machines

MDPI

Article

Gest-SAR: A Gesture-Controlled Spatial AR System for Interactive Manual Assembly Guidance with Real-Time Operational Feedback

Naimul Hasan * and Bugra Alkan

Industrial Systems Intelligence Lab, School of Computer Science and Digital Technologies, London South Bank University, 103 Borough Road, London SE1 0AA, UK; alkanb@lsbu.ac.uk

* Correspondence: hasanm56@lsbu.ac.uk

Abstract

Manual assembly remains essential in modern manufacturing, yet the increasing complexity of customised production imposes significant cognitive burdens and error rates on workers. Existing Spatial Augmented Reality (SAR) systems often operate passively, lacking adaptive interaction, real-time feedback and a control system with gesture. In response, we present Gest-SAR, a SAR framework that integrates a custom MediaPipe-based gesture classification model to deliver adaptive light-guided pick-to-place assembly instructions and real-time error feedback within a closed-loop interaction instance. In a within-subject study, ten participants completed standardised Duplo-based assembly tasks using Gest-SAR, paper-based manuals, and tablet-based instructions; performance was evaluated via assembly cycle time, selection and placement error rates, cognitive workload assessed by NASA-TLX, and usability test by post-experimental questionnaires. Quantitative results demonstrate that Gest-SAR significantly reduces cycle times with an average of 3.95 min compared to Paper (Mean = 7.89 min, $p < 0.01$) and Tablet (Mean = 6.99 min, $p < 0.01$). It also achieved 7 times less average error rates while lowering perceived cognitive workload ($p < 0.05$ for mental demand) compared to conventional modalities. In total, 90% of the users agreed to prefer SAR over paper and tablet modalities. These outcomes indicate that natural hand-gesture interaction coupled with real-time visual feedback enhances both the efficiency and accuracy of manual assembly. By embedding AI-driven gesture recognition and AR projection into a human-centric assistance system, Gest-SAR advances the collaborative interplay between humans and machines, aligning with Industry 5.0 objectives of resilient, sustainable, and intelligent manufacturing.

Keywords: manual assembly; augmented reality; spatial augmented reality; hand gesture control; cognitive workload; operator assistance systems; Industry 5.0

check for updates

Academic Editor: Dan Zhang

Received: 20 June 2025
Revised: 23 July 2025
Accepted: 25 July 2025
Published: 27 July 2025

Citation: Hasan, N.; Alkan, B. Gest-SAR: A Gesture-Controlled Spatial AR System for Interactive Manual Assembly Guidance with Real-Time Operational Feedback. *Machines* **2025**, *13*, 658. https://doi.org/10.3390/machines13080658

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Automation and digitisation are at the core of the Industry 4.0 mission to enhance productivity and operational efficiency. However, despite advances in automation, manual labour remains indispensable in modern manufacturing due to the inherent flexibility, cognitive ability, and problem-solving skills of human workers [1]. With the rising demand for personalised products, the flexibility of mass customisation, and a growing user base, manufacturing companies are increasingly shifting from a traditional build-to-stock approach to a build-to-order system. As a result, the demand for manual labour is expected to increase rather than decline.

This variability in product necessitates the provision of novel assembly information for each configuration, which increases the cognitive burden on workers [2]. Manual assembly tasks, which require memory recall and sequential execution, are susceptible to human limitations such as short attention spans and memory constraints [3]. These limitations often lead to higher error rates and longer assembly cycles, which ultimately impact productivity and increase product costs, especially in regions with high labour expenses [4]. Specialised robotic systems are often not viable for such complex tasks due to their high costs and limited adaptability. Hence, augmenting human operators with assistive technologies can significantly improve accuracy, productivity, and ergonomics [1]. This direction aligns with the vision of Industry 5.0, which emphasises human-centred, resilient, and sustainable industrial practices. It promotes worker empowerment through up-skilling, re-skilling, and the adoption of human-centric technologies to foster balanced human–computer interaction (HCI) [5]. Recent developments in this area include the Lean 5.0 paradigm, which integrates digital transformation with traditional lean principles while preserving human autonomy, adaptability, and knowledge-based decision-making [6]. Complementing this, the Human-Centric Zero Defect Manufacturing (HC-ZDM) approach advocates for placing operators at the core of quality assurance and error prevention systems by leveraging their cognitive and contextual insights within intelligent manufacturing environments [7]. Eriksson et al. [8] further emphasise that social sustainability, skill development, and participatory leadership are essential for embedding digitalisation successfully in industrial settings.

Notably, manual operations remain approximately 9% higher than hybrid human–computer cooperation systems in terms of prevalence according to Patrick Bertram et al. [9]. In tasks such as assembly and order picking, workers follow sequential procedures, where each step depends on the successful completion of the previous one. Mistakes such as using incorrect parts or omitting steps can result in product defects, inefficiencies, and resource waste [10]. Repeated errors may also lead to negative behavioural patterns among workers, including stress, anxiety, and frustration [11].

To mitigate these issues, assembly instruction systems, ranging from paper-based guides to digital displays and in situ projected instructions, are increasingly employed to reduce cognitive workload and improve task accuracy. These systems assist workers by eliminating the need to memorise instructions and enabling real-time guidance. Incorporating real-time feedback and error detection mechanisms further enhances the efficacy of such systems.

Recent case studies have highlighted the growing industrial viability of AR technologies. For instance, Pereira et al. [12] implemented AR within a lean smart factory context to improve operator ergonomics, mitigate musculoskeletal disorder (MSD) risks, and enhance process efficiency—all while promoting human-centric design principles. Bock et al. [13] proposed a decision-support framework tailored for AR deployment in manual assembly settings, addressing practical constraints such as task complexity, cognitive workload, and worker skill diversity. In the fashion industry, Fani et al. [14] integrated AR with data mining to deliver interactive visualisations tailored to dynamic production data, emphasising the technology's cross-domain applicability. Werrlich et al. [15] further demonstrated that HMD-based AR training systems yields improved learning transfer, error reduction, and user satisfaction compared to traditional paper-based methods.

Traditional paper-based instructions are typically single-page documents kept near workstations and workspaces. However, in large-scale manufacturing environments with diverse product lines, managing and accessing the correct paper-based instructions becomes cumbersome and error-prone [16]. In contrast, interactive assembly systems offer a promising alternative, which this study also explores. Assistive technologies such as

smartphones, tablets, smart glasses, head-mounted displays (HMDs), and augmented reality (AR) tools, especially those that employ voice, light, and scan-based selection, have been studied in both industrial and research settings [17].

Although wearable technologies have shown improvements in task performance, they often introduce physical discomfort and fatigue during prolonged usage. In this context, SAR, also known as in situ projection systems, has emerged as an effective alternative. These systems utilise projectors, cameras, and sensors to display digital instructions directly onto the work surface using multimedia formats such as images, videos, and animations. They support hands-free, immersive, and context-aware instruction delivery, allowing workers to remain focused on their task space without diverting their attention to external screens [18].

Compared to HMDs and handheld devices, SAR systems improve spatial cognition and reduce the cognitive disconnect between the instruction medium and the physical workspace [19,20]. However, most SAR systems in the literature operate passively, offering limited interactivity and adaptive feedback [21,22]. This study presents Gest-SAR, an advanced gesture-controlled SAR framework specifically developed to support manual assembly operations. Gest-SAR facilitates intuitive hand-gesture interaction by projecting dynamic pick-to-place instructions directly onto the workspace and provides real-time visual feedback in response to both correct and erroneous operator actions. This closed-loop interaction paradigm incorporates gesture recognition, adaptive instruction sequencing, and real-time error detection to support assembly task completion and reduce operator errors, all combined within a single framework—an advancement not systematically unified in prior AR/SAR systems, particularly for gesture-controlled SAR. To rigorously evaluate the efficacy of Gest-SAR, a comparative study was conducted against paper- and tablet-based work instruction deliveries. Performance was assessed through a range of KPIs—such as assembly cycle time, selection and placement error rates, and task accuracy—as well as subjective measures obtained via the NASA-TLX workload index.

The primary contributions of this research are threefold:

(i) The development and implementation of a novel hand-gesture recognition mechanism tailored for SAR-based operator guidance systems, which enables gesture-driven control and real-time feedback for enhanced assembly quality;
(ii) The introduction of a human-centric SAR workspace incorporating projected virtual controls (e.g., virtual buttons), thereby improving task navigation, instructional clarity, and user engagement during complex assembly tasks;
(iii) The establishment of a systematic benchmarking through empirical field experiments involving human participants, enabling comparative analysis of different work instruction delivery modalities across multiple KPIs.

These contributions collectively advance the integration of AI-driven human–computer interaction techniques within SAR systems, and underscore the potential of gesture-controlled environments to enhance precision, reduce cognitive workload, and promote user autonomy in industrial assembly contexts.

While numerous AR-based solutions employ HMDs or tablets for overlaying digital instructions, they often lack real-time adaptability, blocking real-world view, natural input modalities, or seamless integration into shared workspaces. In contrast, Gest-SAR introduces a projection-based, gesture-controlled SAR framework that enables hands-free interaction, minimises operator distraction, and supports closed-loop real-time instruction delivery. Unlike traditional AR approaches that require wearable devices and predefined workflows, Gest-SAR dynamically responds to operator gestures and task progress, offering context-aware feedback projected directly onto physical workspace. This combination of features—gesture recognition, adaptive projection, and real-time error

detection—represents a novel integration not present in existing AR toolkits for manual assembly tasks.

This paper is organised as follows: Section 2 summarises the state-of-the-art SAR and AR applications in manual assembly, Gap in Gesture-based SAR; Section 3 presents the system design, assembly tasks and experimental design procedure; Section 4 evaluates the results of the experiments; Section 5 evaluates the study; and finally Section 6 concludes with the evaluation of the study with highlighting limitations and future works.

2. Literature Review

In this section, we provide an overview of the previous research investigating the SAR on assembly assistance systems. We considered the research articles which present the past experiments on SAR and particularly gesture-based applications applied on AR- and SAR-based manual assembly and human–robot collaboration for the last 10 years.

2.1. SAR in Manual Assembly

SAR has been extensively explored as a tool to support manual assembly tasks by reducing cognitive workload and improving accuracy and user satisfaction. Funk et al. [23] demonstrated that contour-based visual feedback significantly improves perceived mental workload and performance among cognitively impaired workers during a 12-step clamp assembly task. Their follow-up study [16] compared instruction delivery via in situ projection, HMD, tablet, and paper, finding that in situ projection reduces cognitive load and task completion times, especially benefiting novice users. Extending this, Funk et al. [21] conducted a long-term study in an industrial setting showing that untrained workers benefit significantly from in situ projected AR, while expert users experience performance hindrance, suggesting the need for adaptive systems. Ganesan et al. [24] proposed a mixed-reality SAR system to project real-time visual cues during human–robot collaboration, showing improvements in task efficiency, accuracy, and user trust. Similarly, Uva et al. [25] found that SAR systems are more effective than paper-based manuals for complex maintenance tasks, yielding significant reductions in error rate and task time. Hietanen et al. [26] explored projection-based safety systems to enhance situational awareness in human–robot collaborative assembly through the real-time visualisation of safety zones. Sand et al. [27] demonstrated that smARt.assembly, a projection-based AR system, reduces both task completion time and error rate. Kosch et al. [19] used EEG to show a reduced mental workload when participants used projection-based assistance over paper instructions. Bosch et al. [28] evaluated SAR systems across five industrial case studies, confirming their usefulness for training and guidance, although implementation barriers like system cost and complexity were noted. Swenja et al. [29] developed a pick-by-light system with paper-guided instructions, display instructions via overhead projection and tangible interface guidance, aiming to improve efficiency in manual assembly. Their results showed that the overhead projection significantly reduces the picking error, and improves the picking speed and overall task time compared to paper-based instructions.

