Accuracy of Marker Tracking on an Optical See-Through Head Mounted Display

Michael Brand[1,a[0000-0003-3145-1704]], Lukas Antonio Wulff[1,b[0000-0002-4266-7060]], Yogi Hamdani[1,c] and Thorsten Schüppstuhl[1,d[0000-0002-9616-3976]]

[1]Hamburg University of Technology, Institute of Aircraft Production Technology
Denickestraße 17, 21073 Hamburg, Germany
Email: {[a]michael.brand, [b]lukas.wulff, [c]yogi.hamdani, [d]schueppstuhl}@tuhh.de

Abstract. This work assesses the accuracy of mono and stereo vision-based marker tracking on the Microsoft HoloLens as a representative of current generation AR devices. This is motivated by the need to employ object tracking in industrial AR applications. We integrate mono and stereo vision-based marker tracking with the HoloLens. A calibration procedure is developed that allows users to assess the accuracy of the calibration alignments by walking around the virtual calibration target. This can be generally applied when calibrating additional tracking systems with ready-made AR systems. Finally, the accuracy that can be achieved with the developed system is evaluated (comparing the influence of different parameters).

Keywords: AR, industrial applications, OST-HMD, HoloLens, marker tracking, stereo vision, calibration, accuracy.

1 Introduction

Industrial applications of Augmented Reality (AR) have several domain-specific requirements such as the accuracy of tracking, robustness, safety, ergonomics and ease of use. One especially important requirement is the accurate tracking of real objects in a workplace, for example, to place virtual work instructions on a workpiece. Possible applications include assembly, training, but also more complex applications like AR-assisted programming of industrial robots.

The inside-out tracking methods of current generation AR devices such as the Microsoft HoloLens or the Magic Leap One allow to accurately place virtual objects relative to the world. When the user moves around, the objects stay in their assigned position with little jitter and high accuracy [1]. However, the inside-out tracking in and of itself doesn't allow application developers to place virtual objects relative to real objects in the scene. In order to achieve that, dedicated object tracking is required. It can be accomplished by choosing from different AR tracking techniques available: sensor-based (magnetic, acoustic, inertial, optical and mechanical), vision-based (feature-based or model-based) and hybrid techniques based on sensor fusion [2].

T. Schüppstuhl et al. (Hrsg.), *Annals of Scientific Society for Assembly, Handling and Industrial Robotics*, https://doi.org/10.1007/978-3-662-61755-7_3

All of these techniques have their strengths and weaknesses. Vision-based tracking requires no additional instrumentation, since the aforementioned AR devices already have cameras integrated into them. This allows for very low-cost solutions that can offer high accuracy. Its disadvantages are the limited computing power of the AR device, limited robustness, the reliance on the line of sight and lighting conditions. [2] Fiducial marker tracking is a simple variant of vision-based tracking that offers relatively high robustness at the cost of the effort to place the fiducials accurately. It is readily available through commercial and open-source software. Microsoft endorses the use of Vuforia[1] with the HoloLens. Open-source solutions such as ArUco[2] and ARToolkit[3] are alternatives that can be tailored to the application's needs more freely.

The goal of this paper is to assess the accuracy of mono and stereo vision-based marker tracking on the HoloLens as a representative of current generation AR hardware. Our contributions include:

- Integration of stereo vision-based marker tracking with the HoloLens
- A manual calibration procedure that allows users to assess the accuracy of the calibration alignments by walking around the virtual calibration target, which can be applied for ready-made AR systems like the HoloLens
- Examination of the accuracy that can be achieved with this setup (comparing mono vision to stereo vision)

In Chapter 2 the theoretical background and related research is presented. Following, in Chapter 3 the methods that are used to conduct the assessment of the accuracy are described. In Chapter 4 the results are presented and discussed. Finally, in Chapter 5 a conclusion is drawn and an outlook on future work is presented.

2 Related Work

Many prototypes of applications based on the Microsoft HoloLens employ either a) manual registration by the user [3] or b) marker tracking [4–6]. Option a) offers the advantage of being easy to implement, and can produce accurate registration. Option b) offers the advantage of not requiring user interaction and can dynamically track multiple objects.

