A Human-Computer Collaborative Behavior Measurement Model for Assembly in Constrained Visibility Environments

Hongyuan Zhan[1,2], Zhuo Wang[1,2,*], Yang Wang[3,*], Qingtian Lu[1] and Minglei Zhu[1]

[1] School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai, 200093, China

[2] Institute for Underwater Robotics and Intelligent System, University of Shanghai for Science and Technology, Shanghai, 200093, China

[3] System Integration Department, AVIC Civil Aircraft Airborne Systems Engineering Center Co., Ltd., Shanghai, 200245, China

Revista Internacional
Métodos numéricos
para cálculo y diseño en ingeniería
RIMNI
UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH
UPC
In cooperation with
CIMNE

INFORMATION

Keywords:
Human-machine collaboration
gesture recognition
mixed reality

DOI: 10.23967/j.rimni.2025.10.58350

SCIPEDIA
Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería
RIMNI

A Human-Computer Collaborative Behavior Measurement Model for Assembly in Constrained Visibility Environments

Hongyuan Zhan[1,2], **Zhuo Wang**[1,2,*], **Yang Wang**[3,*], **Qingtian Lu**[1] **and Minglei Zhu**[1]

[1]School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai, 200093, China

[2]Institute for Underwater Robotics and Intelligent System, University of Shanghai for Science and Technology, Shanghai, 200093, China

[3]System Integration Department, AVIC Civil Aircraft Airborne Systems Engineering Center Co., Ltd., Shanghai, 200245, China

ABSTRACT

Collision interference detection is an important concern in the manual installation of cable harnesses. Due to the complex and variable layout of cable harness installations, hand assembly movements are prone to collisions, contact, and other forms of interaction with surrounding equipment. Furthermore, cable harnesses often need to be routed along specific paths in visually constrained environments. Currently, there is a lack of modeling in the extraction of hand motion parameters and the data analysis of hand action intent. To address this challenge, we propose a novel human-machine collaboration behavior measurement model. This model not only provides a rapid solution for extracting hand motion parameters but also delivers efficient and natural visual feedback for the behavioral intent reflected by hand motion features. First, we introduce a hand motion parameter extraction mechanism based on a hand kinematics model. Second, we develop a virtual-to-real spatial registration model specifically designed for visually constrained conditions, enabling accurate recognition and 3D calibration of hand action intent. A user study experiment demonstrates that the proposed model outperforms traditional hand behavior measurement models in terms of manual task efficiency, hand motion recognition accuracy, and the naturalness of hand interactions. This improvement is particularly evident in visually constrained environments, effectively addressing challenges in obstacle avoidance and intent inference during spatially constrained assembly tasks.

OPEN ACCESS

Received: 10/09/2024

Accepted: 08/11/2024

Published: 20/04/2025

DOI
10.23967/j.rimni.2025.10.58350

Keywords:
Human-machine collaboration
gesture recognition
mixed reality

Glossary/Nomenclature/Abbreviations

HCBM	Human-Computer Collaborative Behavior Measurement Model
MR	Mixed reality
AR	Augmented reality

*Correspondence: Zhuo Wang, Yang Wang (wz_jd2013@163.com, wangy134_mj@163.com).

1 Introduction

Cable harnesses, a type of flexible component, serve as the medium for energy and signal transmission, acting as the "link" connecting electronic devices and subsystem modules [1]. Due to the narrow wiring space inside an aircraft's fuselage, the dense distribution of electronic components, the wide variety of connectors, and their large quantity, there are numerous obstructive components along the harness routing that hinder manual operations [2]. A survey by [3] in the past decade on the field of intelligent assembly indicated that assembly/disassembly is indeed one of the primary manufacturing operations where AR/MR assistance is commonly applied. However, due to the lack of effective and accurate methods for classifying manual intent, MR-assisted systems often only use perspective views to provide visual cues for hand movements behind obstructive components. This reliance on physical contact poses a risk of damaging cable harness connectors. Therefore, it is necessary to explore the advantages of new human-machine collaboration behavior models in supporting gesture intent recognition.

In scenarios where visibility is restricted, flexible components impose limitations on gesture recognition, and existing image processing models [4] struggle to achieve low latency and high precision. Consequently, it is crucial to improve the accuracy of computer graphics algorithms. While image preprocessing techniques, such as image enhancement and restoration, can improve the quality of input images, they are not the optimal solution for enhancing gesture recognition quality [5,6]. On the other hand, gesture recognition methods based on convolutional neural networks (CNNs) can improve the accuracy of intent classification in spatially constrained assembly tasks to some extent, but they still face challenges in accurately classifying unstructured information such as images, pictures, and video streams. Therefore, a target feature registration method is required to correct errors in gesture recognition caused by hand occlusion [7]. Given the high-quality demands of assembly processes in restricted visibility environments, the classification of gesture intent data should be more rigorous. Understanding how to ensure the accuracy of gesture classification results and utilizing a new human-machine collaboration model to correct hand operation errors to enhance accuracy is of significant practical importance for advancing aircraft harness assembly toward intelligent recognition of human intent and adaptive assembly guidance.

Our research aims to demonstrate that when an operator's hand enters a visibility-restricted assembly area, a weak AI agent can adaptively detect key installation details of flexible components by analyzing the real-time occlusion relationship between the hand and the geometric space. This process is designed to mitigate the confusion or fragmented perception caused by physical obstructions in the operator's field of view through a natural human-machine interface. Based on this, the human-machine collaboration behavior detection model proposed in this study attempts to explain the data mapping mechanism among occlusion relationships, gesture recognition, and operational intent. While some researchers have explored gesture recognition within MR assembly guidance systems for spatially constrained environments, concluding that accurate gesture recognition indeed enhances the installation quality of precision components, existing blind spot gesture recognition often involves perspective processing of gesture cues and occluding components. To our knowledge, no studies have yet explicitly revealed the behavior collaboration mechanism between humans and wearable smart devices under visibility-restricted conditions.