2.2. Comparative Analysis of Visualisation Methods and Interaction Techniques

Different visualisation modalities and interaction mechanisms play critical roles in the effectiveness of AR/SAR systems. Lucchese et al. [20] presented a comparative experimental study evaluating the effects of different cognitive assistive technologies—projection-based AR, tablet-based instructions, pick-by-scan, and pick-by-light—on human performance and well-being in manual assembly and picking tasks. They performed Lego-based order-picking tasks to present the importance of selecting context-appropriate assistive technologies to improve task efficiency and reduce cognitive strain in industrial settings.

Results from their study showed that SAR was perceived as the most intuitive and supportive for assembly, while pick-by-light led to the best efficiency and lowest error rate in picking tasks. Contour-based and spatially congruent visualisations have shown significant advantages in task performance, particularly for novice users [16,23]. Systems such as smARt.assembly [27] and multi-case SAR evaluations by Bosch et al. [28] reinforce the importance of clear, intuitive visual instructions. Projection-based visualisations outperform traditional instruction methods by minimising the need for visual search and reducing instruction-following errors. However, expert users may find such systems intrusive or redundant without adaptive feedback mechanisms [21]. This highlights a demand for context-sensitive and personalised interfaces. Interaction modalities also vary in effectiveness. Zhang et al. [30] found finger-click interactions to outperform gestures in reliability and speed, while multi-modal interfaces combining gesture, speech, and touch input improve overall performance and user satisfaction. Similarly, systems using gesture recognition for context-sensitive control [22,31] support intuitive and touchless interaction in dynamic industrial settings.

2.3. Gesture Control in Assembly Systems

Gesture is a non-verbal communication method that originates from the physical motion commonly expressed through the face, the centre of the palm, position and shape formed by the fingers of the hand. Communication with deaf–mute people, HCI, home automation, and robot control are different areas where gestures have been applied and used [32]. The gesture recognition process using image processing usually consists of gestures shown through hand shapes and movement to interpret either static or dynamic gestures. This information is captured using an RGB camera [33] and a depth camera [33,34], and then processed with computer vision techniques according to the needs of the experiments. In a gesture control system, the motion of the worker is identified and tracked continuously with different methods, for example, a vision-based approach using RGB cameras, colour-based recognition, and skeleton-based, appearance-based, and deep learning-based approaches [32,35].

2.3.1. Gesture with Augmented Reality

Torres et al. [36] integrated gesture recognition into AR glasses for industrial control, using MediaPipe to recognise both static and dynamic gestures. Their system provided intuitive control over equipment with real-time smoothing and normalisation to enhance recognition accuracy. Wang et al. [37] introduced BeHere, a remote VR/SAR collaboration system that shared gestures and avatars between remote experts and local workers, significantly improving task completion and communication quality. Fang et al. [38] proposed a gesture occlusion-aware AR system to improve realism by predicting occlusion masks from the RGB input, resulting in a lower cognitive load and improved performance for novice users. Dong et al. [39] developed a dynamic gesture-based AR assembly training system capable of recognising and predicting discrete hand actions, enhancing responsiveness and natural interaction during industrial training.

2.3.2. Gesture with Spatial Augmented Reality

SAR systems have increasingly integrated gesture recognition to facilitate natural and touchless interaction. Vogiatzidakis and Koutsabasis [40] evaluated mid-air gesture control for home devices in a SAR environment, reporting high usability despite occlusion and lighting sensitivity. Cao et al. [22] combined SAR and gesture recognition via Leap Motion to control digital interfaces, achieving accurate recognition under good lighting conditions but noting trade-offs with standard RGB cameras. Rupprecht et al. [18,31] demonstrated dynamic SAR systems with gesture-based control using YOLOv3 and synthetic data. These

systems project instructions adaptively based on user behaviour, significantly improving efficiency and reducing cognitive workload. Nazarova et al. [35] presented CobotAR, which combines omnidirectional projection and DNN-based gesture recognition for human–robot collaboration. Users experienced improved situational awareness and lower mental load compared to conventional methods.

2.4. Error Tracking and Real-Time Feedback

Effective error tracking and real-time feedback mechanisms are critical for reliable AR/SAR assembly systems. Kosch et al. [34] performed an empirical study on the use of visual, auditory and tactile feedback systems to notify users of the assembly error using the notification channels—projected red light, directional sound and a wireless vibrotactile glove, respectively—at a projection-based assistance system. Kinectv2 depth camera helps with flagging the picking mistakes automatically, whereas a researcher presses a remote trigger for the mis-assemblies performed by the participants. This study with the cognitively disabled worker resulted in a shorter processing time and faster error handling when using the visual error information compared to other modalities. Chen et al. [33] introduced a dynamic projection system that tracked users' viewpoints for consistent assembly guidance, improving accuracy in part recognition. However, real-world robustness and processing speed remain challenges. Building on prior work, Rupprecht et al. [18] enhanced their SAR system with real-time animated feedback and gesture recognition. They reported a 30.6% reduction in task time and decreased cognitive workload. Future directions include automatic error detection and real-time part tracking. Zhang et al. [30] proposed a PAR system supporting multimodal interaction, which allowed flexible assembly support through gesture, speech, and touch. Multimodal interaction offered improved robustness, especially when gesture input alone was unreliable.

2.5. Research Gap

Despite the growing body of research exploring AR and SAR for assembly assistance, significant limitations persist in current systems that hinder their full potential in dynamic industrial environments. A critical examination of the literature, as summarised in Table 1, reveals that most SAR systems function as passive visual aids, offering limited or no adaptive interaction capabilities. Although systems like smARt.assembly and others have shown notable improvements in reducing task time and errors through projection-based guidance, they often lack real-time adaptability, context sensitivity, and interactive user feedback mechanisms that are essential for modern human-centric manufacturing scenarios. It is evident that real-time assembly error tracking and progress monitoring have not been explored enough, as the majority of SAR systems rely on the manual recording of the experiments to analyse the performance.

Furthermore, although gesture recognition technologies have been investigated in isolated AR or SAR applications, their integration into operational SAR frameworks for manual assembly, particularly in order-picking tasks, remains nascent and underutilised. The few existing gesture-controlled SAR systems tend to suffer from constraints such as limited gesture vocabulary, lack of real-time feedback, susceptibility to environmental variations (e.g., lighting and occlusion), and minimal error correction capabilities. In addition, previous studies often emphasise technological feasibility over holistic system validation in real-world assembly contexts, leaving a critical gap in empirical evidence on cognitive benefits, operational efficiency, and user ergonomics.

Importantly, comparative evaluations across assistive modalities (e.g., SAR versus tablet versus paper) demonstrate that SAR can outperform traditional methods in metrics like error rate and cognitive load, yet they rarely incorporate gesture-based control or closed-

loop feedback. The intersection of SAR, gesture-based interaction, and real-time error tracking, although individually explored, has not been systematically unified and validated in a single comprehensive framework. This fragmentation highlights a pressing need for integrated solutions that not only guide but also respond to user actions, adapt instructions accordingly, and support intuitive, touchless control within the assembly workflow.

Table 1. Overview of studies using AR in industrial settings.

		Assistive Technology				Application		Performance			Cognitive Workload	
Authors	**Year**	**SAR**	**Gest**	**AET**	**PM**	**AP**	**OP**	**TCT**	**AE**	**PE**	**GS**	**NASA-TLX**
Kosch et al. [34]	2016	*			*	*	*	*	*	*	*	
Funk et al. [21]	2017	*		*	*	*		*	*		*	*
Vogiatzidakis & Koutsabasis [40]	2020	*	*					*			*	
Bosch et al. [28]	2020	*					*	*	*	*	*	*
Chen et al. [33]	2021	*	*	*	*	*						
Cao et al. [22]	2021	*	*			*		*			*	
Rupprecht et al. [18]	2022	*	*			*		*		*	*	*
Zhang et al. [30]	2022	*	*			*		*			*	
Swenja et al. [29]	2022	*					*	*		*	*	
Wang et al. [37]	2023	*	*			*		*			*	*
Lucchese et al. [20]	2024	*				*	*	*	*	*	*	*
Presented Work	**2025**	*	*	*	*	*	*	*	*	*	*	*

Note: An asterisk (*) indicates the feature was included or the metric was measured in the study. SAR: Spatial Augmented Reality, Gest: Gesture, AET: Assembly Error Tracking, PM: Progress Monitoring, AP: Assembly Process, OP: Order Picking, TCT: Task Completion Time, AE: Assembly Error, PE: Picking Error, GS: General Survey, NASA-TLX: NASA Task Load Index.

In light of these gaps, the presented study introduces Gest-SAR, a novel gesture-driven SAR framework that uniquely combines spatial projection, mid-air gesture recognition, real-time visual feedback, and closed-loop interaction. This system directly addresses the deficiencies identified in the literature by enabling natural interaction and instant error resolution in each procedural step. By benchmarking Gest-SAR against paper- and tablet-based manuals across objective metrics (e.g., assembly time, error rate, and picking accuracy) and subjective metrics (e.g., Likert Survey and NASA-TLX) metrics, this research advances the empirical understanding of gesture-controlled SAR in manual assembly. Ultimately, this work contributes a significant step toward achieving the vision of Industry 5.0, where human-centric, resilient, and intelligent systems empower workers and enhance operational performance.

3. Materials and Methods

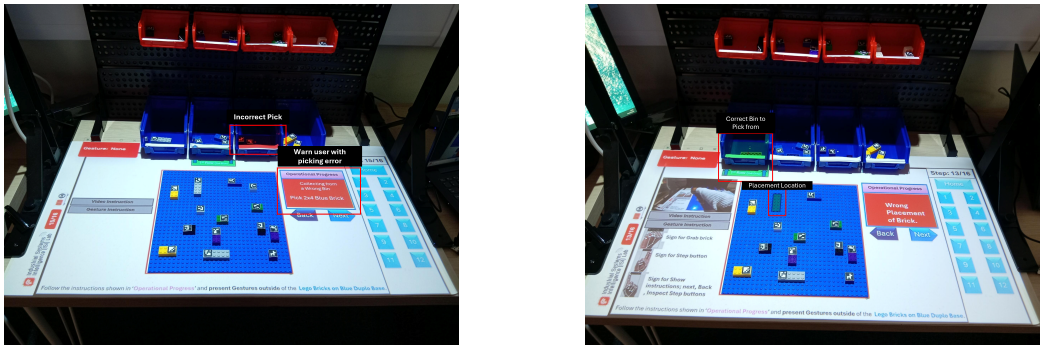
3.1. Proposed In Situ SAR System

The SAR system is designed to improve assembly guidance and the efficient closed-loop monitoring of assembly progress, as well as to enhance the efficiency of product completion by correcting users' mistakes with real-time instruction within the system. In situ instruction follows two key principal strategies to guide participants throughout each step of the assembly procedure. First, in addition to the projected textual information and animated visualisation of the assembly process, a green blinking light is projected on the container to draw the participant's attention to the specific item required for the respective step. This minimises the search time in locating the correct bin and the retrieval of the component.

