Computational and Structural Biotechnology Journal 24 (2024) 393–403

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research Article

Contagious infection-free medical interaction system with machine vision controlled by remote hand gesture during an operation

Van Doi Truong [a,b], Hyun-Kyo Lim [a,b], Seongje Kim [a,b], Than Trong Khanh Dat [c,d], Jonghun Yoon [b,e,f,*]

[a] Department of Mechanical Design Engineering, Hanyang University, 222, Wangsimni-ro, Seongdongsu, Seoul 04763, Republic of Korea
[b] BK21 FOUR ERICA-ACE Center, Hanyang University, 55, Hanyangdaehak-ro, Sangnok-gu, Ansan-si, Gyeonggi-do 15588, Republic of Korea
[c] Faculty of Mechanical Engineering, Ho Chi Minh City University of Technology (HCMUT), 268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City 700000, Viet Nam
[d] Vietnam National University Ho Chi Minh City, Linh Trung Ward, Thu Duc City, Ho Chi Minh City 700000, Viet Nam
[e] Department of Mechanical Engineering, Hanyang University, 55, Hanyangdaehak-ro, Sangnok-gu, Ansan-si, Gyeonggi-do 15588, Republic of Korea
[f] AIDICOME Inc., 221, 5th Eng. Build., 55, Hanyangdaehak-ro, Sangnok-gu, Ansan-si, Gyeonggi-do 15588, Republic of Korea

ARTICLE INFO

Keywords:
Human-machine interaction
Sterile environment
Contactless
Pointing technique
Gesture recognition

ABSTRACT

Background and objective: Medical image visualization is a requirement in many types of surgery such as orthopaedic, spinal, thoracic procedures or tumour resection to eliminate risk such as "wrong level surgery". However, direct contact with physical devices such as mice or touch screens to control images is a challenge because of the potential risk of infection. To prevent the spread of infection in sterile environments, a contagious infection-free medical interaction system has been developed for manipulating medical images.
Methods: We proposed an integrated system with three key modules: hand landmark detection, hand pointing, and hand gesture recognition. A proposed depth enhancement algorithm is combined with a deep learning hand landmark detector to generate hand landmarks. Based on the designed system, a proposed hand-pointing system combined with projection and ray-pointing techniques allows for reducing fatigue during manipulation. A proposed landmark geometry constraint algorithm and deep learning method were applied to detect six gestures including click, open, close, zoom, drag, and rotation. Additionally, a control menu was developed to effectively activate common functions.
Results: The proposed hand-pointing system allowed for a large control range of up to 1200 mm in both vertical and horizontal direction. The proposed hand gesture recognition method showed high accuracy of over 97% and real-time response.
Conclusion: This paper described the contagious infection-free medical interaction system that enables precise and effective manipulation of medical images within the large control range, while minimizing hand fatigue.

1. Introduction

The Fourth Industrial Revolution refers to increasing interconnectivity and smart automation, where human-machine interaction (HMI) has become popular in our lives. HMI is a multidisciplinary field focused on the interaction and communication between users and computers through interfaces. Traditionally, interfaces consist of physical parts such as keyboards, mice, or touch screens. However, with significant developments in the field of machine vision, contactless interactions have demonstrated great potential in HMI systems without the need for specific devices.

Contactless HMI technology, which allows users to naturally and intuitively control machines, is an ideal solution for operating rooms. The operating room normally includes many pieces of equipment such as an operating table, operating room lights, monitor screens to keep track vital signs, a ventilator, sterile instruments for surgery or video screens for laparoscopy to see the surgery area. In the preoperative planning stage, such as in urological surgery [1], orthopaedic surgery [2], accurate diagnosis and surgical planning are important to enhance the detection of pathologies or anatomical relationships. These image

* Corresponding author at: Department of Mechanical Engineering, Hanyang University, 55, Hanyangdaehak-ro, Sangnok-gu, Ansan-si, Gyeonggi-do 15588, Republic of Korea.
E-mail address: yooncsmd@gmail.com (J. Yoon).

https://doi.org/10.1016/j.csbj.2024.05.006
Received 22 November 2023; Received in revised form 30 April 2024; Accepted 6 May 2024
Available online 11 May 2024

data can be X-ray radiography, magnetic resonance image (MRI) or computed tomography (CT), which are used in diagnosis procedures. During surgery, these medical images must available as a reference for the surgeon to eliminate risk such as "wrong level surgery" in spine surgery [3]. The requirements of imaging displays are also critical aspects mentioned in the World Health Organization guidelines for safe surgery 2009 including orthopaedic, spinal, thoracic procedures or tumour resections [4]. However, interacting with medical image data within a sterile environment is a challenging task because there is a potential risk of spreading the infection through direct contact with operating tools and control media, such as a mouse. Therefore, the use of these devices should be avoided, and physicians often need to change their position or obtain support from the surgical assistant. However, this approach is time-consuming, interrupts the workflow, and is ineffective. In this scenario, the application of gesture control to interact with medical data can effectively reduce the risk of infection transmission in sterile environments.

Patient images integrated into the hospital's picture archiving and communication system (PACS) are generally displayed on monitors in the operating room, allowing physicians to navigate through a medical image viewer. The five basic functions include clicking, scrolling slices, zooming, and making pane and contract modifications. Some early studies have mentioned the concept of an HMI system for manipulating medical images [1,5,6]. An HMI system can be divided into three stages: image acquisition, hand or hand pose detection, and gesture recognition. In the image acquisition stage, images can be obtained from different types of cameras such as RGB [2,5], RGB-depth [1,7–11], and IR-depth camera [12–14] as shown in Table 1. The hand can then be segmented using a thresholding method based on depth or colour using deep-learning models. Finally, the hand gestures were determined and connected to a medical viewer. Several methods can be used, such as distance metrics and support vector machines (SVM), as well as hidden Markov models (HMM), artificial neural network (ANN) models, and 3D convolutional neural networks (3D CNN) [15].

To date, some applications of HMI have been reported. The two main commercial packages are Microsoft Kinect (MK) [1,6–9] and Leap Motion Controller (LMC) [12–14,16] as shown in Table 1. The MK is a 3D depth camera launched for the Xbox 360 console by Microsoft Corp., and it is used for manipulating the image viewer in the operating room. Lars et al. [1] applied image-processing methods, including thresholding and blob detection, to detect hands. Ruppert et al. utilised human body-shape tracking and a probabilistic template method to track 15 joints. A mean filter was used to determine the cursor position. The hold method was used for clicking and rotating. However, reports indicated limitations in hand recognition, and mouse clicking was not performed well. Hotker et al. [17] adapted gestures from both Kinect driver and voice system to navigate an OsiriX medical image viewer. On the other hand, LMC is a vision-based position-tracking system developed by Leap Motion, Inc. Mewes et al. [13] and Sanchez-Margallo [18] used a Leap Motion Kit to directly control medical image viewers. Cho et al. [14] proposed a support vector machine (SVM) model and Naïve Bayer classifiers for classifying five gestures. The same group [16] also proposed a capsule network and used IR images from Leap Motion to train the network. The five gestures included: hovering, grabbing, clicking, one peak, and two peaks. The model achieved an accuracy of 86.46%, surpassing both a conventional neutron network (CNN) and a "Very Deep Convolutional Neural Network" (VGG16) in terms of performance. In addition, many researchers have been working to develop hand gesture recognition for various applications such as general hand gestures [19–22] and sign language [23–25] using graph and general deep neuron network. Miah et al. proposed a two-stream multistage graph with an attention mechanism to extract spatial-temporal information for multi-cultural sign language recognition such as Korean sign language [24], Pakistani sign language, and American sign language [23]. The approach achieved 63.25%, and 90,31% accuracy on Top-1, and Top-10 accuracy respectively on Word-Level American Sign Language 100 classes [26] (WLASL-100) by skeleton-based features. It is shown that multiple classes highly influence classification accuracy. Bockhacker et al. [2] introduced five gestures: scrolling through images, zooming, contract modification, and moving the ROI vertically and horizontally. VGG16 was used to classify images from the front user. The proposed method takes an average of 114 s to handle the scrolling of the transverse plane of the CT images to display both pedicles of the fourth lumbar vertebra, and 109 s to manipulate the sagittal plane and zoom in to exclude the thoracic spine. Leap motion tracks have precise control by tracking a user's hand and projecting onto a virtual interactive box. However, it can only operate within a small tracking range of less than 0.75 m. Sanchez-Margallo [18] mentioned that the use of LMC was physically less demanding than the Kinect system during surgery. Rosa et al. [27] also reported that the Kinect system causes faster fatigues due to the wider movements when compared to using the LMC system.

