INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research

International Journal of Manufacturing Research

ISSN online: 1750-0605 - ISSN print: 1750-0591
https://www.inderscience.com/ijmr

Multimodal human-robot collaboration: advancements and future directions

Sichao Liu, Zhihao Liu, Qiang Qin, Xi Vincent Wang, Lihui Wang

DOI: 10.1504/IJMR.2025.10074362

Article History:

Received:	08 May 2024
Last revised:	21 May 2025
Accepted:	08 July 2025
Published online:	14 November 2025

Multimodal human-robot collaboration: advancements and future directions

Sichao Liu*

Department of Production Engineering,
KTH Royal Institute of Technology,
Stockholm, Sweden
and
Institute of Bioengineering,
École Polytechnique Fédérale de Lausanne,
Lausanne, Switzerland
Email: sicliu@kth.se
*Corresponding author

Zhihao Liu, Qiang Qin, Xi Vincent Wang and Lihui Wang

Department of Production Engineering,
KTH Royal Institute of Technology,
Stockholm, Sweden

Abstract: Human-robot collaboration (HRC) envisioned for future factories has been actively explored to facilitate higher overall productivity. The wide applications of HRC in multiple fields, such as manufacturing and production, have seen a series of milestones. In recent years, a shift towards intuitive and natural collaboration between humans and robots has been investigated and discussed for symbiotic scenarios and complex tasks. For this purpose, advancements of multimodality in HRC enable multimodal human-robot interactions and collaboration by utilising different communication channels such as auditory, vision, gestures, haptics, and even brain signals. In addition, understanding human behaviours in terms of intent and motion can be beneficial in achieving mutual human-robot assistance. Within an HRC setting, the digital twin of such a physical collaborative workcell offers a promising tool to implement sim-to-real transformation and on-demand support for real practice. Within the context, this study provides an overview of the past and current status of multimodal HRC and its applications and highlights future research directions.

[Submitted 8 May 2024; Accepted 8 July 2025]

Keywords: robot; assembly; multimodality; large language model; LLM.

Reference to this paper should be made as follows: Liu, S., Liu, Z., Qin, Q., Wang, X.V. and Wang, L. (2025) ‘Multimodal human-robot collaboration: advancements and future directions’, *Int. J. Manufacturing Research*, Vol. 20, No. 5, pp.1–47.

Biographical notes: Sichao Liu received his PhD degree from KTH Royal Institute of Technology, Sweden. He is currently a Research Fellow of the Swedish Research Council – Vetenskapsrådet, and affiliated with the University of Cambridge (UK), EPFL – Swiss Federal Technology Institute of Lausanne (Switzerland), and KTH Royal Institute of Technology. He mainly focuses on neuroengineering, vision AI/AI4Robotics, large language models, autonomous robot systems (AMR), and foundation models.

Zhihao Liu is currently a Postdoc Researcher at KTH Royal Institute of Technology, and a Postdoctoral Fellow at XPRES (https://www.xpres.kth.se/). Before joining KTH, he earned his PhD in Information Technology and Communication Engineering from Wuhan University of Technology in 2023. During his PhD study, he was a Guest Doctoral student at KTH from 2019 to 2021. He finished his Master's and Bachelor's degrees at WUT in 2018 and 2016, respectively. His research interests include Industry 5.0, digital twin and metaverse, embodied AI, human-robot collaboration, robot learning, neural information processing, and human-compatible AI.

Qiang Qin is currently a PhD student at KTH Royal Institute of Technology, Sweden. He received his BEng in Electronic Science and Technology from Nankai University, China in 2018. Subsequently, he obtained his MS in Advanced Robotics from École Centrale de Nantes, France in 2020. His research interests include robotic manipulation, digital twin, and artificial intelligence for robots.

