Eurographics Workshop on Visual Computing for Biology and Medicine (2026)
R. Krueger, E. Mörth and K. Furmanová (Editors) *Conference Paper*

RGB-Based 2.5D Gesture Interaction with Volumetric Medical Data on an Autostereoscopic Display

Lisa Mueller[1,2], Bernhard Preim[1] and Sebastian Krueger[2]

[1]Department of Simulation and Graphics, University of Magdeburg, Germany
[2]Siemens Healthineers, Forchheim, Germany

Abstract

Touchless interaction with volumetric medical data on autostereoscopic displays offers an intuitive way to explore 3D information in sterile and hardware-light environments. We present an RGB-based 2.5D hand tracking system for gesture-based manipulation of medical data, supporting translation, rotation, zooming, and depth translation. A user study with twelve technically experienced participants evaluates usability, workload, and task performance. The results indicate high usability (SUS = 80) and low to moderate workload, suggesting that RGB-based tracking can support intuitive exploratory interaction. However, rotation and depth-related tasks show higher variability due to gesture recognition instability and the lack of absolute depth information. Overall, the approach is suitable for exploratory visualization, while precise 3D control remains limited.

CCS Concepts

• ***Human-centered computing*** → ***Gestural input;*** *User studies; Displays and imagers;* • ***Applied computing*** → *Life and medical sciences;*

1. Introduction

The visualization of volumetric medical data plays an essential role in diagnostics, interventions, surgery, as well as medical education [MMW*18, OLWR23, SSPOJ16]. Advanced rendering techniques, such as cinematic volume rendering [Fel16], generate highly detailed and photorealistic representations of anatomical structures. This data can be viewed in 3D on autostereoscopic displays without any additional equipment in front of the eyes [Dod05], improving spatial perception and depth understanding.

The interaction with volumetric data is still often based on conventional 2D input devices such as mouse and keyboard. These devices are not designed for spatial 3D manipulation, limiting interaction efficiency and intuitiveness. Furthermore, in medical contexts, particularly during interventions or intrasurgically, maximum hygiene is essential. Thus, physical contact with devices should be minimized.

This motivates gesture-based interaction as a more intuitive alternative for 3D data manipulation that can be used in hygiene-critical environments. However, most approaches rely on depth sensors or specialized hardware to capture 3D hand information, increasing system complexity and limiting practical deployment.

RGB cameras offer a widely available and low-cost alternative, especially as many modern autostereoscopic systems already include integrated cameras. Recent methods allow estimation of hand landmarks from monocular RGB input, including so-called 2.5D representations that provide image-space coordinates in combination with inferred relative depth information. Although these are not real depth values, they deliver information about spatial hand motion and therefore create the possibility for interaction that may still feel 3D without the need for specialized depth hardware.

Although RGB-based 2.5D hand tracking has become increasingly capable, its suitability for stable and intuitive interaction with volumetric medical data on autostereoscopic displays remains largely unexplored. This work addresses this gap by presenting and evaluating a touchless RGB-based interaction framework for medical volume visualization.

Our contributions are:

- A real-time RGB-based 2.5D hand tracking system for touchless interaction with volumetric medical data on autostereoscopic displays.
- A lightweight gesture interaction pipeline with temporal filtering and state-based activation for stable 3D manipulation.
- A user study ($N = 12$) evaluating usability, workload, and task performance of RGB-based volumetric interaction.
- An analysis of limitations of RGB-based 2.5D tracking for rotation and depth manipulation.

2. Related Work

Related work is discussed across three areas relevant to this system: touchless interaction, depth-based gesture tracking approaches, and RGB-based hand tracking methods.

2.1. Touchless Interaction in Medical Visualization

Touchless interaction is highly relevant in medical environments due to sterility requirements and workflow constraints, particularly in interventional radiology and surgery. Mewes et al. provide an overview of touchless interaction techniques in medical contexts, highlighting gesture-based control as a frequently used modality for interacting with medical imaging data [MHWH17]. The primary motivation across prior work is to enable efficient and hygienic access to volumetric information without direct contact with physical devices. Typical interaction goals include navigation, manipulation, and exploration of 3D medical data during procedures or training scenarios.