Marker tracking uses computer vision algorithms to estimate the pose of a marker based on camera images. It can be abstracted to the Perspective-n-Point (PnP) problem which describes the task of estimating the pose of a set of n points (with $n = 4$ for square markers, each corner of the marker representing one point). In a first step, the corner points have to be extracted from the raw camera image. In a second step the pose estimation is calculated. [7] There are different algorithms available that solve the PnP problem [8]. ARToolkit 5 uses the Iterative Closest Point algorithm for mono camera marker tracking and Bundle Adjustment for stereo camera marker tracking. ArUco uses

[1] Vuforia: https://developer.vuforia.com/
[2] ArUco: https://sourceforge.net/projects/aruco/files/
[3] ARToolkit 5.4 snapshot, project discontinued: https://github.com/artoolkit/ARToolKit5

homography estimation followed by an iterative Levenberg-Marquardt optimization. There are more advanced algorithms for the special case of 4 coplanar points that can avoid local optima [9]. This is implemented by ARToolkitPlus[4]. There are several studies which characterize the accuracy of different marker tracking solutions [7, 10–12].

AR systems generally need to be calibrated at least once in order to achieve registration of real and virtual content. Calibration is aimed at eliminating systematic errors as well as calculating the correct viewing parameters. There are calibration procedures for the different subsystems, e.g. camera calibration, and the system as a whole needs to be calibrated as well. For Optical See-Through Head Mounted Displays (OST-HMDs) there are manual, semi-automatic, and automatic procedures [13]. Manual procedures such as the classical SPAAM [14], Stereo-SPAAM [15] and SPAAM2 [16] are based on the user creating so called correspondences from manual alignment of real and virtual content. Based on multiple of those correspondences, the viewing parameters are then calculated.

Most of those classical calibration procedures assume that the AR display is uncalibrated, however, with modern devices such as the HoloLens the system as a whole comes pre-calibrated (except for a quick IPD calibration). Qian et al. [17, 18] use this fact as a motivation to present a new manual calibration approach. A linear calibration model that maps from some 3D-tracker space (which can be any tracking system, external or head-mounted) to the 3D-virtual space (the virtual 3D-space created by the display) is proposed. It is calculated from 3D-3D-correspondences created by the user. One of two scenarios being evaluated uses marker tracking based on the RGB camera of the HoloLens. The results indicate a somewhat limited accuracy in the depth axis of the marker tracking camera, when compared to the other two axes. This depth accuracy could be improved by employing stereo vision because it yields a higher information content. Besides, the calibration relies on the depth perception of the display. Our own preliminary experiments have shown that the depth perception of the display can be quite misleading. Furthermore, the working space is confined to arm's length which doesn't fit the requirements for object tracking in most industrial applications. Therefore, we modify the presented calibration procedure, as we believe it is generally well suited for modern AR devices where systems come pre-calibrated.

3 Method

The notion of accuracy can be divided into two aspects, 1) trueness and 2) precision. In the scope of this work, the focus will lie on the trueness. Striving for high trueness (i.e. small systematic error) is what makes a calibration procedure essential. The precision of a tracking system can be improved by employing filtering techniques such as averaging or Kalman filtering, which isn't considered in this work.

[4] ARToolkitPlus, project discontinued: https://github.com/paroj/artoolkitplus

First, the concepts for a) the integration of the marker tracking into the HoloLens, b) the calibration procedure and c) the evaluation procedure, need to be outlined. The concepts need to take the industrial application's requirements as well as the properties of the used AR system into account.

3.1 Marker Tracking Integration

The requirements for object tracking in an industrial application are chosen as follows:

- The object tracking needs to have a high accuracy in all three cartesian axes
- It needs to respect the hardware restrictions of a ready-made mobile AR device.
- The object tracking needs to be robust against losing the line of sight to the marker because, in real-world applications, markers can be occluded or the line of sight can be lost completely for some time. In both cases we assume that it is desirable to use the last known position before the line of sight was lost. It is expected that within some time the line of sight will be restored and the marker tracking can start tracking the current marker position once again.
- The working space needs to range from 60 cm to 150 cm (distance from the marker to the tracking system, see **Fig. 1, left**), so that objects at a medium distance can still be tracked accurately. This can be tailored to a specific application's needs later on.

We propose a method where, first, the marker tracking determines the pose of an object relative to the camera of the HoloLens. It is then combined with the information from the built-in inside-out tracking, so that the marker pose with respect to the device's world coordinate frame can be computed. A virtual object could then be placed relative to the marker in the world coordinate frame, using the inside-out tracking. This is repeated in small time intervals, so an up-to-date pose of the object is always available. This method offers the advantage that even when losing the line of sight, the procedure can still track the pose of the marker where it was last seen, which addresses one of the core requirements. The disadvantage of this approach is the reliance on the inside-out tracking. Furthermore, only the last known position can be tracked if the line of sight is lost. If the marker moves in the meantime the tracking can be severely inaccurate. We propose to deal with this per use-case, e.g. handling the loss of the line of sight, or employing a completely different tracking method.