In this study, we proposed a novel contagious infection-free medical interaction system using hand pointing and gesture recognition

Table 1
Related studies for gestured-based infection-free medical interaction system.

Citation	Sensor	Method	Visualization	Features
Wachs et al. [5]	RGB (Canon VC-C4)	Hand tracking: color segmentation	Gibson Image browser	Gesture: left, right movement, zoom in/out
Bockhacker et al.[2]	RGB camera	Gesture recognition: Very Deep Convolutional Neural Network (VGG16)	mRay DICOM Viewer	Gesture: Go left, up, right, down, change active window
Ruppert et al.[1]	Depth (MK)	Hand tracking: OpenNI (Open Natural Interaction) Cursor detection: Centre of mass method	InVesalius software	Hand distance requirement: ~ 0.5 m from camera Gestures: drag, click
Ebert et al. [17]	RGB-Depth (MK)	Hand tracking: Blob detection	OsiriX	Gestures: pane, scroll, zoom
Hötker et al. [7]	RGB-Depth (MK)	Kinect developer kit	Unknown	Gesture: Scroll up/down, zoom in/out
Jacob et al. [9]	RGB-Depth (MK)	Hand gesture detection: HMMs (Hidden Markov Models)	OsiriX	Gestures: Zoom in/out, brightness changes, rotation
Ogura et al. [12]	Infrared (LMC)	Leap Motion development kit	AZEWIN viewer	Hand distance requirement: ~0.07 – 0.3 m from camera Gestures: swipe, drag, drop. click
Mewes et al. [13]	Infrared (LMC)	Leap Motion development kit	Unknown	Hand distance requirement: ~0.5 m from camera. Gestures: Translate, zoom, slicing
Cho et al. [14]	Infrared (LMC)	Leap Motion development kit, SVM model	Surgeons control clinical software-PACS	Gesture: Hover, grab, click, one peak, two peak
Lee et al. [16]	Infrared (LMC)	Leap Motion development kit, Basic-DCNNs model, VGG-16 Model, CapsNet model	2D/3D PACS	Gesture: Hover, grab, click, one peak, two peak

techniques for precise control, large working range, and less demand for hand movement. Our main contributions are summarized as follows:

- We developed an integrated HMI framework using remote hand control for the operating room with key modules and workflow of the machine vision system.
- We proposed a depth enhancement algorithm and combined it with a deep learning hand-posed detecting model to effectively describe 3D hand skeletons during the hand recognition process.
- We proposed a teleport hand-pointing technique that combined projection and ray-pointing techniques based on the relative position remote hand and camera in a designed system to naturally control and reduce fatigue during manipulation with less hand movement.
- We proposed robust minimal gestures and a control menu concept to activate the functions in user interfaces. Finally, various experiments were conducted to demonstrate the efficiency and performance of the proposed method.

2. System design

The contactless HMI system was developed based on the basic requirements of an operating room, as shown in Fig. 1. Generally, physicians remain around the operating bed, which is approximately 1000 mm in width and 2000 mm in length. Our system includes an Intel Realsense D435 camera, which is attached to a frame on the side of the roof and captures the user's hand movements. To cover the operating bed area, the required height of the camera is approximately 1000 mm from the operating bed, and one or two monitors can be used to display and manipulate the medical data. A popular Radiant DICOM viewer was used for the demonstration. Furthermore, an in-house HMI program and DICOM viewer were installed on the same computer.

When the HMI system recognized a user's hand, it allows the user to manipulate medical images through a DICOM viewer. Instead of selecting a specific function in the tool bar, a control menu concept was proposed to directly connect common functions using a keyboard shortcut. When the user's hand is opened, the menu is displayed at the current cursor position, as shown in Fig. 2a. Considering that the opened cursor position is at the centre, a threshold circle is created to divide the inside and outside areas. The currently selected menu is within the inner circle. Outside the circle, the area is divided into four subareas: right, bottom, left, and top. If the cursor continues to move out of the subarea, the respective function hovers with different illustrations, as shown in Fig. 2b. Finally, the hand is closed to select a function.

3. Methodologies

The data flow is illustrated in Fig. 3. The process begins by continuously acquiring the input data from an RGB-D camera. Subsequently, the acquired image is analysed using a hand landmark detector, enabling real-time hand detection and tracking of hand poses. Subsequently, two algorithms, namely hand pointing and gesture detection, are continuously employed to identify the precise cursor position and specific gestures. Finally, the outputs are combined to generate appropriate actions in real time. The integrated system ensures that the user's gestures and cursor movements are seamlessly connected to the desired actions within the medical viewer interface. This section provides a detailed description of the three primary algorithms: hand landmark construction (Section 3.1), a hand pointing system (Section 3.2) and hand gesture recognition (Section 3.3).

3.1. Hand landmark detection and virtual-world landmark construction

The cursor position and manipulation of gestures at the interface of the medical viewer require accurate positioning of hand landmarks. Many studies have used deep learning models, such as Openpose [28], Mmpose [29], Mediapipe[30], and Leap motion frameworks [31] to detect and track 3D hand poses. Currently, there are two main approaches to estimate 3D hand poses such as model-based and model-free methods [32]. The model-based approach uses a unified hand model such as MANO[33] and fitting hand deep learning model to derive the 3D hand pose. In contrast, a model-free approach directly obtains hand landmarks by regressing heatmaps [34]. Both methods are trying to regress 2D hand landmarks and minimize depth error from a single image, however it is still not good enough and vibration problems often appear during real-time streaming. In this task, instead of using only the depth or RGB image, our approach is to combine the enhanced ground truth depth map and 2D hand landmarks regressed by the Mediapipe model to obtain a better depth accuracy and less vibration. Compared to Openpose and MMpose, Mediapipe has shown high accuracy and runtime performance in hand pose detection [35,36]. Therefore, it was chosen because of its robustness, lightweight nature, and open-source framework for hand-pose estimation and tracking. A general flowchart

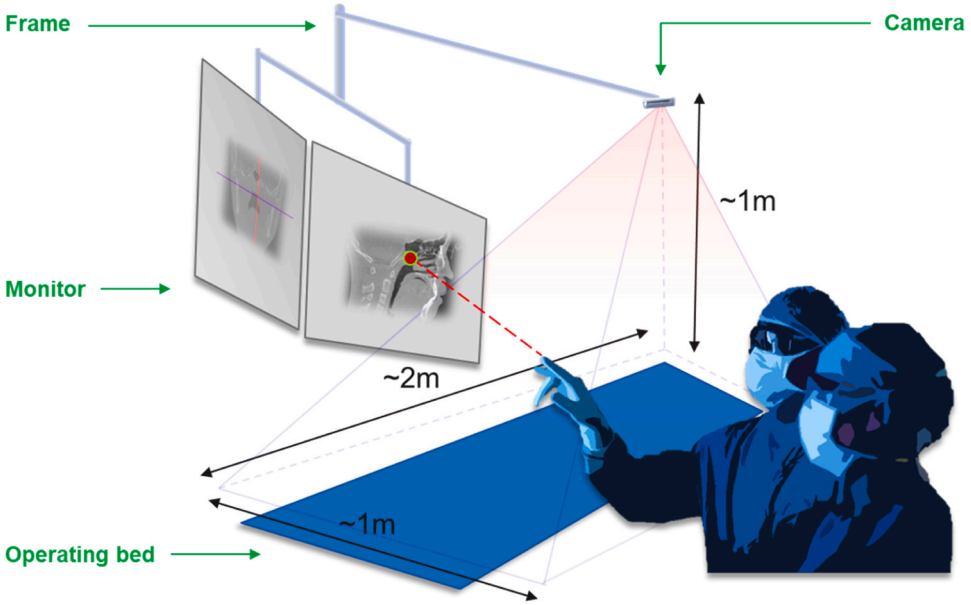


Fig. 1. Schematic configuration of HMI contactless system for the operating room.