2.2. Depth-Based Gesture Tracking

A large amount of existing systems relies on depth sensing technologies to enable robust gesture interaction. Devices such as the Microsoft Kinect, Intel RealSense, and Leap Motion Controller provide explicit 3D hand position estimates, enabling tracking for interaction with volumetric data. These approaches are commonly used in medical visualization systems due to their reliability in capturing hand pose and depth information in real time [ABKB20, CTC*14, SML*15]. For example, Saalfeld et al. demonstrated Leap Motion–based interaction for vascular visualization, achieving intuitive manipulation capabilities but requiring dedicated hardware integration [SML*15]. Despite their robustness, depth-based systems introduce additional cost, hardware dependencies, and integration constraints, which can limit their deployment in clinical environments [MHWH17].

2.3. RGB-Based Hand Tracking

Recent advances in computer vision and machine learning have enabled real-time hand tracking directly from RGB cameras. Frameworks such as OpenPose and MediaPipe Hands estimate hand landmarks from monocular image input, providing 2D image coordinates with inferred relative depth cues (2.5D representations) [CHS*21, ZBV*20]. These approaches significantly reduce hardware requirements and allow simplified integration, especially since modern display setups often already include integrated cameras. Vision-based hand gesture recognition typically consists of hand detection, tracking, feature extraction, and gesture classification, and may employ either rule-based or learning-based recognition approaches [AFHA*22].

Several recent works investigate hybrid approaches that combine RGB-based landmark tracking with additional depth sensing technologies to improve spatial stability and robustness. For example, RGB-based hand tracking has been combined with depth cameras such as Intel RealSense or Microsoft Kinect to enhance reliability and provide more stable 3D positioning in medical or immersive visualization scenarios [TLK*24]. These hybrid systems aim to compensate for limitations of monocular RGB tracking while maintaining comparatively lightweight interaction setups.

However, purely RGB-based tracking remains attractive due to its low cost and ease of deployment, yet it does not provide absolute metric depth, as depth is inferred from learned representations. Recent surveys identify monocular depth ambiguity, occlusions, real-time robustness, and continuous gesture interpretation as persistent challenges for vision-based hand tracking, particularly in RGB-only systems [GZL*24]. Despite recent progress, relatively little work has systematically studied the stability of RGB-only 2.5D interaction for volumetric medical data in autostereoscopic environments. This motivates investigating RGB-based interaction in this work.

3. Design of a Gesture-Based Interaction System for 3D Medical Visualization

This section describes the proposed system, covering the overall pipeline and hardware setup, the RGB-based hand tracking approach, and the gesture recognition and interaction design built on top of it.

3.1. System Overview

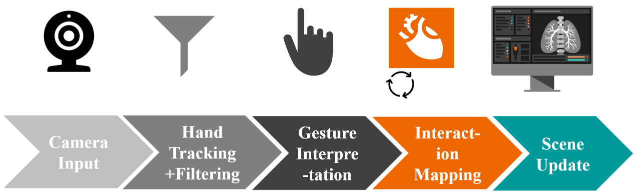

Figure 1: *Overview of the proposed RGB-based gesture interaction pipeline.*

The proposed system enables touchless interaction with volumetric medical data on an autostereoscopic display using RGB-based hand tracking. An RGB webcam (Logitech Brio 4K) placed above the display captures user hand movements in front of the display at a configured resolution of 640×480 pixels and approximately 30 fps, while volumetric medical content is rendered stereoscopically to provide spatial depth perception without additional wearable equipment. Cinematic rendering is implemented using a Vulkan-based path-tracing renderer, while DirectX 11 is used for image compositing and output to the autostereoscopic display.

The interaction pipeline is illustrated in Figure 1. Hand landmarks are extracted in real time using the MediaPipe Hands framework [ZBV*20] and filtered afterwards. Consistent with established vision-based gesture recognition pipelines [AFHA*22], the filtered landmarks are used for rule-based gesture recognition, which identifies predefined gestures and maps them to corresponding manipulation operations, including rotation, translation, zooming, and depth translation along the viewing axis. The resulting transformations are applied directly to the volumetric visualization scene displayed on the autostereoscopic display.

3.2. RGB-Based Hand Tracking

Hand tracking was implemented using the Hands solution of the MediaPipe framework [ZBV*20] (version 0.10.5), which enables real-time estimation of 21 hand landmarks from monocular RGB input. The framework provides a lightweight and hardware-efficient solution that can operate using standard RGB cameras. The estimated landmarks represent anatomical keypoints of the hand, including fingertips, finger joints, and the wrist. Each landmark is represented as a single point corresponding to the estimated location of the respective anatomical keypoint and no landmark size or area estimate is provided. The solution was configured with a minimum detection and tracking confidence of 0.7, balancing stability with responsiveness, and model complexity set to 1 for accurate landmark detection at low latency. The input operates as a continuous video stream (`static_image_mode = false`), and the maximum number of tracked hands is adjusted between one and two depending on the active gesture, reducing latency and ensuring only the relevant hand influences gesture recognition.