The marker tracking is implemented using ARToolkit 5.3 and Unity 2017.1 game engine as a framework. The tracking target is a 3D cube with markers of 10 cm width on each face (see **Fig. 1, right**). For mono vision the RGB camera (resolution of 896x504) in the center of the front of the HoloLens, for stereo vision the central two of the four environment understanding cameras (resolution of 480x640 each, only available in "research mode") of the device are used.

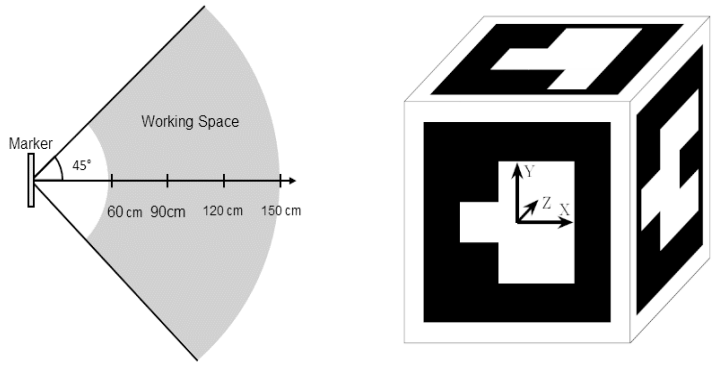

Fig. 1. Left: Working space, right: Marker cube

3.2 Calibration procedure

In the calibration procedure the user needs to create a set of correspondences by aligning a virtual cube with a marker cube manually, following a predefined protocol. It bases on the procedure described by Qian et al. [17]. We give a brief summary of their procedure first, and then explain the modifications we developed.

In Qian's procedure a calibration model that maps from some 3D-tracker space to the 3D-virtual space (the virtual 3D-space relative to the display) is calculated based on 3D-3D-correspondences collected by the user. A linear transformation $T \in \mathbb{R}^{4x4}$ is assumed as a calibration model. It maps points $q_1, \dots, q_n \in \mathbb{R}^3$ from the tracker space to points $p_1, \dots, p_n \in \mathbb{R}^3$ from the virtual space:

$$p_i = T \cdot q_i \tag{1}$$

Three different models with varying degrees of freedom (dof) are used, namely isometric (6 dof), affine (12 dof), perspective (15 dof).

The correspondences are collected as follows: The points p_i are several points that are fixed with relation to the display space. The points q_i are the corresponding measurements from the tracking system, where the tracking target is a hand-held cube on a stick. To collect correspondences the user is required to coordinate hand and head movements in order to align p_i on the virtual cube with q_i on the marker cube, which is repeated multiple times.

We modify this procedure so that the points p_i are gathered based on a virtual cube that can be positioned freely in the device's world coordinate frame by the user via keyboard input. The marker cube is placed on a table. This makes use of the system's ready-made property and allows users to walk around the virtual calibration target freely while adjusting it to exactly align with the marker cube (see **Fig. 2, left**). Once the two cubes are aligned the user can signal the system to collect a correspondence. The correspondences are collected at locations that are combinations of different angles ($\alpha = -45°, 0°, +45°$) and distances ($d = 60\ cm, 90\ cm, 120\ cm, 150\ cm$) between the HMD and the marker, 12 in total, so as to have a good coverage of the chosen

working space (see **Fig. 2, right**). Additionally, for each correspondence 12 individual measurements by the tracking system are carried out and averaged.

This modification has the advantage of allowing for a larger working space beyond arm's length and being able to assess the proper alignment of the cubes more easily, since it doesn't rely on the depth perception of the display which can be very misleading from our experience. Furthermore, this eliminates the need for head-arm coordination which improves ergonomics and accuracy.