2 Related Works

In visibility-restricted assembly environments, the quality of gesture intent recognition is crucial for controlling human-machine collaborative behavior. For instance, when a hand moves behind a physical obstruction, it's essential to overlay appropriate manual operation prompts based on gesture recognition results to mark the installation area. This involves two key challenges: first, accurately and in real-time identifying the positional relationship between the operator's hand and the physical obstruction; second, constructing a human-machine interaction channel that effectively presents gesture intent information. Researchers have conducted exploratory studies in the following three areas:

2.1 Hand Position Detection

Gesture motion capture is used to record the movements of the operator's hand joints while performing specified assembly actions [8]. Traditional motion capture techniques involve using specialized sensors, such as inertial measurement units (IMUs) [9] and synchronized infrared cameras [10], to collect data. These methods depend on sensor components or special markers attached to the wrist, making them unsuitable for manual assembly scenarios with limited visibility. Recently, 3D motion capture technology has evolved to use consumer-grade depth cameras to capture detailed hand joint movements. Chu et al. [11] employed Leap Motion to track operator hand behavior data in blind spots during space-constrained assembly tasks, allowing for the rough and rapid localization of assembly positions in large unknown areas. Similarly, Yang et al. [12] mounted an Intel RealSense R200 depth camera and Leap Motion controller on an HTC VIVE headset and utilized GPU parallel computing to process hand dynamic occlusions, enabling users to operate a full-scale virtual milling machine with bare hands. However, due to the limitations of hardware range and precision, these methods can only provide a rough reconstruction of gesture action [13]. Furthermore, when depth cameras are deployed in visibility-restricted environments, these approaches fail to meet the lightweight gesture motion capture requirements for low-latency, high-precision tracking.

2.2 Operation Gesture Recognition

From the perspective of gesture recognition based on image sequences, these methods often need to process a large amount of data, which proves inadequate in scenarios requiring low latency and high real-time performance, such as cable harness installation. In contrast, methods based on hand skeletal sequences are more robust to background changes and contain less redundant information, making them more suitable for real-time demands in cable harness installation. These methods typically rely on Recurrent Neural Networks (RNN) or Convolutional Neural Networks (CNN) to identify gesture features in skeletal sequences [14,15]. Some researchers have suggested that, due to CNN's ability to extract high-level information, it can effectively capture both spatial and temporal information in hand skeletal sequences [16,17]. Recently, scholars have used Graph Convolutional Networks (GCN) for gesture recognition, extending traditional CNN applications to data structures with arbitrary graphs [18]. They proposed a dynamic hand skeletal model called Spatio-Temporal Graph Convolutional Network (ST-GCN), which automatically learns spatial and temporal patterns in data, showing stronger representational and generalization capabilities. By using ST-GCN for assembly action recognition, researchers successfully reduced redundant information in gesture recognition. However, understanding subtle gestures remains a challenge in manual assembly scenarios because the appearance of the foreground and background can be quite similar, leading to significant differences in recognition performance. Therefore, in vision-constrained and spatially complex environments, accurately recognizing gestures becomes a key issue in improving the efficiency of human-robot collaboration.

2.3 Immersive Human-Computer Interaction

Numerous studies have confirmed the advantages of advanced visual enhancement technologies such as AR and MR in facilitating this specific assembly task. These technologies propose the transparent presentation of occluded hand movements during the assembly process. Wang et al. [19] developed a VR/AR-assisted system that uses Leap Motion for real-time gesture recognition to control a virtual hand model, displaying components and hand movements that the operator cannot see through VR glasses, creating an immersive environment. Feng et al. [20] developed an AR-assisted collaborative assembly system (ARCoA), enabling operators to view their partner's real-time hand gestures behind steel plates. Recently, Chu et al. [11] designed an auxiliary assembly system for confined assembly spaces, which provides the most useful information in an AR environment, prompting users on body movements required during assembly when physical objects occlude the process. However, these methods sometimes oversimplify usage conditions, overlooking behavioral changes and the complexity of tracking during manual operations. Particularly in vision-constrained conditions, when the operator's hands are completely occluded, the lack of precise visual feedback for hand movements can significantly affect the accuracy of manual operations.