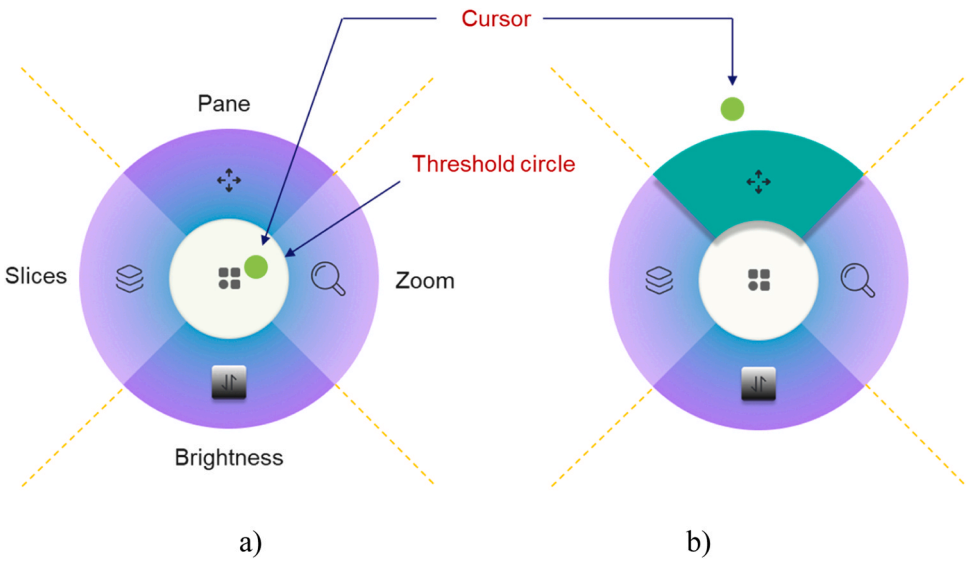


Fig. 2. Control menu concept. a) menu starts, the cursor inside circle, b) cursor moves to the top direction, the top submenu is hovered.

Fig. 3. Data flowchart of HMI system.

depicting this process is shown in Fig. 4. The camera continuously captures raw data, including RGB and depth images. First, the colour image was placed in the Mediapipe hand detector, which comprises two models: palm detection and hand landmark modelling. The palm detector detects the hand position described by the hand-bounding boxes from the RGB image. Subsequently, the hand-bounding boxes are cropped and fed into the hand landmark model. This is a regression model inside the detected hand regions that returns 2.5D landmarks, which include the pixel coordinates (x, y) and relative depth. However, the accuracy of the relative depth was found to be insufficient for implementing the pointing technique; therefore, only 2D coordinates were used. Second, the depth image was used to determine the depth value of each hand landmark. However, the depth at the hand position tends to be unstable, and the hand vibrates. Therefore, a preprocessing process is applied. The preprocessing process is described in the pseudocode shown in Table 2. A temporal filter is used to improve depth data persistence. Subsequently, a hole-filing filter is applied to fill in the missing values within the depth image. When hand-bounding boxes are detected, the corresponding depth area is cropped, and the erode morphology method is applied to enhance the depth map of the hand area. Subsequently, a Gaussian filter is applied to improve the depth of the surface. The intrinsic matrix of the stereo camera is used to obtain the 3D world coordinates of the hand pose. The 3D coordinates (x, y, z) are calculated at the specific pixel coordinates of the hand landmark. Finally, virtual hand skeletons are obtained by combining all hand poses, resulting in visualization, as shown in Fig. 4. The entire procedure of the algorithm is shown in pseudo-code in Table 3.

3.2. Hand pointing system

The positioning of the hand, cursor, and their movements play a crucial role in enhancing the user experience. Regarding pointing technique, the projection and ray pointing technique are commonly used to determine the cursor position. The projection approach considers the rays to be parallel to each other and perpendicular to the screen; therefore, the cursor is the projected point of a given point from the hand onto the monitor. This method allows precise control of the cursor. However, it requires broader movements of the hand, which causes faster fatigue during control. On the other hand, the ray-pointing approach considers a ray passing through a vector at a given point. The cursor is an intersection point between the ray and the monitor plane. This method allows control of the cursor based on hand direction vector without requiring large hand movement; however, it is very sensitive and requires a high accuracy of the hand direction vector to eliminate vibration problems. Both techniques have advantages and disadvantages, a key factor here being the trade-off between hand movement and cursor sensitivity. Therefore, in this HMI system, we proposed implementing a teleportation system to ensure smooth cursor movement that aligns naturally with the user's hand direction without wider movement. The main concept of the teleporting method is based on two pointing techniques: projection and ray pointing, as shown in Fig. 5. In the camera frame, the monitor can be considered as a plane with one point P and a given unit normal vector to the plane $\vec{n}$, while the hand pointing direction can be modelled by two hand landmark coordinates, A and B. Assuming that point A is the ray origin, the coordinate of the projected point A' is calculated as follows:

$$A' = A - d_{AA'}\vec{n} \tag{1}$$

where $d'_{AA} = \overrightarrow{PA} \bullet \vec{n}$ is a distance between A and A′. The intersection coordinate is calculated as follows:

$$I = A + d_{AI}\overrightarrow{AB} \tag{2}$$

$$d_{AI} = \frac{\overrightarrow{PA} \bullet \vec{n}}{\vec{u} \bullet \vec{n}} \tag{3}$$

$$\vec{u} = \frac{\overrightarrow{AB}}{\left|\overrightarrow{AB}\right|} \tag{4}$$