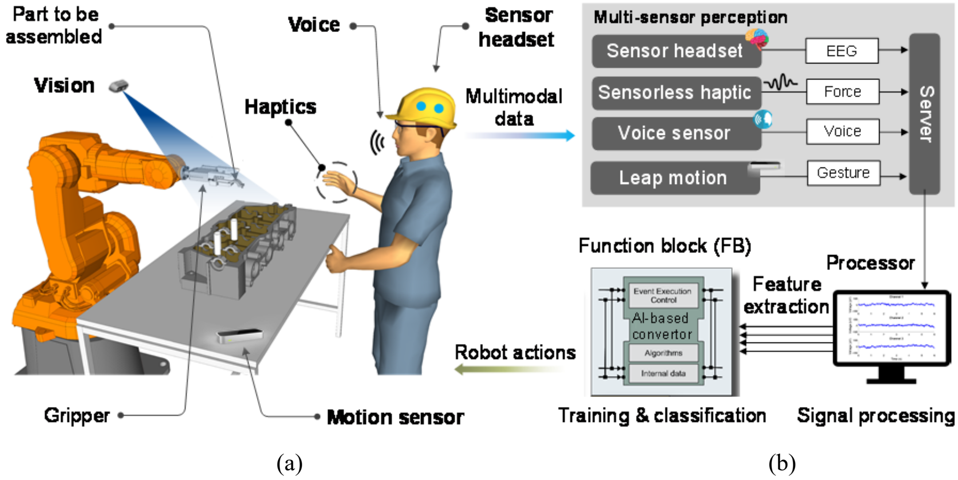
Xi Vincent Wang received his Bachelor's degree from Tianjin University, Tianjin, China in 2008, and PhD degree from the University of Auckland, Auckland, New Zealand in 2013, both in Mechanical Engineering. He is currently an Associate Professor with the Department of Production Engineering, KTH Royal Institute of Technology. His main research interests include cloud-based manufacturing, sustainable manufacturing, robotics, digital twin, computer-aided design, and manufacturing systems. He is the Managing Editor for *International Journal of Manufacturing Research*, Associate Editor for *SME Journal of Manufacturing Systems*, and *Array* by Elsevier, and the editorial board member of other three international journals.

Lihui Wang received his BS in Machine Design from the Academy of Arts and Design, China in 1982, MS in Mechanical Engineering, and PhD in Intelligence Science from Kobe University, Japan in 1990 and 1993. He is a Chair Professor and the Director of Centre of Excellence in Production Research with KTH Royal Institute of Technology, Sweden. His research interests include cyber-physical systems, human-robot collaboration, brain robotics, and manufacturing systems. He is a Fellow of Canadian Academy of Engineering, Society of Manufacturing Engineers, International Academy for Production Engineering, and American Society of Mechanical Engineers.

1 Introduction

Robots are vital catalysts for enhancing competitiveness and flexibility within large-scale manufacturing sectors, notably the automotive industry (Breque et al., 2021). Although robots excel in tasks that demand speed, strength, accuracy, and repeatability, they inherently lack human flexibility, adaptability, and advanced cognitive decision-making abilities (Wang et al., 2019). Hence, human-robot collaboration (HRC) emerges to amalgamate robots' capabilities with humans' intelligence and flexibility (Noohi et al., 2016). Future manufacturing landscapes are expected to involve increased physical HRC in a shared workspace, where human advantages in flexibility and adaptability harmonise with robot precision and strength (Liu et al., 2022b). Exploring robots as collaborative partners in assembly lines has been actively pursued, with companies like ABB and KUKA at the forefront, providing innovative HRC solutions for future production scenarios (Krüger et al., 2009). Collaborative robots (cobots), such as KUKA LBR iiwa and ABB YuMi, are designed to operate alongside nearby humans, often equipped with sensors for intuitive control (Kragic et al., 2018). During the COVID-19 pandemic, there has been a notable increase in the adoption of collaborative robots (cobots), for automation within the medical sector (Taesi et al., 2023). However, heavy-duty industries tend to favour traditional industrial robots for automating manufacturing operations due to the payload and speed limitations associated with cobots (Liu et al., 2021d).

Figure 1 A general framework for multimodal HRC, (a) multimodal inputs (b) perception and control modules (see online version for colours)



Source: Adapted and modified from Liu et al. (2022a)

Conventional industrial robots often lack built-in torque and/or force sensors (Khalil and Dombre, 2002), necessitating the use of additional hardware or assistive systems, such as force sensors, to monitor interaction forces in most current HRC solutions (Liu et al., 2021c). Consequently, a sensorless haptics approach emerges as a promising avenue to facilitate HRC in assembly processes, eliminating the reliance on force and/or torque sensors (Liu et al., 2022a; Sun et al., 2024). For example, adaptive admittance and impedance control methods, which do not require force/torque measurement, have been widely used in physical human-robot interactions (HRIs) (Ngo and Liu, 2024), where the contact force can be accurately estimated and then converted into a robot movement reference. In addition, conventional robots are often governed by rigid native programs that hinder the efficient implementation of HRC (Liu et al., 2021e). To overcome this challenge, multimodal robot programming presents an opportunity for intuitive robot control without the need for specialised expertise in robot programming (Iba et al., 2005; Wang et al., 2019). Within this general framework, as shown in Figure 1, human instructions gathered from a multi-sensor perception system, including voice, gesture, haptics, brainwave data, etc., along with visual information, can serve as multimodal trigger signals for robot control in assembly tasks. These signals are then trained and associated with valid robot control commands by using function blocks (FBs) that embed algorithms for signal processing, to enable seamless HRIs.

