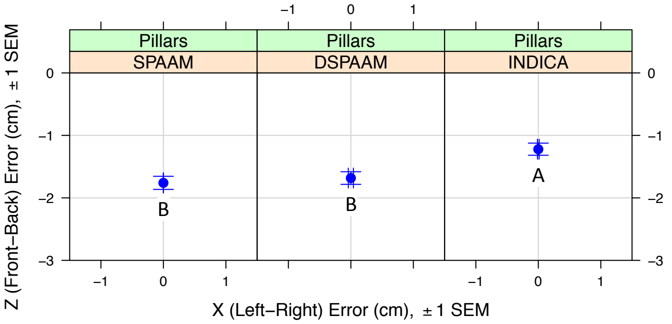

Fig. 11. Mean Pillars task error along the X (Left-Right) and Z (Front-Back) direction relative to the tracking coordinate frame. 0 indicates no error. Error is reported as a distance value, with every 4 cm of error equating to a 1 pillar location difference in the respective direction. Means with the same letter are not significantly different at $p \leq 0.05$ (Ryan REGWQ post-hoc homogeneous subset test).

viewed on the Pillars and the horizontal cubes grid, to be of higher quality over Degraded SPAAM. While SPAAM quality is rated nearly equal to Recycled INDICA in the Pillars task, it rates lowest in horizontal cubes grid trials. All three calibration methods produce nearly identical quality ratings across subjects in vertical cubes grid trials[1].

Registration Error in Pillars Task Figure 11 provides the registration error results for the Pillars task converted into distance measures. Error in both the X, Left-Right, and Z, Front-Back, directions relative to the tracking coordinate frame are provided. ANOVA of X dimension error shows no significant main effect due to calibration method ($F < 1$). Results show subject perceived error is near perfect, 0 error, along the X direction. All three calibration methods produce error in the Z direction, however, with subjects perceiving the registration of virtual objects to be closer than intended for every case. Recycled INDICA results show a shift in distance perception away from the user and closer to the correct location. ANOVA also indicates a highly significant effect of calibration method $(F(2,24) = 14.011, p < 0.001)$ along the Z direction.

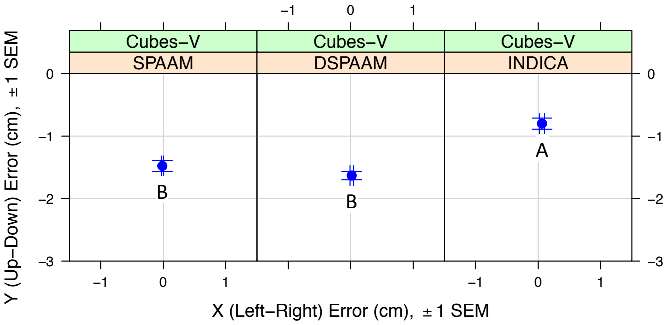

Fig. 12. Mean vertical cubes grid task error along the Y (Up-Down) and X (Left-Right) direction relative to the tracking coordinate frame. 0 indicates no error. Error in each direction is reported as a distance value, with every 2 cm of error equating to a 1 grid square location difference in the respective direction. Means with the same letter are not significantly different at $p \leq 0.05$ (Ryan REGWQ post-hoc homogeneous subset test).

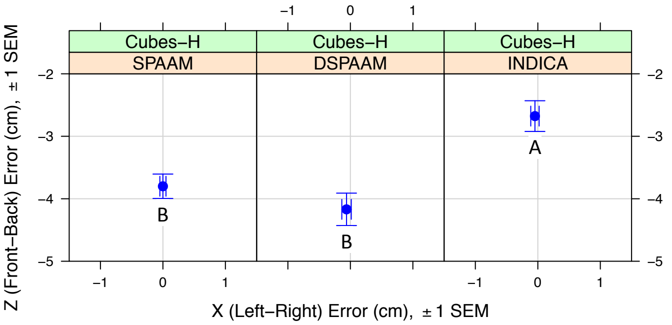

Fig. 13. Mean horizontal cubes grid task error along the Z (Front-Back) and X (Left-Right) direction relative to the tracking coordinate frame. 0 indicates no error. Error in each direction is reported as a distance value, with every 2 cm of error equating to a 1 grid square location difference in the respective direction. Means with the same letter are not significantly different at $p \leq 0.05$ (Ryan REGWQ post-hoc homogeneous subset test).