3 Method

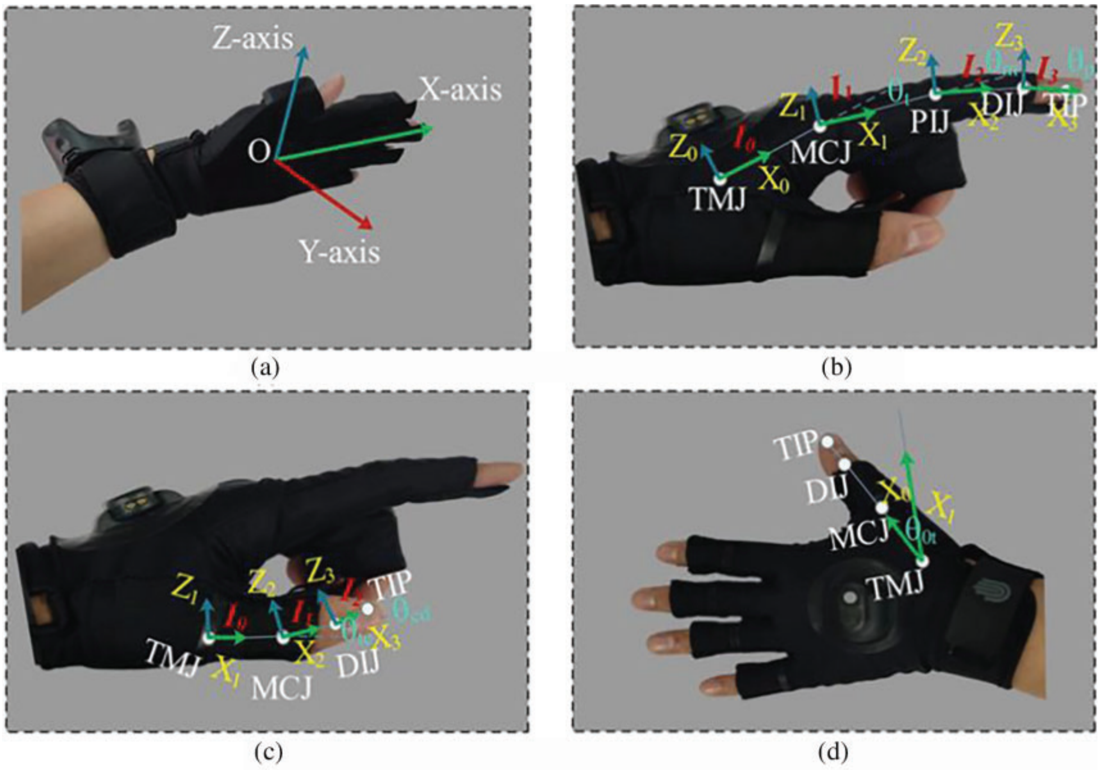
3.1 Hand Pose Perception

The key to gesture posture sensing lies in real-time positioning of any point on the fingers within the palm's coordinate system (reference coordinate system, see Fig. 1a). Specifically, each finger (except the thumb) consists of three phalanges, one metacarpal bone, and three joints. The phalanges are divided into distal phalanx, middle phalanx, proximal phalanx, and metacarpal, from fingertip (TIP) to wrist. The joints from fingertip to wrist include the distal interphalangeal joint (DIJ), proximal interphalangeal joint (PIJ), and metacarpophalangeal joint (MCJ). Unlike the other four fingers (referred to as non-thumb fingers), the thumb lacks a middle phalanx and a proximal interphalangeal joint but has an additional carpometacarpal joint (TMJ). A pair of inertial sensor data gloves is used to obtain the data on the relationship between the position of any point on the fingers and the angles of the finger joints. Kinematic equations are established for each finger, enabling the determination of the position and orientation of any point on the fingers in the coordinate system of the palm. The process of establishing the equations for the non-thumb fingers involves analyzing the relationship between the rotation angles of the joints and the measurements from the inertial sensors. By using the rotation angles of the joints and the lengths of the linkages, the positions and orientations of the points on the non-thumb fingers in the palm coordinate system (reference coordinate system) are determined.

As illustrated in Fig. 1b, the non-thumb fingers can be treated as linkage mechanisms with three degrees of freedom: flexion-extension for MCJ, PIJ, and DIJ. The linkages I_0, I_1, I_2, and I_3 represent a metacarpal bone, proximal phalanx, middle phalanx, and distal phalanx, respectively. The coordinate systems are established with the origin at the proximal end of each phalanx, and the X-axis pointing towards the distal end. The linkage I_0 is fixed, and its coordinate system $\{X_0\ Y_0\ Z_0\}$ serves as the base coordinate system. In the simplified model of the fingers, the lengths of the four linkages, l_0, l_1, l_2, and l_3, are known. The bending angles of the joints are defined as the rotation angles around a Z-axis of their local coordinate systems relative to their parent objects, which are expressed in the world coordinate system.

Due to the presence of only one sensor on each finger, the data glove can only capture the total bending angle of each finger, which is the sum of θM (metacarpophalangeal joint angle), θP (proximal interphalangeal joint angle), and θD (distal interphalangeal joint angle). When the fingers are fully extended, the rotation angles of each joint are 0°. When the fingers are clenched into a fist, the rotation angles of each joint reach their maximum. Based on the flexion and extension angle ranges of finger joints mentioned in Reference [21], we can derive the following approximate ranges:

$$\begin{aligned} &0 \leq \theta_M \leq 90^\circ \\ &0 \leq \theta_P \leq 120^\circ \\ &0 \leq \theta_D \leq 90^\circ \end{aligned} \tag{1}$$



(a) (b) (c) (d)

Figure 1: Gesture motion model construction

To estimate the individual joint angles, we can calculate them based on the proportional relationship between the maximum range of motion for each joint and the maximum range of motion for the finger as a whole.

$$\theta_i = \frac{\theta_{i\max}}{\theta_{\max}} \cdot \theta \tag{2}$$

where θ_{imax} represents the maximum bending angle for the specific joint, θ_{max} represents the maximum bending angle for the finger, which is the sum of the maximum bending angles for all the joints, and θ represents the current measurement value obtained from the sensor. By substituting the maximum bending angle values for each joint into Eq. (2), we can obtain the relationship between the joint angles and the real-time sensor measurements.

$$\begin{cases} \theta_M = 0.3 \cdot \theta \\ \theta_P = 0.4 \cdot \theta \\ \theta_D = 0.3 \cdot \theta \end{cases} \tag{3}$$

Let the coordinate system $\{X_1, Y_1, Z_1\}$ be the coordinate system of the proximal phalanx bone I_1, with its parent coordinate system being the metacarpal coordinate system $\{X_0, Y_0, Z_0\}$. The origin of the coordinate system is located at MCJ and has homogeneous coordinates $[l_0, 0, 0, 1]$. The coordinate system $\{X_1, Y_1, Z_1\}$ is rotated around a Z-axis of $\{X_0, Y_0, Z_0\}$ by an angle θ_M. According to robot kinematics theory, we can obtain a transformation matrix ${}_1^0T$ from the coordinate system $\{X_1, Y_1, Z_1\}$ to the coordinate system $\{X_0, Y_0, Z_0\}$.

$${}_1^0T = \begin{bmatrix} \cos\theta_M & -\sin\theta_M & 0 & l_0 \\ \sin\theta_M & \cos\theta_M & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{4}$$

Similarly, we can obtain the transformation matrix from the coordinate system $\{X_2\ Y_2\ Z_2\}$ of the middle phalanx bone I_2 to the coordinate system $\{X_1\ Y_1\ Z_1\}$, as well as the transformation matrix from the coordinate system $\{X_3\ Y_3\ Z_3\}$ of the distal phalanx bone I_3 to the coordinate system $\{X_2\ Y_2\ Z_2\}$. From these, we can obtain the transformation matrices from the finger segment coordinate systems $\{X_2\ Y_2\ Z_2\}$ and $\{X_3\ Y_3\ Z_3\}$ to the base coordinate system $\{X_0\ Y_0\ Z_0\}$ as follows:

$$\begin{cases} {}_2^0T = {}_1^0T \cdot {}_2^1T \\ {}_3^0T = {}_2^0T \cdot {}_3^2T \end{cases} \tag{5}$$