Concurrently, shifts in the HRC landscape may necessitate the reprogramming and reconfiguration of tasks during assembly processes (Wang et al., 2021). Efforts are underway to enhance HRC efficiency by leveraging human commands for intuitive robot control and streamlined assembly operations. Simultaneously, advancements in sensor technologies are being leveraged to equip robotic and HRC systems with enhanced perception and cognitive capabilities within manufacturing environments (Lin et al., 2020). These perception modalities often rely on a variety of sensors employed concurrently by humans and robots (Xue et al., 2020), highlighting the complementary nature of different sensing techniques and underscoring the importance of multimodal data fusion to optimise HRC assembly (HRCA) efficiency (Li et al., 2023a; Wang et al., 2025c).

Voice commands are one of the most effective communication tools, widely employed to facilitate intuitive programming and control of robots (Liu et al., 2018; Makris et al., 2014). Simultaneously, integrating gesture recognition and natural language understanding into a multimodal human-robot interface design allows human operators to interact with robots using both speech and gestures (Perzanowski et al., 2001). Findings from the SME robot project underscore the potential of speech and gesture-based instructions in intuitively programming industrial robots (Neto et al., 2010). Additionally, combining speech recognition with other control modalities, such as haptic control, proves efficient for intuitive interaction with collaborative robots like Universal Robots and KUKA Robots (Gustavsson et al., 2017). The EU project SYMBIO-TIC's report showcases a symbiotic HRC system controlled by voice and gesture instructions in an assembly production environment, enabling intuitive collaboration between operators and industrial robots. Additionally, combining visual inputs with spoken instructions enables robots to understand task context more effectively and adapt to dynamic changes on the assembly line. Multimodal models can also facilitate real-time decision-making, such as distinguishing between operator gestures and background movements or responding to emergency verbal cues. However, the presence of noise in manufacturing contexts remains a significant limitation for voice command recognition accuracy during online demonstrations (Kardos et al., 2017). To address this challenge, human gestures are explored as a nonverbal communication channel for intuitive programming and robot control (Islam et al., 2019; Wang et al., 2019). Gesture recognition in HRC encompasses hand and arm gestures, head and facial gestures, and body gestures, with a comprehensive study presented in Liu and Wang (2018).

Furthermore, gestures can be classified as communicative or manipulative based on their intent, with communicative gestures conveying intentions and manipulative gestures used for object manipulation tasks such as translation, rotation, and deformation (Pavlovic et al., 1997). Within the realm of HRC, communicative gestures are categorised into actions or symbols commonly used for intuitive robot programming. Symbol gestures convey linguistic functions through modal functions or referential characters. In contrast, action gestures depict specific movements or actions, often employed in conjunction with AI algorithms to predict human motion and action (Wang et al., 2018). To accurately interpret these gestures, integrating skeletal tracking data obtained from depth sensors or motion capture systems can enable real-time estimation of human joint positions and body postures, providing rich contextual information for gesture recognition and intention inference. For example, pointing gestures combined with gaze direction can be used to specify target objects, while full-body movements can indicate collaborative task phases or transitions. In manufacturing and assembly contexts, skeleton-based motion data can help robots anticipate human actions, adjust their trajectories to avoid collisions, and synchronise movements for shared tasks.