Registration Error in Cubes Task Figures 12 and 13 show the registration error results for the Cubes task separated by each grid. Results for trials conducted on the horizontal cubes grid are shown in Figure 13 and provide error in both the X, Left-Right, and Z, Front-Back, directions relative to the tracking coordinate frame. ANOVA performed along each direction shows a significant main effect of calibration method along the Z direction $(F(2,24) = 7.37, p = 0.003)$, with no effect along the X $(F < 1)$. All three calibration methods produce equally, near 0, error along the X direction and Recycled-INDICA produces the least Z error.

Results for trials conducted on the vertical Cubes grid are shown in Figure 12 and provide error in both the X, Left-Right, and Y, Up-Down, directions relative to the tracking coordinate frame. ANOVA shows no main effect of calibration method on results along the X direction $(F < 1)$. A main effect of calibration method is detected along the Y direction $(F(2,24) = 10.96, p = 0.0016, \varepsilon = 0.75)$. Similar to the Pillars and horizontal cubes grid, all three calibration methods produce near 0 errors along X. Y error is less under the Recycled INDICA condition, with both SPAAM and Degraded SPAAM resulting in similar error amounts.

4.2 Quantitative Measures

We perform our quantitative analysis in a similar fashion to Itoh and Klinker [14], considering two metrics: Extrinsic eye position estimates

[1] In addition to ANOVA, we also performed the non-parametric Friedman and Kruskal-Wallis rank sum tests for subjective judgements. The Friedman rank sum test shows a significant main effect for the Pillars $(\chi^2(2) = 5.45, p = 0.066)$ and Cubes-H $(\chi^2(2) = 13.06, p = 0.0015)$ tasks, but not for the Cubes-V task $(\chi^2(2) = 0.15)$. The Kruskal-Wallis rank sum test also shows a significant main effect for the Pillars $(\chi^2(2) = 18.92, p < 0.001)$ and Cubes-H $(\chi^2(2) = 21.21, p < 0.001)$ tasks, but not for the Cubes-V task $(\chi^2(2) = 0.98)$. In contrast to ANOVA, the Friedman rank sum test looses power by ignoring large portions of data; reducing either 624 (Pillars) or 390 (Cubes-H, Cubes-V) data values into 39 to conduct the test. The Kruskal-Wallis rank sum test deviates from ANOVA by producing a large amount of power, because it does not model the within-subjects design of the data. Nevertheless, the interpretation of the results remains the same, regardless of the analysis method used.

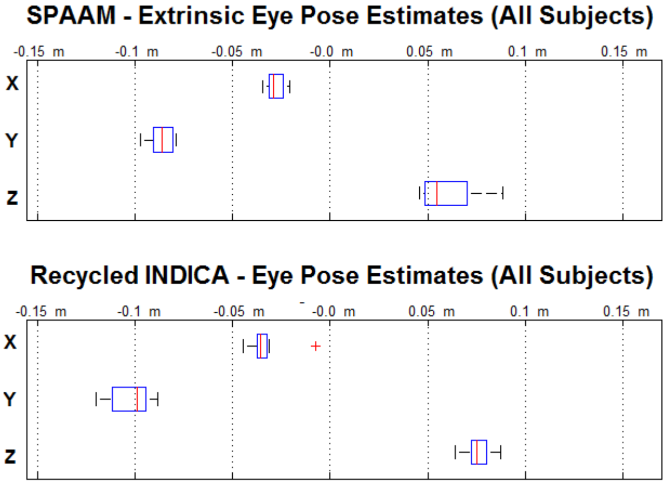

Fig. 14. Eye position estimates across subjects for SPAAM and Recycled INDICA. Axis are relative to the display screen, with X along the horizontal and Y along the vertical screen direction. Positive Z is away from the display screen toward the user. All values are in meters.