MediaPipe represents each hand as 21 keypoints defined by (x,y,z). The x and y coordinates are normalized to the image dimensions and lie in $[0,1]$. The value z describes the order of the relative depth of the joints in relation to the origin located at the wrist, furthermore, smaller values mean closer to the camera. Z uses approximately the same scale as x. No real camera distance is provided. In addition to image-space landmarks, the framework provides relative world landmarks describing the internal 3D structure of the hand in a canonical coordinate system, with the hand center as origin. With varying distances to the camera, the hand size remains approximately the same.

Both representations were used for different aspects of gesture recognition and interaction mapping. However, as the estimated depth values do not represent true metric depth relative to the camera, the system does not have access to absolute spatial hand positions in 3D space.

The lack of explicit depth information introduces challenges regarding interaction stability and spatial consistency, particularly for continuous 3D manipulation tasks. Tracking stability further decreases during rapid motion, occlusion and strongly inclined hand poses, where the delivered depth estimation becomes more ambiguous. These limitations are representative of broader challenges associated with vision-based hand tracking, including depth ambiguity, occlusion, and viewpoint-dependent estimation [GZL*24] To improve robustness and reduce temporal jitter, landmark smoothing using a 1 € filter [CRV12] was applied prior to gesture interpretation and interaction mapping. This adaptive low-pass filter adjusts its cutoff frequency based on movement velocity, balancing jitter reduction during slow poses with responsiveness during fast gestures. The filter was configured with $f_{c,\min} = 0.6\,\text{Hz}$, $\beta = 0.8$, and $f_{c,d} = 1.0\,\text{Hz}$.

3.3. Gesture Set and Recognition

Our system uses a small set of three gestures designed to support common 3D interaction tasks while limiting cognitive load and reducing ambiguity in gesture classification. The gesture vocabulary was selected to favor intuitive, easily distinguishable gestures that can be readily inferred and learned by users, following principles of gesture vocabulary design that emphasize naturalness, guessability, memorability, and user agreement across application domains [VNSVV24, HIea23]. The selected gestures include:

- Fist
- Dual Pinch
- Peace Sign

Together, these gestures enable continuous manipulation of volumetric medical data on the autostereoscopic display.

Gesture recognition is based on a rule-based approach operating on hand landmarks provided by MediaPipe Hands. Threshold values used for gesture recognition were selected empirically during system development and refined through informal testing. Both landmark representations are used. Normalized image-space landmarks (x,y) together with the inferred relative depth component z (2.5D representation) are primarily used for distance- and threshold-based gesture detection and planar interaction. In contrast, world landmarks are used when a more stable 3D representation of hand structure is required, in particular for rotation estimation in the fist gesture and for more robust spatial relationships between key joints.

The *pinch gesture* is detected using normalized 2.5D landmarks by measuring the distance between the thumb and index fingertip. A pinch is confirmed when the relative distance falls below an empirically determined threshold:

$$d_{\text{rel}} = \frac{d_{\text{thumb-index}}}{d_{\text{wrist-MCP}}} < 0.25$$

where $d_{\text{thumb-index}}$ is the Euclidean distance between thumb and index fingertip, and $d_{\text{wrist-MCP}}$ is the wrist-to-middle-finger metacarpophalangeal (MCP) joint distance, used as a hand-size normalization factor.

The *fist gesture* is identified based on the relative proximity of all fingertip landmarks to the palm region, using a normalized compactness metric m computed from 3D world landmarks:

$$m = \frac{||C_{\text{tips}} - P_{\text{palm}}||}{d_{\text{wrist-MCP}}}$$

where C_{tips} is the centroid of all five fingertip landmarks, P_{palm} is the mean position of the wrist and four MCP joints, and $d_{\text{wrist-MCP}}$ is the wrist-to-middle-finger MCP distance, as defined above. The detection tolerates minor deviations from a perfectly closed fist, as one or occasionally two fingers may remain slightly extended during natural interaction, while the overall fingertip configuration remains sufficiently compact. Stability is improved using a 5-frame sliding window requiring at least 3 positive detections and a 3-frame median filter to suppress single-frame outliers. To prevent flickering during ambiguous hand poses, asymmetric hysteresis thresholds are applied: a fist is activated when $m < 0.45$ and deactivated only when $m > 0.52$.