In order to calculate the calibration model T we minimize the sum of squares of the absolute value of the residual errors over all correspondences (p_i, q_i):

$$E = \sum_{i=1}^{n} \|p_i - T \cdot q_i\|_2^2 \tag{2}$$

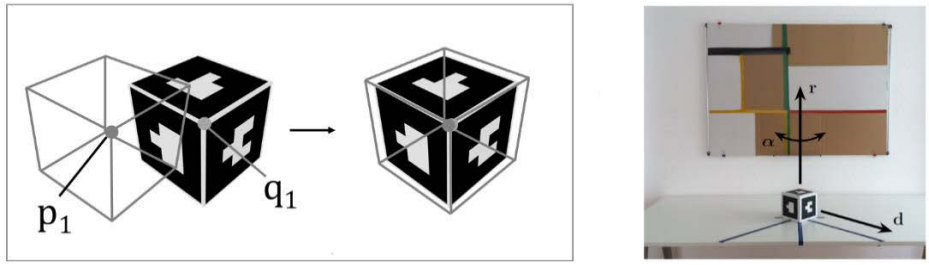

Fig. 2. Left: Aligning p_1 with q_1, right: Calibration Setup

3.3 Evaluation Concept

For the evaluation we adopt Qian's [17] calibrate-and-test procedure. The quality of the computed model will be assessed by collecting a new set of another 12 correspondences and testing how the previously calculated model fits each data point of the test data set. This is done by calculating the residual error e:

$$e = p_i - T \cdot q_i \tag{3}$$

This residual error is used as a metric for the trueness post-calibration. The new correspondences are collected in locations that are combinations of different angles ($\alpha = -45°, 0°, +45°$) and distances ($d = 75\ cm, 105\ cm, 135\ cm, 165\ cm$). Note that the distances differ from those from the calibration phase, so as to assess if the calibration model works as well in different locations.

It needs to be noted that if the display space possessed large errors when compared to the real world, the calculated residual error would be misleading, since it is expressed in relation to the display space and not the real world. It is assumed that the error of the display space is small, since the HoloLens is a pre-calibrated AR system.

The advantage of the chosen evaluation approach bases on the fact that manual alignments can closely represent an ideal alignment as seen by the user. If the user cannot distinguish the position of a real object and a superimposed virtual object, he has an optimal experience.

Finally, the results obtained with mono vision are compared with those obtained with stereo vision.

4 Evaluation

4.1 Results

For the evaluation, the calibration procedure is conducted and subsequently evaluated as outlined in the evaluation concept. These two steps are repeated three times. That way it is assured that the calibration procedure yields reproducible results. The parameters that are evaluated are chosen as follows:

- Camera setup: Mono or Stereo
- Axis: x, y, z or absolute (z represents the depth direction with respect to the display space; x and y are perpendicular to that direction)
- Calibration Model: Isometric, Affine or Perspective

For each configuration (e.g. "z, Stereo, Affine") the trueness over all the 12 correspondences (distributed across the working space) and all three runs (i.e. 36 data points per configuration in total) are averaged.

Fig. 3 shows boxplots for the absolute residual errors of each configuration. With mono vision the absolute residual is $(12.0 \pm 6.7\ \text{mm})$ with the isometric model, $(3.5 \pm 2.0\ \text{mm})$ with the affine model and $(3.6 \pm 1.8\ \text{mm})$ with the perspective model. With stereo vision the absolute residual error that can be achieved is $(40.3 \pm 34.2\ \text{mm})$ with the isometric model, $(20.2 \pm 17.4\ \text{mm})$ with the affine model and $(19.7 \pm 13.9\ \text{mm})$ with the perspective model.

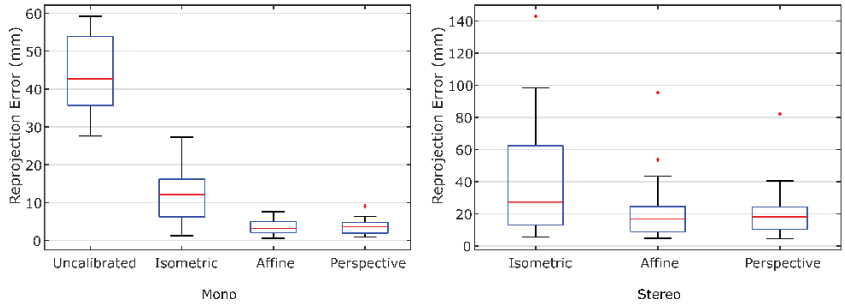

Fig. 3. Boxplots for the *absolute value* of the residual errors

Table 1 and **Table 2** show a more detailed overview of the trueness that stereo vision and mono vision can achieve with the setup developed in this work.