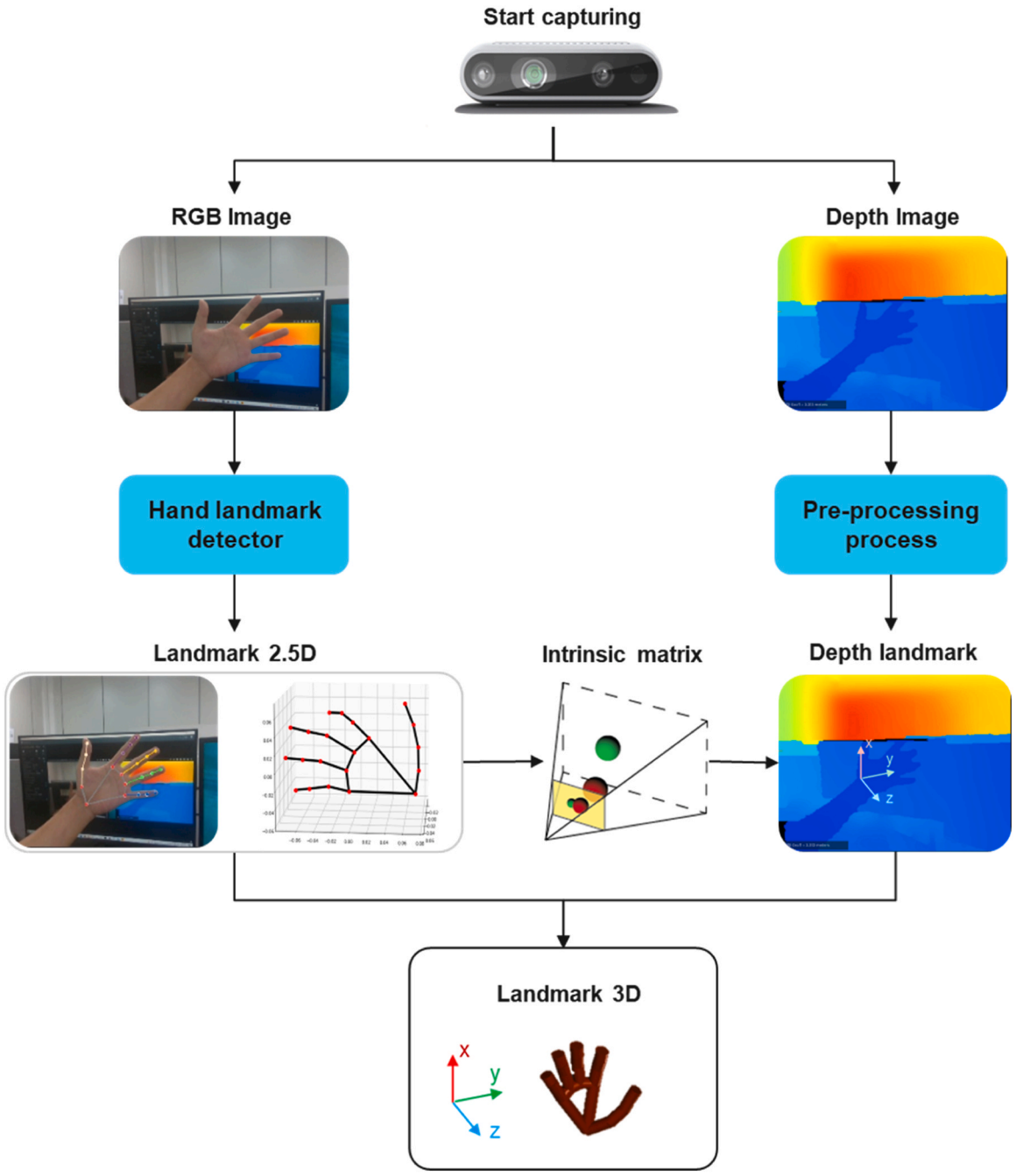


Fig. 4. Flow chart of virtual world landmark construction.

Table 2
Pseudo code of depth pre-processing step.

Algorithm: Depth enhancement
Input: Depth image from camera Output: Enhanced depth image 1. Get depth image from camera. 2. Apply temporal filter to increase persistency 3. Apply hole filling filer to remove missing value 4. Crop hand bounding box ROI 5. Apply erode morphological transformation 6. Apply Gaussian blur to increase smoothness 7. Return enhanced depth image

Table 3
Pseudo code of virtual hand landmark construction.

Algorithm: Virtual hand landmark construction
Input: Depth image from camera Output: 3D hand landmarks 1. Get RGB and depth image from camera. 2. Apply Mediapipe hand detector on RGB image 3. Apply preprocessing process on depth image 4. Get 2D hand landmark and intrinsic matrix of camera 5. Deprojection of 2D pixel coordinate on depth image to obtain world coordinates 6. Gather and return the 3D hand world landmarks

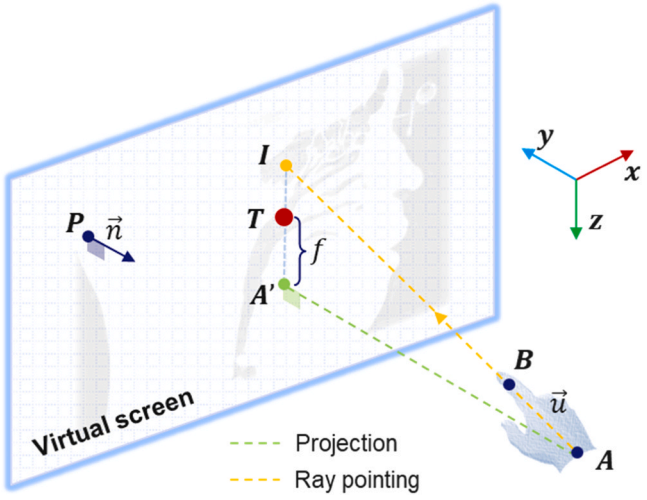


Fig. 5. Configuration of hand pointing system.

where d_{AI} is the distance from A to the intersection point and $\vec{u}$ is the unit hand pointing direction vector. After calculating the projection and intersection points, the teleporting point T is calculated as follows:

$$T = A' + f(I - A') \quad 0 \le f \le 1 \tag{5}$$

where f is a sensitivity coefficient which allows the user to easily control the cursor sensitivity.

To reflect the expected position on the monitor appropriately, the cursor coordinates were converted from real-world coordinates to pixel coordinates. Therefore, a simple calibration process is used to determine the scale values.

$$Scale = h_m / h_r \tag{6}$$

Here, h_m , h_r are the height of the monitor in millimetres and the height resolution of the monitor in pixels, respectively.

When moving the cursor, a vibration problem occurs because of the vibration effect on hand-pose detection. Many filters can be used to reduce the vibration problem, such as the moving average filter, exponential moving average filter, Kalman filter, and One Euro filter. Because the moving process requires a smooth and non-lagging problem, a One Euro filter is applied in both the x and y directions. The One Euro filter is a powerful low-pass filter algorithm for real-time noisy signals [37]. It utilises an adaptive smoothing factor to balance the trade-off between jitter and lag. Therefore, it does not exhibit lag at high speeds and reduces jitter at low speeds.

Regarding the calibration process, we simply point a hand at the centre of the main monitor after setting camera and monitor location and press a certain key to save the hand location as a reference origin. This process only needs to be done once and that information can be reused to determine the cursor position the next time.

3.3. Hand gesture recognition

Based on the hand-tracking method, hand gestures were developed to communicate between the hand and the functions of the DICOM viewer. The viewer requires functions such as clicking, dragging, scrolling images, zooming in, zooming out, rotating, and panning. Recognition of target hand states and hand actions is required to activate these functions as shown in Fig. 6. Two approaches are categorised based on the input data: hand states are defined based on a single image, while hand actions are detected based on a series of historical images.