Voice and gesture commands originate from human brains, prompting exploration into the use of brainwave signals for device control and manipulation (Liu et al., 2024a). Brain-computer interfaces (BCIs) are consequently developed to capture brain signals, enabling operators to interact with objects or the environment (Cheng et al., 2020; He et al., 2020). Electroencephalography (EEG) signals, which reflect brain activity, are typically measured using sensor headsets (Liu et al., 2021e). The commercial development and application of BCI devices have garnered significant interest, resulting in sensor headsets embedded with various channels for human-machine or computer interaction (He et al., 2023; Kawala-Sterniuk et al., 2021). EEG signals often exhibit variations in response to external stimuli, such as visual aids, influencing brain activity (Tang et al., 2020). Various approaches to utilising EEG for robot control have been explored (Djemal and Ko, 2020), with a common method involving deep learning (DL) systems to recognise patterns in raw EEG signals and/or transformed features (Craik et al., 2019). Research efforts on algorithms for EEG signal classification have been extensive, encompassing techniques such as recurrent neural networks (RNNs), graph neural networks (GNNs), convolutional neural networks (CNNs), and their derivatives (Al-Saegh et al., 2021; Liu et al., 2024a; Roy et al., 2019). Additionally, haptic instructions provide contact-based commands and offer multimodal support for HRCA alongside auditory, gesture, and brainwave commands. A promising approach for haptic robot control involves sensorless interaction, relying on accurate contact force estimation and adaptive robot control (Kokkalis et al., 2018; Liu et al., 2020). This has spurred investigations into contact force estimation methods. Following accurate force estimation, force-controlled interactions are executed using admittance control, a widely adopted technique in HRC applications. Admittance control transforms contact force into positions and velocities of the robot, with adaptive adjustment of parameters to enhance control performance (Keemink et al., 2018; Ott et al., 2010). Recently, high-performance encoder-decoder approaches were developed to translate brainwaves into texts, which holds a promise to interface LLMs for robot control in downstream tasks. For example, EEG2TEXT, as a novel method, was developed to enhance decoding performance for open vocabulary-based EEG signal to text translation, where learning of the semnatics of the EEG signals supported by pre-trained models and a multi-view transformer are combined for EEG decoding.

In the realm of component assembly, traditional automation methods have encountered limitations, with much of what can be automated already automated, leaving the remaining tasks to be handled manually by human operators (Wang, 2022). Recognising the advantages of HRC, there has been an active exploration into integrating robots as collaborative partners to aid humans in production and assembly lines. For instance, extensive studies have been conducted on HRC in assembly processes, spanning domains such as astronautics and automotive assembly (Tsarouchi et al., 2017). HRCA typically encompasses a series of operations, ranging from assembly planning and task allocation to task execution and robot actions (Wang et al., 2020). Complex assemblies are divided into individual assembly operations, each defined by assembly features (AFs) such as placing, insertion, and welding (Liu et al., 2021a; Zhang et al., 2025a). The combination of these AFs outlines the plan for the assembly task, while assembly logic dictates the sequence of assembly operations within the task (Adamson et al., 2017; Yi et al., 2024). Consequently, diverse approaches utilising AF-based assembly planning have been developed to streamline HRCA processes (Kardos et al., 2017).

Generative artificial intelligence (AI), such as foundation models and large language models (LLMs), has been widely utilised for robot interaction with humans, owing to its natural language processing, task planning, and robot control capabilities (Gao et al., 2024; Kim et al., 2024a; Wang et al., 2025b). LLMs (e.g., ChatGPT-4) can receive multimodal inputs, such as text, images, and audio, and generate the expected output for downstream tasks. Specifically, multimodal LLMs are rapidly transforming HRI by enabling robots to understand and respond to rich, diverse inputs, including text, speech, images, and video (Li et al., 2023b; Liu et al., 2024b). These models integrate multiple data modalities to form a unified representation, allowing robots to interpret complex environments and communicate more naturally with humans. Notable advances, such as OpenAI GPT-4 (OpenAI et al., 2023), which incorporates vision, and Google's PaLM-E (Driess et al., 2023), demonstrate how MLLMs can ground language in perception and action, thereby bridging the gap between symbolic reasoning and sensorimotor data. MLLMs enhance contextual understanding, support zero-shot task learning, and facilitate instruction following via natural language and visual cues (Kuang et al., 2025), as well as visual information-based defect detection (Zhao et al., 2025b). This significantly reduces the need for task-specific programming, improving adaptability in real-world settings. In addition, industrial foundation models in intelligent manufacturing are large-scale, pre-trained AI models, and they integrate multimodal data (e.g., text, vision, sensor signals) to enable generalisable, adaptable, and scalable solutions across diverse manufacturing tasks such as quality inspection, process optimisation, and HRC (Zhao et al., 2025a). However, challenges remain in precisely grounding language, maintaining robustness across diverse environments, and ensuring real-time performance (Chang et al., 2024). Moreover, ethical and safety concerns must be addressed, particularly regarding hallucinations, interpretability, and alignment with human intent (Ferdaus et al., 2024; Huang et al., 2025).

