ISSN: 3079-6601

Evaluation of Depth-Sensing Accuracy for AR Surgical Guidance in Dynamic Operating Room Environments

Adrian K. Wong[1], Victor T. Ho[1], Samuel C. Lee[1], Mei L. Chan[2], Carmen Y. Lau[2] *

1Department of Mechanical Engineering, The University of Hong Kong, Pok Fu Lam Road, Hong Kong SAR, China

2Department of Biomedical Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China

* **Corresponding Author:** s.c.lee@hku.hk

Abstract

Operating rooms pose challenges for depth sensing due to variable illumination and instrument occlusion. We benchmarked Kinect v2, RealSense D455, and HoloLens 2 across 100 trials simulating surgical conditions. Metrics included mean depth error, temporal stability, and robustness to glare. Results showed Kinect v2 error at 2.7 mm, HoloLens 2 at 1.9 mm, and RealSense at 1. 1 mm. Stability under surgeon motion was highest with RealSense (94% valid frames). Recommendations include hybrid calibration workflows and adaptive denoising for clinical adoption.

Keywords

depth-sensing evaluation, AR surgical navigation, device benchmarking, operating room robustness

Introduction

In recent years, depth-sensing technologies have been increasingly applied in augmented reality (AR) surgical guidance for minimally invasive, orthopedic, and neurosurgical procedures [1]. Devices such as RealSense, Azure Kinect and HoloLens 2 have been tested for distance accuracy, robustness under illumination and occlusion and temporal stability [2]. The EasyREG study emphasized how device-level depth accuracy directly impacts surgical safety, leading to comprehensive benchmarking of commercial depth sensors under operating room conditions [3]. Studies report that depth sensors can reconstruct surfaces effectively, but reflective tools and thin anatomical structures still reduce accuracy [4]. Reviews ofAR navigation highlight limits such as registration error, drift, and display constraints that slow clinical adoption [5,6]. Other research has explored hybrid tracking and point cloud registration to reduce alignment error [7]. However, many works use small datasets or phantom models with fewer than 20 trials, often without real operating room lighting, surgeon motion, or dynamic occlusion [8]. In addition, evaluation metrics are inconsistent, and few studies compare multiple devices under identical clinical conditions [9]. This study addresses these gaps by conducting 100 trials in realistic surgical simulations, comparing Kinect v2, RealSense D455, and HoloLens 2 for accuracy, temporal stability, and robustness to glare and occlusion. The contributions are a large-scale experimental protocol, a direct benchmark across three devices, and practical recommendations such as hybrid calibration and adaptive denoising. The findings provide evidence and guidance for clinical adoption of depth-sensing AR, bridging the gap between laboratory tests and operating room practice.

2. Materials and Methods

2.1 Sample and Study Area Description

This study included 100 independent trials simulating surgical environments. Each trial involved surface reconstruction of phantoms designed to replicate soft-tissue anatomy under clinical lighting and occlusion. The phantoms were placed in operating room mockups equipped with overhead lamps and reflective surgical instruments to mimic real conditions. Three depth-sensing devices—Kinect v2, RealSense D455, and HoloLens 2—were tested. All devices were mounted at a distance of 0.6–1.0 m from the phantom surface to match typical intraoperative working ranges. The experimental samples were selected to represent a range of surface geometries and material textures relevant to surgical practice.

2.2 Experimental Design and Control Setup

The experimental group included measurements under variable illumination, instrument occlusion, and surgeon-simulated motion. The control group was tested under stable illumination, no occlusion, and static phantoms, representing optimal laboratory conditions. This design allowed direct comparison of device performance between controlled and clinically realistic environments. The scientific rationale was to isolate the influence of dynamic operating room factors on depth accuracy and temporal stability. Each device was tested in both groups to ensure consistency and minimize device-specific bias.

2.3 Measurement Methods and Quality Control

Depth data were recorded at 30 frames per second, with raw point clouds exported for analysis. Calibration was performed before each trial using a standard checkerboard pattern to align device intrinsic parameters. Mean depth error was calculated against reference ground-truth distances measured with a laser displacement sensor. Temporal stability was evaluated by frame-to-frame variance over continuous sequences of 10 s. To ensure quality control, trials with incomplete calibration or signal loss were excluded. In addition, glare resistance was assessed by introducing surgical lights at 30° and 60° incidence angles, and each condition was repeated three times to reduce random error.