Runtime System Design Overview

However, even if a participant retrieves a component from an incorrect container, the designed runtime system warns the participant with a projected red light within the

container and a textual instruction on the workspace which assists them in correcting the mistake (Figure 1a, Incorrect Pick). Second, after a correct component pick-up, the system highlights a green light to the correct location of the component to be placed on the Duplo base (Figure 1b, Correct Pick). Upon placing the component, the click gesture triggers the LEGO brick marker detection model, which identifies whether the expected component is placed correctly and notifies the users with the outcome, with a green light for correct identification and a red light for incorrect identification. These two stages are placed in a closed-loop monitoring integrated with the background runtime system, with a repetition for each step of the instruction until the assembly is completed. A video demonstration of the pick-and-place component was integrated within the SAR system, which additionally assists users in correcting the step with the expected picking and placement of the component for each step. It should be noted that the participant has the option to hide or show the video instruction with a click gesture pressing the designated virtual button. Similarly, gestures that control the component selection, placement, moving back, next, and specific assembly steps of the instructions are shown to ensure that users are aware of the correct gestures (discussed in Section 3.5).



(**a**) Incorrect Pick (**b**) Correct Pick

Figure 1. Incorrect and correct component picking by red and green lights.

The runtime system is the backbone of the proposed SAR system and has been implemented with Python programming with version 3.9 and computer vision techniques. Note that the runtime system is scalable and can be extended to future extensions of software or platform design spontaneously. We used the RGB frames captured using the Intel RealSense depth camera, which were then integrated with the run-time system to perform real-time hand tracking and gesture detection during the assembly tasks. Gesture-based interaction is a component of 3D interaction, which offers greater degrees of freedom compared to 2D interaction. Therefore, we decided to use MediaPipe [41], a framework developed by Google which can serve as a foundational component for constructing efficient on-device machine learning pipelines and provides a robust and efficient foundation for real-time hand landmark detection using only RGB video input. Compared to frameworks such as OpenPose and Leap Motion Controller, MediaPipe offers an ideal balance of speed, accuracy, and hardware independence. While OpenPose is known for high-fidelity multi-person pose estimation, it is computationally intensive and better suited for offline processing or systems with dedicated GPUs [42]. In contrast, Leap Motion offers sub-millimetre accuracy in static hand tracking scenarios [43] but requires specific lighting and environmental conditions, and is restricted in terms of working volume and gesture classification flexibility. Moreover, Leap Motion is based on IR sensors with a top–down field of view, which complicated its integration into in situ SAR environments. MediaPipe emerged as the most appropriate choice given the requirements of our in situ projection system—low-latency response, markerless single-user hand tracking, freedom from IR interference, and the

need for model customisation. Furthermore, MediaPipe's cross-platform compatibility and proven robustness in various HCI contexts made it a suitable choice for our goal of developing a lightweight and real-time closed-loop SAR system. Our initial interest was to use the Gesture Recogniser model that comes with MediaPipe. However, its built-in Gesture Recogniser model supports only seven predefined gestures. To address this, we implemented a custom gesture classification pipeline using the 21 landmark coordinates provided by MediaPipe, enabling the system to support a wider vocabulary of gestures to support domain-specific gestures tailored to the task requirements in our assembly use case.

Using the Cartesian coordinate framework, specific Regions of Interest (ROIs) are introduced into the system, where each ROI is associated with distinct control actions such as advancing to the next step, returning to the previous step, showing or hiding video and gesture instruction, navigating to the home screen, or accessing individual assembly stages. Leveraging the MediaPipe framework with the customised gesture classification model, the system continuously tracks both hands but concentrates on detecting the spatial coordinates of the index fingertip of one single hand at any given time.

When this fingertip enters a designated ROI with specific gestures shown in Table 2 detected for 1 s, provided that a minimum cooldown interval of 1.5 s has elapsed, the system initiates the corresponding action. The minimum cooldown interval is added to prevent inadvertent activations rapidly. Upon satisfying these conditions, the appropriate command is issued to proceed with the assembly process, executed through the PyAutoGUI library [44], which emulates keyboard input for system interaction.

Table 2. Gesture classes, labels, and associated actions.

ID	Gesture Class	Actions
0	Neutral	No action
1	Grab	Component pick
2	Right	Next step
3	Left	Previous step
4	DoubleClick	All other buttons including the Home button
5	Click	Click gesture for Back, Next buttons, Video and Gesture Information buttons
6	Pick	Component pick

The SAR runtime system employs a closed-loop interaction model shown in Figure 2, wherein user actions are continuously monitored, verified, and responded to in real time to ensure the correctness and completion of tasks. This control strategy ensures that each step of the assembly process, component identification, retrieval, and placement, is validated before the system allows progression to the next instruction. Specifically, the closed-loop involves a cyclical process where user gestures (e.g., grab and click) are continuously tracked using a custom MediaPipe-based gesture classification model to control key actions. For instance, a grab gesture confirms a picking action, which is then cross-validated against the expected predefined bin location within the workspace. Once a brick is picked from the correct component bin, users are informed via a sound notification and a textual update in the operational progress segment in the workspace. Upon successful selection of the component, the marker-based brick recognition triggers automatically to verify the correct placement. In the case of an incorrect pick or misplacement error, the system delivers real-time feedback with a projected red light, textual warnings and prevents progression until the error is corrected. Only upon successful completion of the current step does the system enable advancement, maintaining task integrity throughout. Following a correct placement, the system informs the user about the completion of the step and allows the user to click the next button with the click gesture to move to the next step and repeat the process. This closed-loop design improves accuracy, provides just-in-time correction for

each step, and supports self-guided learning within the assembly environment, making the system particularly effective for novice users or training contexts. The click gesture further allows one to control the video information and the gesture instruction buttons to hide or display the information during any stage of the assembly.

Both the back and the next buttons have a four-second interval between two consecutive button clicks. The step buttons placed to the far right of the workspace are controlled with the double click gesture and allow users to jump to a specific step of the assembly to verify or make additional changes when needed. A detailed view of the completed in situ projected workspace is shown in Figure 3.

The runtime system can detect from which box the worker is picking the component based on the position of the pick or grab gesture in the predefined box number for each of the steps. Continuous monitoring of the position of the hand is tracked, and Figure 4 shows each of the boxes marked on the blue rectangle with the respective box number that defines the region of interest of the selection zone of components. Each box contains a distinct type and colour of components. In addition, the Duplo base is where the placement zone for the assembly process is firmly mounted in the workspace for each assembly instruction type. All the interaction zones of the virtual button are on the right, with two buttons to the left of the placement zone. Note that the runtime system shows green ROI when the grab gesture is detected to pick up the correct component in the predefined bin and changes to red when the grab gesture is detected at an incorrect picking location as shown in Figure 4.

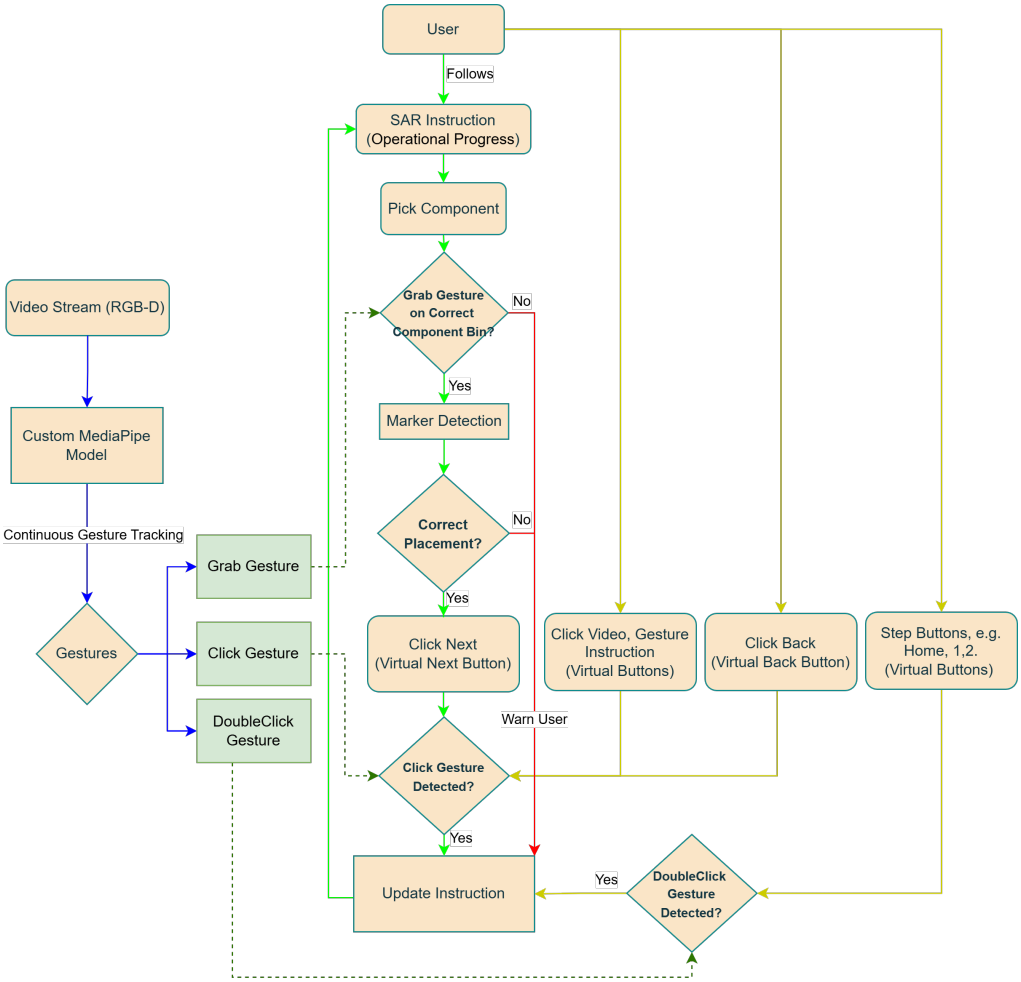


Figure 2. Schematic architecture of the proposed SAR runtime assembly assistance system.

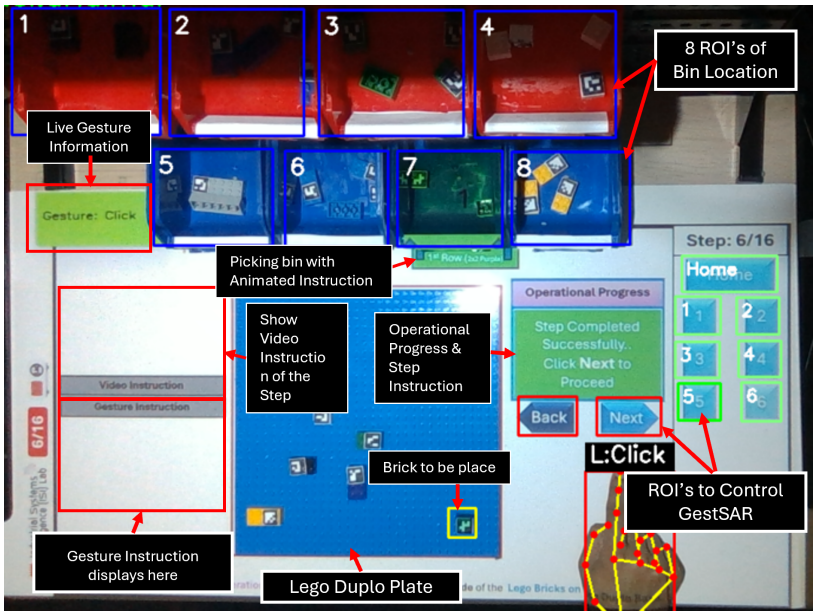


Figure 3. Detailed view of the SAR workspace components.