Four natural hand gestures were developed following the minimalism concept: fist, palm, one, and two. An intuitive landmark geometry constraint algorithm was developed to determine the hand status based on the detected hand landmarks of each frame as shown in Fig. 7a. First, the palm position is selected as the pivot joint and the relative distances of the metacarpophalangeal joint (MCP), proximal interphalangeal joint (PIP), distal interphalangeal joint (DIP), and fingertip are calculated as d1, d2, d3, and d4, respectively. By applying the constraints defined in Eq. 7 to the order of these distances, each finger can be differentiated as either closed or open.

$$finger\ status = \begin{cases} 1 \quad (open) & : if\ d_1 < d_2 < d_3 < d_4 \\ 0 \quad (close) & : other \end{cases} \tag{7}$$

A template is then created as a list consisting of four single-finger statuses to determine the status of the entire hand as follows:

$$hand\ status = \begin{cases} fist : if\ template = [0,0,0,0,0] \\ palm : if\ template = [1,1,1,1,1] \\ one : if\ template = [0,1,0,0,0] or [1,1,0,0,0] \\ two : if\ template = [0,1,1,0,0] or [1,1,1,0,0] \end{cases} \tag{8}$$

The second approach uses historical images to determine the pushing and rotating actions, as shown in Fig. 8. The pushing detection algorithm is based on the change in fingertip position along the pointing direction using the following equation:

$$d_{push} = \overrightarrow{DD'} \bullet \overrightarrow{u_{CD}} \tag{9}$$

$$push\ status = \begin{cases} push\ forward : if\ d_{push} \ge \mu d_{CD} \\ not\ pushing : if\ d_{push} < \mu d_{CD} \end{cases} \tag{10}$$

where $\overrightarrow{DD'}$ is the change vector of the fingertip, as shown in Fig. 8a, and $\overrightarrow{u_{CD}}$ denotes the unit normal pointing direction vector. The empirical coefficient $\mu = 1/3$ is multiplied by a finger distance to determine if the user's hand is being pushed forward. The push status is accumulated to make a clicking decision by comparing it with a user-defined pushing sensitivity coefficient. Regarding rotation detection, the cursor when the user performs a counterclockwise and clockwise rotation action is as shown in Fig. 8b. Historical data are gathered to recognise a rotation gesture including the x- and y-coordinates of the cursor. These data are then converted into one-dimension datasets and trained using a long short-term memory (LSTM) model. A list of historical cursor positions was fed into the trained model to predict each frame.

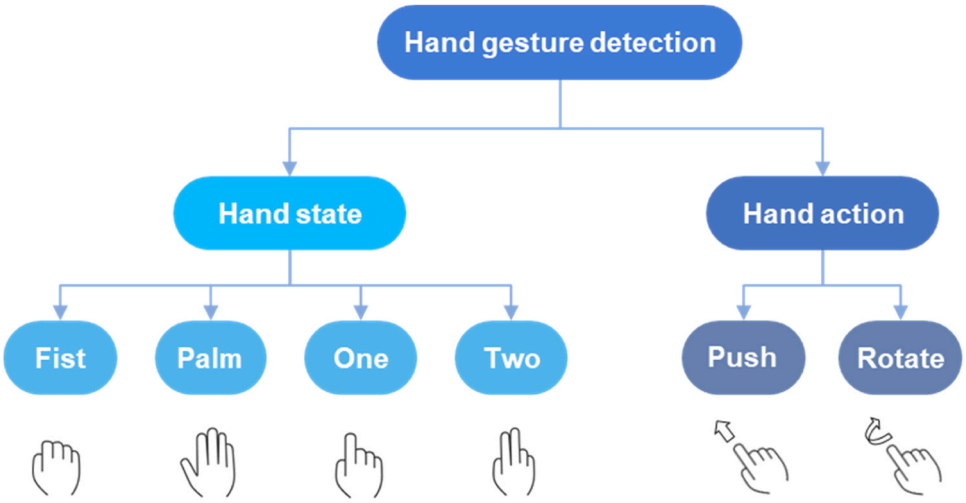


Fig. 6. Gesture recognition targets.

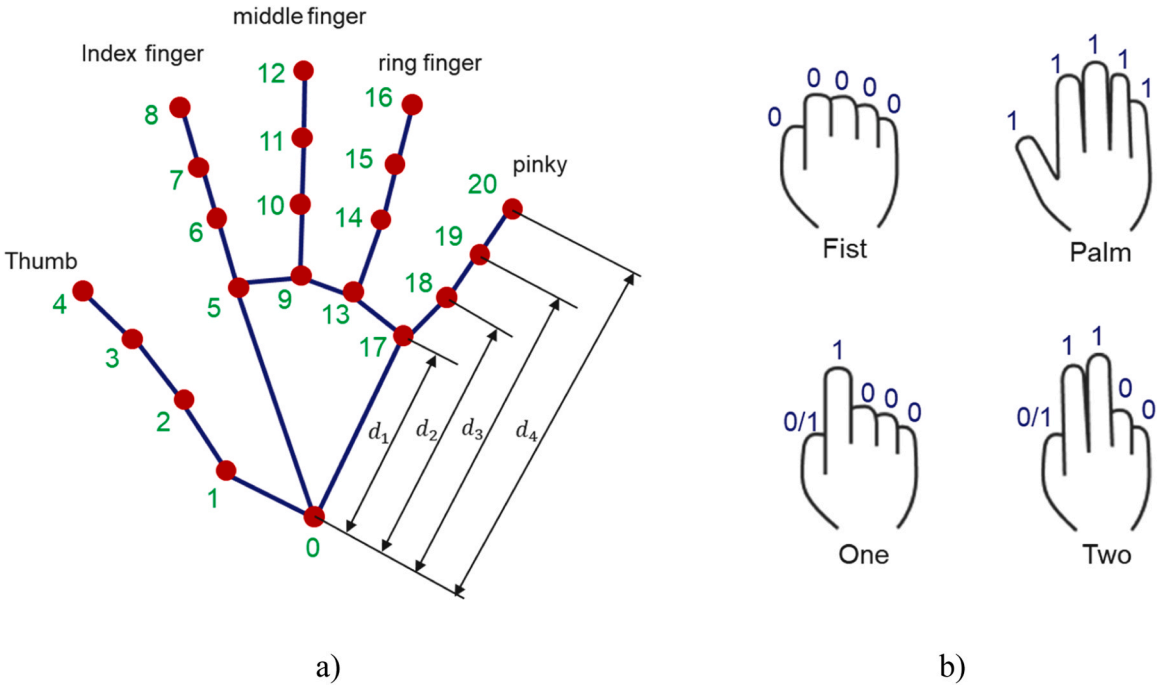


Fig. 7. Hand state recognition a) Landmark geometry constraint algorithm and b) finger state template.

4. Results and discussions

4.1. HMI integrated system

The HMI system integrated individual modules into a compact package. The field-of-view (FOV) of the pointing system was followed by the depth FOV of the camera at 87° × 58°. The preferred vertical range from the camera to the user's hand was 300–1200 mm, and the preferred horizontal range was 600–2200 mm. Regarding the hand reconstruction accuracy, our proposed method showed better accuracy than using the original depth value, as shown in Table 4. This is because the accuracy of the depth value depends on the distance between the camera and the user's hand, which decreases as the distance increases. If the distance is less than 600 mm, there is no problem with the depth value as the hand area is still large. However, if the depth exceeds 600 mm, the accuracy of the depth is reduced. In such situations, the tip depth value cannot be observed, resulting in a poor reconstruction of virtual hand and incorrect pointing ray. Conversely, our proposed techniques show good performance. This large working range is suitable for positioning the surgeon. In addition, the system allows the control of the sensitivity coefficient to customise the sensitivity for each user.

In terms of gesture recognition, we utilised a hand gesture recognition image dataset (HaGRID) sample of 30k 380p [38] to validate the performance of the combined hand tracking model and landmark geometry constraint algorithm. This was an RGB sample dataset that included 18 gestures under various conditions such as different light conditions, distances from the camera to the hand, and hand positions. Although the dataset was not specifically collected from the operating room, it showed similar gestures; therefore, four proposed gestures were extracted from the original data and used to make a prediction. All four gestures exhibited a good accuracy exceeding 97%, as listed in Table 5.