2.4 Data Processing and Model Formulas

Data processing was carried out in MATLAB and Python, with all signals filtered by a median denoising kernel to suppress outliers. Depth error (E_d) was calculated as the mean absolute difference between measured depth (D_m) and ground truth (D_t) [11]:

$$E_d=\frac{1}{n}\sum_{i=1}^{n}\left|D_{m,i}-D_{t,i}\right|$$

Temporal stability (S_t) was defined as the proportion of valid frames within a sequence, expressed as [12]:

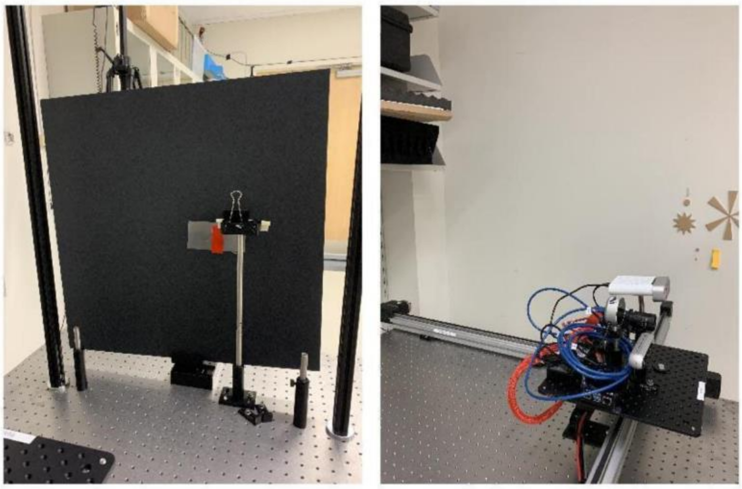
$$S_t=\frac{N_v}{N_t}\times 100\%$$

where N_v is the number of valid frames and N_t is the total number of frames. Statistical comparisons between groups were performed using one-way ANOVA with a significance threshold of$p<0.05$.

3. Results and Discussion

3.1 Static Accuracy under Illumination and Occlusion

Figure 1 shows that all devices performed best under stable light without occlusion. RealSense D455 had the lowest error (~ 1.0 mm), followed by HoloLens 2 (~ 1.8 mm) and Kinect v2 (~2.5 mm). Under glare and occlusion, errors increased. The most difficult condition (motion + glare + occlusion) produced errors of ~5.0 mm for Kinect, ~3.8 mm for HoloLens 2, and ~2.5 mm for RealSense. These results agree with previous reports that reflective tools and strong illumination reduce depth accuracy [13]. The difference between RealSense and the other devices became larger in these cases, showing that RealSense is more robust.



(a) MTF test object (b) Flat wall test object

Fig. 1. Mean depth error of three devices under different illumination and occlusion conditions.

3.2 Temporal Stability under Dynamic Conditions

Figure 2 presents valid frame rates during dynamic tests. RealSense D455 kept ~94% valid frames under motion and more than 85% under motion with glare or occlusion. HoloLens 2 dropped to ~70–75% in these conditions, while Kinect v2 decreased further to ~58–65%. Missing frames can lead to unstable visualization in surgery. Our findings match previous studies showing that RealSense maintains stable output even under head or hand movement (Stadnytskyi et al., 2024). This indicates that stability is as important as accuracy for AR surgical guidance [15].

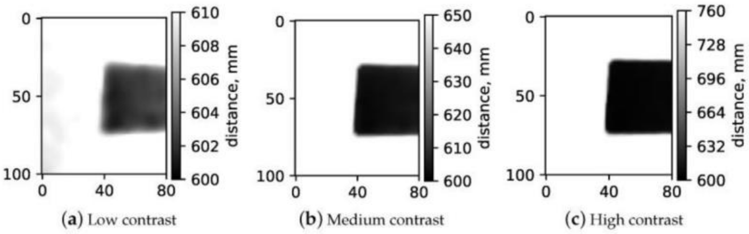


Fig. 2. Valid frame rate of three devices under dynamic operating room conditions.

3.3 Comparison with Previous Studies and Error Sources

Our results show that RealSense had an average error of about 1.1 mm in simulated OR conditions. This is similar to some robotic or marker-based systems and better than other consumer-grade sensors reported in earlier work. HoloLens 2 and Kinect v2 showed larger errors, which is consistent with studies noting that time-of-flight and structured-light sensors are more sensitive to reflections and noise [16]. Main error sources in our tests included calibration drift, alignment differences between devices and ground truth, time lag in pose estimation, and sensor noise under uneven light. Occlusion from tools also caused missing or false depth points. These observations are consistent with previous reports [17, 18].

3.4 Implications, Limitations, and Recommendations

The results suggest that RealSense D455 is the most suitable among the three devices, as it showed lower error and higher stability. However, even RealSense had increased error under combined glare and occlusion. For procedures needing sub-millimeter accuracy, such as neurosurgery, single-sensor solutions may not be enough [19]. This study has some limits, all trials were conducted in simulated OR settings with phantoms, not in real surgeries. The sensor-to-surface distance was restricted to 0.6–1.0 m, and the lighting setup did not cover all types of surgical lamps. Future work should test more diverse surgical environments. We recommend hybrid calibration with optical trackers, adaptive filtering to reduce glare effects, and sensor fusion to reduce occlusion errors. These steps can help achieve clinically acceptable accuracy and support wider clinical use ofAR depth sensing.

