A Survey of Calibration Methods for Optical See-Through Head-Mounted Displays

Jens Grubert (*Member, IEEE*), Yuta Itoh (*Member, IEEE*), Kenneth Moser, and J. Edward Swan II (*Senior Member, IEEE*)

Abstract—Optical see-through head-mounted displays (OST HMDs) are a major output medium for Augmented Reality, which have seen significant growth in popularity and usage among the general public due to the growing release of consumer-oriented models, such as the Microsoft Hololens. Unlike Virtual Reality headsets, OST HMDs inherently support the addition of computer-generated graphics directly into the light path between a user's eyes and their view of the physical world. As with most Augmented and Virtual Reality systems, the physical position of an OST HMD is typically determined by an external or embedded 6-Degree-of-Freedom tracking system. However, in order to properly render virtual objects, which are perceived as spatially aligned with the physical environment, it is also necessary to accurately measure the position of the user's eyes within the tracking system's coordinate frame. For over 20 years, researchers have proposed various calibration methods to determine this needed eye position. However, to date, there has not been a comprehensive overview of these procedures and their requirements. Hence, this paper surveys the field of calibration methods for OST HMDs. Specifically, it provides insights into the fundamentals of calibration techniques, and presents an overview of both manual and automatic approaches, as well as evaluation methods and metrics. Finally, it also identifies opportunities for future research.

Index Terms—augmented reality, head-mounted displays, optical see-through calibration

1 Introduction

Augmented reality (AR) is an interactive, real-time technology, which gives the user the sense that virtual objects exist among real objects, in the physical world. For example, the user might see a virtual glass sitting next to a physical glass on a tabletop. A major goal of AR is for the location of the virtual glass to appear as equally real, solid, and believable as the physical one.

In this paper, we refer to this concept as *locational realism*. We contrast locational realism with the better-known term *photorealism*, which is the traditional computer graphics goal of rendering objects and scenes that are visually indistinguishable from reality. In AR, the primary goal may not be to render the glass photorealistically, but we are usually interested in the locational realism of the glass—while it may obviously be a cartoon glass, with incorrect illumination and color, we still want its location to be perceived in a manner that is indistinguishable from the location of the physical glass.

In order for any degree of locational realism to be possible, the AR system must know the 6-degree-of-freedom (6DoF) *pose*—the position (x, y, z) and orientation (*roll*, *pitch*, *yaw*)—of the virtual rendering camera within the physical world. From this information, the system can determine which 2D screen pixels will be required to display a virtual object at a corresponding 3D location (Robinett and Holloway [73]). The more accurately this pose can be known, the greater the locational realism.

The rendering camera's pose is typically measured using a *tracking system*, which needs to be calibrated in order to report accurate pose estimates. It is possible for the tracking system to directly use a physical video camera within the AR system (Kato and Billinghurst [42]); otherwise, the tracking system tracks a fiducial that is attached to the AR system. In this latter case, even though the tracking system needs to report the pose of the AR system's rendering camera, the tracker instead reports the pose of the fiducial. This leads to the additional requirement that a secondary *calibration* needs to be performed, which yields the transformation between the tracked fiducial and the rendering camera.

In addition, there are two major ways of displaying AR content. In video see-though AR (VST AR), the user sees the physical world mediated through a video camera within the AR system. The system receives a constant stream of image frames from the real world, and combines virtual content to these frames. VST AR can be used with standard video monitors, handheld devices such as tablets or phones, as well as opaque, VR-style head-worn displays, also referred to as Mixed Reality (MR) displays. In contrast, optical see-through AR (OST AR) gives the user a view of the physical world directly, while virtual objects are simultaneously imposed into the user's view through optical combiners. OST AR is almost always accomplished through a transparent head-worn display; although microscopes (Edwards et al. [11]) and other optical devices are also possible. While both forms of AR have their respective strengths (and weaknesses), as well as various applications (Billinghurst et al. [8]), this paper focuses on OST AR.

Although in VST AR it is possible to use a single camera for both the video stream and the tracking camera (Kato and Billinghurst [42]), this is never possible in OST AR, because the "video stream" comes from the user's eye. Instead, in OST AR the pose of the head-worn display is tracked, and the AR system needs to know the transformation between

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2017.2754257, IEEE Transactions on Visualization and Computer Graphics

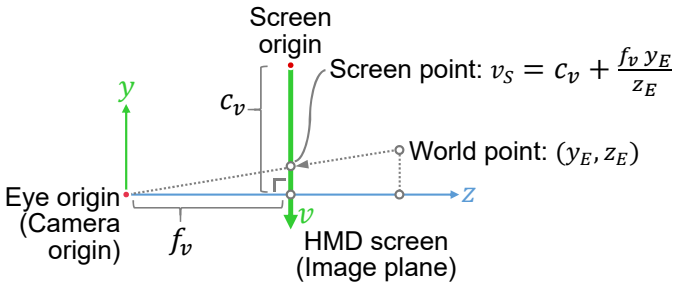

Fig. 1: The y-z plane of an off-axis pinhole camera model. See also Fig. 2.

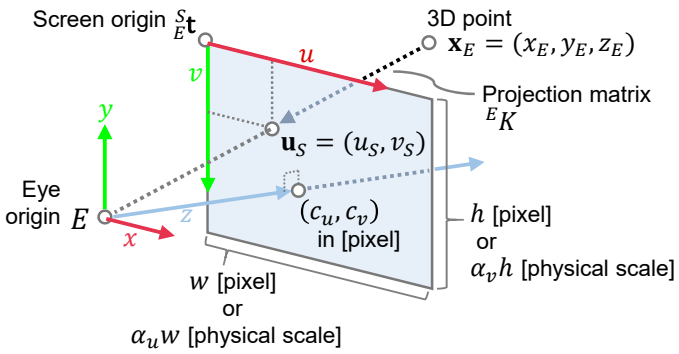

Fig. 2: A 3D representation of the image plane, and related intrinsic properties of the pinhole camera model.

the display and the user's eyes. Therefore, in OST AR a calibration procedure is always necessary.

This paper surveys and summarizes calibration procedures published until September 2017. First, it provides an overview of calibration fundamentals for head-mounted OST AR displays. It then presents an overview of calibration methods, which are categorized according to *manual*, *semi-automatic*, and *automatic* approaches. Next, it discusses how these calibration methods have been evaluated as well as the metrics used for analysis. Finally, the paper concludes by discussing opportunities for future research.

2 Fundamentals

2.1 Nomenclature

Through the paper, we use the following nomenclature. Lower-case letters denote scalar values, such as a focal length f_u. Upper-case letters denote coordinate systems, such as an eye coordinate frame E. Lower-case bold letters denote vectors, such as a 3D point in eye coordinates $\mathbf{x}_E \in \mathbb{R}^3$, or a 2D image point $\mathbf{u} \in \mathbb{R}^2$. Upper-case typewriter letters denote matrices, such as a rotation matrix $\mathtt{R} \in \mathbb{R}^{3\times3}$.

We now define a 6DoF transformation from one coordinate system to another. Given coordinate systems A and B, we define the transformation from A to B by $({}^{A}_{B}\mathtt{R}, {}^{A}_{B}\mathbf{t})$, where ${}^{A}_{B}\mathtt{R}$ is a rotation matrix, and ${}^{A}_{B}\mathbf{t}$ is a translation vector. For example, we can transform $\mathbf{x}_A$, a 3D point in A, from A to B by

$$\mathbf{x}_B = {}^{A}_{B}\mathtt{R}\mathbf{x}_A + {}^{A}_{B}\mathbf{t}. \tag{1}$$

2.2 The Off-axis Pinhole Camera Model

In computer vision, the *intrinsic matrix* $\mathtt{K} \in \mathbb{R}^{3\times3}$ defines the projection transformation from 3D to 2D coordinate spaces. The elements of this matrix describe the properties of the pinhole camera, and its derivation is well described in a plethora of academic texts and research publications [12], [24], [30], [51], [79], [80].

Readers desiring to gain a complete and thorough understanding of the physical and mathematical principles behind projection, transformation, or computer graphics in general are encouraged to read the cited publications. Nevertheless, here we provide a brief overview, with the goal of enhancing the reader's understanding of the eye-HMD transformation.

The eye-HMD system is commonly modeled as an off-axis pinhole camera. We define its intrinsic matrix as:

$${}^{E}\mathtt{K} = \begin{bmatrix} f_u & 0 & c_u \\ 0 & f_v & c_v \\ 0 & 0 & 1 \end{bmatrix}. \tag{2}$$

The parameters of ${}^{E}\mathtt{K}$ are derived directly from the pinhole camera model illustrated in Figures 1 and 2. The focal distances f_u and f_v denote the distances between the imaging plane and the camera center. In the ideal pinhole camera model, the f_u and f_v components from Equation (2) are identical, meaning the pixels of the image are perfectly square.

For example, given a 3D point in the eye coordinate system $\mathbf{x}_E$, the point is projected to a 2D point $\mathbf{u}_S$ in the HMD screen space S by

$$\mathbf{u}_S = {}^{E}\mathtt{K}\mathbf{x}_E. \tag{3}$$

In practice, however, we first obtain $\mathbf{x}_E$ as the 3D point $\mathbf{x}_H$, in the HMD coordinate system[1]. Therefore, we first transform $\mathbf{x}_H$ into $\mathbf{x}_E$ by

$$\mathbf{x}_E = {}^{H}_{E}\mathtt{R}\mathbf{x}_H + {}^{H}_{E}\mathbf{t}, \tag{4}$$

where the rotation matrix ${}^{H}_{E}\mathtt{R} \in \mathbb{R}^{3\times3}$, and the translation vector ${}^{H}_{E}\mathbf{t} \in \mathbb{R}^3$, represent a transformation from the HMD coordinate system H, which is attached to the HMD, to the user's eye's coordinate system E.

By integrating this transformation into the camera model ${}^{E}\mathtt{K}$, we obtain a 3×4 projection matrix ${}^{H}_{E}\mathtt{P}$, from the display (HMD) coordinates to the user's eye's coordinates:

$${}^{H}_{E}\mathtt{P} = {}^{E}\mathtt{K}\begin{bmatrix} {}^{H}_{E}\mathtt{R} & {}^{H}_{E}\mathbf{t} \end{bmatrix} \in \mathbb{R}^{3\times4}. \tag{5}$$

Figure 3, top left, is another illustration of these coordinate systems.