Table 1. Residual error (mean and standard deviation) by model and axis (mono vision)

	x (mm)		y (mm)		z (mm)		absolute (mm)	
Model	mean	std	mean	std	mean	std	mean	Std
Uncalib.	10.2	2.3	19.2	1.2	37.1	12.7	43.7	10.3
Isometric	-0.1	2.4	-0.4	3.7	-4.1	12.5	12.0	6.7
Affine	-0.8	1.3	-0.2	1.3	0.0	3.6	3.5	2.0
Perspective	-0.7	1.3	-0.1	1.2	0.0	3.6	3.6	1.8

Table 2. Residual error (mean and standard deviation) by model and axis (stereo vision)

	x (mm)		y (mm)		z (mm)		absolute (mm)	
Model	mean	std	mean	std	mean	std	mean	Std
Uncalib.	-28.0	4.2	-223.1	62.3	-4.0	35.8	227.7	62.3
Isometric	-2.8	4.0	2.7	5.8	-34.6	39.3	40.3	34.2
Affine	-2.1	2.4	4.8	3.8	-18.2	18.2	20.2	17.4
Perspective	-2.2	2.9	4.6	3.1	-16.9	15.9	19.7	13.9

4.2 Discussion

First, it has to be discussed how the values for the mean and standard deviation can be interpreted. As stated above, 36 data points are combined for each configuration. The mean gives a summary of the overall trueness of each configuration across all locations. The standard deviation has to be interpreted with some care. A varying trueness between the different locations would contribute to the standard deviation, as well as a fluctuation in the three runs. However, these two factors can't be differentiated with the data presented.

Bearing that in mind, the results indicate that mono vision yields clearly superior trueness when compared to stereo vision. This is attributed to the fact that the mono camera has got a higher resolution than the two stereo cameras individually, which plays a big role especially at large distances between the HMD and the marker that were included in the experiments. Furthermore, a stereo vision system is geometrically more complex so it apparently can't be calibrated sufficiently with a linear model, let alone an isometric one. This has to be investigated further.

It is concluded that the presented calibration procedure yields good results for simpler systems like the mono vision system. It is notable that the trueness in z-direction is good, which we attribute to the design of the calibration procedure that is not reliant on the depth perception.

It is observed that the affine and perspective models yield better results than the isometric model for both mono and stereo vision. All in all, there is no apparent difference between affine and perspective, which is why we prefer to use the affine model with less degrees of freedom in a future application.

With mono vision the mean trueness of the absolute value is at 3.5 mm with the affine model which satisfies our requirement for a high accuracy. However, we need to take a closer look at the trueness in different locations in the future, since the standard deviation of the trueness indicates fluctuation among the different locations and runs. The residual errors are generally thought to be due to non-linear errors that cannot be compensated by a linear model.

5 Conclusion

We conclude that the proposed integration concept and calibration procedure are feasible for object tracking in industrial applications. The accuracy that could be achieved by mono vision is sufficient for many potential applications.

However, stereo vision-based marker tracking on the HoloLens yields no improvement over using mono vision in the working space that was examined.

We have achieved 1) the integration of mono and stereo vision-based marker tracking with the HoloLens, 2) the development of a new manual calibration procedure that allows users to assess the accuracy of the calibration alignments by walking around the virtual calibration target which can be applied for ready-made AR systems like the HoloLens and 3) an assessment of the accuracy of both mono and stereo vision-based marker tracking.

In the future we want to further examine and improve the accuracy of mono and stereo-based marker tracking with the HoloLens by fine-tuning different parameters and analyzing the data we collected in more depth. Furthermore, we want to explore how marker tracking can be integrated into prototype applications.

Acknowledgement.
This work was created as part of the research project "MiReP" and is supported by the Federal Ministry for Economic Affairs and Energy as part of the Federal Aeronautical Research Programme "LuFo V-3".

Supported by:
Federal Ministry for Economic Affairs and Energy
on the basis of a decision by the German Bundestag