The remainder of this study is organised as follows. Section 2 discusses the contact-based force-controlled collaborative assembly. Section 3 presents natural language-driven HRIs. Section 4 demonstrates the use of brainwave signals in robot control. Section 5 discusses digital twin (DT)-enabled human-robot collaborative assembly (HRCA), and Section 6 describes skeleton-gesture-driven HRC, followed by a discussion of future directions in Section 7. Finally, Section 8 concludes this study.

2 Contact-based force-controlled collaborative assembly

2.1 Dynamic modelling of robots

The dynamic model of the robot with N DoF (Liu et al., 2021b) is defined below:

$$M(q)\ddot{q} + C(q, \dot{q})\dot{q} + G(q) + \tau_f(\dot{q}) = \tau \tag{1}$$

where $M(q) \in \mathbb{R}^{N \times N}$ and $q \in \mathbb{R}^{N}$ are the mass matrix and the vector of joint variables, respectively. $G(q) \in \mathbb{R}^{N}$ and $C(q, \dot{q}) \in \mathbb{R}^{N \times N}$ are the gravity effect and the centrifugal and Coriolis effects, respectively. $\tau \in \mathbb{R}^{N}$ and $\tau_f(\dot{q}) \in \mathbb{R}^{N}$ are the joint torque and friction torques of the joint, respectively.

Contact force and moment acting within the robot's configuration space generate the external torques experienced at its joints. These torques $\tau_e \in \mathbb{R}^{N}$ are linked to the endeffector wrench $f \in \mathbb{R}^{M}$ through a relationship defined in the configuration space

$$\tau_e = J^T(q) f \tag{2}$$

where $J(q) \in \mathbb{R}^{M \times N}$ is the Jacobian matrix.

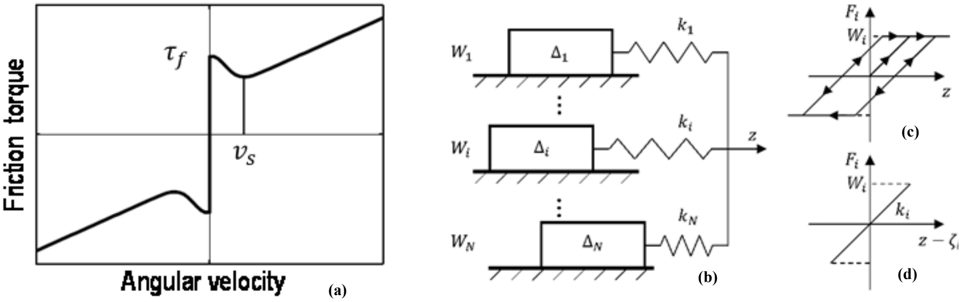
2.2 Stribeck model and generalised Maxwell-slip

In this context, both Stribeck and the generalised Maxwell-slip (GMS) models, illustrated in Figures 2(a), 2(b), 2(c) and 2(d), are employed to elucidate sliding and pre-sliding friction. The Stribeck model comprehensively captures various frictional characteristics (Swevers et al., 2000), whereas the GMS model excels in accurately estimating friction within the pre-sliding regime (Al-Bender et al., 2005). The Stribeck model is represented as follows:

$$F_f(v) = F_c + (F_s - F_c) e^{-\left|\frac{v}{v_s}\right|^{\sigma}} + F_v v \tag{3}$$

where F_c, F_s and F_v are the Coulomb, static and viscous friction, respectively. v and v_s the joint and Stribeck velocities, respectively. σ is the exponent of the Stribeck nonlinearity.