Therefore, all calibration methods must be able to produce ${}^{H}_{E}\mathtt{P}$, either by solving for all of the matrix components at once, or by systematically determining the parameters in Equation (5).

Generally, when solving for all of the components of ${}^{H}_{E}\mathtt{P}$ at once, the most common approach is the direct linear transformation (DLT) [1], [23], [74]. This method estimates

1. The HMD coordinate system is typically defined by an inside-out looking camera or an outside-in looking tracking system that determines the pose of a fiducial

${}^{H}_{E}\mathtt{P}$ by solving a linear equation, which is constructed from a minimum of 6 3D-2D correspondences. Given the linear solution as an initial estimate, a non-linear optimization method, such as Levenberg-Marquardt [23], [59], can then be applied.

2.3 Modeling the Intrinsic Matrix of OST HMDs

While most rendering engines used for computer graphics presume ${}^{H}_{E}\mathtt{P}$ as an ideal pinhole camera, in physical implementations, this model may be unequal as a result of distortion, imperfections on the imaging plane, non-uniform image scale, etc.. In this case an alternative model using a single focal length value and the image aspect ratio may be more appropriate [69].

The "principal axis", in this alternative, lies perpendicular to the imaging plane and extends to the aperture. The intersection of the principal axis and the imaging plane occurs at the "principal point". Ideally, the principal point would occur at the origin of the image coordinate system. However, when this is not the case, the parameters c_u and c_v, illustrated in Figure 2, represent the offset from the origin. τ represents a skew factor when the axes of the image plane are not orthogonal, which would produce an image plane resembling a parallelogram instead of a rectangle or square.

When the camera is located at, and is orthogonal to, the origin of the 3D coordinate space, then the transformation of objects into the camera frame of reference is implicit. However, should the camera move to another viewing location in the world, as is often the case, then an extrinsic transformation is required to transform the coordinates of the objects in the world into the camera frame. This transform is the $({}^{H}_{E}\mathtt{R}, {}^{H}_{E}\mathbf{t})$ component of Equation (5), referred to as the extrinsic component.

The ${}^{H}_{E}\mathtt{R}$ describes the rotation of the camera with respect to the world coordinate axes and the ${}^{H}_{E}\mathbf{t}$ denotes the translational offset from the origin along the X, Y, Z cardinal directions. This transformation, with respect to OST HMD calibration, represents the transformation of the tracked coordinate frame of the HMD relative to the eye. Unfortunately, the location of the user's optical center, or alternatively the nodal point, is not easily determined at run-time. Nonetheless, given the extrinsic and intrinsic parameters, calculation of the 12 values in the final camera projection matrix ${}^{H}_{E}\mathtt{P}$ in Equation (5) is achieved through simple matrix multiplication.

In an OST HMD system, we can further break down the intrinsic matrix by using the position of the virtual screen of an OST HMD with respect to the eye.

Given an eye tracker T attached on an HMD with a pose $({}^{H}_{T}\mathtt{R}, {}^{H}_{T}\mathbf{t})$ from the HMD to the tracking camera, we can get the position of the eye with respect to the screen as ${}^{S}_{E}\mathbf{t}$ if we also know the pose of the screen $({}^{S}_{H}\mathtt{R}, {}^{S}_{H}\mathbf{t})$ (Figure 2 and Figure 3 bottom right).

Assuming that we know that position as the translation vector ${}^{S}_{E}\mathbf{t} = [x, y, z]^{\mathrm{T}}$ then the intrinsic matrix in Equation (2) can be defined as the following [34], [35] (Figure 3 bottom right):

$$ {}^{E}\mathtt{K} = \begin{bmatrix} \alpha_u & & c'_u \\ & \alpha_v & c'_v \\ & & 1 \end{bmatrix} \begin{bmatrix} z & & -x \\ & z & -y \\ & & 1 \end{bmatrix}, \tag{6} $$

where S is the virtual screen coordinate system and $\mathbf{a} = [\alpha_u, \alpha_v]^{\mathrm{T}}$ is the scaling factor that converts 3D points on the screen to pixel points. $c'_u = (w-1)/2$ and $c'_v = (h-1)/2$ define the image center with the pixel width w and height h of the screen.

Note that ${}^{S}_{E}\mathbf{t}$ is dependent on the current position of the user's eye with respect to the display, thus the intrinsic matrix varies when the display is repositioned on one's head.

If we know an intrinsic matrix based on an old eye position as ${}^{E_0}\mathtt{K}$, we can *update* the projection matrix as follows (Figure 3 bottom left):

$$ {}^{E}\mathtt{K} = {}^{E_0}\mathtt{K} \begin{bmatrix} 1 + z'/z & & -x'/z \\ & 1 + z'/z & -y'/z \\ & & 1 \end{bmatrix}, \tag{7} $$

where ${}^{E_0}_{E}\mathbf{t} = [x', y', z']^{\mathrm{T}}$ is a translation from the old eye position to the new eye position.

Equation 6 (full setup) does not rely on knowledge about a previous eye position ${}^{H}_{E_0}\mathbf{t}$. Instead, it requires the virtual screen pose $({}^{H}_{S}\mathtt{R}, {}^{H}_{S}\mathbf{t})$ and the scaling vector $\mathbf{a}$ $[\mathrm{pixel/meter}]$. On the other hand, Eq. 7 (recycled setup) does not rely on these parameters, except for $[{}^{H}_{S}\mathbf{t}]_z$, because it reuses the old intrinsic matrix ${}^{E_0}\mathtt{K}$.

2.4 Estimating Projection

Unfortunately, it is rarely, if ever, possible to explicitly possess the exact intrinsic and extrinsic parameters for a specific HMD and user configuration at run time. Therefore, OST HMD calibration procedures often utilize manual user interaction techniques in order to produce an approximation, or estimate, of the final projection matrix parameters. Initial manual calibration modalities, for example, adapt existing computer vision camera calibration mechanisms, which utilize pixel to world correspondences for determining the viewing parameters. These adapted techniques do not obtain all correspondences at once, as would be possible in an image captured from a camera, but instead reduce the strategy to simple bore-sighting through which each separate correspondence is recorded in sequence [9], [45]. The correspondence data obtained from this process includes both the 2D pixel location of the on–screen reticle and the 3D location of the physical alignment point. Values from multiple alignments can then be combined into a system of linear equations describing the projection of the 3D point into the 2D space and solved using standard methods in the context of DLT discussed above. The solution to this linear equation system is the complete set of parameters describing the projection matrix, or virtual camera, from Equation (5).

The bore–sighting schema though, forces a number of requirements, including placement of the HMD such that the user's view is perpendicular to the display screen and that the user is able to reliably align the on-screen indicator with a high level of precision. In order to satisfy these conditions, the user's head must be rigidly secured, preventing

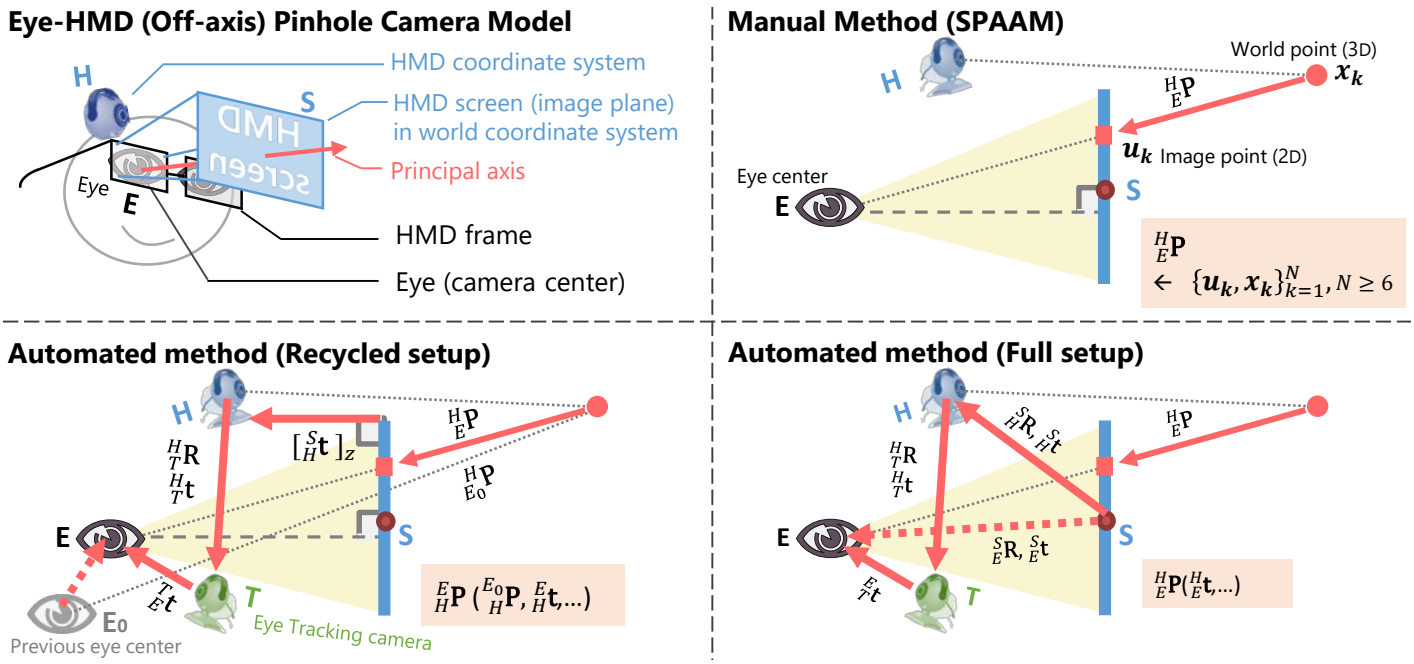

Fig. 3: Illustration of pinhole camera projection. (top left) coordinate systems of an eye-HMD system. (top right) Manual method. (bottom left) Automated method with the recycled setup. (bottom right) Automated method with the full setup.

movements which may shift the display screen or disrupt the alignment process. Inhibition of user movement makes this methodology not only uncomfortable and tedious, but also impractical for use outside of a laboratory setting. Successive iterations and adaptations have fortunately enabled a relaxation of the fixation constraint by affording a compromise with other requirements as well.