(**a**) Correct Grab (**b**) Incorrect Grab

Figure 4. Runtime GUI interface tracking correct and incorrect grab gestures for step 6 of the assembly.

3.2. Experimental Setup

We conducted a comparative analysis of three related research instruction systems to deliver instructional content in assemblies. All assembly and order-picking tasks were performed in the same workspace as shown in Figure 5 for the three visualisation methods. The assembly platform is designed using modular steel profiles and workstations, accompanied by a component bin storage unit containing eight bins arranged in two rows of four bins, all positioned within an ergonomically accessible range for the participants. Participants are seated in a comfortable chair facing the assembly workspace. The setup incorporates a low-blue-light projector (ELEPHAS 3800 Lumens, London, UK) and an Intel d435 RealSense 3D depth camera [45]. Although we focus on the RGB frame in this experiment, this device can record both RGB video with 30 frames per second (FPS) and depth data with 90 FPS, enabling comprehensive monitoring without interruption in data collection. The projector was mounted vertically above the workspace to facilitate precise alignment and projection of the assembly instructions directly onto the work surface, with the depth camera mounted vertically with an additional steel frame along the projector within close proximity (81 cm) of the workspace surface area to reliably track hand movements, trigger actions, and monitor the workspace, bins, picks, and placement areas. A second RGB camera module (Logitech Webcam) was included within the workspace to capture live footage with timestamps to enable the accurate capture of task durations for

each assembly step in each of the instruction techniques. This data is also used to perform the quantitative performance analysis for each participant.

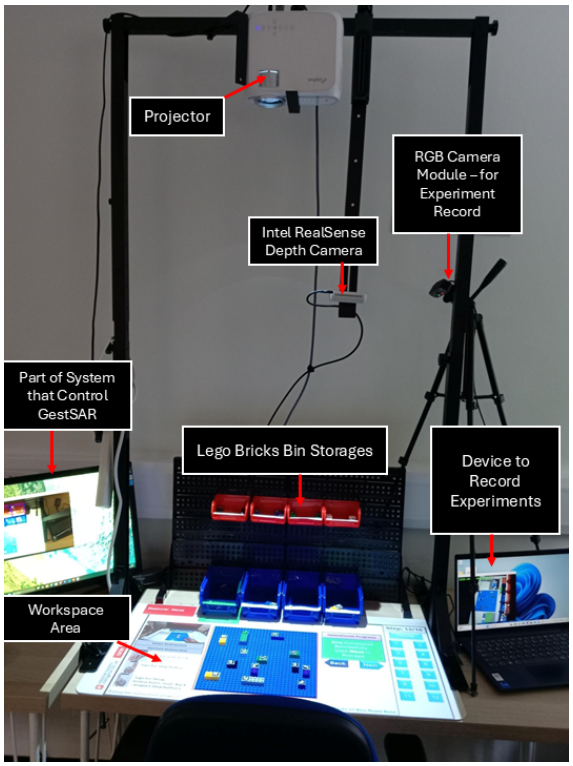
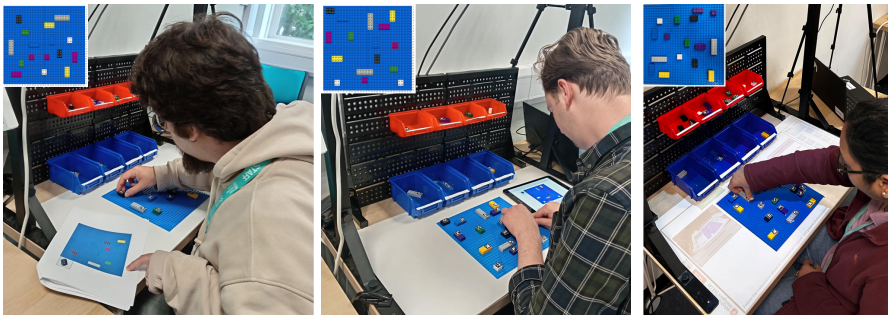


Figure 5. High-level overview of the workspace setup.

3.3. Paper-Based Instructions

For the paper-based instructional method (see Figure 6a, paper assembly), each step of the construction process was printed individually on an A4 sheet, maintaining a consistent layout throughout. Instructions were presented uniformly on a single side of the paper, ensuring that the size and positioning of the instructional content, including the Duplo base image, remained constant across all steps. Each instructional step featured a 2D representation of the brick to be selected, with the number of bricks to be picked displayed in the upper left corner relative to the LEGO Duplo construction. The target placement area for the brick was clearly indicated by a red border, which facilitates quick visual identification and minimises the cognitive demand associated with recalling prior construction stages. Additionally, to enhance spatial orientation and precision in brick placement, numerical labels ranging from 1 to 32 were included along both the x-axis and the y-axis of the Duplo base. This grid system served to support the more accurate and efficient execution of the construction tasks.



(**a**) Paper Assembly (**b**) Tablet Assembly (**c**) SAR Assembly

Figure 6. Assembly instruction modalities.

3.4. Tablet-Based Instruction

As a digital substitute for traditional paper-based assembly instructions, we utilised a tablet device to display the instructional content (see in Figure 6b, Tablet Assembly). Specifically, a Samsung Galaxy S2 tablet (Samsung, London, UK) was used to present the image-based instructions featured in the printed instruction manual, ensuring consistency across different instruction modalities. The design of our tablet-based interface was largely inspired by the approach proposed by Funk et al. [16], with some key modifications. In our implementation, participants were not required to continuously hold the tablet during the assembly task, thus reducing potential ergonomic strain and allowing for more flexible interaction. Furthermore, unlike in the study by Funk et al., the participants in our setup did not use a wireless presenter to navigate the instructions; instead, the instructional flow was controlled by the participants. Numerical labelling of the Duplo studs was also added to the tablet-based instructions.

3.5. Gesture Classification Model

The Gesture Recogniser model operates using a composite model bundle comprising two pre-integrated components, a hand landmark detection model and a gesture classification model. The hand landmark model is responsible for identifying the presence of hands and capturing their geometric structure in 3D space, while the gesture classification model interprets these geometric features to recognise specific hand gestures [46,47]. The hand landmark detection model in MediaPipe can localise the 21 key points with a 3D hand-knuckle coordinate regardless of the handedness. In Figure 7, we can see that coordinates [4, 8, 12, 16, 20] are the finger tips with the rest of the knuckle points of the hand.

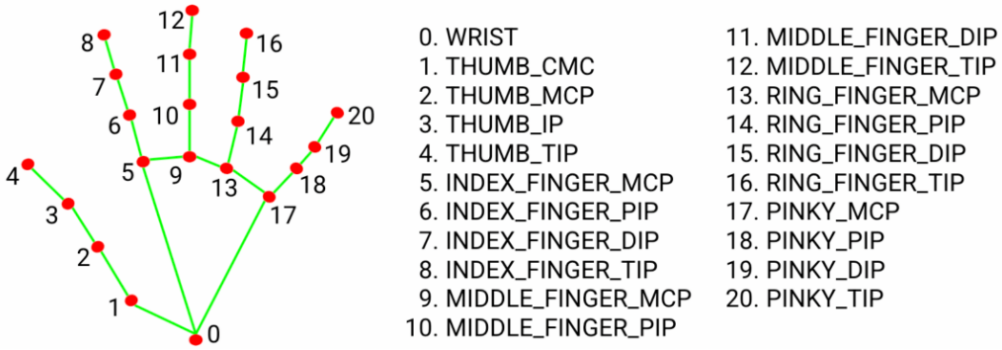


Figure 7. The 21-point hand landmarks.

We generated a dataset of the coordinate values of these 21 landmarks against respective gesture classes with different variations of lighting and handedness. We considered the 2D (width and height) information of each of the coordinates and normalised the landmark data using Translation to move all landmarks so that the wrist (landmark 0) is at the origin and scaled by measuring the distance between the wrist and the middle fingertip. This is because the raw coordinates vary on the basis of the hand size and position. Following normalisation, we flattened these normalised landmarks into a single vector to use in the classifier. We trained a simple Multi-Layer Perceptron (MLP) model to classify 7 different classes, where 0 is Neutral, 1 is Grab, 2 is Right, 3 is Left, 4 is DoubleClick, 5 is Click, and 6 is Pick. The overall architecture design of the modified gesture classification model can be seen in Figure 8.

This lightweight classification model is suitable for real-time gesture recognition with minimal inference latency and fast execution of the model. The input layer in this feed-forward neural network contains 42 elements of the x and y values for each landmark point. The two dense layers contain 20 and 10 neurons, respectively, with a dropout layer

for regularisation before passing to a softmax output layer for the multiclass classification. It achieves an overall accuracy of 99% and consistently high precision, recall, and F1-scores across all classes. The weighted average of the F1-score is 0.99, indicating a strong, balanced performance. Figure 9 shows an overview of the gesture recognition structure from hand detection to landmark tracking, gesture identification and the respective gesture actions associated with each gesture.

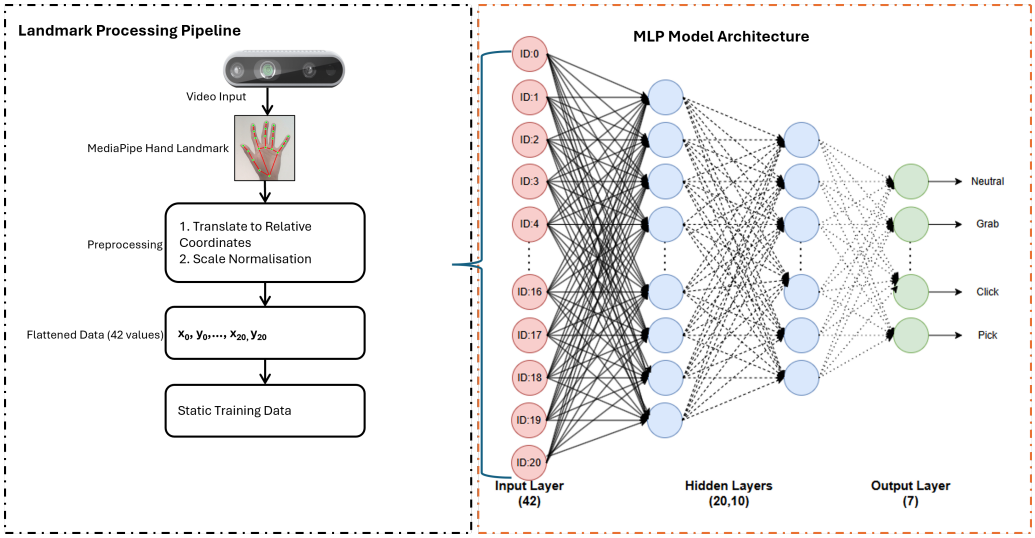


Figure 8. Architecture for customised gesture classification model.

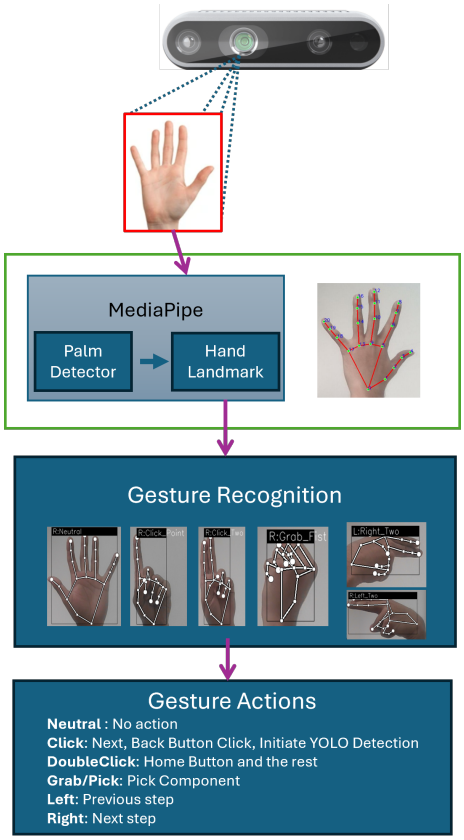


Figure 9. Gesture recognition structure.