Let's assume that the homogeneous coordinates of an arbitrary point P on finger segment $i\,(i = 1, 2, 3)$ in coordinate system are represented as $P^i(P^i = [x_i, y_i, z_i, 1]^T)$. We can calculate the coordinates of this point P^0 in the base coordinate system as follows:

$$P^0 = {}_i^0T \cdot P^i \tag{6}$$

From Fig. 1b, we can determine the homogeneous coordinates of MCJ in the coordinate system$\{X_0\ Y_0\ Z_0\}$ as $[I_0, 0, 0, 1]^T$, the homogeneous coordinates of PIJ in the coordinate system$\{X_1\ Y_1\ Z_1\}$ as $[I_1, 0, 0, 1]^T$, the homogeneous coordinates of DIJ in the coordinate system$\{X_2, Y_2, Z_2\}$ as $[I_2, 0, 0, 1]^T$, and the homogeneous coordinates of TIP in the coordinate system$\{X_3\ Y_3\ Z_3\}$ as $[I_3, 0, 0, 1]^T$. By substituting the homogeneous coordinates of PIJ, DIJ, and TIP into Eq. (5), we can obtain the coordinates of these three points in the base coordinate system $\{X_0\ Y_0\ Z_0\}$. Once we obtain the coordinates of the joints in the fingertip (TIP) and reference coordinate systems, we can map the real state of the hand onto virtual fingers in the digitally reconstructed space in real time.

Unlike the other fingers, the thumb can be simplified as a linkage mechanism with three degrees of freedom: MCJ, DIJ, and TMJ. MCJ and DIJ rotate in the X–Y plane, while TMJ rotates in the X–Z plane. In the data glove, three sensors are deployed at a thumb, which can measure the rotation angles θ_M, θ_D, and θ_T for MCJ, DIJ, and TMJ, respectively. The linkages I_0, I_1, and I_2 represent the metacarpal bone, proximal phalanx, and distal phalanx, respectively. Coordinate systems are established with each joint as the origin and the direction of each phalanx as the X-axis (see Fig. 1c,d).

To analyze the flexion-extension motion of a finger in a X–Z plane, we define the coordinate system of the proximal phalanx as $\{X_1\ Y_1\ Z_1\}$ and that of the metacarpal as $\{X_0\ Y_0\ Z_0\}$, with both sharing the same origin. The thumb coordinate system is rotated by an angle θ_T around the Y-axis. According to the principles of robotics, we can establish the transformation relationship between the two coordinate systems as follows:

$$ {}^{0}_{1}T = \begin{bmatrix} \cos\theta_T & 0 & \sin\theta_T & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\theta_T & 0 & \cos\theta_T & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{7} $$

We define the coordinate system of the intermediate phalanx as $\{X_2\ \ Y_2\ Z_2\}$. The transformation matrix from $\{X_2\ \ Y_2\ Z_2\}$ to $\{X_1\ \ Y_1\ Z_1\}$ can be given as follows:

$$ {}^{1}_{2}T = \begin{bmatrix} \cos\theta_M & -\sin\theta_M & 0 & l_1 \\ \sin\theta_M & \cos\theta_M & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{8} $$

We define the coordinate system of the distal phalanx as $\{X_3\ \ Y_3\ Z_3\}$. The transformation matrix from $\{X_3\ \ Y_3\ Z_3\}$ to $\{X_2\ \ Y_2\ Z_2\}$ can be given as follows:

$$ {}^{2}_{3}T = \begin{bmatrix} \cos\theta_D & -\sin\theta_D & 0 & l_2 \\ \sin\theta_D & \cos\theta_D & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{9} $$

Once we have obtained the transformation matrices between the coordinate systems of each bone, similar to the analysis principle for the other fingers, we can determine the transformation relationship between each bone's coordinate system and the base coordinate system. This allows us to determine the pose of any point on the thumb in the base coordinate system. The specific calculation process is similar to Eqs. (4) and (5). By analyzing the kinematic equations of the thumb and other fingers, we can obtain real-time coordinates of each joint of the fingers in the hand coordinate system. These coordinates will be transmitted to a digital scene, where the spatial position of the hand in the world coordinate system is registered using a tracking device. This enables real-time tracking of the hand's posture in the digital environment.

3.2 Space-Constrained Assembly Guidance

In spatially constrained human-machine environments, even with pre-processing to enhance target images, issues like incomplete capture of hand images and inaccurate extraction of gesture shapes may still occur. To address these challenges, we propose an automatic mechanism for recognizing gesture intent between the hand and obstructing components, catering to the needs of both partially obstructed and fully obstructed gesture recognition, as shown in Fig. 2.

This mechanism employs a 3D reconstruction and point cloud segmentation technique based on edge computing to real-time extract unstructured behavior data such as the physical distances of 3D targets in human-machine collaboration environments, including hands, cable harnesses, brackets, fasteners, and mounting plates. As shown in Fig. 3a, HTC VIVE Tracker is a position tracking sensor designed to be used in conjunction with two tracking base stations. The Tracker utilizes precise laser signals emitted by the two base stations to perform spatial localization, determining its own pose within the coordinate system defined by the base stations. Noitom Hi5 2.0 motion capture glove, in addition to being used with the Tracker, also incorporates 12 inertial sensors located at the center of the palm and the joints of each of the five fingers. Each inertial sensor can detect rotational angles along three axes. In position tracking, multiple bones of the palm are treated as rigid bodies, and their sizes can be measured. $\{X_0\ Y_0\ Z_0\}$ represents the operator's hand coordinate system. $\{X_t\ Y_t\ Z_t\}$ represents the tracker's origin coordinate system, which is determined when the VIVE tracker is powered on. $\{X_H\ Y_H\ Z_H\}$ represents the coordinate system of HoloLens 2 glasses, which is set by the device itself at startup. Assembly guidance instructions are rendered in the glasses based on this coordinate system. $\{X_W\ Y_W\ Z_W\}$ represents the world coordinate system of the entire scene, which is used to establish the connection between the real environment and the virtual environment. In this paper, the coordinate system of fixed artificial markers in the scene is considered as the world coordinate system. Once the scene is determined, this coordinate system remains unchanged. Besides, this system addresses the issue of coordinate system unification for pre-reconstructed 3D targets in virtual reality space using Bundle Fusion-based manual identification detection and PnP positioning technology. Building on this, a background point cloud segmentation algorithm based on octree spatial retrieval and a foreground adhesion point cloud segmentation strategy based on supervoxels are proposed. These methods accurately segment independent point clouds of various targets from constrained visibility scenes, achieving high-quality reconstruction of physical scenes into corresponding digital mirror scenes. Subsequently, when the operator wears HoloLens 2 and performs real-time scanning and reconstruction of the target task scene, the digital mirror scene created in a VR space will also be activated (see Fig. 3b). During this process, the system uses ray-casting detection to determine whether there is a real-time reconstructed 3D point cloud model between the target hand and the VR camera. If the ray encounters a physical object along its path, the area of the hand's surface where this object is projected is marked as 1, while areas not triggering the collision detection remain marked as 0. For scenarios where hand movements are too small to capture detailed operations from the physical perspective, the operator can move a HoloLens 2 to adjust VR camera within the digital mirror scene to a more suitable angle for collecting additional information.