Figure 2 (a) Stribeck model (b) (c) (d) GMS model



Source: Adapted and modified from Swevers et al. (2000)

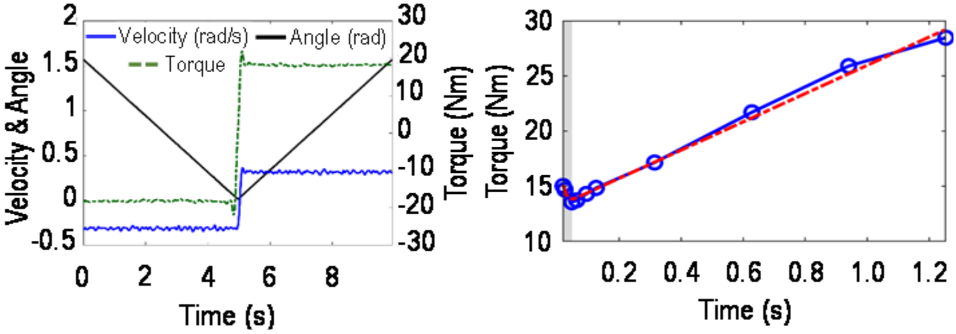
The GMS model is represented by a series of N parallel elasto-slide elements (Al-Bender et al., 2005). The hysteresis force F_i generated by the deformation of the contact point between Maxwell-slip elements is illustrated as:

$$\text{If } |z-\zeta_i| < \frac{W_i}{k_i}, \text{ then } \begin{cases} F_i = k_i(z-\zeta_i) \\ \zeta_i = const \end{cases}$$
$$\text{else} \begin{cases} F_i = \operatorname{sgn}(z-\zeta_i)W_i \\ \zeta_i = z - \operatorname{sgn}(z-\zeta_i)\dfrac{W_i}{\zeta_i} \end{cases} \tag{4}$$

where z is the common displacement, ζ_i, k_i, and W_i are the position, stiffness, and maximum force of element i, respectively. Therefore, the pre-sliding friction is defined as the sum of the hysteresis force of the N elasto-slide elements and formulated as follows.

$$F_h = \sum_{i=1}^{N} F_i \tag{5}$$

Figure 3 (a) Trajectories for measuring the Stribeck friction of Joint 3 at 0.314 rad/s (b) The measured and predicted Stribeck friction (see online version for colours)



(a) (b)

Source: Adapted and modified from Liu et al. (2021b)

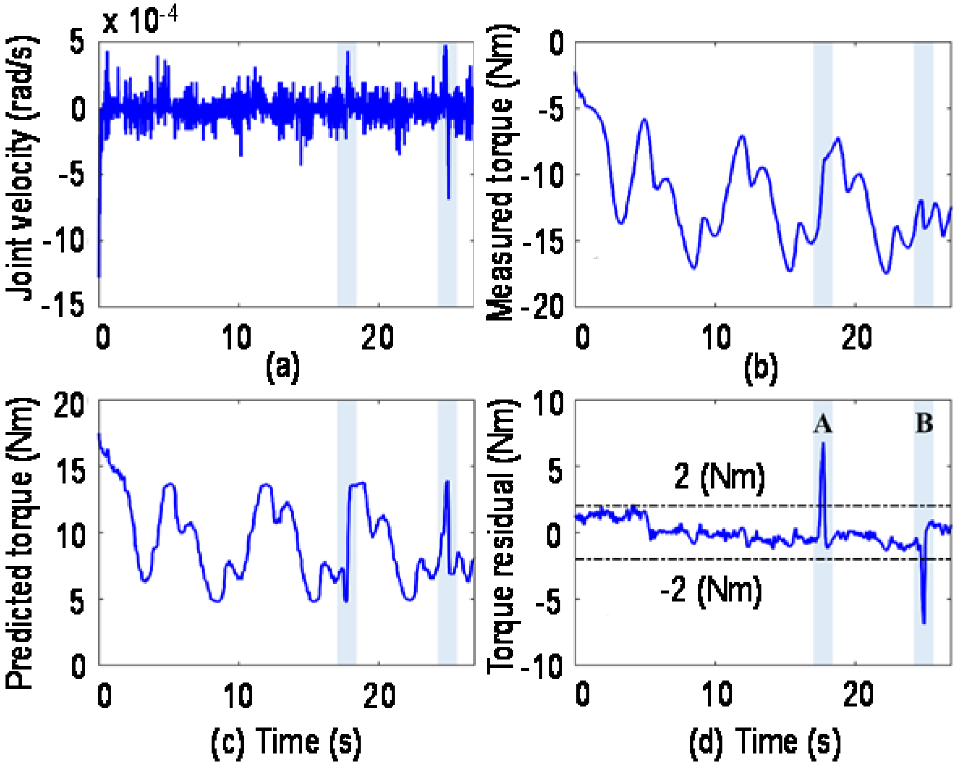
To acquire the parameters of the Stribeck model, each robot joint executes an excitation trajectory specifically designed to measure Stribeck friction. This trajectory involves constant velocity-based movement of the joint at a specified angle in both directions. As shown in Figure 3(a), Joint 3 follows the experimental trajectory with a velocity of 0.314 rad/s. The measured friction of Joint 3 is used for Stribeck parameter identification, indicated by the blue lines in Figure 3(b), and serves as input for identifying the Stribeck parameters. This identification process utilises a sequential quadratic programming method, known for its nonlinear solving capabilities and superlinear convergence characteristics. The red line indicates the predicted Stribeck friction calculated by the identified Stribeck friction model.