3.5.1. Assembly Tasks

The assembly and order-picking tasks were performed in the same workspace for all three visualisation methods. Studies in [34,48–51] have shown that Lego Duplo tasks

can work as an abstraction for industrial pick-and-place tasks due to their low-cost setup, reproducibility, representability, and scalability [52]. Therefore, we used the Lego Duplo task, where each of the assembly types consists of 16 steps. The position of the 32 × 32 blue Lego Duplo base plate was kept within the fixed location across the experiments. The 8 boxes were arranged in a 2 × 4 grid contain Lego Duplo bricks (size 2 × 2: white, purple; size 2 × 3 green, black; size 2 × 4 purple, blue, yellow; size 2 × 6 grey). Each experiment was carried out with a different assembly type to prevent a learning curve across the assembly instructions. Further to prevent a systemic learning effect, we counterbalanced the order of the instruction systems across the participants according to the Balanced Latin Square.

3.5.2. Experimental Design Procedure

To begin the study, we explained the study, followed by signing the consent form and collecting demographic information from the participant. We ensured that the participants clearly understood the experimental procedures and instructions by providing an introduction to each type of instruction different from the one used in the study before starting the experiments, which ensures greater consistency. In particular, they were trained with gesture classes before the experiments, as gesture identification requires additional memory for participants to recall relevant actions to each gesture type. When participants were able to conduct gestures effortlessly, meeting the need of performing the task, they were allowed to move to the main assembly tasks. Notably, no time limit was imposed to complete each task during the experiments. When participants felt familiar with the instructions and assembly workspace, they were instructed to start the recording by pressing the start button on the Tkinter interface, which records the assembly to determine the exact times for t_{lb}, t_{p}, t_{lp}, t_{a} in Equation (1):

$$T_{\mathrm{cycletime}} = T_{lb} + T_p + T_{lp} + T_a \tag{1}$$

In all types of experiments, the measurement was initiated as soon as the first step of the instruction was shown. The error was counted from the first starting point. Each pick and placement location of the brick was randomly positioned in all the assemblies. Participants were told to follow the instructions exactly as shown in the instructions from the very first step. During the assembly, one researcher sat 2 m behind the participant to avoid interactions during the experiments. The total time taken to complete each assembly task was analysed using the recorded data from the video footage collected by the device to record experiments (see in Figure 5). This task completion time (TCT) method was followed according to the General Assembly Task Model (GATM) [52], where t_{lb} is the time to locate the correct component picking bin's position and placing the hand in the expected bin, t_{p} is the time to pick the component and take the hand out of the component bin, including the gesture detection time by the runtime, t_{lp} is the time the participant takes to understand the instruction and locate the placing position of the collected component, and t_{a} is the time taken to perform the assembly including any modification of errors made within the step. To mitigate potential learning and order effects inherent in within-subject designs, a Balanced Latin Square was employed to counterbalance the sequence of instruction modalities. This ensured that each instruction type appeared equally often in each position (first, second, and third) across participants, thus distributing practice effects evenly and enhancing the internal validity of the results. An error was defined as any deviation from the correct assembly sequence or final product configuration. PE was considered when the participant attempted or picked the wrong component. AE is identified when a brick is assembled in the wrong position on the Duplo plate. Following each task, participants completed the NASA Task Load Index (NASA-TLX) [53], a validated tool for measuring

perceived workload. The NASA-TLX evaluates six dimensions: mental demand, physical demand, temporal demand, effort, performance, and frustration. This metric provides insight into the mental and emotional effort required by participants for each instruction type. To assess overall usability, participants responded to a post-task questionnaire comprising Likert-scale items. These items captured subjective assessments of ease of use, clarity, satisfaction, and preferences related to the instruction system. Open-ended questions were also included to gather qualitative feedback and suggestions for improvement.

3.5.3. Participants

In this experimental study, 10 participants were selected with an average age of participants of 30.8. All participants were students, researchers, lecturers, and PhD students within the university. Participants with prior experience with projector-based assembly had an average of 3.0 (SD = 1.33), which was the highest, while familiarity with tablet-based assembly tasks was the lowest, with an average of 2.10 with SD = 1.10. Only one participant agreed to having prior experience with the Gesture-based Lego Duplo task.

4. Results

We recorded the task completion times $T_{\text{cycletime}}$, that are divided into t_{lb}, t_{p}, t_{lp}, and t_{a} in Equation (1), to present the contribution of the proposed Gest-SAR system not just for overall TCT but specific phases of the assembly process. Moreover, the pick and place errors with the total number of errors, cognitive workload measurement using the NASA-TLX questionnaire for each experiment and finally, different general questionnaires were recorded to present the usability and feature contributions of the SAR system. All the objective measures were statistically compared using the non-parametric Friedman test for repeated measures, followed by Wilcoxon signed-rank tests for post hoc pairwise comparisons in cases where Friedman indicated a significant difference between the visualisation methods. This approach was followed due to the violations of the normality assumption in the data. We used the Shapiro–Wilk test to assess normality, and most of the variables were found to significantly deviate from a normal distribution ($p < 0.05$); therefore, we used non-parametric methods.

4.1. Task Completion Times

We were interested in finding the effectiveness of the SAR system on the order-picking task with minimal to no errors and the assembly task; therefore, it was important to divide the task completion time with the phases ($t_{\text{locate_bin}}$, t_{pick}, $t_{\text{locate_place}}$, t_{assemble}) [52]. According to Figure 10, the in situ projector was the fastest to locate the correct position of the bin (Mean = 2.94 s, SD = 1.13 s) compared to the paper and tablet methods, where the paper-based method took the longest time with an average of 3.59 s (SD = 1.47 s). The Friedman test did not reveal any significant differences among the three methods for the time it takes to locate a correct bin location. The average time for brick selection in the in situ (Mean = 4.26 s, SD = 1.28 s) was dramatically higher than for the paper (Mean = 0.78 s, SD = 0.13 s) and tablet (Mean = 0.66 s, SD = 0.11s). This is due to the additional one second added for the pick/grab gesture duration while picking the brick and the time taken for a correct gesture detection by the classification model. The Friedman test shows a very highly significant difference ($p < 0.001$) for the time of component pick up. The Wilcoxon test further shows a very significant difference ($p < 0.01$) between (Gest-SAR, paper) and (Gest-SAR, tablet), with no significant difference for (paper, tablet) which is shown with the black lines and asterisks at the top for the t_{pick} graph in Figure 10. The position to place the brick was proved to be the most important phase among the four phases of TCT, which makes the SAR method significant, as the average time for the

in situ (Mean = 2.04 s, SD = 1.38 s) is massively lower than the counter methods, paper (Mean = 15.36 s, SD = 5.46 s) and tablet (Mean = 15.53 s, SD = 5.57 s). The post hoc test revealed a very significant difference when comparing the SAR method with both paper and tablet ($p < 0.01$). As the assembly time contains the positional error corrections, made within or prior steps, it influences the overall t_{assemble} time for paper (Mean = 9.87 s, SD = 8.04 s) and tablet (Mean = 6.61 s, SD = 3.45 s), which are higher than the in situ (Mean = 5.57 s, SD = 1.69 s). However, the Friedman test did not reveal any significant differences between the approaches. $t_{\text{total-cycle}}$ shows the total cycle time in minutes instead of seconds. From Figure 10, it is evident that the proposed SAR system requires the lowest time to complete the assembly (Mean = 3.95 min, SD = 0.97 min). The paper (Mean = 7.89 min, SD = 3.32 min) and tablet (Mean = 6.99 min, SD = 2.49 min) methods are higher than the in situ and are not significantly different from each other according to the Friedman and post hoc tests. For the TCT, Gest-SAR shows a highly significant (**, $p < 0.01$) difference against the paper and tablet methods.

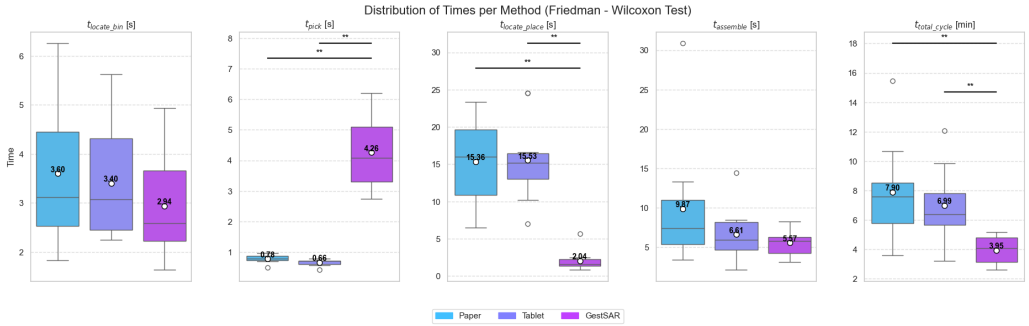


Figure 10. Task completion times for the Paper, Tablet, and GestSAR methods. Task completion times for the Paper, Tablet, and GestSAR methods. Asterisks indicate statistically significant differences found using a Wilcoxon signed-rank test (** $p < 0.01$).

4.2. Error Distribution

We calculated both picking and placing errors for each visualisation method completed by the participants. Figure 11 shows the error distribution of each method with the total errors recorded during the experiments. Starting with the picking error, the graph shows that the in situ error count (Mean = 0.70, SD = 0.67) is much lower than the average error count of paper (Mean = 2.60) and tablet (Mean = 3.60). The Friedman test shows a significant ($p = 0.0076$) difference among the methods, which is further clarified by the Wilcoxon post hoc test, as the difference between (paper, Gest-SAR) and (tablet, Gest-SAR) are both significant ($p = 0.333$, 0.0439), which is shown with a single asterisk '*' in the picking error graph. The average placement errors in the paper (Mean = 2.30, SD = 1.83) and tablet (Mean = 2.20, SD = 1.48) methods are much higher than the error count of the in situ (Mean = 0.00). This shows that the participant did not make any mistakes placing the Lego bricks during the in situ assembly. Friedman's test shows a very significant difference ($p = 0.0014$) for all the methods. The post hoc test further establishes that in situ is significantly better than the paper (*, $p = 0.0213$) and tablet (*, $p = 0.0216$). The total error count of Gest-SAR, with only an average of 0.70 (SD = 0.67), shows its dominance over the paper and tablet methods, which contain 5.00 (SD = 2.00) and 5.90 (SD = 3.48) average errors, respectively. The Friedman test shows a very highly significant ($p = 0.0006$) difference among all the modalities. The Wilcoxon test provides further validation that the in situ method causes significantly lower error count than both paper (**, $p = 0.0059$) and tablet (*, $p = 0.0226$).