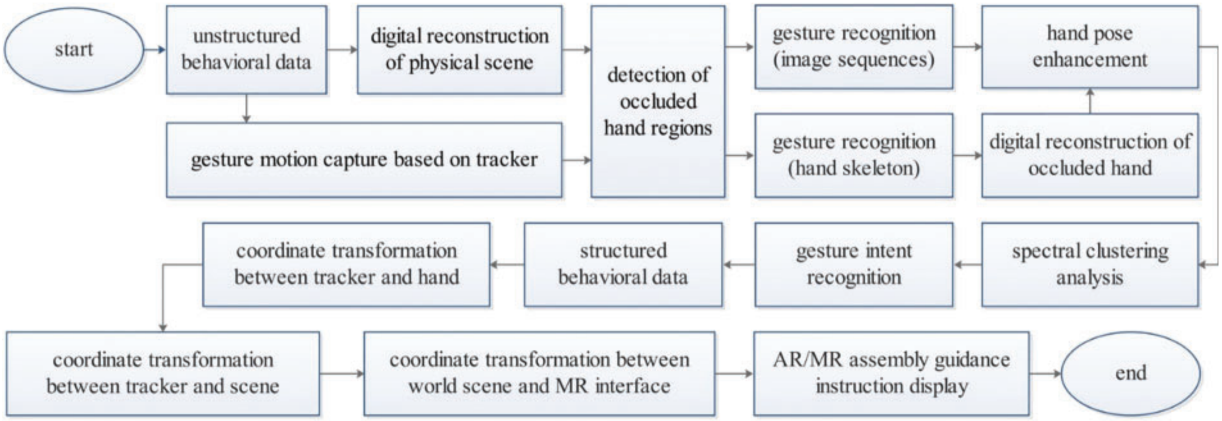

start
unstructured behavioral data
digital reconstruction of physical scene
detection of occluded hand regions
gesture recognition (image sequences)
hand pose enhancement
gesture motion capture based on tracker
gesture recognition (hand skeleton)
digital reconstruction of occluded hand
coordinate transformation between tracker and hand
structured behavioral data
gesture intent recognition
spectral clustering analysis
coordinate transformation between tracker and scene
coordinate transformation between world scene and MR interface
AR/MR assembly guidance instruction display
end

Figure 2: Gesture intention recognition mechanism

As shown in Fig. 4a, we developed HCBM using HoloLens 2 and HTC VIVE Pro2. By setting up positional tracking through artificial markers and trackers, the system enables real-time interaction between the physical scene and the digital mirror scene. Hand motion images are captured by the VR camera placed in the digital mirror scene, and these images are sent to a cloud server via the shared service module of the VR device. The cloud server performs calculations and processing on the images, outputting gesture intent recognition results. If there is any deviation in the gesture intent recognition results, the virtual and real spaces can be corrected using a spatial registration algorithm. On the other hand, the system architecture of HCBM is shown in Fig. 4b.