Within the next sections we will discuss the evolution of approaches targeted at estimating these intrinsic and extrinsic parameters, as well as methods which propose calibration models which diverge from the pinhole camera model.

3 Manual Calibration Methods

This section summarizes the methods where the calibration requires the human operator to perform manual tasks. The top portion of Table 1 summarizes these methods, and for each method Figure 5 gives a key thumbnail image.

Calibration through parameter estimation has yielded a number of procedures that rely upon user interaction to collect 3D-2D correspondences by manually aligning a world reference point to 2D points displayed on the screen of an OST HMD. Azuma and Bishop [7] propose estimating the extrinsic virtual camera parameters by manually aligning virtual squares and a cross-hair with a wooden box. For estimating the field-of-view of the virtual camera the user must simultaneously align two virtual lines with the box edges.

Janin et al. [39] proposed two methods to estimate parameters associated with the external sensor, a virtual screen and user-specific parameters (eye location): direct measure of the parameters and optimization-based. As the authors note, it is difficult to accurately measure relevant parameters, and, hence recommend parameter estimation through optimization. Similar, to Azuma and Bishop [7] they propose to align a virtual cursor with a physical registration device with known geometry. The authors note, that the joint estimation of the 17 parameters of their model is susceptible to noise but do not quantify their calibration results.

Oishi et al. [64] use an elaborate "shooting gallery" calibration setup, which presents LEDs fixed on a large plate 0.5 to 4 meters from the user, and a predetermined virtual projection model, which matches the physical calibration environment and the physical HMD. The head of the user has to be fixed during the procedure in the world coordinate origin, and virtual points have to be aligned with physical targets manually. The calibration process has the following steps: First, the HMD is positioned at the world coordinate origin. Then a physical calibration pattern has to be matched with a virtual target indicator (using a control joystick, 13 times per eye). If the recorded matches are below a predetermined threshold the system is calibrated. If the mismatch is too large, further correspondences are collected and the projection model is updated.

Tuceryan and Navab [78] introduced SPAAM (Single Point Active Alignment Method), as an improvement to the manual alignment data collection scheme (Figure 4). They propose collecting individual 2D-3D point correspondences, one at a time, and then solving for all projection parameters simultaneously. To this end, the user is asked to align a single 2D symbol (e.g., cross or circle) with a 3D object. Both the HMD and the 3D object are spatially tracked. After collecting sufficient correspondences (at least 6), the correspondences are used to create and solve a system of linear equations according to the DLT method introduced in section 2 [1], [23], [74]. The biggest advantage of SPAAM is its weak requirements of hardware: needing only a tracking system for calibration to be done by a user ad-hoc. Thus

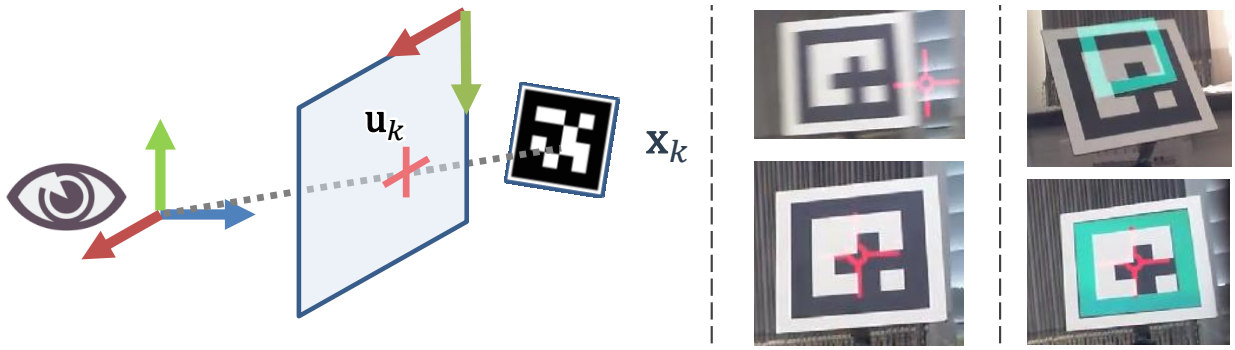

Fig. 4: Data collection in SPAAM. Left: A single 2D point u_k is manually aligned with a 3D point x_k. Middle: Ego-centric view through an OST HMD aligning a virtual 2D cross hair with a 3D tracked marker. Right: Green virtual square overlaid on the physical marker before and after the calibration.

SPAAM can easily be integrated into most OST HMD applications.

Unfortunately, the manual procedure inevitably induces human related errors during the data collection, due to imperfect alignments from user posture [3] and input actions [52]. Furthermore, these errors may be sufficiently high, for users not familiar with SPAAM, to render the calibration a failure, requiring the need to repeat the procedure multiple times. Despite the potential drawbacks, SPAAM has proved to be a popular and influential calibration method, onto which a number of improvements to the original approach have been proposed.

Instead of performing a completely new calibration every time a user puts on an HMD, Genc et al. [18] proposed Two-Stage SPAAM (SPAAM2), which reuses existing calibrations. Their process works as follows: Initially, all 11 parameters (extrinsics + intrinsics) are estimated. If the user removes the HMD and then later puts it back on, only a subset of those parameters are re-estimated. The intrinsics of the virtual camera are assumed to not change over time, only the position of the virtual camera center (i.e., the position of the eye's nodal point relative to the display screen). Therefore, linear scale and shift parameters are estimated, which correct for a potential image shift and scale change due to the new projection center.

Using their updated model, the user only needs to collect two point correspondences. However, the justification of SPAAM2 is that the 3D shift of the virtual camera center can be modeled by a linear scale and shift transformation on the image plane. Their assumption is rather "redefining" the scale and position of the display's image plane under the assumption that the orientation of the plane stays the same. We elaborate the theory behind this in Section 4.

When using a vision-based inside-out camera system for 3D tracking, Genc et al. [17] also propose to avoid computing the pose between the external camera system and 3D object directly. Instead, and under the assumption that the camera and HMD are rigidly attached to each other, they present a formulation that uses the projection matrix of the inside-out tracking camera to estimate the projection matrix of the virtual camera. Unfortunately, they do not present results that are significantly better than the base algorithm (SPAAM + explicit pose computation).

Fuhrmann et al. [14] propose to determine the parameters of the virtual viewing frustum by collecting 8 2D-3D point correspondences per eye that define the viewing frustum corner points. They propose to further reduce the number of needed point correspondences to two per eye for adopting the projection for individual users. This can be achieved under the assumption that the projection of the 3D point intersects the virtual image plane at a known distance. Now, only the eye position has to be determined and the user only needs to provide point samples for two opposite corners of the display. For distortion correction, the authors propose to fallback on a camera-based detection of a distorted line pattern (or alternatively let the user specify many points of intersecting lines).

Another data collection scheme was proposed by Kellner et al. [43] in 2012. They propose to "aim" at a distant 3D target with another handheld target, resulting in 2D point-to-3D line correspondences. Subsequently, they first determine the display rotation and translation, and then the focal length and principal point. While the method results in a shorter acquisition time compared to SPAAM, it also results in larger calibration errors.

Instead of aiming with the head, several approaches proposed to move a handheld target instead. O'Loughlin and Sandor [65] proposed to use a handheld marker for alignment, and Moser et al. [62] investigated the contextual impact of user-centric tracking markers, finding that simple stylus alignment is preferable to finger tracking for 3D point input. The latter approach is also employed in the calibration of the Microsoft HoloLens HMD[2].

In contrast to SPAAM, Grubert et al. [20], [21], propose to collect multiple 2D-3D point correspondences simultaneously in a Multiple Point Active Alignment scheme. Here, the user aligns a grid of 9 3D points aligned on calibration board placed at a distance of about 150 cm. While the calibration procedure significantly speeds up the data collection phase, it also results in larger calibration errors.

The manual calibration techniques for Optical See-Through calibration has also seen application in other HMD domains types, particularly to head-mounted projective dis-

2. Please note, that the Microsoft HoloLens does not offer a complete user-based OST calibration procedure, but solely determines the interpupilary distance - see https://developer.microsoft.com/en-us/windows/mixed-reality/calibration - last accessed September 2nd, 2017.

plays (HMPDs). Hua et al. developed a HMPD and accompanying calibration procedures [15], [28]. Their calibration relies on Tsai's calibration technique [77]. Hence, for data collection, they show a printed grid 14x13, on which the user has to align a virtual cross. This procedure is repeated at least two times. The biggest disadvantage of using the Tsai method is that, in practice, a large number of point correspondences have to be collected and that for at least two grid poses. For example, in the work of Hua et al. the authors conducted the calibration with 10 different grid poses resulting in 14x13x10 = 1820 correspondences, which needed to be aligned [28]. The authors argue, that the data collection could be automated by placing a camera in the exit pupil [15]. However, this could lead to additional errors as the camera position during calibration is not identical with the eye position during use. SPAAM was also used for other custom HMD designs, such as [83], in which a natural feature tracking target was used for collecting 27 point correspondences.

In 2016, Jun and Kim [41] proposed a calibration method for stereo OST-HMDs equipped with a depth camera. They presented a simplified HMD-eye model assuming collimated displays with no focal length and perceptual pinhole centers for the eyes (i.e. the perceptual eye projection). Their model solves for the extrinsic parameters of the depth camera, the interpupillary distance of the user and the position of the users' eyes. The authors claim that a full calibration can be achieved with 10 point correspondences (collected by pointing with a finger on a 2D circle). After initial full calibration, only the user parameters (interpupillary distance, eye position) are estimated in subsequent simplified calibrations.

In 2017, Zhang et al. [81], [82] proposed a dynamic SPAAM method that considers eye orientation to optimize the projection model. Their method, RIDE (region-induced data enhancement), splits the user's FoV into 3-by-3 segments and update the main projection matrix to adapt the shift of the eye center (the nodal point) due to eye orientation.

While manual calibration methods can achieve accurate results, the burden on users in terms of time and workload can be substantial. Empirically, we found that many users would calibrate an HMD (at most) once per work session or when the display is used for the first time. The calibration process is of an open-loop nature, requiring substantial effort from the users to successfully complete the calibration task. Hence, the need for automated, closed-loop methods arises, which will be discussed in the following section.