4.3. Cognitive Workload—NASA TLX

We consider five measures (mental demand, temporal demand, performance, effort, and frustration) to calculate the perceived cognitive workload during the experiments (Figure 12). Paper (Mean = 3.60, SD = 1.43) and tablet (Mean = 3.70, SD = 0.95) are almost similar on average; however, the paper method shows mental demand with higher variance than the tablet method. In situ (Mean = 2.10, SD = 0.99) is the least mentally demanding method. The Friedman test shows a very significant difference of 0.0060. The post hoc test shows a significant difference in mental demand (p = 0.0493) for (paper, Gest-SAR) and a significant p-value (p = 0.0293) when compared to (Gest-SAR, tablet). Participants felt less hurried or rushed with the pace of the task when using in situ (Mean = 2.20, SD = 1.23) compared to the tablet (Mean = 2.80) and paper (Mean = 3.20). The Friedman test does not show any significance for temporal demand. The average performance for all three methods is very close, with the paper having the lowest average at 4.30. The average effort to accomplish the performance is very high with paper (Mean = 3.80, SD = 1.32) compared to the in situ result (Mean = 2.00, SD = 1.15). The Friedman test shows a very significant p-value of 0.0119 for the effort made in each method. However, the post hoc tests show no significant differences among the pair comparisons of the modalities. Participants felt the least frustration with the SAR (Mean = 1.80, SD = 1.14). The paper and tablet methods are equally frustrated, with an average of 3.10, where the tablet (SD = 1.29) is slightly higher in variance than the paper (SD = 1.20). The overall Friedman test does not show any significant difference (p = 0.1575).

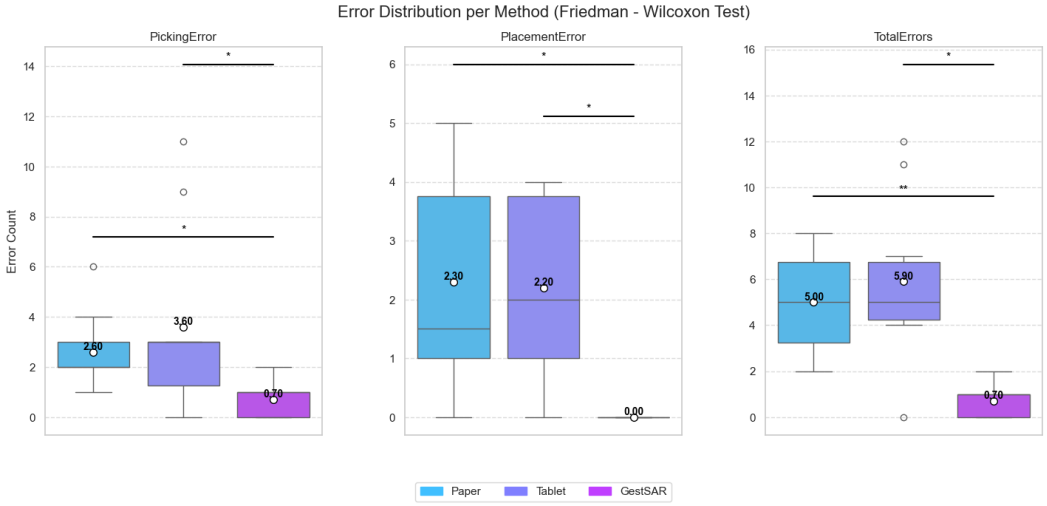


Figure 11. Error counts for the Paper, Tablet, and GestSAR methods. Asterisks indicate statistically significant differences (* $p < 0.05$, ** $p < 0.01$) from a Wilcoxon signed-rank test.



Figure 12. NASA-TLX workload ratings for the Paper, Tablet, and GestSAR methods. Asterisks indicate a statistically significant difference (* $p < 0.05$) from a Wilcoxon signed-rank test.

4.4. Instruction Method Comparisons

The results of user feedback on the assembly instructions for each method show that participants found the instructions easy to understand across all the methods (Figure 13). Similarly, the steps were clearly outlined for all the methods, as the average agreement is 4.50 for paper, tablet, and projector. Users felt that the in situ (Mean = 4.90, SD = 0.32) method is the quickest to assemble Lego bricks compared to the paper (Mean = 2.10, SD = 0.74) and tablet (Mean = 2.40, SD = 1.07). Following a very significant difference shown by the Friedman test ($p = 0.0001$), the post hoc test further confirms the significant p-value when comparing SAR with paper ($p = 0.0059$) and tablet ($p = 0.0178$). Paper shows the lowest satisfaction rate (Mean = 3.20, SD = 1.14), with the average of tablet being 0.30 higher than that of paper. The average satisfaction of in situ is 4.70, which is the highest and clearly shows higher satisfaction and a very close to strong agreement while using the projector-based instruction. The Friedman test shows a very highly significant ($p = 0.0010$) satisfaction across the methods. The Wilcoxon test shows significant difference when Gest-SAR is compared with paper ($p = 0.0178$) and tablet ($p = 0.0473$).

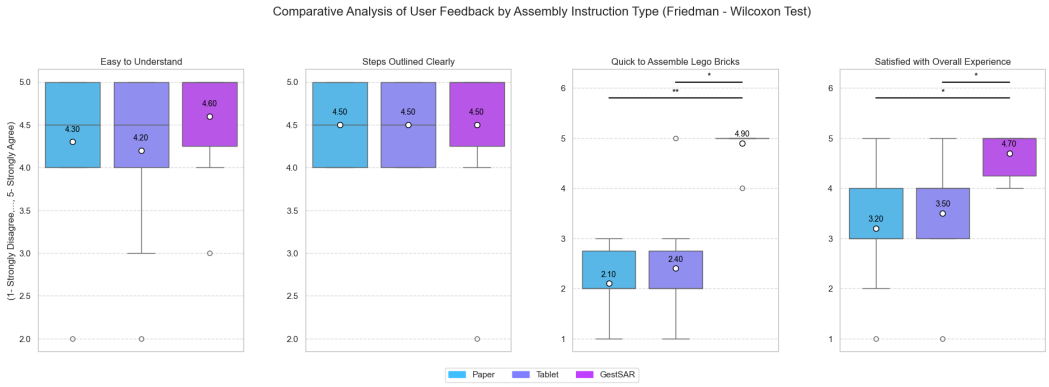


Figure 13. User feedback ratings for the Paper, Tablet, and GestSAR assembly instructions. Asterisks indicate statistically significant differences found using a Wilcoxon signed-rank test (* $p < 0.05$, ** $p < 0.01$).

4.5. User Feedback on SAR and Its Features

Participants were asked a few usability questions on the in situ projector system at the end of the experiment. Users strongly agreed that the SAR system made their work simpler, which reflected on the average rating of 4.9 (Figure 14). Participants stated that "counting the Duplo plate studs for paper and tablet was irritating and confusing due to the base plate colour" and "SAR was one of the best among the three experiments..." (P1, P2, P4, P9, P10). Participant 3 stated that "Locating the brick position automatically" was the part they liked the most in the SAR system. The SAR system was rated as very helpful (Mean = 4.8, SD = 0.42) and easy to use (Mean = 4.7, SD = 0.67). When asked which method the users found more effective for completing the assemblies, 90% of them agreed with SAR-based instructions. Participant 6 further stated that "The instructions [in SAR] were easy to follow and made it worthwhile [conducting the experiments]", and Participant 8 added that "This kind of assembly is the first one I have ever seen. I really liked the projector based assembly. I can imagine an improved version is being used in real-life tasks." The question of whether the system causes discomfort had the lowest average score (3.9) with SD = 1.29. This is still a positive outcome, leaning towards "Agree" that SAR does not cause discomfort. One of the reasons could be the room lighting condition, which reduces the contrast of the projected information within the workspace. Participant also mentioned a similar outcome in their feedback, stating "A better quality projector so that the text is clearer to read" (P9), "The projector-based system cause some strain in the eyes

if worked for a prolonged period. This may be due to the fact that the projector is not as bright and some of the text needs to have better contrast with the background of the box where the text is placed..." (P2). Users showed a strong preference to work with the in situ projector-based system over traditional, both paper- and tablet-based, assembly instructions, with a high average score of 4.7. For most of the questions, the median score was 5.0, excluding the question on physical discomfort (Median = 4.0), which shows that at least half of the users gave the highest possible rating ("Strongly Agree"). This indicates a very strong positive consensus on the SAR system usability feedback.

The boxes on the graph shown in Figure 15 are mostly positioned far to the right, between 4 and 5, reinforcing the positive user sentiment on the in situ work assistance system. When asked about different aspects and features of the SAR system that the participants found helpful, the result in Figure 15 shows that the top three features, pick-and-place by light, step-by-step guidance, and animation of bins location comprise 62.2%. Audio feedback on a correct brick pick and real-time gesture display within the workspace features were both reported 4 times by the participants. Video demonstration of the steps was the least helpful feature among all the SAR system features.

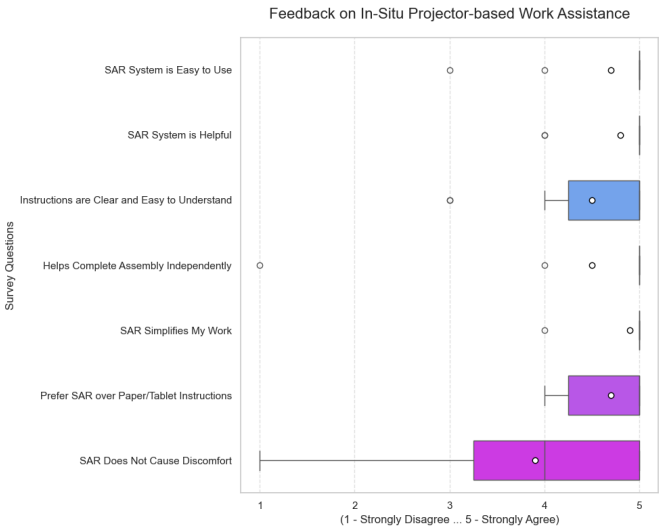


Figure 14. Usability feedback for the Gest-SAR system.

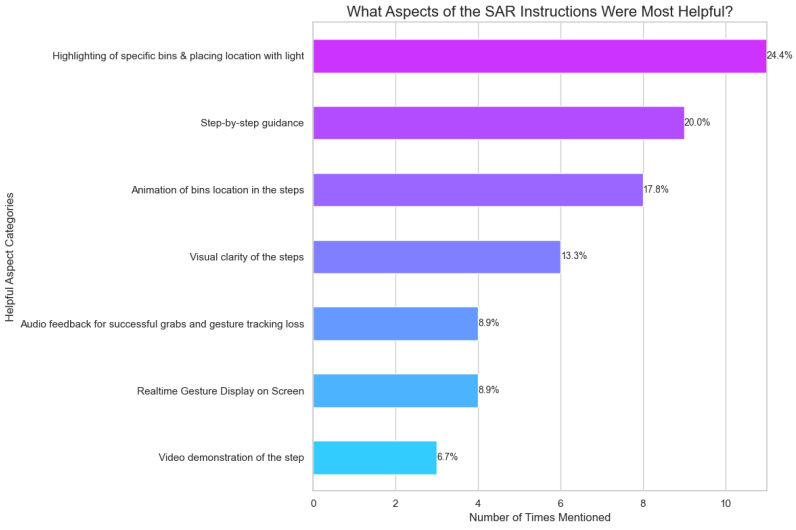


Figure 15. Ranking of SAR Features.

5. Discussion

Both the objective and the subjective analyses shown in the result section provide strong empirical evidence supporting the effectiveness of the proposed Gest-SAR framework in improving manual assembly tasks. The in situ system consistently outperformed traditional instruction methods such as paper-based and tablet-based approaches, across all key performance indicators—task completion time, error rates, and perceived cognitive workload.