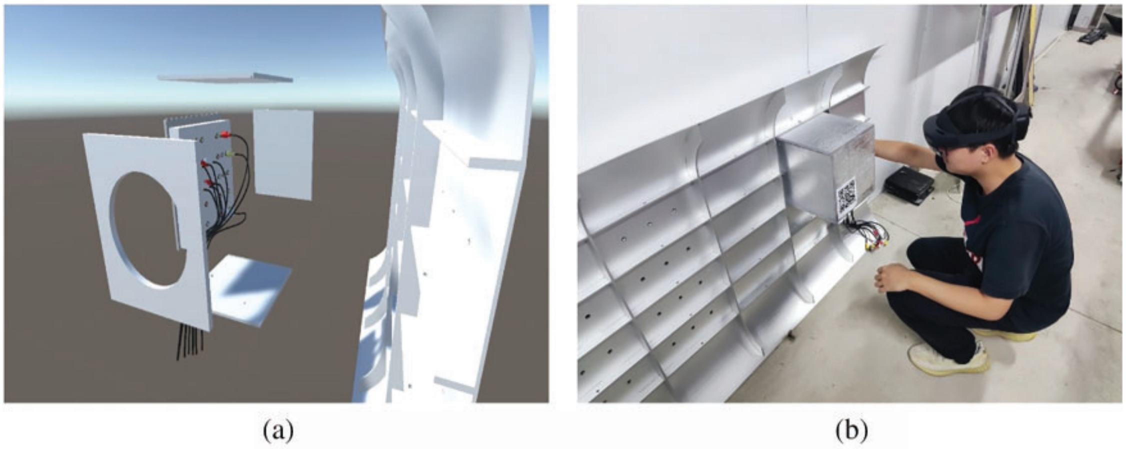
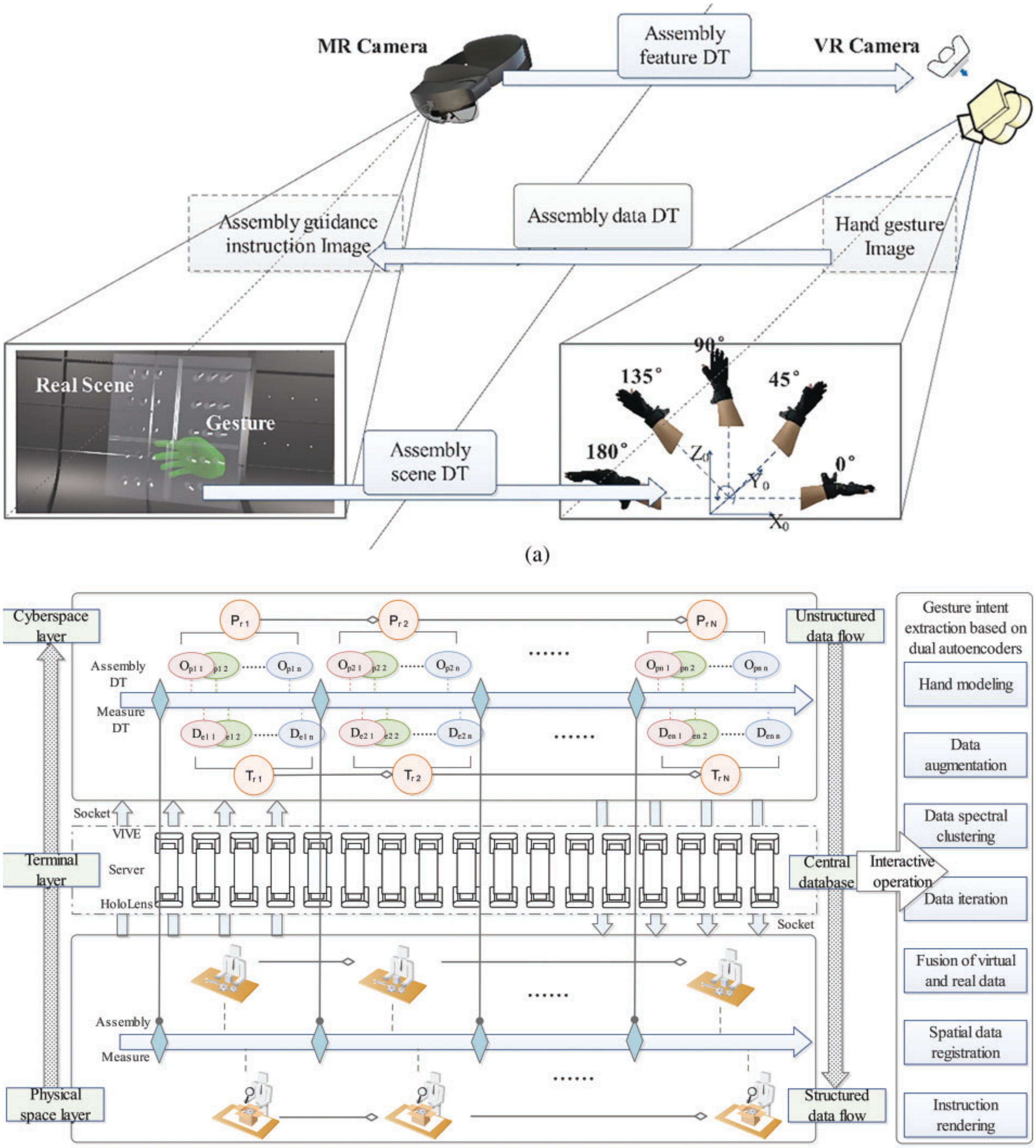


Figure 3: Human-computer collaboration in assembly scenarios

4 Experimental Methods and Results

The main problem encountered by a human-machine model without full occlusion gesture recognition capability (T model) during blind assembly is the inability to obtain feedback on the operator's hand position in blind spots, which we have already confirmed in our previous research on the mixed reality system for unobstructed gesture recognition. We attribute this to two main risks: firstly, in adjacent areas, there may be line-holes similar to the target object, making it easy to mistakenly insert wiring harnesses into incorrect positions, causing damage to corresponding electrical equipment. Secondly, when the target object is located in the inner edge area of the assembly, the difficulty of graphic rendering increases significantly. Even after repeated attempts, it is difficult or impossible to accurately determine whether the target position has been reached effectively. Especially in cases of visual obstruction, operators need to rely on trial and error to locate the assembly holes, compensating for the lack of spatial positioning information.

To verify the applicability of the proposed model in space-constrained assembly tasks, we invited 102 developers from a smart technology company in Shanghai, China, to participate in the experiment. The participants included 64 males and 38 females, aged between 21 and 48 years (M = 50.5 years, SD = 26.1). Their professional backgrounds varied, including industrial design, intelligent science and technology, electrical engineering and automation, and artificial intelligence. Prior to the experiment, we provided all participants with a detailed explanation of the study's academic ethics, personal privacy security, and related issues. We also trained all participants to ensure they understood the task objectives and operational procedures. We used a between-subjects experimental design, with 48 participants evenly divided into two groups. Each group was randomly assigned to test either T model or HCBM. The experiment was conducted in a controlled environment. To avoid learning effects, participants randomly selected 10 out of 68 visually identical cable harnesses, differing only in power connectors, for the assembly task. After the experiment, participants completed the IEMQ questionnaire (see Table 1).



(a)

(b)

Figure 4: Human-computer collaboration behavior measurement model

Table 1: Dimensions and measurement items of IEMQ

Feedback (Q1–Q5)	Immersion (Q6, Q7)	Challenge (Q8–Q12)	Interaction (Q13, Q14)	Control (Q15, Q16)
-The instructions I provide can receive real-time feedback from the current system.	-I find the task content very appealing.	-During task execution, effective hints will be provided to guide me in overcoming challenges.	-The task execution system is adaptable to different gesture poses (grasping, gripping, etc.).	-In the task, I can use strategies according to my own preferences.
-I receive timely feedback on the completion status of each operation during the task.	-My thinking is influenced by the changes in the task progress.	-After completing some subtasks, my skills and understanding of the task goals are enhanced.	-The task execution system is adaptable to different environments, including un-occluded, partially occluded, fully occluded.	-In case of task interruption, it is possible to quickly restore to the normal state.
-New tasks are promptly provided to me within the task.		-I am motivated when I see an improvement in my technical skills during the task.		
-I can observe my status information at any time within the task.		-The task interface is simple and easy to understand.		
-I can receive notifications in a timely manner when the game provides me with positive or negative feedback.		-The interactive operations are easy to get started with.		