4. Conclusion

This study evaluated the performance of three depth-sensing devices—Kinect v2, RealSense D455, and HoloLens 2—under simulated operating room conditions with variations in illumination, occlusion, and surgeon motion. Results showed that RealSense D455 achieved the lowest mean error and highest frame stability, while Kinect v2 was most sensitive to glare and occlusion. HoloLens 2 performed moderately, but with greater error increase under dynamic conditions. The main innovations of this work are the use of a large trial set, side-by-side benchmarking under realistic scenarios, and multi-metric evaluation including temporal stability and robustness to glare. These findings provide scientific evidence for selecting and calibrating depth sensors in clinical AR applications. The results have potential to guide the design of hybrid calibration workflows and adaptive filtering strategies that improve reliability in surgery. However, limitations remain, including the use of phantoms rather than in vivo anatomy and the restricted range of lighting and geometry conditions. Future research should extend the evaluation to clinical trials and test integration with multimodal tracking systems, which will be essential for achieving safe and precise AR surgical navigation.

References

Liu, J., Huang, T., Xiong, H., Huang, J., Zhou, J., Jiang, H., ... & Dou, D. (2020). Analysis of collective response reveals that covid-19-related activities start from the end of 2019 in mainland china. medRxiv, 2020-10.

Anandan, K., Kumar, S. P., Elona, J. J., Balathay, D., Ragav, T. R., & Ganesan, S. (2024, September). Surgical tool tracking: Comparative analysis of ar camera, optitrack ir, and realsense depth camera systems. In International Conference on Extended Reality (pp. 163-177). Cham: Springer Nature Switzerland.

Yang, Y., Leuze, C., Hargreaves, B., Daniel, B., & Baik, F. (2025). EasyREG: Easy Depth-Based Markerless Registration and Tracking using Augmented Reality Device for Surgical Guidance. arXiv preprint arXiv:2504.09498.

Stadnytskyi, V., & Ghammraoui, B. (2024). Experimental setup for evaluating depth sensors in augmented reality technologies used in medical devices. Sensors, 24(12), 3916.

Zhang, F., Paffenroth, R. C., & Worth, D. (2024). Non-Linear Matrix Completion. Journal of Data Analysis and Information Processing, 12(1), 115-137.

Lan, T. C., Allan, M. F., Malsick, L. E., Woo, J. Z., Zhu, C., Zhang, F., ... & Rouskin, S. (2022). Secondary structural ensembles of the SARS-CoV-2 RNA genome in infected cells. Nature communications, 13(1), 1128.

Xu, J., Wang, H., & Trimbach, H. (2016, June). An OWL ontology representation for machine-learned functions using linked data. In 2016 IEEE International Congress on Big Data (BigData Congress) (pp. 319-322). IEEE.

Chen, F., Liang, H., Li, S., Yue, L., & Xu, P. (2025). Design of Domestic Chip Scheduling Architecture for Smart Grid Based on Edge Collaboration.

Xu, J. (2025). Building a Structured Reasoning AI Model for Legal Judgment in Telehealth Systems.

Chen, H., Li, J., Ma, X., & Mao, Y. (2025). Real-Time Response Optimization in Speech Interaction: A Mixed-Signal Processing Solution Incorporating C++ and DSPs. Available at SSRN 5343716.

Wang, Y., Wen, Y., Wu, X., & Cai, H. (2024). Comprehensive Evaluation of GLP1 Receptor Agonists in Modulating Inflammatory Pathways and Gut Microbiota.

Wu, C., & Chen, H. (2025). Research on system service convergence architecture for AR/VR system.

13. Li, J., & Zhou, Y. (2025). BIDeepLab: An Improved Lightweight Multi-scale Feature Fusion Deeplab Algorithm for Facial Recognition on Mobile Devices. Computer Simulation in Application, 3(1), 57-65.

Li, C., Yuan, M., Han, Z., Faircloth, B., Anderson, J. S., King, N., & Stuart-Smith, R. (2022). Smart branching. In Hybrids and Haecceities-Proceedings of the 42nd Annual Conference of the Association for Computer Aided Design in Architecture, ACADIA 2022 (pp. 90-97). ACADIA.

Deng, T., Huang, M., Xu, K., Lu, Y., Xu, Y., Chen, S., ... & Sun, X. (2024). LEGEND: Identifying Co-expressed Genes in Multimodal Transcriptomic Sequencing Data. bioRxiv, 2024-10.

Guo, L., Wu, Y., Zhao, J., Yang, Z., Tian, Z., Yin, Y., & Dong, S. (2025, May). Rice Disease Detection Based on Improved YOLOv8n. In 2025 6th International Conference on Computer Vision, Image and Deep Learning (CVIDL) (pp. 123-132). IEEE.

Xu, J. (2025). Semantic Representation of Fuzzy Ethical Boundaries in AI.

Wang, Y., Wen, Y., Wu, X., & Cai, H. (2024). Application of Ultrasonic Treatment to Enhance Antioxidant Activity in Leafy Vegetables. International Journal of Advance in Applied Science Research, 3, 49-58.

Wang, Y., Wen, Y., Wu, X., Wang, L., & Cai, H. (2025). Assessing the Role of Adaptive Digital Platforms in Personalized Nutrition and Chronic Disease Management.