Results for the average task completion time indicate that the SAR system takes significantly less time than the paper and tablet instruction modalities. This proves that conducting order and assembly tasks is faster than the traditional approaches. A similar trend was observed within the phase-wise time distribution of the Total Completion Time (TCT). It is noteworthy that only the time to pick up a Lego brick during the steps was taking significantly longer in the SAR-based system. While picking the brick from the correct box, the gesture classification mechanism often fluctuated between gesture states, causing delays in the picking process. Additionally, the fixed 1-sec duration required to confirm the pick/grab gesture contributed to the increased brick-picking (t_{p}) time.

However, the increased t_{p} time did not impact the overall TCT due to the massive contribution of time toward the location of the placement brick. Furthermore, the average picking error is significantly lower when the SAR modality is used. This proves that though the picking time is slightly higher, it can ensure that the error made during the assembly is reduced. Furthermore, future improvements in robust and complex gesture classification models can enhance the gesture detection, thus helping reduce the picking time in the SAR system. Additional factors like object on hand detection can help true positive picking; thus, the one-second constant time can be reduced to help balance the t_{p} during the SAR system. The average time to locate the placement position in paper and tablet modalities is more than seven times higher than the SAR system, which outlines the clear impact of the proposed system over the traditional approaches. One of the most notable outcomes of the Gest-SAR is the drastic reduction in placement error, with a mean of 0.0 error compared to significantly higher rates with conventional methods. This suggests that real-time visual guidance and adaptive feedback not only improve task execution time and accuracy but also enhance confidence in users. Additionally, the system's capability of detecting incorrect actions, like incorrect pick and place of brick and guiding users towards resolution of errors in real-time, reinforces the benefits of a closed-loop interaction paradigm during the assembly time, which was largely absent in prior SAR systems. Gest-SAR advances in the field of SAR-based assembly by providing a dynamic, responsive interface which supports mid-air gesture recognition and live instructional adaptation based on the user's actions, compared to some of the earlier studies on SAR and AR. Unlike the static projection systems, Gest-SAR does not suffer from limited interactivity; rather helps the users engage with the assembly tasks directly through touchless control, enabling fluid progression through the assembly steps without relying on external input devices like a keyboard or sliders. This also aligns with the principles of Industry 5.0 that advocate for human-centric, intelligent, and collaborative technologies.

User feedback through the NASA-TLX and usability surveys further underscores the system's strengths. Users rated Gest-SAR the highest for ease of use, supportive towards the assembly task, satisfaction, and perceived effectiveness. Many users reported that the pick-and-place light projection guidance and real-time gesture recognition within the workspace, with added sound to notify of loss of gesture tracking, significantly reduced the need for cognitive mapping between instructions and physical components. NASA-TLX showed significant mental demand and effort needed to accomplish the task when the traditional approach was followed, and overall positive outcomes when the Gest-SAR system was

followed. Though 90% of the users preferred the SAR system, a few of them noted a few issues with projection readability under certain lighting conditions, highlighting a potential area for hardware enhancement in future implementations.

6. Conclusions

This study introduced Gest-SAR, a novel closed-loop gesture-driven SAR framework designed to improve real-time operational guidance and feedback in manual assembly tasks. Through the combination of custom hand-gesture recognition for controlling the SAR system with the closed-loop feedback mechanisms, Gest-SAR demonstrated substantial improvements over the traditional paper-based and tablet-based instruction systems. Despite the promising results, there are a few areas where future improvement is necessary. The participant sample was relatively small and drawn from an academic environment, comprising students, researchers, and academic staff. While this sample provides meaningful insight into usability and task performance, it may not fully represent the operational characteristics of trained industrial workers. Experience and professional shopfloor operators, for instance, may exhibit different levels of proficiency in component handling, faster picking and placing time, and possibly different subjective experiences.The assembly was based on an Aruco marker, which limits the assembly task to being a single layer. The detection of correct placement within the expected placement location was performed based on scanning the marker ID, which limits the assembly task to be made for a multi-layer complex assembly task for the SAR system. In the future experiment, the use of image processing or deep learning algorithms like You Only Look Once (YOLO) models for object detection can replace the marker-based detection, thus allowing multi-layer Lego-based assembly tasks as performed in prior research. Moreover, the custom multi-layer gesture classification model could be improved and tailored for the assembly task by expanding the gesture data collection under varied environmental conditions (e.g., lighting and occlusion), which will be critical for the deployment of SAR systems in real-world industrial settings.

The mediapipe hand detection loses detection tracking when the hand movement is very fast under low-light conditions, while increased light can improve the frame rates and robust detection of hand landmarks and gesture detection, but it brings the issue with the projected information in the workspace. Higher ambient lighting conditions reduce the contrast and visibility of the information shown in the workspace, which was reported by a few participants. They also reported in their feedback that because of the increased illumination during SAR assembly, they needed additional time to look for the green light projected on the picking box location. As used in previous research, using a laser projector could be useful to assist in projecting high-contrast and sharp visual information on the workspace. This may help in reducing eye fatigue or strain during prolonged usage of SAR systems in assembly tasks. The use of deep learning algorithms like the Convolutional Neural Network (CNN) for gesture classification for the assembly task can be an alternative to hand landmark-based gesture classification. Our current study allows the users to use only a single hand to be tracked for hand gesture detection, thus lacking the flexibility of using both hands at the same time during the assembly. Future improvements can include the detection of both hands to increase the robustness of Gest-SAR. Furthermore, the current study includes only a single worker to conduct the assemblies at a time; therefore, future improvements of Gest-SAR can be explored towards the potential of using gesture-based assemblies having multiple workers collaboratively working together within a single workspace.

Gest-SAR advances the state of the art in SAR-based assembly support by integrating mid-air gesture control and real-time feedback into a unified closed-loop framework. Unlike prior passive SAR systems, this work demonstrates a dynamic and responsive system that

adapts to user actions and provides immediate error correction. This contribution fills a critical gap in the literature by systematically combining gesture recognition, adaptive projection, and error tracking within a single operational pipeline aligned with Industry 5.0 goals of human-centric, intelligent manufacturing.

The practical potential of GestSAR is significant, particularly for industries involved in high-mix, low-volume production, where frequent changes in assembly instructions increase operator cognitive load. The system's flexibility, low-cost components (e.g., standard projector and RGB camera), and intuitive interface make it viable for deployment in small- and medium-sized enterprises. Its projected instruction modality avoids wearable discomfort, while touchless control reduces contamination risks in sensitive environments. These factors support the replicability and adaptability of Gest-SAR across various industrial domains, such as electronics assembly, packaging, and quality inspection.

The proposed system significantly reduced overall assembly time and errors while lowering the users' perceived cognitive workload. The integration of adaptive projection with intuitive gesture control presents a scalable solution for industrial applications where human dexterity and real-time responsiveness are essential for enhancing productivity and providing a cognitively supportive work environment.

Future work will focus on expanding the system's gesture robustness, improving projection visibility under diverse lighting and occlusion conditions, enabling dual-hand interaction, deep learning-based object detection (e.g., YOLO). Further experiments involving experience shopfloor operators in industrial environments will be conducted to validate scalability, usability, and long-term adoption in real-world settings. Overall, Gest-SAR proves a successful step toward an intelligent, human-centric manufacturing system aligned with the goals of Industry 5.0. This research bridges the gap between manual labour and digital augmentation, thus offering a promising direction for the future of smart collaborative assembly technologies.

Author Contributions: Conceptualisation, N.H. and B.A.; methodology, N.H.; software, N.H.; validation, N.H. and B.A.; formal analysis, N.H.; investigation, N.H.; resources, N.H.; data curation, N.H.; writing—original draft preparation, N.H.; writing—review and editing, B.A.; visualisation, N.H.; supervision, B.A.; project administration, B.A.; funding acquisition, B.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The research protocol for this study received formal approval from the Ethics Committee at London South Bank University (Approval Reference: ETH2324-0112).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Survey and Project Details Can Be Found in the GitHub Repository: https://github.com/ISILablsbu/GestSAR.

Acknowledgments: The authors gratefully acknowledge the support and assistance provided by Safia Barikzai from the Division of Computer Science and Digital Technologies, College of Technology and Environment, LSBU.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Gest-SAR	Gesture Spatial Augmented Reality
Gest	Gesture
KPI	Key Performance Indicator
SAR	Spatial Augmented Reality

HCI	Human Computer Interaction
HMD	Head Mounted Displays
AR	Augmented Reality
AET	Assembly Error Tracking
AP	Assembly Process
OP	Order Picking
PM	Progress Monitoring
AE	Assembly Error
PE	Picking Error
GS	General Survey
NASA-TLX	NASA Task Load Index
PAR	Projected Augmented Reality
DNN	Deep Neural Network
ROI	Region of Interest
FPS	Frame Per Second
MLP	Multi-Layer Perceptron
GATM	General Assembly Task Model
TCT	Task Completion Time
HC-ZDM	Human-Centric Zero Defect Manufacturing

References

1. Bortolini, M.; Faccio, M.; Galizia, F.G.; Gamberi, M.; Pilati, F. Adaptive Automation Assembly Systems in the Industry 4.0 Era: A Reference Framework and Full–Scale Prototype. *Appl. Sci.* **2021**, *11*, 1256. [CrossRef]
2. Funk, M.; Schmidt, A. Cognitive Assistance in the Workplace. *IEEE Pervasive Comput.* **2015**, *14*, 53–55. [CrossRef]
3. Alkan, B.; Vera, D.; Ahmad, M.; Ahmad, B.; Harrison, R. A model for complexity assessment in manual assembly operations through predetermined motion time systems. *Procedia CIRP* **2016**, *44*, 429–434. [CrossRef]
4. Alkan, B. An experimental investigation on the relationship between perceived assembly complexity and product design complexity. *Int. J. Interact. Des. Manuf. (IJIDeM)* **2019**, *13*, 1145–1157. [CrossRef]
5. European Commission. What Is Industry 5.0? Available online: https://research-and-innovation.ec.europa.eu/research-area/industrial-research-and-innovation/industry-50_en#what-is-industry-50 (accessed on 30 May 2025).
6. Moraes, A.; Carvalho, A.M.; Sampaio, P. Lean and Industry 4.0: A Review of the Relationship, Its Limitations, and the Path Ahead with Industry 5.0. *Machines* **2023**, *11*, 443. [CrossRef]
7. Wan, P.K.; Leirmo, T.L. Human-centric zero-defect manufacturing: State-of-the-art review, perspectives, and challenges. *Comput. Ind.* **2023**, *144*, 103792. [CrossRef]
8. Eriksson, K.M.; Olsson, A.K.; Carlsson, L. Beyond Lean Production Practices and Industry 4.0 Technologies toward the Human-Centric Industry 5.0. *Technol. Sustain.* **2024**, *3*, 286–308. [CrossRef]
9. Bertram, P.; Birtel, M.; Quint, F.; Ruskowski, M. Intelligent Manual Working Station through Assistive Systems. *IFAC-PapersOnLine* **2018**, *51*, 170–175. [CrossRef]
10. Alkan, B.; Vera, D.; Ahmad, B.; Harrison, R. A method to assess assembly complexity of industrial products in early design phase. *IEEE Access* **2017**, *6*, 989–999. [CrossRef]
11. Bovo, R.; Binetti, N.; Brumby, D.P.; Julier, S. Detecting Errors in Pick and Place Procedures: Detecting Errors in Multi-Stage and Sequence-Constrained Manual Retrieve-Assembly Procedures. In Proceedings of the 25th International Conference on Intelligent User Interfaces, Cagliari, Italy, 17–20 March 2020; pp. 536–545. [CrossRef]
12. Pereira, A.C.; Alves, A.C.; Arezes, P. Augmented Reality in a Lean Workplace at Smart Factories: A Case Study. *Appl. Sci.* **2023**, *13*, 9120. [CrossRef]
13. Bock, L.; Bohné, T.; Tadeja, S.K. Decision Support for Augmented Reality-Based Assistance Systems Deployment in Industrial Settings. *Multimed Tools Appl.* **2025**, *84*, 23617–23641. [CrossRef]
14. Fani, V.; Antomarioni, S.; Bandinelli, R.; Ciarapica, F.E. Data Mining and Augmented Reality: An Application to the Fashion Industry. *Appl. Sci.* **2023**, *13*, 2317. [CrossRef]
15. Werrlich, S.; Daniel, A.; Ginger, A.; Nguyen, P.A.; Notni, G. Comparing HMD-Based and Paper-Based Training. In Proceedings of the 2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Munich, Germany, 16–20 October 2018; pp. 134–142. [CrossRef]