We conducted a statistical analysis of the time taken for visual guidance during cable harness installation using two human-machine collaboration models (T model and HCBM). The results showed that T model had the longest average assembly time (102.41 s), while the HCBM had the shortest average assembly time (74.02 s). Compared to the T model, HCBM further reduced the average assembly time by 27.7%. Additionally, the results of a one-way analysis of variance (ANOVA) indicated that there was a statistically significant difference in overall average assembly time at the $\alpha = 0.05$ significance level (F(1, 46) = 12.421, $p = 0.022 < 0.050$). Additionally, based on Hikin's questionnaire development methodology, we designed a questionnaire to measure the interaction experience for the human-machine interface. This questionnaire is based on the eight dimensions of Gameflow model and has been modified to account for the specific features of the HCBM interface. With input from experts in the field of user experience, we identified the dimensions for evaluating the interactive experience of the interface's visual encoding, including feedback, sense of immersion, challenge, interactivity, and sense of control. Each question item was rated on a 5-point Likert scale

(1 representing "strongly disagree" and 5 representing "strongly agree"). Analysis using the Wilcoxon signed-rank test ($\alpha = 0.05$) indicated statistically significant differences ($p < 0.01$) between the paper-based checklist and a MR system across these five dimensions.

5 Discussion

Cognitive resource efficiency was evaluated from two aspects: effective target fixation and subjective mental workload. In the analysis of time-on-interest data, the use of HCBM significantly reduced fixation errors caused by difficulties in understanding task intent, thereby reducing the accumulation of short-term mental workload. This was confirmed through operator feedback in face-to-face interviews. From the perspective of subjective mental perception, the T model used only visual element layouts (such as wireframes or lines) to represent specific 3D spatial areas, leading to confusion among operators regarding the physical occlusion relationships of certain components. This may have been due to the lack of visual encoding standards for spatial areas, causing local operators to overlook certain details in the task intent, such as color-coded annotations in blind assembly operation trajectories that indicate the importance of distance. Additionally, operators struggled to easily identify the logical relationship between the operation trajectory and the surrounding wireframes, resulting in a rapid increase in cognitive load. In contrast, HCBM supported a more intuitive visual encoding format, where annotation instructions were encoded through the MR interface into assembly guidance instructions that were easier to distinguish. This reduced the cognitive load for local personnel in understanding the spatial range of invisible areas. By integrating visual element layouts with visual depth cues from special graphic hints, operators were encouraged to observe subtle changes in surrounding components related to the target component, prompting them to allocate attention to other visual areas to anticipate changes in task conditions. Although this initially increased their mental workload, the positive effects of scaling cues quickly mitigated this issue.

In the context of aircraft harness installation, HCBM outperformed the T model in terms of both time consumption and user experience, showing significant differences and advantages. Specifically, we found that HCBM effectively avoided multiple manual operations, reduced the number of errors, and minimized feedback discrepancies. Some participants mentioned that using the T model for searching and validation was time-consuming, leading to significant accumulation of physical and mental fatigue in subsequent tasks, reduced immersion, and a lack of suitable challenge. HCBM significantly improved the execution efficiency of space-constrained assembly tasks, providing necessary visual guidance at the right moments, lowering the barrier to mastering new interaction skills, and making the interaction process more convenient and user-friendly. The study results indicate that the implementation of the enhanced aircraft harness assembly model is effective. The digital twin-driven system integration solution facilitates the development of intelligent assembly workshops by enabling the convergence of cyber-physical systems, industrial big data computation, and human-machine interaction. Ensuring consistency in data formats within the manufacturing execution system effectively reduces the diversity of data carriers and the complexity of data visualization. This supports the flow of manufacturing data throughout the upstream and downstream processes of the workshop, making real-time tracking and traceability across the product lifecycle possible. It is worth noting that, considering different industrial application scenarios, the corresponding human-machine collaboration assembly methods should be adjusted based on the specific context. This requires more extensive data analysis to evaluate the overall efficiency of human-machine collaboration systems and the effectiveness of human-machine interaction. Although we demonstrated the broad applicability of MR in harness assembly under full-occlusion gesture conditions, MR may not be the optimal solution for non-space-constrained assembly tasks. Manufacturing workshops should choose the most suitable assembly assistance methods according to the characteristics of their product manufacturing processes to avoid unnecessary waste of human, material, and financial resources.

6 Conclusions and Future Works

This study proposes a HCBM that integrates deep vision technology to address issues in traditional gesture recognition and the perception of virtual and real scenes. It employs mixed reality technology to guide users in performing correct manual operations. To accurately identify and calibrate hand occlusion relationships, a specialized virtual and real space registration model for fully occluded gestures was developed. Additionally, an assembly data enhancement algorithm that integrates mixed reality and multi-channel interaction technology was proposed to standardize manual operations. Compared to a T model, HCBM significantly reduced working time and effectively improved the level of collaboration between humans and robots during the assembly process. In summary, fully occluded gesture recognition not only reduces workload but also enhances recognition accuracy and work efficiency, providing a new solution for human-robot motion collaboration under limited visibility conditions.

Acknowledgement: We sincerely thank University of Shanghai for Science and Technology for the support and resources that made this research possible. Special thanks to the Human-Machine Integration Manufacturing Laboratory for their invaluable contributions in providing technical assistance and expertise throughout the project. We also extend our gratitude to 102 participants from Weilingke Intelligent Technology (Shanghai) Co., Ltd., for their time and effort in assisting with the experiments. Finally, we appreciate the constructive feedback provided by the anonymous reviewers, which greatly improved the quality of this paper.