4 Automating Calibration

Unlike those manual calibrations reviewed in Section 3, some works propose (semi-)automatic calibration methods. A common idea behind these automated methods is to formulate an OST HMD system as the combination of a display model and an eye model. Indeed, a projection matrix from SPAAM implicitly models the system as a planar display screen floating in midair, with the eye center position relative to the screen. This leads to an off-axis pinhole camera model as discussed in Section 2. Given an OST HMD model, these automated methods measure the parameters on-line, and/or estimate them prior to the actual calibration. Overall, these methods simplify the calibration procedure.

4.1 Semi-Automatic Calibration Methods

This section introduces methods, which by estimating a subset of parameters in separate calibration stages, attempt to minimize the number of point correspondences that need to be manually collected by the human operator. The middle portion of Table 1 summarizes these methods, and for each method Figure 5 gives a key thumbnail image.

In early works, Genc, Tuceryan, and Navab [18] and Navab et al. [63] developed the Easy SPAAM method, which updates an old projection matrix from a previous SPAAM calibration using a simple manual adjustment. This method assumes that the matrix change can be modeled with a 2D warping of the screen image, including scaling. Therefore, fewer parameters are needed, and users need to collect only two or more 2D-3D correspondences.

After Easy SPAAM, Owen et al. [66] propose Display-Relative Calibration (DRC). Their work is one of the first attempts to explicitly split an OST HMD system into a display model and an eye model. In DRC, the authors proposed a two-step calibration process. They first create an off-line calibration for the display model using a mechanical jig, and then propose 5 different options for the on-line estimation of the eye model. The options involve varying degrees of simplifying assumptions, ranging from not performing any on-line calibration, to performing a Easy SPAAM-like simple warping, and finally to a full 6 DoF eye pose estimation.

Similar to Owen et al. [66], Gilson et al. [19] propose replacing the user's eye with a camera and exploit established camera calibration techniques for determining the virtual camera parameters. They differ from Owen's work in that they do not need measurements conducted by a human operator, instead they take images directly through the HMD optics. They also found that no user adaption was needed for their calibration techniques.

In 2013, Makibuchi et al. [55] proposed a vision-based robust calibration (ViRC) method. It first uses a view-point camera for off-line parameter estimation. Then, a camera attached to the HMD tracks a fiducial marker as the user aligns the marker with a crosshair on the screen. Using the correspondences, the perspective-n-point algorithm (PnP) optimizes both offline and on-line parameters at the same time. They find that, compared to the direct linear transform (DLT) method, the ViRC method requires fewer user input trials, and achieves up to 83% more accurate reprojection errors.

The methods proposed so far, allow human operators to lower the number of point correspondences required for a successful calibration. However, this partially comes at the cost of separate and elaborate calibration phases, which can require additional hardware such as cameras or a calibration rig [19], [55], [66].

4.2 Automatic Calibration Methods

Finally, this section covers methods which attempt to completely free the human operator from having to manually perform any calibration procedures. The bottom portion of

TABLE 1: Overview of calibration methods. **MC**: minimum number of 2D–3D correspondences. **Parameters**: estimated parameters. **Alignment Mode**: **FIX**: fixed head or camera-rig, **H**: through head-movement, **F**: through finger or hand movement. **Data Collection**: **i**: individually (1 correspondence at a time), **m**: multiple correspondence at once. Figure 5 shows representative thumbnail images from each method.

	Method	MC	Parameters	Alignment Mode	Data Collection
Manual Methods	Azuma and Bishop [7]	8 points + 4 lines	eye location, FoV	H	points i, lines m (2)
	Oishi and Tachi [64]	13 per eye	eye location	FIX	i
	SPAAM (DLT): Tuceryan et al. [78] Genc et al. [17]	min 6	projection matrix (full)	H (F in [62], [65])	i
	SPAAM2 / EasySPAAM: Genc et al. [18] Navab et al. [63]	2	scale, shift	H	i
	Tsai [77]: Hua et al. [28] Gao et al. [15]	14×13 × (2..10) = 364˜1820 [28]	all intrinsics + extrinsics	FIX [15] or F [28]	m [15] or i [28]
	Fuhrmann et al. [14]	8 (full) 2 (update)	all intrinsics + extrinsics (full), eye position (update)	F	i
	Kellner et al. [43]	5	all intrinsics + extrinsics	H+F	i
	Jun and Kim [41]	3 (full), 2 (update) (10 and 5 recommended)	offline only: extrinsic orientation of depth camera, interpupillary distance, eye positions	F	i
	Zhang et al. [81], [82]	9 regions x 6 samples = 54 (3x3 grid with standard SPAAM)	multiple projection matrices for individual regions (full), single projection matrix for update	H	i
	MPAAM: Grubert et al. [20], [21]	6	projection matrix (full)	H	m
Semi-automatic	DRC: Owen et al. [66]	offline: 20 online: 1	offline: all intrinsics + extrinsics + radial distortion + spherical aberration. online: eye location, focal length	offline: FIX, online: H	offline: m, online: i
	Gilson et al. [19]	30	all intrinsics + extrinsic	FIX	m
	Makibuchi et al. [55]	offline: 4 online: 4	offline: virtual screen pose and approximate eye location. online: current eye location	offline: FIX, online: H	m
Automatic	Priese et al. [70]	not applicable (NA)	eye location	NA	NA
	INDICA: Itoh and Klinker [34], [35]	offline: 6 online: NA	offline: projection matrix (full) or virtual screen pose. online: eye location	offline: FIX or H	offline: i or m, online: NA
	CIC: Plopski et al. [67]	online: 2 × 3 frames	offline: display screen pose and eyeball parameters. online: eye location	FIX	m
	Figl et al. [13]	unknown	offline only: eye location, focal length	FIX	m

Table 1 summarizes these methods, and for each method Figure 5 gives a key thumbnail image.

Luo et al. [50] developed an on-axis camera for eyeglass-like OST HMDs, which in theory eliminates the need for manual calibration. However, because of the optical design's small size, the camera must be placed 20 mm behind the user's eye location, which can lead to registration errors at close distances.

In 2007, Priese et al. [70], after an initial full calibration, proposed estimating the eye location using eye tracking. However, they tested their approach only using static images of an eye and did not verify the system with actual users.

Figl et al. [13] presented a method for the determination of focal lengths and eye location for a binocular medical HMD (Varioscope M5), using a fully automated setup, including stepping motors for changing the distance of a calibration pattern. However, after this initial camera-based calibration, they do not consider the calibration of the user's actual eye position.

In 2014, Itoh and Klinker [34] proposed the INteraction-free DIsplay CAlibration (INDICA) method, which utilizes an eye-tracker installed on an OST HMD. Their method measures the eye center on-line and automatically generates a projection matrix. They use the same pinhole camera model as SPAAM. The display parameters are decomposed

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2017.2754257, IEEE Transactions on Visualization and Computer Graphics

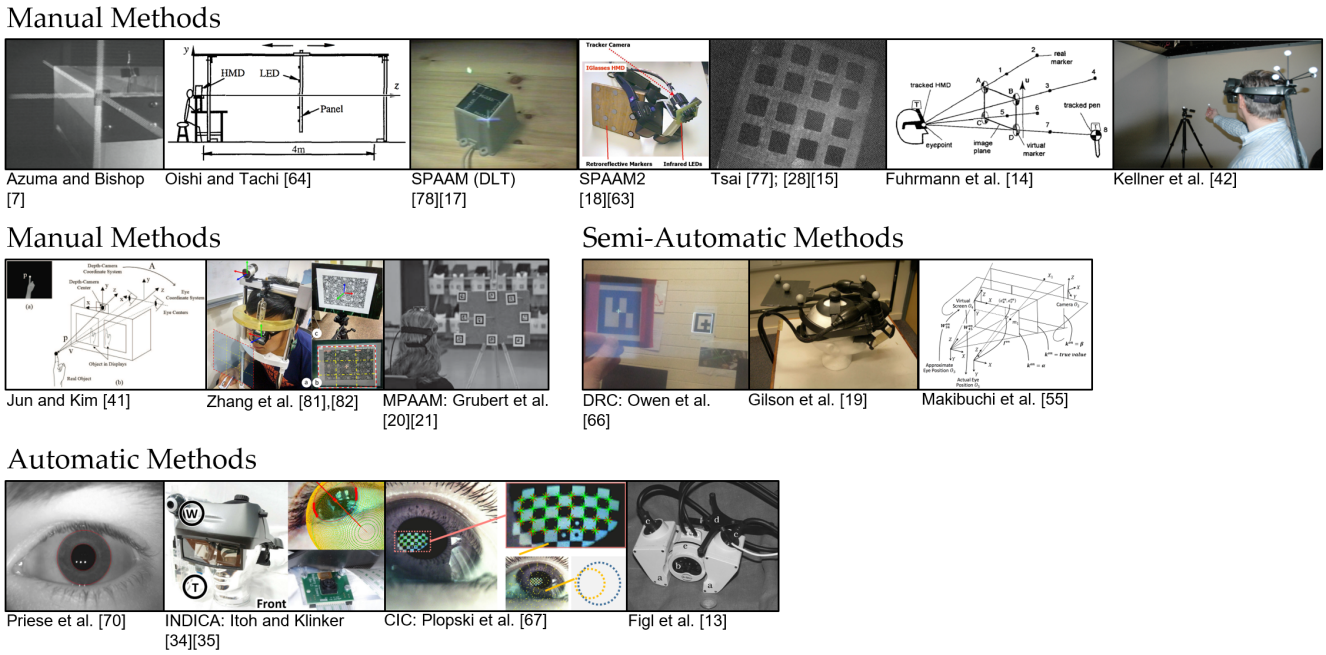

Fig. 5: Thumbnail images from therows of Table 1.

from a projection matrix, which is obtained from a SPAAM calibration performed once off-line beforehand (Itoh and Klinker [35]). Their follow-up work evaluated INDICA with display parameters calibrated off-line via a camera, which means the method operates totally without the need of additional user input [34].