Besides that, although the system can show a large working area, it also has limitations because of hardware specifications. Normally, image resolution has limitations in the number of pixels, if an object is far from

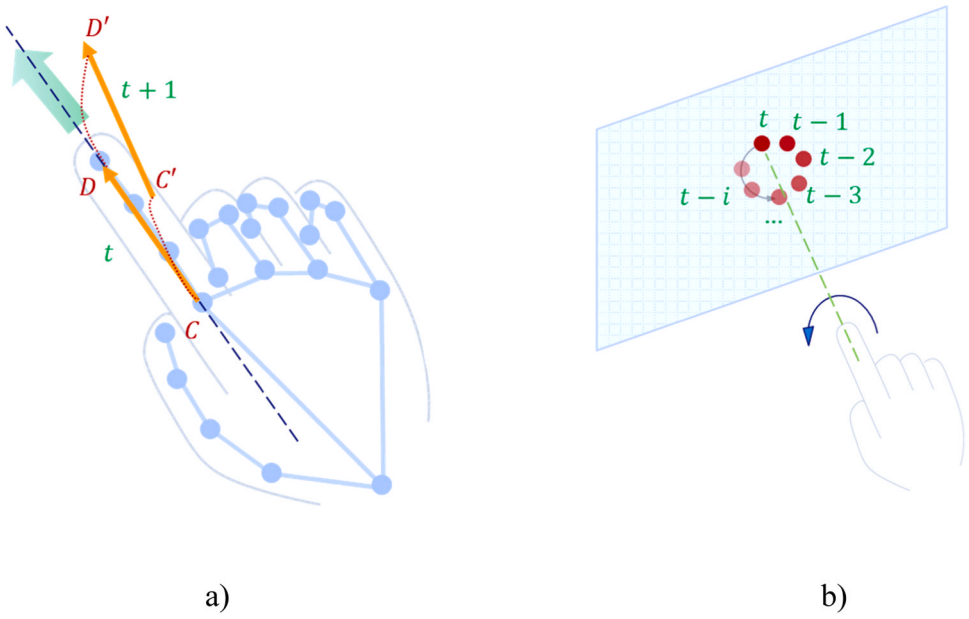


Fig. 8. Action recognition: a) pushing detection and b) rotation detection.

Table 4
Comparison of the Mediapipe + original depth and Mediapipe + our proposed method with respect to distance from camera to user hand.

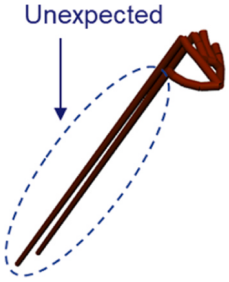
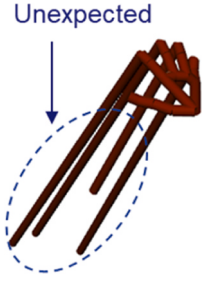
Distance from camera to user hand (mm)	300	600	900	1200
Mediapipe (2D) + depth map			 Unexpected	 Unexpected
Mediapipe (2D) + **proposed method**				

Table 5
Gesture recognition accuracy.

Gesture	Fist	Palm	One	Two
Number of testing images	1324	4154	1252	5413
Accuracy	98.34%	99.93%	97.36%	98.31%

the camera, the depth map cannot include all the positions of the object and usually causes a large error. If we think about the improvement of hardware specifications, it will raise another problem regarding capture speeds. It is always a trade-off; therefore, we believe that it is not a good approach. Recently, there have been some potential methods using deep learning architecture for depth reconstruction from monocular cameras. It means that we can apply this method to predict the depth of information instead of directly capturing it. This can be a novel approach to extend the detection scope.

4.2. Usability study validation

To validate the performance of the proposed HMI system, we performed five common functions, namely, click, image browser control, zoom, drag, and brightness control. The experiment is similar to the designed system in that it includes two monitors, one monitor shows the DICOM viewer and monitoring windows on another monitor as shown in Fig. 9. The camera is set on the topside and in front of monitors and the hand inside the region of the camera is shown on the right-hand side in Fig. 9. In this study, the CBCT head dataset was used to simulate control process such as zoom, drag, and change brightness. Regarding the computational demand for system, we suggest the following hardware computational cost based on our system: GPU: Intel core i7–7700, 32 GB RAM memory and GPU GeForce RTX 2070. These functions were used to manipulate the CBCT data in the Radiant DICOM viewer. The distance from the camera to hand is approximately 500 mm, and the sensitivity coefficient is set to 0.5. Figs. 9 and 10 show examples of the zoom and

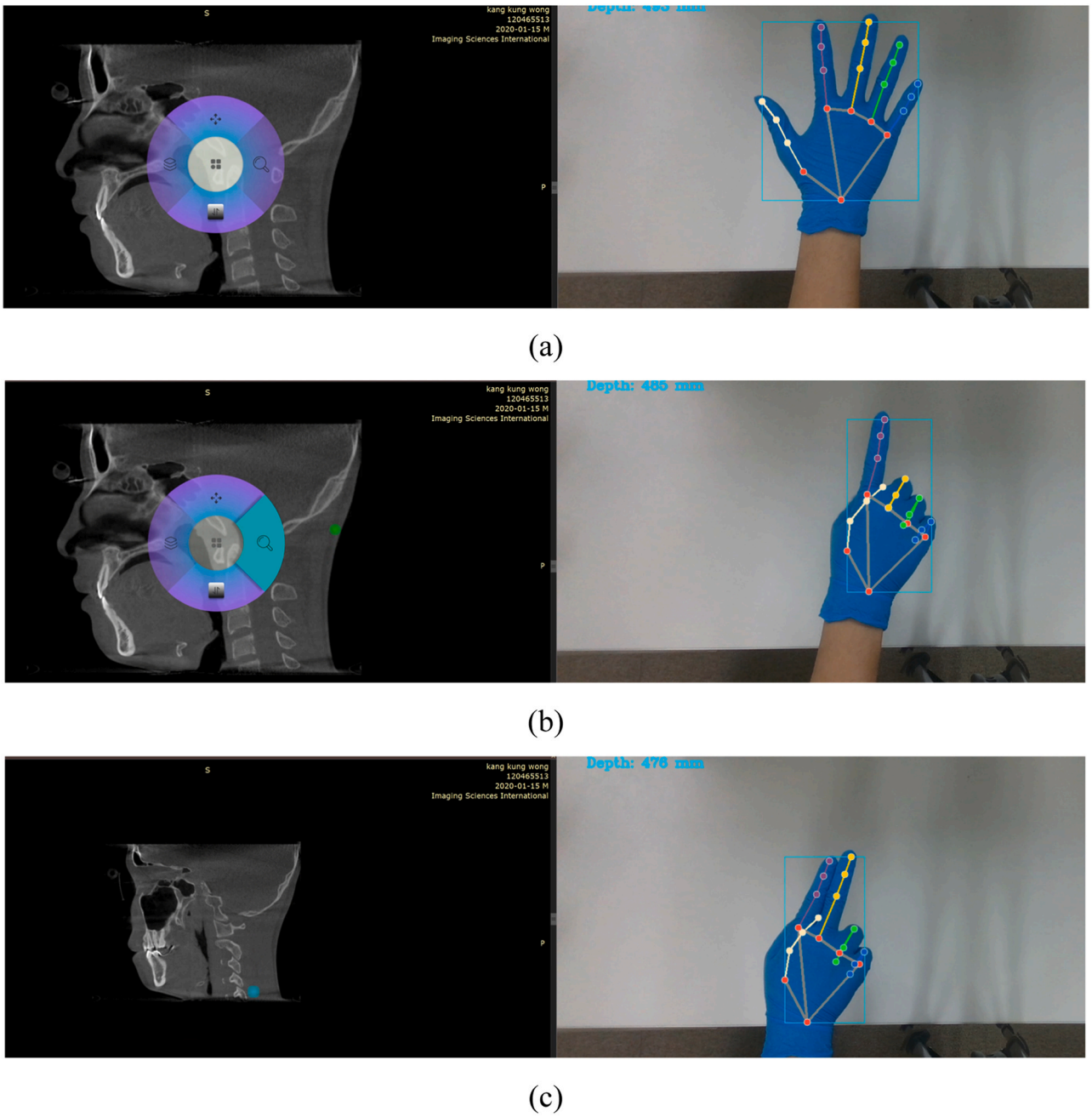


Fig. 9. Zoom function. a) open menu, b) select zoom sub menu, and c) zoom control.