The *peace sign gesture* is detected by evaluating the extension state of the index and middle fingers while ensuring that the remaining fingers are curled. The gesture is confirmed when index and middle fingertip distances to the palm center exceed a relative threshold ($d_{\text{idx}}, d_{\text{mid}} > 1.0$), while thumb, ring, and pinky remain curled ($d_{\text{thumb}} < 0.7$, $d_{\text{ring}}, d_{\text{pinky}} < 0.5$).

Proceedings published by Eurographics - The European Association for Computer Graphics.

To improve interaction stability, temporal filtering and a state-based activation strategy are applied. The temporal thresholds follow the same empirical selection. A gesture is only activated if it is continuously detected for 150 ms, while deactivation requires absence for 300 ms, to allow responsive activation while maintaining stable state transitions. To further improve robustness and reduce computational load, the system dynamically switches between single- and dual-hand tracking depending on the active gesture, adjusting the MediaPipe maximum hand limit accordingly.

Moreover, during transitions between tracking modes (from two to one hand), a simple spatial hand-masking strategy is applied for 1.2 s, darkening the non-active hand region in the input image with a 30-pixel safety margin. This reduces incorrect hand selection and stabilizes interactions.

3.4. Interaction Design

Figure 2: *Gesture-to-interaction mapping for volumetric medical data manipulation. Each gesture is mapped to a corresponding spatial transformation: fist (rotation and translation), dual pinch (zooming), and peace sign (depth translation).*

The defined gesture set maps to three core interaction functions (covering four spatial transformations):

- Combined rotation and translation
- Zooming
- Depth Translation

Gestures and mappings are displayed in Figure 2. Hand movements are continuously mapped to corresponding spatial transformations of the volume, enabling users to control position, scale, and depth through mid-air interaction. The proposed interaction design is tailored to autostereoscopic displays, where direct touch would occlude the displayed volumetric content and interfere with depth perception. In addition to supporting hygiene-critical medical environments, the gesture mappings are designed to complement the 3D visualization capabilities of autostereoscopic displays by providing a more natural interaction than conventional 2D input devices.

The fist gesture enables combined rotation and planar translation of the volume. Rotation is derived from a hand orientation basis constructed via the cross product of $\vec{v}_1 = P_{\text{index}} - P_{\text{wrist}}$ and $\vec{v}_2 = P_{\text{pinky}} - P_{\text{wrist}}$, where P_{index}, P_{pinky}, and P_{wrist} correspond to the 3D world coordinates of the index MCP, pinky MCP, and wrist landmarks respectively. The resulting orientation is converted to a quaternion, and rotation is applied as $q_{\text{obj,new}} = q_{\text{obj,0}} \cdot q_{\text{hand,0}}^{-1} \cdot q_{\text{hand,current}}$, ensuring smooth rotation independent of the absolute initial hand pose. This mapping leads to a more intuitive interaction as users can grasp and reposition the dataset in a way that closely mirrors how objects are manipulated in the physical world. The fist gesture was additionally selected due to its robustness under hand rotation and inclination, as a stable visible hand region remains detectable during tracking, improving landmark consistency compared to more open hand configurations. This contributes to more stable interaction behavior during continuous manipulation.

The dual pinch gesture is used for zooming the volume. The scale factor is computed as $s = d_{\text{current}}/d_0$, where d_0 is the initial hand distance at gesture start, giving a proportional scaling that is clamped to a fixed range to prevent extreme scaling. The two-handed pinch mapping is analogous to zoom on touchscreen devices, making it immediately familiar. This choice is inspired by previous work which report high user agreement for two-handed scaling gestures across application domains [HIea23].

The peace sign gesture controls depth-related translation of the volume along the viewing direction. This allows users to reposition the volume within the perceived 3D space of the autostereoscopic display and supports depth exploration of anatomical structures. Although less intuitive than planar gestures, the peace sign was selected because its forward and backward hand movement provides a natural association with depth translation, while its distinct hand configuration enables reliable recognition. Depth translation is computed from the change in 2D wrist-to-pinky-MCP distance d, which approximates forward/backward hand motion:

$$T_z = T_{z,0} + k_z \cdot \Delta d, \quad \Delta d = d_0 - d_{\text{current}}$$

where d_0 is the reference distance at gesture start and k_z is an empirically determined scaling constant. This indirect mapping is the primary reason depth interaction is sensitive to hand orientation, as discussed in Section 5.