In Section 3, we mentioned that the assumption of SPAAM2 (Genc et al. [18]) leads to a different interpretation. Based on this assumption, we get

$$ {}^{H}_{E}\mathtt{P}' = \underbrace{\begin{bmatrix} \alpha'_x & & c'_x \\ & \alpha'_y & c'_y \\ & & 1 \end{bmatrix}}_{{}^{E}\mathtt{K}'} {}^{H}_{E}\mathtt{P}' \tag{8} $$

$$ = ({}^{E}\mathtt{K}'\,{}^{E}\mathtt{K}) \begin{bmatrix} {}^{H}_{E}\mathtt{R} & {}^{H}_{E}\mathbf{t} \end{bmatrix}, \tag{9} $$

where ${}^{E}\mathtt{K}'$ denotes the scale and shift parameters. This means that SPAAM2 redefines the screen parameter matrix as ${}^{E}\mathtt{K}'\,{}^{E}\mathtt{K}$. Since the screen parameters should stay the same, this interpretation is incorrect. An implicit assumption of SPAAM2 is that only the eye center position changes, which actually leads to Eq. 7 instead. And, the three parameters ${}^{E_0}_{E}\mathbf{t}$ could be estimated via two 2D-3D data correspondences.

Eye Models: Plopski et al. [67] propose another automated method: Corneal-Imaging Calibration (CIC). Unlike INDICA, which uses an iris-based method for eye-tracking, CIC estimates the eye position by utilizing a reflection of an image on the cornea of a user's eye—an effect known as the corneal reflection. In CIC, a fiducial pattern is displayed on an HMD screen, and an eye camera captures its corneal reflection. CIC then computes the rays that are reflected on the eye cornea and pass through corresponding display pixels. Given the 3D pose of the display in the HMD coordinate system, the diameter of the cornea sphere under the dual circle eye model, and a minimum of two rays, the method computes the position of the corneal sphere of the eyeball. Then, given three corneal sphere positions while the eyeball is rotating, CIC estimates the 3D center of the eyeball. This eye position estimation, based on the reflected features and a simplified model of the eyes structure, yields more accurate 3D localization estimates than direct iris detection.

However, the 3D eye model that both INDICA and CIC use can be improved. The model assumes that the eyeball can be schematically modeled as two intersecting 3D spheres, where the first sphere models the spherical part of the eyeball that consists of the sclera, and the second sphere models the cornea curvature. Under this model, the optical center of the eye camera is assumed to be located at the center of the sclera (eyeball) sphere. However, the nodal point of the eye—the point where light rays entering through the pupil will intersect—would be a more appropriate location for the optical eye center (c.f. Jones et al. [40]).

Display Models: Most of the methods we mentioned so far treat the image screen of an OST HMD as a planar panel. However, this model ignores the fact that the combining optics could distort the incoming light rays before they reach the eye, in a manner similar to corrective glasses. This distortion can affect both the virtual image of the display (the *augmented view*), as well as the view of the real world as seen through the combining optics (the *direct view*).

For correcting the augmented view, Lee and Hua [47] propose a camera-based calibration method, that learns a corrective 2D distortion map in screen image space. For correcting the direct view, Itoh and Klinker [36] propose modeling the distortion as the shift in a bundle of 4D light rays (light field) passing through the optics, and then estimating a 4D-to-4D mapping between the original and distorted light fields. Because it uses light fields, this method can handle viewpoint-dependent distortions. Itoh and Klinker [37] then extend this method to correct

distortions of the augmented view. Their evaluation with an OST HMD shows that removing both direct-view and augmented-view distortions provides overall registration accuracy comparable to 20/50 visual acuity.

Beyond the distortion estimation addressed by Itoh and Klinker [36], the same authors [32] further propose modeling the view-dependent color aberration (point spread function) of OST HMDs. This method models the image blur as Gaussian functions integrated in the 4D-to-4D distortion mapping, and estimate it by measuring the impulse response of the display from different viewpoints.

Summary: Clearly, automatic calibration methods are the future of OST HMDs. In addition to freeing the human operator from having to manually perform calibration procedures, automatic methods could also operate in a closed-loop manner, continuously adjusting the calibration, and, therefore, correcting for small movements of the HMD on the user's head. In addition, integrating eye trackers into an OST HMD allows many useful interaction techniques, such as gaze-based interaction, and also allows optimized rendering methods, such as foveated rendering. However, as discussed, automatic calibration methods still face challenges, especially related to both eye models and display models.

5 Evaluation

It is, of course, important to evaluate calibration procedures. However, especially compared to video see-through AR, evaluation in an OST HMD is particularly challenging, because, in the end, only the user can assess the locational realism of the result. This section summarizes existing evaluation methods. Specifically, investigations have examined various data collection schemes, which have considered the presence or absence of postural sway, the effects of confirmation methods on dependent variables such as reprojection error, intrinsic and extrinsic parameter estimation, task completion time, and workload. Table 2 summarizes these methods.

In 2000, Genc et al. [16] evaluated a stereo calibration method briefly with two users, but used a video-see-through system. Hence, these results are not easily transferable to optical see-through systems.

In 2001, McGarrity et al. [57], [58] presented a method for providing registration accuracy feedback, where, using a stylus on a tablet, users indicate perceived positions of virtual objects. They also propose using both 3D input points and their projected 2D point correspondences to adjust the calibration, but they do not provide an accuracy analysis of their approach. Navab et al. [63] later applied the idea to personal digital assistants (PDAs), and suggested using a point-and-shoot game to motivate users to complete calibration tasks.

In 2003, Tang et al. [75] compared 4 variants of SPAAM: SPAAM, DepthSPAAM (modified SPAAM to collect different depth values by moving the whole body relative to the 3D target), Stereo-SPAAM, and Stylus-Mark calibration (DepthSPAAM using a tracked stylus). They focus on task completion time and geometric error (measured using the procedure described in McGarrity et al. [57], [58]), and presented results for the decomposed principal point. They found that SPAAM resulted in the fastest task completion time but had the largest calibration error, while the Stylus-mark calibration had the lowest error. However, the authors also note that "none of the four procedures can achieve a reliable and accurate result for naive users".

In 2008, Grubert et al. [20], [21] compared a Multiple Point Active Alignment scheme (MPAAM) with SPAAM, and found that although the MPAAM calibration procedure significantly speeds up the data collection phase, it also results in larger calibration errors.

In 2010, Axholt et al. [4] used Monte-Carlo simulation to investigate the effects of human alignment noise on view parameter estimation. Compared to a camera on a tripod, which can be perfectly still, a standing human will exhibit involuntarily postural sway, even if they attempt to stand perfectly still. They found that the relatively large alignment noise induced by humans (>5px), compared to the lower alignment error typically reported in the computer vision literature for camera calibration (ca. 1px), primarily led to estimation variance in the extrinsic parameters along the user's line of sight (z direction). To mitigate this effect, they found that distributing the 3D correspondence points over a greater range of depths was more effective than simply adding additional correspondence points.

Subsequently, Axholt et al. [5] investigated the effects of 3D point distribution patterns. They compared *static* z (a single z depth distance), *sequentially increasing* z (resulting in an upward curved trapezoidal shape), and *magic square* (systematic variance in x, y such that z depth changes are maximized) acquisition patterns. The authors found that the magic square pattern resulted in the least parameter variance. They also found that orientation and lateral principal point offset are not primarily affected by the correspondence point distributions, but depend on the number of correspondence points.

In his 2011 dissertation, Axholt [2] further summarized the main findings of several studies on the influence of human alignment noise on OST HMD calibration. He had several main findings: First, for standing users completing a calibration task based on visual alignment, postural stability gives a translational head-aiming precision of 16 mm, which improves to 11 mm after 12–15 seconds, and can be modeled with a Weibull distribution. Second, for standing users, head aiming precision is 0.21° straight ahead and 0.26° in directions $\geq 30°$ azimuth, but the precision can be improved by considering postural sway and compensatory head rotation together, resulting in a precision of 0.01°. For seated users, the precisions are 0.09° in directions $\leq 15°$ azimuth, and can be approximated with a circular distribution. Third, pinhole camera parameter estimation variance increases linearly as a function of both alignment noise and diminishing correspondence point depth distribution. It decreases optimally if 25 or more correspondence points are used. It also decreases for all parameters with increasing correspondence point depth distribution variance, except for rotation, which primarily depends on the number of correspondence points. Finally, for seated subjects using SPAAM and a pinhole camera model, the eye point estimation accuracy is only 5cm on average, and depends on the camera matrix decomposition method used (none; closed form solution as described by Faugeras [12] (p. 52) as well

as Trucco and Verri [76] (p. 134); RQ decomposition using Givens rotations (Hartley and Zisserman [23], p. 579)).

In 2011, Maier et al. [52] investigated how different confirmation methods affect calibration quality. They compared keyboard, hand held, voice, and waiting methods, and found that the waiting method was significantly more accurate than the other methods. They also found that averaging the data collection over time improved the accuracy of all methods. However, their experiment used a video see-through HMD, and so the results could differ for an OST HMD.

In 2014, Moser et al. [61] conducted an experiment to generate baseline accuracy and precision values for OST HMD calibration, without human postural sway error. To this end, the authors mounted a camera inside an OST HMD, used SPAAM, and compared the same three depth distributions as Axholt et al. [5]: static z, sequentially increasing z, and magic square. Replicating Axholt et al. [5], they found that the magic square pattern, which yields greater depth variance for the same number of correspondence points, produces the most accurate and precise results.

In 2014, Itoh and Klinker [34], [35] analyzed the error sensitivity of SPAAM, DSPAAM (a degraded version of SPAAM where actual display use is simulated by removing and then replacing the display on the head), and the recycled / full INDICA method. For each calibration method, they simulated how errors in calibration parameters propagate to the final calibration result. Their analysis shows that, for both INDICA methods, the display orientation with respect to the HMD coordinate system has the largest impact on the reprojection error. In addition, they confirmed that SPAAM tends to provide erroneous eye positions along the z viewing direction. Note that the DSPAAM method simulates a common scenario in consumer applications, where non-expert users rely on a factory calibration or only perform the calibration once. This use pattern creates several errors; for example, every time the user puts on the HMD, the alignment between the display screens and the eyes varies slightly.