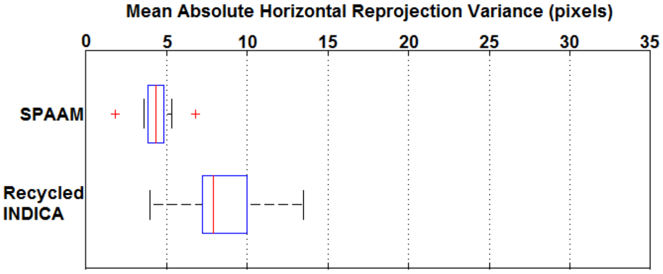

Fig. 15. Absolute reprojection variance in horizontal screen space for SPAAM and Recycled INDICA. Error values are the difference in pixels between the correct 2D screen coordinate used for each alignment of the SPAAM procedure and the reprojection of the corresponding 3D world point using the calibration result. Error values are in pixels.

and reprojection of the SPAAM screen to world correspondence pairs. We, therefore, do not consider the Degraded SPAAM condition in the quantitative analysis, since it used the same projection matrix as the SPAAM condition.

Quantitative Eye Position Estimates Figure 14 provides a comparison of eye position estimates for both SPAAM and Recycled INDICA. The plots show the mean and variance of values across all 13 subjects. We denote the axis positions relative to the display screen, with X along the horizontal and Y along the vertical screen direction. The Z axis is distance from the display screen toward the user.

Eye estimates for the SPAAM calibration are derived by decomposing the projection matrix result to obtain the extrinsic values. We use the eye imaging results to produce eye position estimates for Recycled INDICA. Our results show that both SPAAM and Recycled INDICA produce similar eye position estimates. Values along the Z direction are less varied, similar across all subjects, using the Recycled INDICA eye imaging method. SPAAM produces less varied positions in Y, and both methods perform similarly in estimating eye position along X.

Quantitative Reprojection Values We use the projection matrix results from both SPAAM and Recycled INDICA to reproduce the 2D-3D correspondence pairs, from each subject's SPAAM calibration. Every 3D point from each SPAAM alignment is reprojected into screen space using the projection matrix acquired from the SPAAM and Recycled INDICA results. We then calculate error by taking the difference in pixel location between the result of reprojection and the actual 2D screen location used during the SPAAM alignment. Figures 15 and 16 show the reprojected pixel errors for SPAAM and Recycled INDICA across all subjects. Our results show that Recycled INDICA produces much larger reprojection errors, compared to the true values, in vertical screen space. We also find that SPAAM produces lower horizontal reprojection error as well. Our reprojection error analysis does not take into account movement of the HMD between the SPAAM and Recycled INDICA procedures, however.

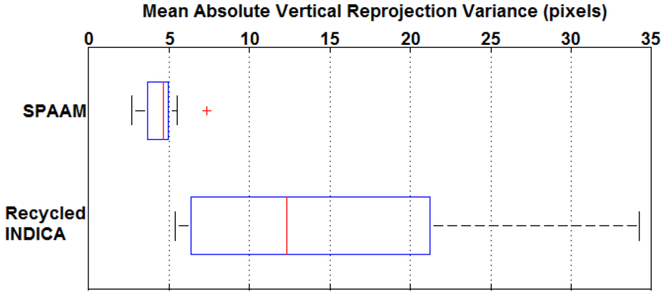

Fig. 16. Absolute reprojection variance in vertical screen space for SPAAM and Recycled INDICA. Error values are the difference in pixels between the correct 2D screen coordinate used for each alignment of the SPAAM procedure and the reprojection of the corresponding 3D world point using the calibration result. Error values are in pixels.