Visual feedback. Visual feedback plays an important role in maintaining interaction clarity. The system provides near real-time response to gesture activation through continuous updates of the volume transformation, allowing users to perceive a direct connection between hand motion and scene changes. This reinforces the mapping between physical gestures and virtual actions and supports spatial awareness during manipulation.

In addition to object motion, a virtual hand representation (Figure 3) is used to support interaction feedback. The tracked hand is visualized in the scene using the estimated landmark structure, allowing users to understand whether the system correctly detects the current gesture state and hand position. This is particularly important in a touchless setting, where no physical contact provides confirmation.

Figure 3: *Visualization of the tracked hand using MediaPipe landmarks integrated into the interaction scene. The virtual hand provides visual feedback of the recognized hand pose and helps users understand the relationship between their hand configuration and the resulting interaction.*

To further improve interpretability, text-based feedback is displayed. This includes the currently active gesture and interaction mode (e.g., move/rotation, or zoom). This simple feedback helps users verify the performing gesture and reduces ambiguity during gesture transitions, especially in cases where tracking may briefly fluctuate.

Overall, hand motion is translated into corresponding spatial transformations of the volumetric medical data. The design aims to maintain consistency between physical gesture movement and virtual interaction behavior while supporting intuitive and efficient exploration of medical volumes.

4. Evaluation

This section describes the user study conducted to evaluate the proposed system, including the participant group, task design, and measures used to assess usability.

4.1. Participants and Procedure

A user study was conducted with N = 12 participants (5 female, 7 male; $M_{age} = 38.25$ years, SD = 10.59, range = 24–57 years). Participants were recruited from a medical imaging and healthcare technology company. All participants had technical backgrounds and prior experience with medical visualization tools and gesture-based interaction, primarily in virtual reality environments. While not representative of clinical end users, the participants were suitable for evaluating interaction design and system usability. The evaluation therefore focuses on usability rather than clinical applicability.

Each session began with a short introduction to the system and its interaction concept. Participants were asked about prior experience with medical imaging and gesture-based interfaces and were introduced to the gesture set and corresponding interaction mappings. A brief training phase allowed participants to familiarize themselves with the system before performing structured tasks.

The study was conducted using two 27-inch autostereoscopic displays (Sony Spatial Reality Display and Barco), which were assigned based on availability and were used in approximately equal proportions. Participants interacted with the display at a comfortable viewing distance of approximately 40 cm. Participants were encouraged to think aloud during interaction. The resulting comments were documented by the moderator and used to identify recurring usability issues reported in the qualitative findings.

4.2. Tasks

Participants completed a set of interaction tasks using a 3D-rendered human head on an autostereoscopic display. The tasks were designed to cover all core interaction types:

1. **Translation:** Move the object to each of the four corners of the display area.
2. **Rotation:** Rotate the object by 360 degrees around the vertical axis to inspect it from all sides.
3. **Zooming:** Adjust the scale of the object to focus on the nose region of the model.
4. **Depth Manipulation:** Move the object forward and backward along the viewing axis to explore spatial depth.
5. **Combined Interaction:** Starting from an initial clipped view, rotate and zoom the object to locate and inspect the Circle of Willis in the brain.

These tasks represent typical operations in volumetric data exploration and were designed to evaluate both isolated and combined gesture usage. Between tasks, the object was approximately restored to its initial position, orientation, and scale.

4.3. Measures

Usability was assessed using the System Usability Scale (SUS) [Bro96], providing a standardized measure of perceived system usability. Cognitive workload was evaluated using the NASA Task Load Index (NASA-TLX) [HS88], capturing perceived mental, physical, and temporal demand during interaction.

Task performance was evaluated through moderator observation of task success. Outcomes were categorized as successful, successful but with obvious problems, successful with moderator guidance, or not successful / stopped by the participant. This measure provides an indication of the practical effectiveness of the interaction techniques. Qualitative feedback was collected through open-ended questions after the study and participant comments, focusing on usability issues, interaction behavior, and perceived system limitations.

5. Results and Discussion

This section presents the evaluation results regarding usability, workload, task performance, and interaction behavior of the proposed RGB-based interaction system with volumetric medical data on an autostereoscopic display.

5.1. Usability and Workload

The system achieved a mean System Usability Scale (SUS) score of 80 (SD = 16.61) across 12 participants, indicating good perceived usability according to established SUS benchmarks [BKM09]. Despite overall positive ratings, variability between users suggests differences in perceived control and interaction stability.