16. Funk, M.; Kosch, T.; Schmidt, A. Interactive Worker Assistance: Comparing the Effects of in-Situ Projection, Head-Mounted Displays, Tablet, and Paper Instructions. In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Heidelberg, Germany, 12–16 September 2016; pp. 934–939. [CrossRef]
17. Grosse, E.H. Application of supportive and substitutive technologies in manual warehouse order picking: A content analysis. *Int. J. Prod. Res.* **2024**, *62*, 685–704. [CrossRef]
18. Rupprecht, P.; Kueffner-Mccauley, H.; Trimmel, M.; Hornacek, M.; Schlund, S. Advanced Adaptive Spatial Augmented Reality Utilizing Dynamic In-Situ Projection in Industrial Site Assembly. *Procedia CIRP* **2022**, *107*, 937–942. [CrossRef]
19. Kosch, T.; Funk, M.; Schmidt, A.; Chuang, L.L. Identifying Cognitive Assistance with Mobile Electroencephalography: A Case Study with In-Situ Projections for Manual Assembly. *Proc. ACM Hum.-Comput. Interact.* **2018**, *2*, 11. [CrossRef]
20. Lucchese, A.; Panagou, S.; Sgarbossa, F. Investigating the Impact of Cognitive Assistive Technologies on Human Performance and Well-Being: An Experimental Study in Assembly and Picking Tasks. *Int. J. Prod. Res.* **2024**, *63*, 2038–2057. [CrossRef]
21. Funk, M.; Bächler, A.; Bächler, L.; Kosch, T.; Heidenreich, T.; Schmidt, A. Working with Augmented Reality?: A Long-Term Analysis of In-Situ Instructions at the Assembly Workplace. In Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments, Island of Rhodes, Greece, 21–23 June 2017; pp. 222–229. [CrossRef]
22. Cao, Z.; He, W.; Wang, S.; Zhang, J.; Wei, B.; Li, J. Research on Projection Interaction Based on Gesture Recognition. In *HCI International 2021—Late Breaking Posters*; Stephanidis, C., Antona, M., Ntoa, S., Eds.; Springer International Publishing: Cham, Switzerland, 2021; Volume 1498, pp. 311–317. [CrossRef]
23. Funk, M.; Bächler, A.; Bächler, L.; Korn, O.; Krieger, C.; Heidenreich, T.; Schmidt, A. Comparing Projected In-Situ Feedback at the Manual Assembly Workplace with Impaired Workers. In Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments, Corfu, Greece, 1–3 July 2015; pp. 1–8. [CrossRef]
24. Kalpagam Ganesan, R.; Rathore, Y.K.; Ross, H.M.; Ben Amor, H. Better Teaming Through Visual Cues: How Projecting Imagery in a Workspace Can Improve Human-Robot Collaboration. *IEEE Robot. Autom. Mag.* **2018**, *25*, 59–71. [CrossRef]
25. Uva, A.E.; Gattullo, M.; Manghisi, V.M.; Spagnulo, D.; Cascella, G.L.; Fiorentino, M. Evaluating the Effectiveness of Spatial Augmented Reality in Smart Manufacturing: A Solution for Manual Working Stations. *Int. J. Adv. Manuf. Technol.* **2018**, *94*, 509–521. [CrossRef]
26. Hietanen, A.; Changizi, A.; Lanz, M.; Kamarainen, J.; Ganguly, P.; Pieters, R.; Latokartano, J. Proof of concept of a projection-based safety system for human-robot collaborative engine assembly. In Proceedings of the 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), New Delhi, India, 14–18 October 2019; pp. 1–7. [CrossRef]
27. Sand, O.; Büttner, S.; Paelke, V.; Röcker, C. smARt.Assembly—Projection-Based Augmented Reality for Supporting Assembly Workers. In *Virtual, Augmented and Mixed Reality*; Lackey, S., Shumaker, R., Eds.; Springer International Publishing: Cham, Switzerland, 2016; Volume 9740, pp. 643–652. [CrossRef]
28. Bosch, T.; Van Rhijn, G.; Krause, F.; Könemann, R.; Wilschut, E.S.; De Looze, M. Spatial augmented reality: A tool for operator guidance and training evaluated in five industrial case studies. In Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments, Corfu, Greece, 30 June–3 July 2020; pp. 1–7. [CrossRef]
29. Swenja, S.; Maximilian, P.; Thomas, S. Evolution of Pick-by-Light Concepts for Assembly Workstations to Improve the Efficiency in Industry 4.0. *Procedia Comput. Sci.* **2022**, *204*, 37–44. [CrossRef]
30. Zhang, J.; Wang, S.; He, W.; Li, J.; Cao, Z.; Wei, B. Projected Augmented Reality Assembly Assistance System Supporting Multi-Modal Interaction. *Int. J. Adv. Manuf. Technol.* **2022**, *123*, 1353–1367. [CrossRef]
31. Rupprecht, P.; Kueffner-McCauley, H.; Trimmel, M.; Schlund, S. Adaptive Spatial Augmented Reality for Industrial Site Assembly. *Procedia CIRP* **2021**, *104*, 405–410. [CrossRef]
32. Oudah, M.; Al-Naji, A.; Chahl, J. Hand Gesture Recognition Based on Computer Vision: A Review of Techniques. *J. Imaging* **2020**, *6*. [CrossRef] [PubMed]
33. Chen, C.; Tian, Z.; Li, D.; Pang, L.; Wang, T.; Hong, J. Projection-Based Augmented Reality System for Assembly Guidance and Monitoring. *Assem. Autom.* **2021**, *41*, 10–23. [CrossRef]
34. Kosch, T.; Kettner, R.; Funk, M.; Schmidt, A. Comparing tactile, auditory, and visual assembly error-feedback for workers with cognitive impairments. In Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS), Reno, NV, USA, 23–26 October 2016; pp. 53–60. [CrossRef]
35. Nazarova, E.; Sautenkov, O.; Altamirano Cabrera, M.; Tirado, J.; Serpiva, V.; Rakhmatulin, V.; Tsetserukou, D. CobotAR: Interaction with Robots Using Omnidirectionally Projected Image and DNN-based Gesture Recognition. In Proceedings of the 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Melbourne, Australia, 17–20 October 2021; pp. 2590–2595. [CrossRef]
36. Torres, W.; Santos, L.; Melo, G.; Oliveira, A.; Nascimento, P.; Carvalho, G.; Neves, T.; Martins, A.; Araújo, Í. A Framework for Real-Time Gestural Recognition and Augmented Reality for Industrial Applications. *Sensors* **2024**, *24*, 2407. [CrossRef]
37. Wang, P.; Wang, Y.; Billinghurst, M.; Yang, H.; Xu, P.; Li, Y. BeHere: A VR/SAR Remote Collaboration System Based on Virtual Replicas Sharing Gesture and Avatar in a Procedural Task. *Virtual Real.* **2023**, *27*, 1409–1430. [CrossRef]

38. Fang, W.; Hong, J. Bare-Hand Gesture Occlusion-Aware Interactive Augmented Reality Assembly. *J. Manuf. Syst.* **2022**, *65*, 169–179. [CrossRef]
39. Dong, J.; Xia, Z.; Zhao, Q. Augmented Reality Assisted Assembly Training Oriented Dynamic Gesture Recognition and Prediction. *Appl. Sci.* **2021**, *11*, 9789. [CrossRef]
40. Vogiatzidakis, P.; Koutsabasis, P. Mid-Air Gesture Control of Multiple Home Devices in Spatial Augmented Reality Prototype. *Multimodal Technol. Interact.* **2020**, *4*, 61. [CrossRef]
41. Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.L.; Yong, M.G.; Lee, J.; et al. MediaPipe: A framework for building perception pipelines. *arXiv* **2019**, arXiv:1906.08172. [CrossRef]
42. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.E.; Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *arXiv* **2019**, arXiv:1812.08008. [CrossRef]
43. Weichert, F.; Bachmann, D.; Rudak, B.; Fisseler, D. Analysis of the Accuracy and Robustness of the Leap Motion Controller. *Sensors* **2013**, *13*, 6380–6393. [CrossRef]
44. Sweigart, A. PyAutoGUI Documentation. Available online: https://pyautogui.readthedocs.io/en/latest/ (accessed on 15 June 2025).
45. Intel. Intel® RealSense™ Depth Camera D435. Available online: https://www.intelrealsense.com/depth-camera-d435/ (accessed on 1 June 2025).
46. Indriani, M.H.; Agoes, A.S. Applying Hand Gesture Recognition for User Guide Application Using MediaPipe. In Proceedings of the 2nd International Seminar of Science and Applied Technology (ISSAT 2021), Online, 12 October 2021. [CrossRef]
47. Google Research. On-Device, Real-Time Hand Tracking with MediaPipe. Available online: https://research.google/blog/on-device-real-time-hand-tracking-with-mediapipe/ (accessed on 14 May 2025).
48. Bannat, A.; Wallhoff, F.; Rigoll, G.; Friesdorf, F.; Bubb, H.; Stork, S.; Müller, H.J.; Schubö, A.; Wiesbeck, M.; Zäh, M.F. Towards optimal worker assistance: A framework for adaptive selection and presentation of assembly instructions. In Proceedings of the 1st International CoTeSys Workshop, Munich, Germany, 6–8 October 2008; p. 7.
49. Funk, M.; Mayer, S.; Schmidt, A. Using in-situ projection to support cognitively impaired workers at the workplace. In Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS), Lisbon, Portugal, 26–28 October 2015; pp. 185–192. [CrossRef]
50. Tang, A.; Owen, C.; Biocca, F.; Mou, W. Comparative effectiveness of augmented reality in object assembly. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03), Ft. Lauderdale, FL, USA, 5–10 April 2003; pp. 73–80. [CrossRef]
51. Sakata, N.; Kurata, T.; Kuzuoka, H. Visual Assist with a Laser Pointer and Wearable Display for Remote Collaboration. In Proceedings of the 2nd International Conference on Collaboration Technologies (CollabTech 2006), Tsukuba, Japan, 13–14 July 2006; pp. 66–71.
52. Funk, M.; Kosch, T.; Greenwald, S.W.; Schmidt, A. A Benchmark for Interactive Augmented Reality Instructions for Assembly Tasks. In Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia, Linz, Austria, 30 November–2 December 2015; pp. 253–257. [CrossRef]
53. Hart, S.G. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **2006**, *50*, 904–908. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.