5 Discussion

We begin our discussion with a review of the quantitative results, starting with comparison of eye position estimates. Our eye position estimates for SPAAM, Figure 14, show larger variance relative to frontal screen distance. These findings match closely with those from previous studies [1, 2, 14]. Similarly, our eye estimates along Z for Recycled INDICA show smaller variance compared to SPAAM. Our findings differ from Itoh and Klinker's [14], however, in that we find larger variance horizontally relative to the display, for Recycled INDICA compared to SPAAM. An explanation for this difference lies in the fact that our analysis considers eye estimates taken from 13 subjects, whereas Itoh and Klinker's results derive from a single user. It is reasonable to infer that the eye location of each subject differed from the others, and the larger variance in eye locations should be expected. It is possible that the variance may be due in part to inaccuracies in the eye position estimation algorithms used in INDICA. However, significant error in the eye location estimates should have resulted in poor performance of INDICA, which is not the case for our findings. We believe that another possible influence on the eye position estimates, for both methods, comes from the optical design of our HMD used in the experiment. The NVIS ST50 incorporates pupil forming lenses, meaning that images are only visible within a certain volume behind the display screen, a property which may have influence on the position of the HMD relative to the user's eyes. We verified that subjects could clearly see all portions of the display screen before each task, thus each subject's eye would have been placed within the exit pupil viewing volume. A follow-up study, using an HMD with non-pupil forming optics is required to identify this influence.

Our reprojection estimates from Figures 15 and 16 also differ from those in Itoh and Klinker's analysis [14]. Their results indicate that Recycled INDICA should produce errors with similar variance to SPAAM. However, our analysis shows that Recycled INDICA reprojection error is higher in vertical screen space. Similar to the eye position values, this result may also be a product of the data being from 13 unique subjects, all of which reported no prior familiarity with HMDs. This is in contrast to [14, 15] in which all results are from performance of a single expert user. The sensitivity analysis reported in [15] also shows that Recycled INDICA results are influenced most heavily by the values pulled from the existing calibration result. Thus, it is reasonable to conclude that the SPAAM performance of our subjects contributed to the reprojection error we see in our findings. [15] does note that sensitivity to parameters is system dependent and we did not perform a similar system sensitivity analysis.

A final consideration, and the logical cause of the discrepancy between our reprojection analysis and that performed by [14], is the removal and replacement of the HMD on the subject between the SPAAM and Recycled INDICA calibrations. In [14], the eye imaging step was conducted during the SPAAM calibration itself. Since both eye position estimates are taken in the same location relative to the screen, it makes sense that their reprojection errors for both methods are similar. The 2D-3D correspondences in our experimental design, however, are only relative to the subject's eye position during the SPAAM condition. Refitting the display, before the Recycled INDICA trials, will have shifted the 2D pixel locations used in the SPAAM alignments to different locations relative to the user's eye. Therefore, it is not valid to compare the reprojection of points taken from one eye location with those produced relative to another, and we cannot use the results of the reprojection error as a prediction of registration quality. Due to this issue, the use of similar reprojection error analysis in future studies should be avoided, unless the movement of the screen between conditions is considered and corrected for in the calculations.