In 2015, Moser et al. [60] compared SPAAM, DSPAAM, and recycled INDICA, using both objective and subjective evaluations of two tasks: (1) indicating the location of a virtual pillar, 15.5cm high, seen against a 4x4 grid of physical pillars, and (2) indicating the location of a 2x2x2cm virtual cube on a physical 20x20x20cm grid. They found no significant differences between SPAAM and DSPAAM, and no difference between any of the methods in the left / right x axis. They found that recycled INDICA resulted in the best accuracy in the depth z and up / down y axes. Also, recycled INDICA resulted in the highest subjective user preference, likely because it requires minimal user effort and took the least time to perform. Finally, the authors note that there was a substantial disagreement between subjective and objective measures, because depth errors are less easily perceived than left / right or up / down errors.

In 2016, Moser et al. [62] evaluated the feasibility of performing SPAAM calibration using a Leap Motion controller[3] as a 3D input tool, similar to Tang et al.'s [75] use of a 3D tracked stylus. Moser et al. [62] compared four tracked objects: the user's finger matched to a virtual cross, box, and finger-shaped reticle, and a wooden stylus matched to a virtual cross. SPAAM calibrations were performed in both monoscopic and stereo conditions. A single expert user performed 20 calibrations for each of the 8 object-by-stereopsis conditions, where calibration required matching 25 calibration points. For dependent measures they evaluated both reprojection error and eye location estimates, and for the stereo calibrations they additionally evaluated binocular x, y, z disparities. Note that the x binocular disparity measures the expert user's inter-pupillary distance, and because this number is independently measurable, the inter-pupillary distance is an excellent metric for evaluating the accuracy of a stereo calibration procedure. For all dependent measures, Moser et al. [62] found that stylus calibrations were were much more accurate than all of the finger methods. They attributed this finding to the Leap Motion controller's relatively low accuracy finger tracking.

Also in 2016, Jun and Kim [41] evaluated their calibration method for stereo calibration using a depth-camera against stereo-SPAAM [16]. They found their model to perform better (in terms of positional error) with a fewer number of point correspondences.

Zhang et al. [82] compared their RIDE method with standard SPAAM within a grid of 5 x 5 = 25 sampling points and found their method to result in a lower reprojection error (3.36 pixel for RIDE vs. 5.29 pixel for SPAAM in a 800x600 px display. However, they only used a single user who performed three repetions per method.

Additionally, Qian et al. [71] proposed additional constraints for Stereo SPAAM calibration utilizing known properties and physical presumptions about the physical structures of the user's eyes. These constraints include the assumption of identical pixel density in both the x and y axis along the screen for each eye, no skew in the perceived image, identical viewing direction of both eyes perpendicular to the imaging plane, and that the interpupillary distance can be measured and known. Reprojection error is used for the comparative metric, and their results show that the inclusion of these additional constraints show promise in reducing calibration errors for binocular systems that are able to conform to the necessary restrictions. A larger study is still needed, however.

A second work by Qian et al. [72] examines the use of physical head constraint during SPAAM, akin to boresighting, to reduce the number of free parameters needed for estimation. In this study, the user's head is fixed to translation and rotational error with alignments being controlled by a mouse pointer on a physical display screen. As with the previous study, this work also shows a reduction in calibration error, in terms of reprojection error, with the trade-off of head restriction.

Azimi et al. [6] in 2017, proposed a new metric for both quantifying and improving calibration accuracy with regards to reprojection error. They propose the use of Mahalonobis distance instead of the traditional Euclidean distance in measuring reprojection offset of SPAAM results. Their reasoning lies in the anisotropic nature of the user alignment data during the calibration. Through the application of Mahalonobis distance measures, a post calibration

3. www.leapmotion.com/ - last accessed September 2nd, 2017.

iterative error reduction process is applied to identify and reduce the number of user alignment outliers. The results of their study show statistically significant improvements in both final reprojection error and reduction in required user alignments using this new metric.

6 Opportunities for Future Research

Future directions for research in optical see-through calibration methods can be identified in the areas of error metrics, display models, eye trackers and methods that go beyond spatial calibration.

Improving Error Metrics: Almost all of the reviewed papers report calibration reprojection errors in pixels (px). However, because HMDs have different resolutions, and because the distance between the user's eyes and the virtual screen plane varies by both HMD and user, reporting errors in pixels makes it difficult to meaningfully compare different methods. To address this, we propose expressing calibration errors in degrees of visual angle. In addition to being a more comparable unit, degrees of visual angle is the standard unit for many results in the vision science community. In addition, most evaluation methods only report performance metrics. However, manual calibration can be a strenuous task, and we therefore advise using subjective workload measurements (e.g., NASA TLX [22]), as well as additional measures focusing specifically on eye strain (c.f., oculumotor component of the Simulator Sickness Questionnaire [44]). In addition, objective stress measures could be employed. Finally, as discussed in Section 5 above, when possible, we recommend reporting a calibration technique's measurement of inter-pupillary distance, because this metric is user-centered, varies between users, is an important graphical parameter, and for each user can be independently measured with a high degree of accuracy.

Advanced Display Models: The vast majority of the methods reviewed here model the OST HMD as an off-axis pinhole camera, introduced in Section 2. Although this assumption is plausible for OST HMDs that use an optical combiner and virtual screen plane, additional accuracy is likely possible by using more complex graphical and eye models (Axholt [2], Jones et al. [40]). In addition, other display types call for extending the current display model. For example, focus-tunable OST HMDs can move the screen to an arbitrary focus depth (Liu et al. [49], Hu and Hong [26], [27], and Dunn et al. [10]). In such displays, the display model must always represent the current screen focus depth. Light field displays create virtual images with variable accommodation (Maimone and Fuchs [53]), but to update the virtual image, these displays also need to know the user's eye position.

Other recent HMD systems use phase-only liquid crystal on silicon (P-LCOS) displays. P-LCOS displays are essentially a programmable lens mirror that can change the refraction index of each pixel independently. The surface focal displays use P-LCOS to create a dynamic depth field in VR HMDs (Matsuda et al. [56]). A true holographic OST HMDs was developed by using P-LCOS displays (Maimone et al. [54]). These are promising approaches once we make P-LCOS small enough to be integrated in HMDs.

Integrating Eye Trackers: As discussed in Section 4, many current methods seek to automate the calibration process, but these methods will require eye trackers seamlessly integrated into the OST HMD. Therefore, integrating an eye tracker with an acceptable form factor is an important issue. Hua et al. [29] prototyped an OST HMD with an IR eye tracker that is integrated in the optics, and partially shares the optical path with the display.

Even after integrating an eye tracker (eye camera), for automatic calibration, one still needs to calibrate its pose in the display coordinate system. The pose estimation, however, can be challenging. If the coordinate system is defined on a scene camera, one needs to calibrate the pose between the scene camera and the eye camera, where the two cameras are looking in opposite directions, at the world and at the eye, with extremely different focal lengths, in the range of meters and centimeters (Itoh and Klinker [34]). Some researchers report estimating the pose via visual marker tracking, with a custom multi-marker jig (Itoh and Klinker [34], [35]) .

In an outside-in tracking setup, where the display coordinate system is defined on a set of optical markers attached on the display, the calibration procedure could even be more complex. In such systems, the external outside-in tracking system may not be able to track the jig, and, therefore, a marker jig may not work. Instead, a hand-eye calibration must be applied between the marker set and the eye camera (Horaud and Dornaika [25]).

Furthermore, the eye tracker pose might change during the use of the OST HMD because the user may touch the eye camera or the camera needs to be re-oriented to capture the eyes properly. As a result, the eye tracker could require frequent re-calibration. Plopski et al. [68] propose to automatically calibrate the pose via the corneal reflection of LED arrays attached on both the eye tracker and the display frame. The integrated design from Hua et al. [29] mentioned above may also be another hardware solution. Since the eye camera image is frontal to the eye, the camera inside the HMD frame could be fixed and only calibrated once, possibly at the manufacturer side.

Beyond Spatial Calibration: Throughout the paper, we looked at various existing calibration methods that aim at improving the alignment accuracy of AR images against the physical world in the user's field of view. A question, which arises, is, how accurate do calibrating methods need to become.

Logically thinking, the maximum accuracy would end up to the *retinal-precise* accuracy where an OST HMD can align each pixel of a displayed AR image to desired retina cells. In other words, the display can stimulate arbitrary retinal cells selectively with desired light stimulation.

Such accuracy might be overkill for most AR applications. However, if such calibration accuracy could be achieved, OST HMDs may go beyond the realm of AR displays—they could be devices that can arbitrarily manipulate human vision. A potential application of such direction is vision augmentation, where AR displays enhance human vision by retinal-precise image processing. There already exist a few applications that demonstrate such vision augmentation concepts with OST HMDs (Itoh and Klinker [38],

TABLE 2: Overview of evaluation methods. **Methods Evaluated**: **DC**: Data collection approach (**O**: objective measures, **SQN**: Subjective quantitative, **SQL**: subjective qualitative), **TCT** (task completion time), **WL** (workload), **GE** (geometric errors: reprojection or viewing angle), **PRE** (parameter reconstruction error from projection matrix, e.g., eye location).

		Measures (dependent variables)					
Work	**Methods Evaluated**	**DC**	**TCT**	**WL**	**GE**	**PRE**	**Other**
McGarrity et al. [57], [58]	SPAAM	SQN	x	x	✓	x	x
Tang et al. [75]	SPAAM, DepthSPAAM head pointing, DepthSPAAM stylus pointing, Stereo-SPAAM	SQN	✓	x	✓	✓	x
Grubert et al. [20], [21]	DepthSPAAM, MPAAM	SQN	✓	x	✓	x	x
Axholt et al. [4]	DLT (simulated point correspondences)	O	x	x	x	✓	x
Axholt et al. [5]	SPAAM, DepthSPAAM Sequential, DepthSPAAM Magic Square	O	x	x	x	✓	condition number
Maier et al. [52]	SPAAM with 4 confirmation methods: keyboard button, handheld button, voice input, waiting	O	x	x	✓	x	x
Moser et al. [61]	SPAAM, DepthSPAAM Sequential, DepthSPAAM Magic Square	O	x	x	x	✓	x
Moser et al. [60]	SPAAM, Degraded SPAAM, Recycled INDICA	SQN, SQL	x	x	✓	✓	x
Itoh and Klinker [35]	SPAAM, Degraded SPAAM, Recycled/Full INDICA	O	x	x	✓	x	perturbation sensitivity
Moser and Swan [62]	SPAAM with head pointing, finger pointing, stylus pointing (mono + stereo)	SQN	x	x	✓	✓	x
Qian et al. [72]	SPAAM with fixed head, mouse pointer alignment	SQN	x	x	✓	x	x
Qian et al. [71]	Stereo SPAAM, head pointing	SQN	x	x	✓	x	x
Zhang et al. [82]	RIDE, SPAAM	SQN	x	x	✓	x	x
Azimi et al. [6]	SPAAM with head pointing and Mahalonobis distance error correction	O	x	x	✓	x	Mahalonobis distance
Jun and Kim [41]	Proprietary methods (full, simplified) vs. Stereo SPAAM	O, SQL	x	x	x	x	3D positional error, qualitative image results

Hwang and Peli [31]).