The NASA-TLX results indicate low to moderate workload across all dimensions. Mental demand (M = 5.0, SD = 3.16), physical demand (M = 7.33, SD = 5.37), performance (M = 6.83, SD = 4.24), effort (M = 4.25, SD = 2.92), frustration (M = 4.25, SD = 3.09), and temporal demand (M = 2.67, SD = 1.31) were rated on a scale from 0 (low) to 20 (high). Physical demand showed the highest variability, reflecting the constant effort required for mid-air interaction and continuous arm movement, mostly related to the movement of the fist, according to qualitative feedback. An elevated performance score indicates that participants perceived lower task success during certain interactions, as well as potentially reduced perceived control during more demanding gestures.

5.2. Task Performance and Interaction Characteristics

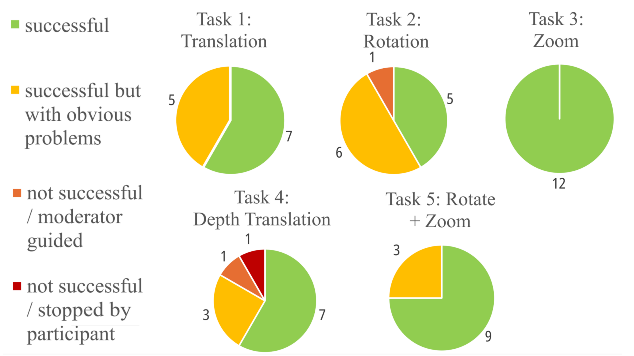

Figure 4: *Task success distribution across interaction types (translation, rotation, zooming, depth manipulation, and combined interaction).*

Task performance varied across interaction types (Figure 4). Zooming was successfully completed by all participants. Translation was also completed successfully by all participants, although minor difficulties were observed, particularly reduced precision near the edges of the camera view and during faster movements. Rotation was mostly successful, but more than half of the participants experienced difficulties, including delayed activation, reduced control accuracy, and in some cases the need for guidance.

Depth translation showed the highest variability in performance. While more than half of the participants were able to complete the task successfully, several encountered difficulties related to gesture recognition stability and hand orientation sensitivity. One participant required assistance, and one participant was unable to complete the task due to gesture recognition issues. Overall, depth-related interaction proved to be the most challenging due to its dependence on consistent hand orientation and less stable recognition behavior.

Image-plane-based interactions were significantly more stable than depth-related manipulation. This confirms that RGB-based 2.5D tracking is more suitable for relative 2D transformations than for consistent 3D depth control due to limitations of the depth estimation.

A key issue was observed in rotation-based interaction. While the fist gesture itself was generally robust during execution, inconsistencies occurred during gesture transitions, particularly when releasing the fist gesture. In some cases, continued landmark motion during gesture release or delayed gesture state updates led to unintended object rotation after gesture termination. This indicates that the system occasionally fails to sharply distinguish between active manipulation and idle states.

Additional tracking instability was observed during fast motion, strong hand inclination, and partial occlusions. These conditions occasionally led to transient misclassifications, particularly during gesture changes, resulting in reduced interaction continuity.

5.3. Qualitative Findings

Participants generally described the interaction as intuitive after a short familiarization phase. The fist gesture was perceived as natural for grasping and manipulation, while the dual pinch gesture was consistently rated as the most stable and reliable interaction technique.

Depth interaction using the peace sign was frequently described as less intuitive and strongly dependent on hand orientation. Several participants reported uncertainty regarding the mapping between hand motion and resulting depth movement, reflecting the indirect nature of this interaction.

Stability issues were commonly reported, including occasional unintended gesture activations, recognition delays, and orientation-dependent tracking variations, particularly affecting rotation and depth-related tasks.

The virtual hand representation was considered helpful for understanding system state but was also reported to occlude parts of the volumetric data and introduce depth mismatches that led to user confusion. Textual feedback was seen as useful for confirming the active gesture, although not consistently used.

Overall, feedback indicates that the system is usable and intuitive for basic interactions, while depth control and tracking stability remain the primary limitations.

5.4. Discussion

The results reveal a clear separation: image-plane interactions such as translation and zooming benefit from stable 2D landmark relationships and clear geometric interpretation, while depth related interactions such as rotation and depth translation are fundamentally limited by the absence of absolute metric depth, uncertainty in hand pose estimation, and unstable gesture transitions. These observations are consistent with the broader challenges of vision-based hand gesture interpretation reported in recent literature [GZL*24] This distinction has direct design implications. In particular, gestures relying on 3D landmark geometry are sensitive to changes in hand orientation, which produce inconsistent model outputs and can combine with partial tracking errors to destabilize gesture state transitions.