Our subjective results show a more accurate depiction of the registration accuracy produced by each calibration method. Surprisingly, all three calibration techniques produce registrations that are perceived as being closer to the subject than intended. This may be a bi-product of the poor Z eye estimates produced by SPAAM and its influence on Recycled INDICA. We can see from Figure 13, though, that the updated eye estimates for Recycled INDICA had a correcting effect on perceived registration distance. Since we also restricted users to viewing images through only the left eye piece, it is possible that the lack of stereo depth cues influenced the underestimation of registration location. A follow-up study should extend the eye imaging required by INDICA to both eyes and compare subjective results against a Stereo-SPAAM implementation to include any advantages offered by binocular HMD's. Subjects judged the registered location of objects using Recycled INDICA, in not only distance but also in the vertical field of view, as being much closer to the correct values. It is also interesting to note that all three calibration techniques produce nearly perfect registration in the horizontal direction. It is yet unclear whether this correlates to the similar eye location estimates in the X direction seen in Figure 14. It is also possible that the object position in the X direction may have been easier to isolate due to the availability of multiple viewing angles, from subjects leaning forward, backward, and sideways during the tasks. Subsequent experimental setups should endeavor to isolate the dimensional constraints more thoroughly in order to reduce possible interaction between the perceived object locations along each axis.

An additional item of note is the difference in significance produced by the ANOVA analysis between the subjective and objective results. The subjective quality values for trials on the vertical cubes grid (Figure 10) show no significant difference even though the objective measures along the Y direction (Figure 12) show significance. A similar result can be seen for quality values on the horizontal cubes grid and objective measures along the X direction. This discrepancy between the two measures is due primarily to the inclusion of both directions for the quality values, whereas the objective plots show results for each direction in isolation. The experimental design did not facilitate the recording of independent qualities for each direction of the grid, and, therefore, it must be inferred that the quality evaluations are based on the perceived registration along the X and Y direction together. This inconsistency highlights the requirement of a more rigorous evaluation method capable of fully detaching measures for each direction.

Our analysis of the subjective measures also show that subjects felt the overall quality of the Recycled INDICA registrations are higher in comparison to SPAAM and Degraded SPAAM, Figure 10. We can safely presume that the higher subjective quality given to Recycled INDICA directly correlates to the higher registration accuracy observed in the tasks. This implies that non-expert users rely heavily on perceived registration location for information, an important item of consideration for AR designers. This result further declares the need for accurate OST calibration methods accessible to a wide user base.

6 Conclusions and Future Work

Our experimental design can be improved upon, for future studies, in a number of ways. The subjective quality scales, in our design, are different for each task. In best practice, these scales should cover identical ranges in order to facilitate direct comparison without the need for normalization. The SPAAM calibration procedure should also use a recording method similar to that in Maier et al. [21], having the subject hold the alignment for some time. This would help reduce error and alleviate any concerns of movement due to our hand clicker. Further development of the task design is also needed to more thoroughly isolate the measures taken in each trial to values along a single axis. This will reduce any perceptual interactions caused by object position judgments made in reference to multiple directions. A final consideration must also be made to the number of subjects. Given that the aim is to evaluate calibration performance among novice users, increasing the number of subjects will be critical to ensure the findings adequately represent the broader demographics of the common populous.

We have shown that in a registration dependent task, the Recycled INDICA OST HMD calibration method produces registration that is both more accurate and of subjectively higher quality than common calibration techniques. An additional surprising finding is that Recycled INDICA produces registration that is perceived as more accurate in depth over SPAAM registration. It can be further noted that the performance of Recycled INDICA will degrade far slower than that of interaction dependent methods. This inherent robustness is required for any calibration method to achieve wide acceptance. However, even though the Recycled INDICA performance was high, our results also show it was not completely without error. Further development of the technique is required to reduce the introduction of error from intrinsic parameters taken from previous calibrations. We also suggest a follow-up study using a binocular HMD and Stereo-SPAAM implementation be conducted. This investigation would confirm if the increased performance of Recycled INDICA in depth persists over other methods when stereoscopic depth cues are available, and whether the tendency for perceiving virtual objects as closer would be corrected. A drawback to implementing INDICA, though, is the need for eye imaging hardware. A large number of the currently available OST HMD's are not factory equipped with the required cameras, and thus it is up to the investigator to suitably mount the necessary equipment. However, due to the increasing availability and decreasing cost of miniature imaging devices, it is reasonable to conceive that HMD manufacturers will have the ability to incorporate eye focused cameras into future designs.