7 Conclusion

This paper surveyed the field of calibration methods for OST HMDs. Specifically, it reviewed approaches accessible until September 2017. It provided insights into the fundamentals of calibration techniques, manual and automatic calibration approaches, as well as evaluation methods. These calibration methods are focused achieving on locational realism, that is, the correct spatial alignment of virtual content in a physical environment. Besides locational realism, further consistency domains in addition to the spatial domain, including color calibration (Itoh et al. [33], Langlotz et al. [46]) and latency (Lincoln et al. [48]), should be considered to achieve a as high degree of perceived realism as possible.

Acknowledgments

This material is based upon work supported by the National Science Foundation under awards IIS-1018413 and IIS-1320909, to J. E. Swan II, and a NASA Mississippi Space Grant Consortium fellowship and Japan Society for the Promotion of Science (JSPS) through an East Asia and Pacific Summer Institutes Fellowship, award IIA-141477, to K. Moser, and JSPS KAKENHI Grant Numbers 16H07169, 17H04692, and 17K19985, and European Unions 7th Framework Programmes for Research and Technological Development under PITN-GA-2012- 316919 EDUSAFE, to Y. Itoh.

References

[1] Y. Abdel-Aziz, "Direct linear transformation from comparator coordinates in close-range photogrammetry," in *ASP Symposium on Close-Range Photogrammetry in Illinois*, 1971.

[2] M. Axholt, "Pinhole camera calibration in the presence of human noise," Ph.D. dissertation, Linköping University Institute of Technology, 2011.

[3] M. Axholt, S. Peterson, and S. R. Ellis, "User boresight calibration precision for large-format head-up displays," in *ACM Symp. on Virtual Reality Software and Technology (VRST)*. New York, NY, USA: ACM, Oct. 2008, pp. 141–148.

[4] M. Axholt, M. Skoglund, S. D. Peterson, M. D. Cooper, T. B. Schön, 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 *Human Factors and Ergonomics Society Annual Meeting (HFES)*, vol. 54, no. 28. SAGE Publications, 2010, pp. 2427–2431.

[5] M. Axholt, M. A. Skoglund, S. D. O'Connell, M. D. Cooper, S. R. Ellis, and A. Ynnerman, "Parameter estimation variance of the single point active alignment method in optical see-through head mounted display calibration," in *IEEE Virtual Reality (VR)*. Piscataway, NJ, USA: IEEE, 2011, pp. 27–34.

[6] E. Azimi, L. Qian, P. Kazanzides, and N. Navab, "Robust optical see-through head-mounted display calibration: Taking anisotropic nature of user interaction errors into account," in *2017 IEEE Virtual Reality (VR)*, March 2017, pp. 219–220.

[7] R. Azuma and G. Bishop, "Improving static and dynamic registration in an optical see-through hmd," in *Annual Conf. on Computer Graphics and Interactive Techniques (SIGGRAPH)*. ACM, 1994, pp. 197–204.

[8] M. Billinghurst, A. Clark, and G. Lee, "A Survey of Augmented Reality," *Foundations and Trends in Human-Computer Interaction*, vol. 8, pp. 73–272, 2014.

[9] T. P. Caudell and D. W. Mizell, "Augmented reality: An application of heads-up display technology to manual manufacturing processes," in *Hawaii Inter. Conf. on System Sciences*, vol. 2. IEEE, 1992, pp. 659–669.

[10] D. Dunn, C. Tippets, K. Torell, P. Kellnhofer, K. Akşit, P. Didyk, K. Myszkowski, D. Luebke, and H. Fuchs, "Wide field of view varifocal near-eye display using see-through deformable membrane mirrors," *IEEE Trans. on Visualization and Computer Graphics*, vol. 23, no. 4, pp. 1322–1331, 2017.

[11] P. J. Edwards, A. P. King, C. R. Maurer, D. a. de Cunha, D. J. Hawkes, D. L. Hill, R. P. Gaston, M. R. Fenlon, a. Jusczyzck, A. J. Strong, C. L. Chandler, and M. J. Gleeson, "Design and evaluation of a system for microscope-assisted guided interventions (MAGI)." *IEEE Trans. on Medical Imaging*, vol. 19, no. 11, pp. 1082–1093, 2000.

[12] O. Faugeras, *Three-dimensional computer vision: a geometric viewpoint*. MIT Press, 1993.

[13] M. Figl, C. Ede, J. Hummel, F. Wanschitz, R. Ewers, H. Bergmann, and W. Birkfellner, "A fully automated calibration method for an optical see-through head-mounted operating microscope with variable zoom and focus," *IEEE Trans. on Medical Imaging*, vol. 24, no. 11, pp. 1492–1499, 2005.

[14] A. Fuhrmann, D. Schmalstieg, and W. Purgathofer, "Fast calibration for augmented reality," in *ACM Symp. on Virtual Reality Software and Technology (VRST)*. ACM, 1999, pp. 166–167.

[15] C. Gao, H. Hua, and N. Ahuja, "Easy calibration of a head-mounted projective display for augmented reality systems," in *IEEE Virtual Reality (VR)*. IEEE, 2003, pp. 53–60.

[16] 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 *IEEE and ACM Intern. Symp. on Augmented Reality (ISAR)*. IEEE, 2000, pp. 165–174.

[17] Y. Genc, M. Tuceryan, A. Khamene, and N. Navab, "Optical see-through calibration with vision-based trackers: Propagation of projection matrices," in *IEEE and ACM Intern. Symp. on Augmented Reality (ISAR)*. IEEE, 2001, pp. 147–156.

[18] Y. Genc, M. Tuceryan, and N. Navab, "Practical solutions for calibration of optical see-through devices," in *Intern. Symp. on Mixed and Augmented Reality (ISMAR)*. IEEE, 2002, p. 169.

[19] S. J. Gilson, A. W. Fitzgibbon, and A. Glennerster, "Spatial calibration of an optical see-through head-mounted display," *J. of Neuroscience Methods*, vol. 173, no. 1, pp. 140–146, 2008.

[20] J. Grubert, J. Tuemle, R. Mecke, and M. Schenk, "Comparative user study of two see-through calibration methods," in *IEEE Virtual Reality (VR)*. IEEE, 2010, pp. 269–270.

[21] J. Grubert, J. Tümler, and R. Mecke, "Untersuchungen zur optimierung der see-through-kalibrierung fuer mobile augmented reality assistenzsysteme," *Michael Schenk, Hrsg*, vol. 6, no. 7, 2008.

[22] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," *Advances in psychology*, vol. 52, pp. 139–183, 1988.

[23] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge University Press, 2003.

[24] ——, "Multiple view geometry in computer vision," *Robotica*, vol. 23, no. 2, pp. 271–271, 2005.

[25] R. Horaud and F. Dornaika, "Hand-eye calibration," *The international journal of robotics research*, vol. 14, no. 3, pp. 195–210, 1995.

[26] X. Hu and H. Hua, "Design and assessment of a depth-fused multi-focal-plane display prototype," *Journal of Display Technology*, vol. 10, no. 4, pp. 308–316, 2014.

[27] ——, "High-resolution optical see-through multi-focal-plane head-mounted display using freeform optics," *Optics Express*, vol. 22, no. 11, pp. 13 896–13 903, 2014.

[28] H. Hua, C. Gao, and N. Ahuja, "Calibration of a head-mounted projective display for augmented reality systems," in *Intern. Symp. on Mixed and Augmented Reality (ISMAR)*. IEEE, 2002, pp. 176–185.

[29] H. Hua, X. Hu, and C. Gao, "A high-resolution optical see-through head-mounted display with eyetracking capability," *Optics express*, vol. 21, no. 25, pp. 30 993–30 998, 2013.

[30] J. F. Hughes, A. Van Dam, J. D. Foley, and S. K. Feiner, *Computer graphics: principles and practice*. Pearson Education, 2013.

[31] A. D. Hwang and E. Peli, "An augmented-reality edge enhancement application for google glass," *Optometry and vision science: official publication of the American Academy of Optometry*, vol. 91, no. 8, p. 1021, 2014.

[32] Y. Itoh, T. Amano, D. Iwai, and G. Klinker, "Gaussian light field: estimation of viewpoint-dependent blur for optical see-through head-mounted displays," *IEEE Trans. on Visualization and Computer Graphics*, vol. 22, no. 11, pp. 2368–2376, 2016.

[33] Y. Itoh, M. Dzitsiuk, T. Amano, and G. Klinker, "Semi-parametric color reproduction method for optical see-through head-mounted displays," *IEEE Trans. on Visualization and Computer Graphics*, vol. 21, no. 11, pp. 1269–1278, 2015.

[34] Y. Itoh and G. Klinker, "Interaction-free calibration for optical see-through head-mounted displays based on 3d eye localization," in *IEEE Symp. on 3D User Interfaces (3DUI)*. IEEE, 2014, pp. 75–82.

[35] ——, "Performance and sensitivity analysis of INDICA: Interaction-free display calibration for optical see-through head-mounted displays," in *Intern. Symp. on Mixed and Augmented Reality (ISMAR)*. IEEE, 2014, pp. 171–176.

[36] ——, "Light-field correction for spatial calibration of optical see-through head-mounted displays," *IEEE Trans. on Visualization and Computer Graphics*, vol. 21, no. 4, pp. 471–480, 2015.

[37] ——, "Simultaneous direct and augmented view distortion calibration of optical see-through head-mounted displays," in *Intern. Symp. on Mixed and Augmented Reality (ISMAR)*. IEEE, 2015, pp. 43–48.