pane functions, respectively. To perform each function, users are required to follow three steps. First, they need to open the menu by a palm gesture, then select the submenu by moving the cursor in the respective direction and changing to any other gesture, and finally using two fingers to activate and control the function. The time required to activate each function was less than 1 s, although it may vary depending on the user skill. In addition, the rotation gesture can be customised to control the image layer or to zoom in or out in either a clockwise or counterclockwise direction. These functions can be used to select a region of interest from CT scans or X-ray images. Users can scroll through a series of images and select corresponding images. Subsequently, the image should be zoomed in or panned to find an abnormal position on the CT image. Besides that, the click function allows the user to select any button in software with touch graphical user interface design. With this HMI system, users can efficiently and conveniently manipulate medical images using natural gestures without experiencing any discomfort.

5. Conclusion

In this paper, we presented a method for developing a contagious infection-free medical interaction system to manipulate a DICOM viewer in an operating room. We proposed the depth enhancement algorithm and combined it with deep learning hand recognition model to reconstruct virtual hand landmarks. A hand-pointing system is developed to define the cursor based on the proposed teleport method with the proposed sensitivity coefficient for reducing hand fatigue. We developed a hand gesture recognition algorithm by combining the hand landmark detector and landmark geometry constraint algorithm. Five common functions are used in the prototype system to validate the performance of the proposed HMI system. The following conclusions were drawn:

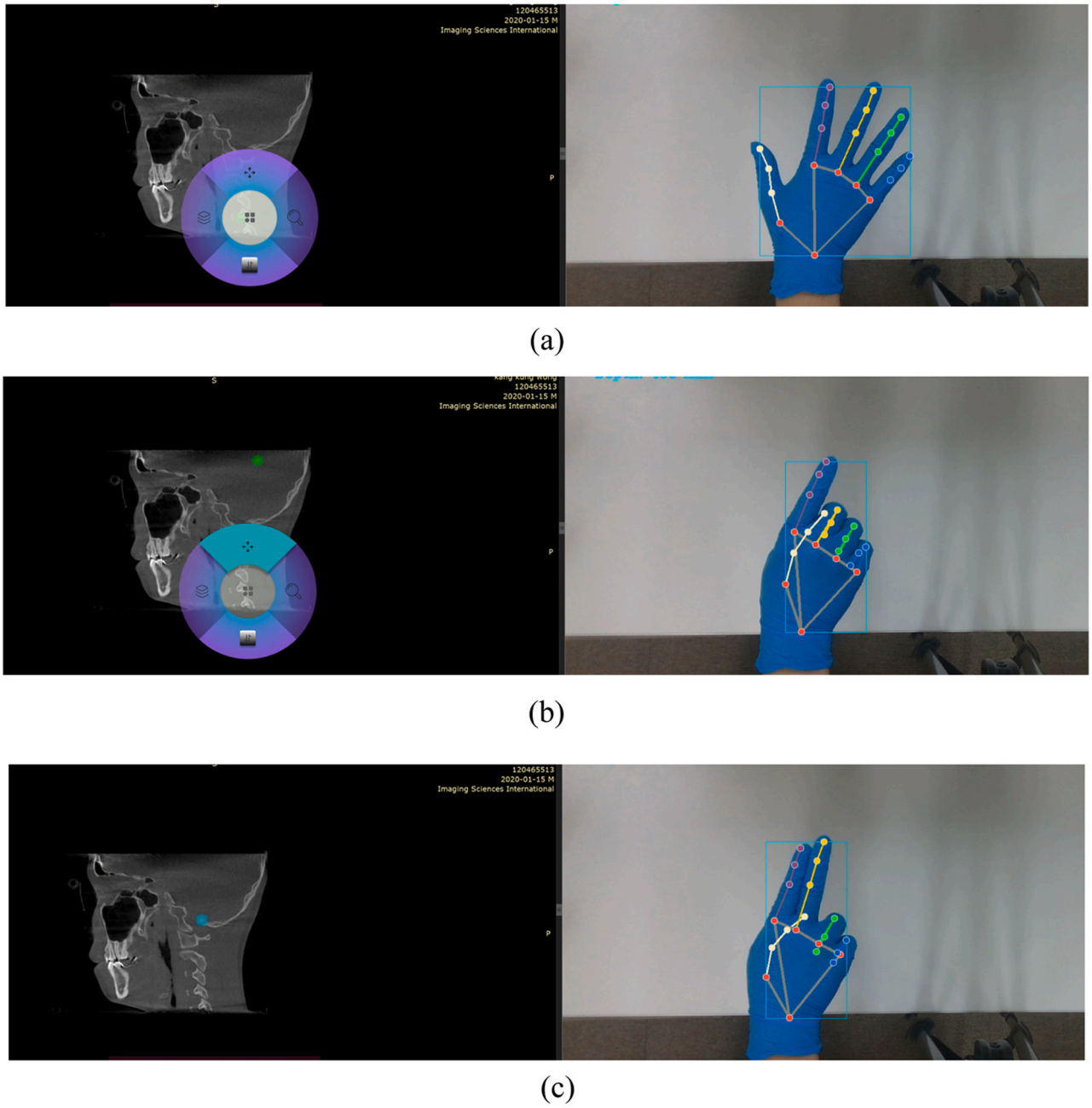


Fig. 10. Pane function a) open menu, b) select pane sub menu and c) pane control.

- The performance of depth enhancement algorithm shows that pre-processing of depth is a crucial step that reduces jitter during hand pointing and enhances 3D hand landmark reconstruction.
- Hybrid pointing technique can be applied to increase movement range of cursor with less hand movement compared to projection method. This can reduce hand fatigue during the interaction process.
- Many steps can affect the inference performance of system, therefore a minimalism concept for hand gestures can be applied to optimize the integrated system, and the landmark geometry constraint algorithm can be used to boost the recognition speed of the static hand gestures.
- Functions may vary from medical imaging viewers; the control menu concept can be used as a customised way to allow users to manipulate medical images naturally and effectively.

Limitation and potential future work: Current work is only limited for one hand, the main reason of this limit is to optimize the processing time and tracking instead of multi hand, besides that the multi-hand occlusion remains a challenge for our approach. In terms of multi-hand situation, one potential solution is applying a hand tracking method to recognize the new hand or lost hand during manipulation. In an occlusion situation, deep learning with a transformer architecture can be a potential solution to improve hand detection accuracy. By borrowing ideas from a multimodal approach with different input sources, our system can potentially be used to integrate the voice control system to allow natural interactions between users and machines. This study is limited to medical applications but can be expanded to other fields such as hologram control, interactions in future automobiles or exhibition models.

CRediT authorship contribution statement

Jonghun Yoon: Writing – review & editing, Resources, Project administration, Methodology, Conceptualization. **Van Doi Truong:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Hyun-Kyo Lim:** Writing – review & editing, Conceptualization. **Than Trong Khanh Dat:** Writing – review & editing, Conceptualization. **Seongje Kim:** Writing – review & editing, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2022R1A4A3031263).

This research was financially supported by the Ministry of Trade, Industry, and Energy (MOTIE), Korea, under the "170k closed section roll forming and free curvature bending technology development for electric vehicle body" (reference number 20022814) supervised by the Korea Institute for Advancement of Technology (KIAT).