The observed rotation inconsistencies highlight an important design trade-off. While the fist provides a highly intuitive gesture for grasping, its continuous mapping makes it sensitive to gesture transition errors. This raises a fundamental question for 3D interaction design: whether future systems should prioritize intuitive manipulation, or instead rely on a more constrained but more precise interaction.

Medical environments often require high precision and predictable behavior, where reducing uncertainty can be more important than making gestures feel completely natural. A more constrained interaction design, such as trackball-like rotation or axis-limited movement, could improve accuracy but make the system less intuitive. In contrast, the current design focuses on easy learning and natural mappings, which is better suited for exploratory visualization tasks.

These findings suggest that RGB-based interaction is well suited for exploratory workflows in medical visualization, but less suited to precision-critical manipulation. Its suitability for manipulation tasks depends on the required level of precision, with image-plane interactions being more robust than those relying on inferred depth or hand orientation.

6. Conclusion and Outlook

This work investigated RGB-based 2.5D hand tracking for interaction with volumetric medical data on an autostereoscopic display, focusing on usability and perceived workload. The results show that the system supports intuitive exploratory interaction, with high usability (SUS = 80) and low to moderate workload. Translation and zooming were perceived as reliable and easy to use, while rotation and depth interaction showed higher variability due to tracking instability and gesture ambiguity.

Overall, RGB-based hand tracking is suitable for exploratory medical visualization tasks but less effective for precise 3D manipulation due to limited depth information and interaction robustness. The evaluation was conducted in a controlled setting with a small, technically experienced participant group, which limits generalizability to clinical environments.

Future work should focus on reducing gesture transition instability, improving depth-consistent interaction mappings, supporting customizable gesture sets for users with diverse needs and hand anatomies, and evaluating RGB-depth approaches in clinical environments. Furthermore, the evaluation focused on usability rather than quantitative gesture recognition performance. Future work should assess recognition accuracy, false activations, missed detections, and robustness under varying hand orientations and occlusions. Multimodal input and hybrid sensing systems combining RGB and depth information could enhance reliability. More advanced interaction techniques, such as clipping planes and crop-box interaction, could further support detailed volumetric exploration. Finally, evaluation under realistic clinical conditions, including prolonged use and interaction with medical gloves, is necessary. Prior work has visually demonstrated MediaPipe-based tracking on colored gloves (e.g., blue gloves) [TLK*24], although robustness may depend on visual contrast and environmental conditions.

Acknowledgements. We used Claude (Anthropic) and ChatGPT (OpenAI) for assistance with academic writing, including sentence rephrasing, partial structural editing, and language polishing of author-written content. GitHub Copilot was used for coding assistance and code completion. All scientific contributions, system design, experimental results, and conclusions are entirely the authors' own.

References

[ABKB20] AMEUR S., BEN KHALIFA A., BOUHLEL M. S.: Hand-gesture-based touchless exploration of medical images with leap motion controller. In *Proc. of Multi-Conference on Systems, Signals & Devices (SSD)* (2020), pp. 6–11. doi:10.1109/SSD49366.2020.9364244. 2

[AFHA*22] AL FARID F., HASHIM N., ABDULLAH J., BHUIYAN M. R., MOHD ISA W. N. S., UDDIN J., HAQUE M. A., HUSEN M. N.: A structured and methodological review on vision-based hand gesture recognition system. *Journal of Imaging 8*, 6 (2022), 153. doi:10.3390/jimaging8060153. 2

[BKM09] BANGOR A., KORTUM P., MILLER J.: Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of Usability Studies 4* (2009), 114–123. 6

[Bro96] BROOKE J.: SUS: A quick and dirty usability scale. In *Usability Evaluation in Industry*, Jordan P. W., Thomas B., McClelland I. L., Weerdmeester B., (Eds.). Taylor & Francis, 1996, pp. 189–194. 5

[CHS*21] CAO Z., HIDALGO G., SIMON T., WEI S.-E., SHEIKH Y.: OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence 43*, 1 (2021), 172–186. doi:10.1109/TPAMI.2019.2929257. 2

[CRV12] CASIEZ G., ROUSSEL N., VOGEL D.: 1 euro filter: A simple speed-based low-pass filter for noisy input in interactive systems. In *Proc. of ACM SIGCHI Conference on Human Factors in Computing Systems* (2012), pp. 2527–2530. doi:10.1145/2207676.2208639. 3