References

1. Vassallo R, Rankin A, Chen ECS et al. (2017) Hologram stability evaluation for Microsoft HoloLens. In: Kupinski MA, Nishikawa RM (eds) Medical Imaging 2017: Image Perception, Observer Performance, and Technology Assessment. SPIE, p 1013614
2. Zhou F, Duh HB-L, Billinghurst M (2008) Trends in augmented reality tracking, interaction and display: A review of ten years of ISMAR. In: Livingston MA (ed) 7th IEEEACM International Symposium on Mixed and Augmented Reality, 2008: ISMAR 2008 ; Sept. 15 - 18, 2008, Cambridge, UK. IEEE Service Center, Piscataway, NJ, pp 193–202
3. Mitsuno D, Ueda K, Hirota Y et al. (2019) Effective Application of Mixed Reality Device HoloLens: Simple Manual Alignment of Surgical Field and Holograms. Plast Reconstr Surg 143(2): 647–651. doi: 10.1097/PRS.0000000000005215
4. Eschen H, Kötter T, Rodeck R et al. (2018) Augmented and Virtual Reality for Inspection and Maintenance Processes in the Aviation Industry. Procedia Manufacturing 19: 156–163. doi: 10.1016/j.promfg.2018.01.022
5. Hoover M (2018) An evaluation of the Microsoft HoloLens for a manufacturing-guided assembly task. Master Thesis, Iowa State University
6. Evans G, Miller J, Iglesias Pena M et al. (2017) Evaluating the Microsoft HoloLens through an augmented reality assembly application. In: Sanders-Reed JN, Arthur JJ (eds) Degraded Environments: Sensing, Processing, and Display 2017. SPIE, 101970V
7. Garrido-Jurado S, Muñoz-Salinas R, Madrid-Cuevas FJ et al. (2014) Automatic generation and detection of highly reliable fiducial markers under occlusion. Pattern Recognition 47(6): 2280–2292. doi: 10.1016/j.patcog.2014.01.005

8. Marchand E, Uchiyama H, Spindler F (2016) Pose Estimation for Augmented Reality: A Hands-On Survey. IEEE Trans Vis Comput Graph 22(12): 2633–2651. doi: 10.1109/TVCG.2015.2513408
9. Schweighofer G, Pinz A (2006) Robust pose estimation from a planar target. IEEE Trans Pattern Anal Mach Intell 28(12): 2024–2030. doi: 10.1109/TPAMI.2006.252
10. Bergamasco F, Albarelli A, Torsello A (2011) Image-Space Marker Detection and Recognition Using Projective Invariants. In: International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2011: 16 - 19 May 2011, Hangzhou, China ; proceedings. IEEE, Piscataway, NJ, pp 381–388
11. Abawi DF, Bienwald J, Dorner R (2004) Accuracy in Optical Tracking with Fiducial Markers: An Accuracy Function for ARToolKit. In: Third IEEE and ACM International Symposium on Mixed and Augmented Reality, 2004: ISMAR 2004 ; 02 - 05 Nov. 2004, [Arlington, VA, USA ; proceedings …]. IEEE Computer Society, Los Alamitos, Calif., pp 260–261
12. Malbezin P, Piekarski W, Thomas BH (2002) Measuring ARTootKit accuracy in long distance tracking experiments. In: Katō H, Billinghurst M (eds) IEEE ART02: The First IEEE International Augmented Reality Toolkit Workshop : 29 September 2002, Darmstadt, Germany. [IEEE], Piscataway, N.J., p 2
13. Grubert J, Itoh Y, Moser K et al. (2018) A Survey of Calibration Methods for Optical See-Through Head-Mounted Displays. IEEE Trans Vis Comput Graph 24(9): 2649–2662. doi: 10.1109/TVCG.2017.2754257
14. Tuceryan M, Navab N (2000) Single point active alignment method (SPAAM) for optical see-through HMD calibration for AR. In: Proceedings, IEEE and ACM International Symposium on Augmented Reality (ISAR 2000): October 5-6, 2000, Munich, Germany. IEEE Computer Society, Los Alamitos, Calif, pp 149–158
15. Genc Y, Sauer F, Wenzel F et al. (2000) Optical see-through HMD calibration: a stereo method validated with a video see-through system. In: Proceedings, IEEE and ACM International Symposium on Augmented Reality (ISAR 2000): October 5-6, 2000, Munich, Germany. IEEE Computer Society, Los Alamitos, Calif, pp 165–174
16. Genc Y, Tuceryan M, Navab N (2002) Practical solutions for calibration of optical see-through devices. In: International Symposium on Mixed and Augmented Reality: ISMAR 2002, September 30-October 1, 2002, Darmstadt, Germany : proceedings. IEEE Computer Society, Los Almitos, Calif, pp 169–175
17. Qian L, Azimi E, Kazanzides P et al. (2017) Comprehensive tracker based display calibration for holographic optical see-through head-mounted display. arXiv preprint arXiv:1703.05834
18. Qian L, Deguet A, Kazanzides P (2018) ARssist: augmented reality on a head-mounted display for the first assistant in robotic surgery. Healthcare technology letters 5(5): 194–200. doi: 10.1049/htl.2018.5065