This work was supported by the Industrial Strategic Technology Development Program-A program for win-win type innovation leap between middle market enterprise and small & medium sized enterprise (P0024516, Development and commercialization of a customized dental solution with intelligent automated diagnosis technology based on virtual patient data) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea) and the Korea Institute for Advancement of Technology (KIAT).

This research was funded by the Korea Institute of Industrial Technology's research project (KITECH, EH-24-0002).

References

[1] Ruppert GCS, Reis LO, Amorim PHJ, de Moraes TF, da Silva JVL. Touchless gesture user interface for interactive image visualization in urological surgery. World J Urol 2012;vol. 30:687–91.
[2] Bockhacker M, Syrek H, von Elster ME, Schmitt S, Roehl H. Evaluating usability of a touchless image viewer in the operating room. Appl Clin Inform 2020;vol. 11 (01):088–94.
[3] Hsiang J. Wrong-level surgery: a unique problem in spine surgery. Surg Neurol Int 2011;vol. 2.
[4] G.W.H. Organization, *WHO Guidelines for Safe Surgery 2009-Safe Surgery Saves Lives*. 2009.
[5] Wachs JP, et al. A gesture-based tool for sterile browsing of radiology images. J Am Med Inform Assoc 2008;vol. 15(3):321–3.
[6] Gallo L, Placitelli AP, Ciampi M. Controller-free exploration of medical image data: experiencing the Kinect. 2011 24th international symposium on computer-based medical systems (CBMS). IEEE; 2011. p. 1–6.
[7] Hötker AM, Pitton MB, Mildenberger P, Düber C. Speech and motion control for interventional radiology: requirements and feasibility. Int J Comput Assist Radiol Surg 2013;vol. 8:997–1002.
[8] Iannessi A, Marcy P-Y, Clatz O, Fillard P, Ayache N. Touchless intra-operative display for interventional radiologist. Diagn Interv Imaging 2014.
[9] Jacob MG, Wachs JP. Context-based hand gesture recognition for the operating room. Pattern Recognit Lett 2014;vol. 36:196–203.
[10] Yoshimitsu K, Muragaki Y, Maruyama T, Yamato M, Iseki H. Development and initial clinical testing of "OPECT": an innovative device for fully intangible control of the intraoperative image-displaying monitor by the surgeon. Oper Neurosurg 2014;vol. 10(1):46–50.
[11] Nouei MT, Kamyad AV, Soroush AR, Ghazalbash S. A comprehensive operating room information system using the kinect sensors and RFID. J Clin Monit Comput 2015;vol. 29:251–61.
[12] Ogura T, Sato M, Ishida Y, Hayashi N, Doi K. Development of a novel method for manipulation of angiographic images by use of a motion sensor in operating rooms. Radiol Phys Technol 2014;vol. 7:228–34.
[13] Mewes A, Saalfeld P, Riabikin O, Skalej M, Hansen C. A gesture-controlled projection display for CT-guided interventions. Int J Comput Assist Radiol Surg 2016;vol. 11:157–64.
[14] Cho Y, Lee A, Park J, Ko B, Kim N. Enhancement of gesture recognition for contactless interface using a personalized classifier in the operating room. Comput Methods Prog Biomed 2018;vol. 161:39–44.
[15] De Smedt Q, Wannous H, Vandeborre J-P. Skeleton-based dynamic hand gesture recognition. Proc IEEE Conf Comput Vis Pattern Recognit Workshops 2016:1–9.
[16] Lee A-r, Cho Y, Jin S, Kim N. Enhancement of surgical hand gesture recognition using a capsule network for a contactless interface in the operating room. Comput Methods Prog Biomed 2020;vol. 190:105385.
[17] Ebert LC, Hatch G, Thali MJ, Ross S. Invisible touch—control of a DICOM viewer with finger gestures using the Kinect depth camera. J Forensic Radiol Imaging 2013;vol. 1(1):10–4.
[18] Sánchez-Margallo FM, Sánchez-Margallo JA, Moyano-Cuevas JL, Pérez EM, Maestre J. Use of natural user interfaces for image navigation during laparoscopic surgery: initial experience. Minim Invasive Ther Allied Technol 2017;vol. 26(5): 253–61.
[19] Miah ASM, Hasan MAM, Tomioka Y, Shin J. Hand gesture recognition for multi-culture sign language using graph and general deep learning network. IEEE Open J Comput Soc 2024.
[20] Miah ASM, Hasan MAM, Shin J. Dynamic hand gesture recognition using multi-branch attention based graph and general deep learning model. IEEE Access 2023; vol. 11:4703–16.
[21] Islam MN, et al. A multilingual handwriting learning system for visually impaired people. IEEE Access 2024.
[22] Zhou B, et al. Decoupling and recoupling spatiotemporal representation for rgb-d-based motion recognition. Proc IEEE/CVF Conf Comput Vis Pattern Recognit 2022: 20154–63.
[23] Miah ASM, Hasan MAM, Nishimura S, Shin J. Sign language recognition using graph and general deep neural network based on large scale dataset. IEEE Access 2024.
[24] Shin J, Miah ASM, Suzuki K, Hirooka K, Hasan MAM. Dynamic Korean sign language recognition using pose estimation based and attention-based neural network. IEEE Access 2023.
[25] Chen Y, Zuo R, Wei F, Wu Y, Liu S, Mak B. Two-stream network for sign language recognition and translation. Adv Neural Inf Process Syst 2022;vol. 35:17043–56.
[26] Li D, Rodriguez C, Yu X, Li H. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2020. p. 1459–69.
[27] Rosa GM, Elizondo ML. Use of a gesture user interface as a touchless image navigation system in dental surgery: case series report. Imaging Sci Dent 2014;vol. 44(2):155–60.
[28] Cao Z, Simon T, Wei S-E, Sheikh Y. Realtime multi-person 2d pose estimation using part affinity fields. Proc IEEE Conf Comput Vis Pattern Recognit 2017:7291–9.
[29] Eastman P, et al. OpenMM 7: rapid development of high performance algorithms for molecular dynamics. PLoS Comput Biol 2017;vol. 13(7):e1005659.
[30] F. Zhang *et al.*, Mediapipe hands: On-device real-time hand tracking, *arXiv preprint arXiv:2006.10214,* 2020.
[31] "Ultraleap." https://www.ultraleap.com/. (accessed.
[32] Jiang C, et al. A2J-Transformer: anchor-to-Joint Transformer Network for 3D Interacting Hand Pose Estimation from a Single RGB Image. Proc IEEE/CVF Conf Comput Vis Pattern Recognit 2023:8846–55.
[33] J. Romero D. Tzionas M.J. Black Embodied hands: modeling and capturing hands and bodies together arXiv Prepr arXiv:2201 02610 2022.
[34] Cai Y, Ge L, Cai J, Yuan J. Weakly-supervised 3d hand pose estimation from monocular rgb images. Proc Eur Conf Comput Vis (ECCV) 2018:666–82.
[35] Docekal J, Rozlivek J, Matas J, Hoffmann M. Human keypoint detection for close proximity human-robot interaction. 2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids). IEEE; 2022. p. 450–7.
[36] M. De Coster E. Rushe R. Holmes A. Ventresque J. Dambre Towards the extraction of robust sign embeddings for low resource sign language recognition arXiv Prepr arXiv 2306 2023 17558.
[37] Casiez G, Roussel N, Vogel D. 1€ filter: a simple speed-based low-pass filter for noisy input in interactive systems. Proc SIGCHI Conf Hum Factors Comput Syst 2012:2527–30.
[38] A. Kapitanov A. Makhlyarchuk K. Kvanchiani HaGRID-HAnd gesture recognition image dataset arXiv Prepr arXiv:2206 08219 2022.