[CTC*14] CHAO C., TAN J., CASTILLO E. M., ZAWAIDEH M., ROBERTS A. C., KINNEY T. B.: Comparative efficacy of new interfaces for intra-procedural imaging review: The microsoft kinect, hillcrest labs loop pointer, and the apple ipad. *Journal of Digital Imaging 27*, 4 (2014), 463–469. doi:10.1007/s10278-014-9687-y. 2

[Dod05] DODGSON N.: Autostereoscopic 3D displays. *Computer 38*, 8 (Aug. 2005), 31–36. doi:10.1109/MC.2005.252. 1

[Fel16] FELLNER F. A.: Introducing cinematic rendering: A novel technique for post-processing medical imaging data. *Journal of Biomedical Science and Engineering 9*, 3 (2016), 170–175. doi:10.4236/jbise.2016.93013. 1

[GZL*24] GAO K., ZHANG H., LIU X., WANG X., XIE L., JI B., YAN Y., YIN E.: Challenges and solutions for vision-based hand gesture interpretation: A review. *Computer Vision and Image Understanding 248* (2024), 104095. doi:10.1016/j.cviu.2024.104095. 2, 3, 7

[HIea23] HOSSEINI M., IHMELS T., ET AL.: Towards a Consensus Gesture Set: A Survey of Mid-Air Gestures in HCI for Maximized Agreement Across Domains. In *Proc. of ACM CHI Conference on Human Factors in Computing Systems* (2023). doi:10.1145/3544548.3581420. 3, 4

[HS88] HART S. G., STAVELAND L. E.: Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Human Mental Workload*, Hancock P. A., Meshkati N., (Eds.), vol. 52. North-Holland, 1988, pp. 139–183. doi:10.1016/S0166-4115(08)62386-9. 5

[MHWH17] MEWES A., HENSEN B., WACKER F., HANSEN C.: Touchless interaction with software in interventional radiology and surgery: A systematic literature review. *International Journal of Computer Assisted Radiology and Surgery 12*, 2 (2017), 291–305. doi:10.1007/s11548-016-1480-6. 2

[MMW*18] MAGALHÃES D. S. F., MANSOOR S., WENG Y., GHIZONI E., BARBOSA T., SILVEIRA F. D. A., TOLEDO R. S., LI L. M.: Multi-view autostereoscopic system for 3D visualization in anatomy. *Research on Biomedical Engineering 34* (2018), 279–283. doi:10.1590/2446-4740.180066. 1

[OLWR23] OTOO E.-M. A., LEIBOWITZ H., WONG O., RHODE K.: Using a portable autostereoscopic screen to improve anatomy teaching and learning. *Anatomia 2*, 1 (2023), 88–98. doi:10.3390/anatomia2010008. 1

[SML*15] SAALFELD P., MEWES A., LUZ M., PREIM B., HANSEN C.: Comparative evaluation of gesture and touch input for medical software. In *Proc. of Mensch und Computer* (2015), pp. 143–152. doi:10.1515/9783110443929-016. 2

[SSPOJ16] SAALFELD P., STOJNIC A., PREIM B., OELTZE-JAFRA S.: Semi-immersive 3D sketching of vascular structures for medical education. In *Proc. of EG Workshop on Visual Computing in Biology and Medicine* (2016), pp. 123–132. 1

[TLK*24] TRUONG V. D., LIM H.-K., KIM S., DAT T. T. K., YOON J.: Contagious infection-free medical interaction system with machine vision controlled by remote hand gesture during an operation. *Computational and Structural Biotechnology Journal 24* (2024), 393–403. doi:10.1016/j.csbj.2024.05.006. 2, 7

[VNSVV24] VILLARREAL-NARVAEZ S., SLUÝTERS A., VANDERDONCKT J., VATAVU R.-D.: Brave new ges world: A systematic literature review of gestures and referents in gesture elicitation studies. *ACM Computing Surveys 56*, 5 (2024), 128:1–128:55. doi:10.1145/3636458. 3

[ZBV*20] ZHANG F., BAZAREVSKY V., VAKUNOV A., TKACHENKA A., SUNG G., CHANG C.-L., GRUNDMANN M.: MediaPipe Hands: On-device real-time hand tracking. arXiv preprint arXiv:2006.10214, 2020. doi:10.48550/arXiv.2006.10214. 2, 3