Our quantitative analysis has also identified possible problems using a reprojection error metric to predict registration quality. The goal of the INDICA methods is to allow a user to put on an HMD without the need to perform additional alignments for registration correction. Therefore, it is reasonable to conclude that the display will not be reworn in an identical manner each time. On-screen pixel locations will, therefore, change relative to the new eye location of the user. Any analysis examining reprojection error should consider this and compensate for the shift in the calculations.

While the current Recycled INDICA method relies on previous results for intrinsic parameters, the Full INDICA setup, outlined by Itoh and Klinker, is a further step toward a fully user independent OST HMD calibration method. We plan to conduct a similar experimental examination, to the one we present in this work, to compare the subjective results of Full versus Recycled INDICA. If the Full INDICA procedure is able to produce results similar to our findings of the recycled setup, then it would be an ideal candidate for use in a continuous calibration technique. Development of a closed-loop calibration procedure, able to perform constant registration correction and account for display shift during use, would significantly expedite the broader acceptance of OST AR. Removing the user interaction requirements will make the use of OST HMDs more accessible to the general public, and also to investigators within the AR community as a whole.

Acknowledgments

This material is based upon work supported in part by the National Science foundation (NSF) and the Japan Society for the Promotion of Science (JSPS) through an East Asia and Pacific Summer Institutes Fellowship, award IIA-1414772, to K. Moser. Funding was also provided through the NASA Mississippi Space Grant Consortium, to K. Moser, as well as NSF awards IIS-1320909 and IIS-1018413, to J. E. Swan II. The project also received funding from the European Union's Seventh Framework Programme for Research and Technological Development under PITN-GA-2012-316919-EDUSAFE awarded to Y. Itoh. The authors would like to especially thank Dr. Hirokazu Kato and Dr. Goshiro Yamamoto, the members of the Interactive Media Design Lab and staff of the Nara Institute of Science and Technology, as well as all of the anonymous subjects.

References

[1] M. Axholt, M. A. Skoglund, S. D. O'Connell, M. Cooper, and S. R. Ellis. Parameter estimation variance of the single point active alignment method in optical see-through head mounted display calibration. In *Proceedings of IEEE Virtual Reality*, pages 27–34, 2011.

[2] M. Axholt, M. A. Skoglund, S. Peterson, M. D. Cooper, T. B. Schon, F. Gustafsson, A. Ynnerman, and S. R. Ellis. Optical see-through head mounted display direct linear transformation calibration robustness in the presence of user alignment noise. In *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 54, pages 27–34, 2010.

[3] R. Azuma. Predictive tracking for augmented reality. *PhD thesis*, 1995.

[4] A. Boud, D. Haniff, C. Baber, and S. Steiner. Virtual reality and augmented reality as a training tool for assembly tasks.

[5] A. N. C. Nitschke and H. Takemura. Corneal imaging revisited: An overview of corneal reflection analysis and applications. In *Information and Media Technologies*, volume 8, page 389, 2013.

[6] T. Caudell and D. Mizell. Augmented reality: An application of heads-up display technology to manual manufacturing processes. In *In 25th Hawaii International Conference on Systems Sciences*, pages 659–669, 1992.

[7] J. L. Gabbard and J. E. S. II. Usability engineering for augmented reality: Employing user-based studies to inform design. *IEEE Transactions on Visualization and Computer Graphics*, 14(3):513–525, May/June 2008.

[8] Y. Genc, F. Sauer, F. Wenzel, M. Tuceryan, and N. Navab. Optical see-through hmd calibration: A stereo method validated with a video see-through system. In *In Proceedings of IEEE and ACM International Symposium on Augmented Reality*, pages 165–174, 2000.

[9] Y. Genc, M. Tuceryan, and N. Navab. Practical solutions for calibration of optical see-through devices. In *In Proceedings of the 1st International Symposium on Mixed and Augmented Reality*, pages 169–175, 2002.