[38] ——, "Vision enhancement: defocus correction via optical see-through head-mounted displays," in *Proceedings of the 6th Augmented Human International Conference*. ACM, 2015, pp. 1–8.

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

[40] J. A. Jones, D. Edewaard, R. A. Tyrrell, and L. F. Hodges, "A schematic eye for virtual environments," in *IEEE Symp. on 3D User Interfaces (3DUI)*. IEEE, 2016, pp. 221–230.

[41] H. Jun and G. Kim, "A calibration method for optical see-through head-mounted displays with a depth camera," in *Virtual Reality (VR), 2016 IEEE*. IEEE, 2016, pp. 103–111.

[42] H. Kato and M. Billinghurst, "Marker tracking and HMD calibration for a video-based augmented reality conferencing system," in *IEEE and ACM Intern. Workshop on Augmented Reality (IWAR)*, Jun. 1999, pp. 85–94.

[43] F. Kellner, B. Bolte, G. Bruder, U. Rautenberg, F. Steinicke, M. Lappe, and R. Koch, "Geometric calibration of head-mounted displays and its effects on distance estimation," *IEEE Trans. on Visualization and Computer Graphics*, vol. 18, no. 4, pp. 589–596, 2012.

[44] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness," *The international journal of aviation psychology*, vol. 3, no. 3, pp. 203–220, 1993.

[45] G. Klinker, D. Stricker, and D. Reiners, "Augmented reality: A balance act between high quality and real-time constraints," *Mixed Reality–Merging Real and Virtual Worlds*, pp. 325–346, 1999.

[46] T. Langlotz, M. Cook, and H. Regenbrecht, "Real-time radiometric compensation for optical see-through head-mounted displays," *IEEE Trans. on Visualization and Computer Graphics*, vol. 22, no. 11, pp. 2385–2394, 2016.

[47] S. Lee and H. Hua, "A robust camera-based method for optical distortion calibration of head-mounted displays," in *IEEE Virtual Reality (VR)*. IEEE, 2013, pp. 27–30.

[48] P. Lincoln, A. Blate, M. Singh, T. Whitted, A. State, A. Lastra, and H. Fuchs, "From motion to photons in 80 microseconds: Towards minimal latency for virtual and augmented reality," *IEEE Trans. on Visualization and Computer Graphics*, vol. 22, no. 4, pp. 1367–1376, 2016.

[49] S. Liu, H. Hua, and D. Cheng, "A novel prototype for an optical see-through head-mounted display with addressable focus cues," *IEEE Trans. on Visualization and Computer Graphics*, vol. 16, no. 3, pp. 381–393, 2010.

[50] G. Luo, N. Rensing, E. Weststrate, and E. Peli, "Registration of an on-axis see-through head-mounted display and camera system," *Optical Engineering*, vol. 44, no. 2, pp. 024 002–024 002, 2005.

[51] Y. Ma, S. Soatto, J. Kosecka, and S. S. Sastry, *An invitation to 3-d vision: from images to geometric models*. Springer Science & Business Media, 2012, vol. 26.

[52] P. Maier, A. Dey, C. A. Waechter, C. Sandor, M. Tönnis, and G. Klinker, "An empiric evaluation of confirmation methods for optical see-through head-mounted display calibration," in *Intern. Symp. on Mixed and Augmented Reality (ISMAR)*. IEEE, 2011, pp. 267–268.

[53] A. Maimone and H. Fuchs, "Computational augmented reality eyeglasses," in *Intern. Symp. on Mixed and Augmented Reality (ISMAR)*. IEEE, 2013, pp. 29–38.

[54] A. Maimone, A. Georgiou, and J. S. Kollin, "Holographic near-eye displays for virtual and augmented reality," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 85:1–85:16, Jul. 2017. [Online]. Available: http://doi.acm.org/10.1145/3072959.3073624

[55] N. Makibuchi, H. Kato, and A. Yoneyama, "Vision-based robust calibration for optical see-through head-mounted displays," in *IEEE Inter. Conf. on Image Processing (ICIP)*. IEEE, 2013, pp. 2177–2181.

[56] N. Matsuda, A. Fix, and D. Lanman, "Focal surface displays," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 86:1–86:14, Jul. 2017. [Online]. Available: http://doi.acm.org/10.1145/3072959.3073590

[57] E. McGarrity, Y. Genc, M. Tuceryan, C. Owen, and N. Navab, "A new system for online quantitative evaluation of optical see-through augmentation," in *IEEE and ACM Intern. Symp. on Augmented Reality (ISAR)*. IEEE, 2001, pp. 157–166.

[58] E. McGarrity, M. Tuceryan, C. Owen, Y. Genc, and N. Navab, "Evaluation of optical see-through systems," in *Inter. Conf. on Augmented, Virtual Environments and 3D Imaging*, 2001, pp. 18–21.

[59] J. J. Moré, "The levenberg-marquardt algorithm: implementation and theory," in *Numerical analysis*. Springer, 1978, pp. 105–116.

[60] K. Moser, Y. Itoh, K. Oshima, J. E. Swan, G. Klinker, and C. Sandor, "Subjective evaluation of a semi-automatic optical see-through head-mounted display calibration technique," *IEEE Trans. on Visualization and Computer Graphics*, vol. 21, no. 4, pp. 491–500, 2015.

[61] K. R. Moser, M. Axholt, and J. E. Swan, "Baseline spaam calibration accuracy and precision in the absence of human postural sway error," in *IEEE Virtual Reality (VR)*. IEEE, 2014, pp. 99–100.

[62] K. R. Moser and J. E. Swan, "Evaluation of user-centric optical see-through head-mounted display calibration using a leap motion controller," in *IEEE Symp. on 3D User Interfaces (3DUI)*. IEEE, 2016, pp. 159–167.

[63] N. Navab, S. Zokai, Y. Genc, and E. M. Coelho, "An on-line evaluation system for optical see-through augmented reality," in *IEEE Virtual Reality (VR)*. IEEE, 2004, pp. 245–246.

[64] T. Oishi and S. Tachi, "Methods to calibrate projection transformation parameters for see-through head-mounted displays," *Presence: Teleoperators and Virtual Environments*, vol. 5, no. 1, pp. 122–135, 1996.

[65] M. O'Loughlin and C. Sandor, "User-centric calibration for optical see-through augmented reality," Honours Thesis, University of South Australia, 2013.

[66] C. B. Owen, J. Zhou, A. Tang, and F. Xiao, "Display-relative calibration for optical see-through head-mounted displays," in *Intern. Symp. on Mixed and Augmented Reality (ISMAR)*. IEEE, 2004, pp. 70–78.

[67] A. Plopski, Y. Itoh, C. Nitschke, K. Kiyokawa, G. Klinker, and H. Takemura, "Corneal-imaging calibration for optical see-through head-mounted displays," *IEEE Trans. on Visualization and Computer Graphics*, vol. 21, no. 4, pp. 481–490, 2015.

[68] A. Plopski, J. Orlosky, Y. Itoh, C. Nitschke, K. Kiyokawa, and G. Klinker, "Automated spatial calibration of hmd systems with unconstrained eye-cameras," in *Intern. Symp. on Mixed and Augmented Reality (ISMAR)*. IEEE, 2016, pp. 94–99.

[69] J. Ponce, D. Forsyth, E.-p. Willow, S. Antipolis-Méditerranée, R. dactivité RAweb, L. Inria, and I. Alumni, "Computer vision: a modern approach," *Computer*, vol. 16, no. 11, 2011.

[70] L. Priese, F. Schmitt, and P. Lemke, "Automatische see-through kalibrierung," Universität Koblenz Landau, Tech. Rep. 7/2007, 2007.

[71] L. Qian, A. Winkler, B. Fuerst, P. Kazanzides, and N. Navab, "Modeling physical structure as additional constraints for stereoscopic optical see-through head-mounted display calibration," in *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, Sept 2016, pp. 154–155.

[72] ——, "Reduction of interaction space in single point active alignment method for optical see-through head-mounted display calibration," in *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, Sept 2016, pp. 156–157.

[73] W. Robinett and R. Holloway, "The visual display transformation for virtual reality," *Presence: Teleoperators and Virtual Environments*, vol. 4, no. 1, pp. 1–23, 1995.

[74] I. E. Sutherland, "Three-dimensional data input by tablet," *Proc. of the IEEE*, vol. 62, no. 4, pp. 453–461, April 1974.

[75] A. Tang, J. Zhou, and C. Owen, "Evaluation of calibration procedures for optical see-through head-mounted displays," in *Intern. Symp. on Mixed and Augmented Reality (ISMAR)*. IEEE, 2003, p. 161.

[76] E. Trucco and A. Verri, *Introductory techniques for 3-D computer vision*. Prentice Hall Englewood Cliffs, 1998, vol. 201.

[77] R. Tsai, "A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses," *IEEE J. on Robotics and Automation*, vol. 3, no. 4, pp. 323–344, 1987.

[78] M. Tuceryan, Y. Genc, and N. Navab, "Single-point active alignment method (spaam) for optical see-through hmd calibration for augmented reality," *Presence: Teleoperators and Virtual Environments*, vol. 11, no. 3, pp. 259–276, 2002.

[79] M. Tuceryan and N. Navab, "Single point active alignment method (spaam) for optical see-through hmd calibration for ar," in *Intern. Symp. on Mixed and Augmented Reality (ISMAR)*. IEEE, 2000, pp. 149–158.

[80] J. Vince, *Mathematics for computer graphics*. Springer Science & Business Media, 2013.

[81] Z. Zhang, D. Weng, J. Guo, Y. Liu, and Y. Wang, "An accurate calibration method for optical see-through head-mounted displays based on actual eye-observation model," in *Intern. Symp. on Mixed and Augmented Reality (ISMAR)*. IEEE, 2017.

[82] Z. Zhang, D. Weng, Y. Liu, Y. Wang, and X. Zhao, "Ride: Region-induced data enhancement method for dynamic calibration of optical see-through head-mounted displays," in *Virtual Reality (VR), 2017 IEEE*. IEEE, 2017, pp. 245–246.

[83] Z. Zhang, D. Weng, Y. Liu, and L. Xiang, "3d optical see-through head-mounted display based augmented reality system and its application," in *International Conference on Optical and Photonic Engineering (icOPEN2015)*. International Society for Optics and Photonics, 2015, pp. 952 428–952 428.