Funding Statement: This work was supported by National Natural Science Foundation of China (No. 52205534), Industry & University Collaborative Education Project by Ministry of Education of China (No. 230900487263953) and Shanghai Higher Education Young Teacher Training Support Program Project (No. 10-24-304-006), awarded to Zhuo Wang, whose work focuses on AR/MR manufacturing. The financial support provided by the foundation was essential to the successful completion of this study.

Author Contributions: Zhuo Wang conceived the research idea and designed the study. Hongyuan Zhan conducted the experiments and collected the data. Yang Wang performed data analysis and interpretation. Hongyuan Zhan and Minglei Zhu contributed to writing the original draft of the manuscript. Qingtian Lu provided critical revisions and approved the final version of the manuscript. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Due to the nature of this research, participants of this study did not agree for their data to be shared publicly, so supporting data is not available.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Guo J, Zhang J, Wu D, Gai Y, Chen K. An algorithm based on bidirectional searching and geometric constrained sampling for automatic manipulation planning in aircraft cable assembly. J Manuf Syst. 2020 Oct 1;57(7):158–68. doi:10.1016/j.jmsy.2020.08.015.
2. Wang B, Zhang Y, Su G. An integrated approach for electromagnetic compatible commercial aircraft engine cable harnessing. J Ind Inf Integr. 2022 May 1;27(1):100344. doi:10.1016/j.jii.2022.100344.
3. Wang Z, Bai X, Zhang S, Billinghurst M, He W, Wang P, et al. A comprehensive review of augmented reality-based instruction in manual assembly, training and repair. Robot Comput Integr Manuf. 2022 Dec 1;78:102407. doi:10.1016/j.rcim.2022.102407.
4. Yin X, Fan X, Zhu W, Liu R. Synchronous AR assembly assistance and monitoring system based on egocentric vision. Assem Autom. 2019;39(1):1–16. doi:10.1108/AA-03-2017-032.
5. Laviola E, Gattullo M, Evangelista A, Fiorentino M, Uva AE. *In-situ* or side-by-side? A user study on augmented reality maintenance instructions in blind areas. Comput Ind. 2023 Jan 1;144:103795. doi:10.1016/j.compind.2022.103795.
6. Zhang J, Wang S, He W, Li J, Wu S, Huang J, et al. Augmented reality material management system based on post-processing of aero-engine blade code recognition. J Manuf Syst. 2022 Oct 1;65:564–78. doi:10.1016/j.jmsy.2022.10.006.
7. Fang W, Hong J. Bare-hand gesture occlusion-aware interactive augmented reality assembly. J Manuf Syst. 2022 Oct 1;65(2):169–79. doi:10.1016/j.jmsy.2022.09.009.
8. Papandreou G, Zhu T, Kanazawa N, Toshev A, Tompson J, Bregler C, et al. Towards accurate multi-person pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017; p. 4903–4911.
9. Stiefmeier T, Roggen D, Ogris G, Lukowicz P, Tröster G. Wearable activity tracking in car manufacturing. IEEE Pervasive Comput. 2008;7(2):42–50. doi:10.1109/MPRV.2008.40.
10. Thewlis D, Bishop C, Daniell N, Paul G. Next-generation low-cost motion capture systems can provide comparable spatial accuracy to high-end systems. J Appl Biomech. 2013;29(1):112–7. doi:10.1123/jab.29.1.112.
11. Chu C-H, Ko C-H. An experimental study on augmented reality assisted manual assembly with occluded components. J Manuf Syst. 2021 Oct 1;61(3):685–95. doi:10.1016/j.jmsy.2021.04.003.
12. Yang C-K, Chen Y-H, Chuang T-J, Shankhwar K, Smith S. An augmented reality-based training system with a natural user interface for manual milling operations. Virtual Real. 2020 Sep 1;24(3):527–39. doi:10.1007/s10055-019-00415-8.
13. Elhayek A, de Aguiar E, Jain A, Thompson J, Pishchulin L, Andriluka M, et al. MARCOnI—ConvNet-based MARker-less motion capture in outdoor and indoor scenes. IEEE Trans Pattern Anal Mach Intell. 2016;39(3):501–14. doi:10.1109/TPAMI.2016.2557779.
14. Urgo M, Tarabini M, Tolio T. A human modelling and monitoring approach to support the execution of manufacturing operations. CIRP Annals. 2019;68(1):5–8. doi:10.1016/j.cirp.2019.04.052.
15. Hu L, Xu J. Learning discriminative representation for skeletal action recognition using LSTM networks. In: Computer Analysis of Images and Patterns: 17th International Conference, CAIP 2017, 2017 Aug 22–24; Ystad, Sweden: Springer; p. 94–104.
16. Naveenkumar M, Domnic S. Deep ensemble network using distance maps and body part features for skeleton based action recognition. Pattern Recognit. 2020;100:107125.
17. Al-Amin M, Qin R, Moniruzzaman M, Yin Z, Tao W, Leu MC. An individualized system of skeletal data-based CNN classifiers for action recognition in manufacturing assembly. J Intell Manuf. 2021;34(1):633–49. doi:10.1007/s10845-021-01815-x.
18. Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition. Proc AAAI Conf Artif Intell. 2018;32(1):7444–52. doi:10.1609/aaai.v32i1.12328.

19. Wang Z, Zhang S, Bai XL. Augmented reality based product invisible area assembly assistance. In: Proceedings of the 2018 3rd International Conference on Control, Automation and Artificial Intelligence (CAAI 2018), 2018 Aug; p. 109–113. doi:10.2991/caai-18.2018.26.

20. Feng S, He W, Zhang Q, Billinghurst M, Yang L, Zhang S, et al. ARCoA: using the AR-assisted cooperative assembly system to visualize key information about the occluded partner. Int J Hum Comput Interact. 2022;39(18):1–11. doi:10.1080/10447318.2022.2099237.

21. Bain GI, Polites N, Higgs BG, Heptinstall RJ, McGrath AM. The functional range of motion of the finger joints. J Hand Surg Eur. 2015;40(4):406–11. doi:10.1177/1753193414533754.