[10] S. J. Gilson, A. W. Fitzgibbon, and A. Glennerster. An automated calibration method for non-see-through head mounted displays. *Journal of Neuroscience Methods*, 199(2):328–335, 2011.

[11] J. Grubert, J. Tuemler, R. Mecke, and M. Schenk. Comparative user study of two see-through calibration methods. In *Proceedings of IEEE Virtual Reality*, pages 269–270, March 2010.

[12] D. C. Howell. In *Statistical Methods for Psychology*. Duxbury, Pacific Grove, CA, 2002.

[13] M. Huber, D. Pustka, P. Keitler, F. Echtler, and G. Klinker. A system architecture for ubiquitous tracking environments. In *Proceedings of the 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 1–4, November 2007.

[14] Y. Itoh and G. Klinker. Interaction-free calibration for optical see-through head-mounted displays based on 3d eye localization. In *IEEE Symposium on 3D User Interfaces*, pages 75–82, march 2014.

[15] Y. Itoh and G. Klinker. Performance and sensitivity analysis of indica : Interaction-free display calibration for optical see-through head-mounted displays. In *IEEE International Symposium on Mixed and Augmented Reality*, Sept. 2014.

[16] A. L. Janin, D. W. Mizell, and T. P. Caudell. Calibration of head-mounted displays for augmented reality applications. *IEEE Virtual Reality Annual International Symposium*, pages 246–255, Sept. 1993.

[17] H. Kato and M. Billinghurst. Marker tracking and hmd calibration for a video-based augmented reality conferencing system. *In Proceedings of the 2nd IEEE and ACM International Workshop on Augmented Reality*, pages 85–94, 1999.

[18] G. Klinker, D. Stricker, and D. Reiners. Augmented reality: A balancing act between high quality and real-time constraints. In *Mixed RealityMerging Real and Virtual Worlds*, pages 325–346, 1999.

[19] M. A. Livingston and Z. Ai. The effect of registration error on tracking distant augmented objects.

[20] M. A. Livingston, L. J. Rosenblum, S. J. Julier, D. Brown, Y. Baillot, J. E. Swan, J. L. Gabbard, and D. Hix. An augmented reality system for military operations in urban terrain.

[21] P. Maier, A. Dey, C. A. L. Waechter, and C. Sandor. An empiric evaluation of confirmation methods for optical see-through head-mounted display calibration. In *IEEE International Symposium on Mixed and Augmented Reality*, pages 267–268, October 2011.

[22] E. Mcgarrity, Y. Genc, M. Tuceryan, C. Owen, and N. Navab. A new system for online quantitative evaluation of optical see-through augmentation. In *Proceedings of IEEE and ACM International Symposium on Augmented Reality*, pages 157–166, October 2001.

[23] N. Navab, S. Zokai, Y. Genc, and E. Coelho. An on-line evaluation system for optical see-through augmented reality. In *In Proceedings of IEEE Virtual Reality*, pages 245–246, 2004.

[24] C. B. Owen, J. Zhou, A. Tang, and F. Xiao. Display-relative calibration for optical seethrough head-mounted displays. In *Third IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 70–78, 2004.

[25] L. Swirski, A. Bulling, and N. Dodgson. Robust real-rime pupil tracking in highly off-axis images. In *Eye Tracking Research and Applications*, pages 173–176, March 2012.

[26] A. Tang, J. Zhou, and C. Owen. Evaluation of calibration procedures for optical see-through head-mounted displays. In *Proceedings of the 2nd IEEE/ACM International Symposium on Mixed and Augmented Reality*, page 161, 2003.

[27] M. Tuceryan and N. Navab. Single point active alignment method (spaam) for optical see-through hmd calibration for ar. In *IEEE and ACM International Symposium on Augmented Reality*, pages 149–158, October 2000.