$$\max\left(\sum_{k=1}^{N}\left(\tau_f - F_f\right)^2 <\in_f\right) \tag{6}$$

where $\in_f$ is a predefined constant approximation error.

To determine the robot's dynamic parameters, Joint 3 initiates a low-speed clockwise movement with a slight joint angle, saturating its pre-sliding friction, evidenced by the maximal deflection of the asperity junctions. Subsequently, a slight position adjustment is induced on Joint 3 by the trajectory movement of Joint 6, ensuring that the resultant joint velocity remains below the defined threshold, as shown in Figure 4. This ensures that Joint 3 operates within the pre-sliding regime, generating vibration torque. The applied contact force on the robot's end-effector causes Joint 3 to move forward and backwards at velocities slower than the threshold. Experimental results showcasing the measured and predicted friction, as well as the torque residual of Joint 3, are illustrated in Figures 4(b), 4(c) and 4(d), respectively. Figure 4(a) shows that Joint 3 remains within the presliding regime throughout the process, with the maximum deflection of the asperity junctions initially reached.

Figure 4 Results of external force estimate for Joint 3, (a) joint velocity (b) measured joint torque (c) predicted torque (d) torque residual (see online version for colours)



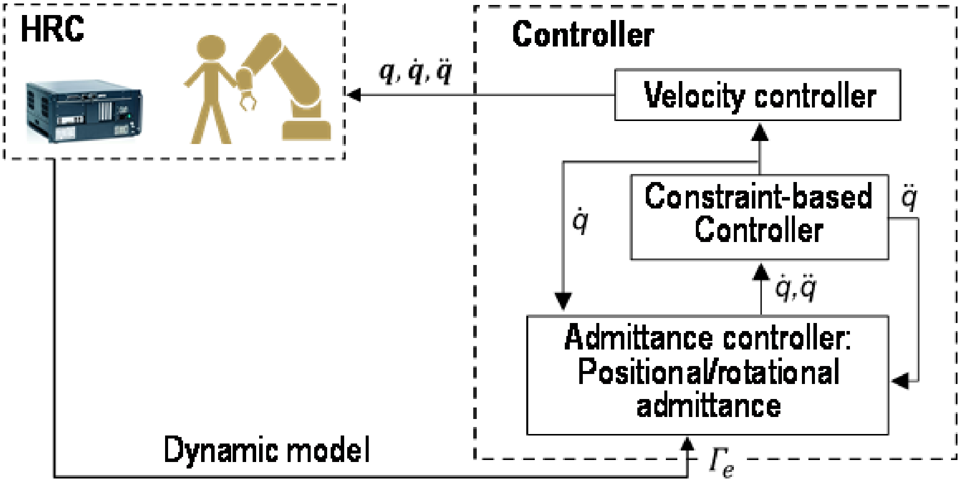
Source: Adapted and modified from Liu et al. (2021c)

Furthermore, the torque residual, as shown in Figure 4(d), indicates the application and detection of contact force on the robot at approximately 17 s and 24 s, denoted by Areas A and B, respectively. The black dash-dot lines represent the threshold of detectable external force, set at 2 Nm. Despite unmodelled parameters and backlash, the developed approach demonstrates commendable performance in detecting external forces.

2.3 Admittance-based compliant control

Admittance control (Liu et al., 2021c) converts the detected contact force into reference displacement and velocity, enabling human control over the robot through applied contact force, as shown in Figure 5. Admittance parameters with low values are utilised during collaborative assembly tasks to guide the robot toward targets swiftly. In contrast, higher values are employed for precise manipulation and accurate positioning at lower speeds. Within this framework, adaptive admittance control is implemented, allowing for online adjustment of admittance parameters to accommodate human preferences during assembly, thereby enhancing overall efficiency in robot control.

Figure 5 Admittance control for natural HRIs (see online version for colours)



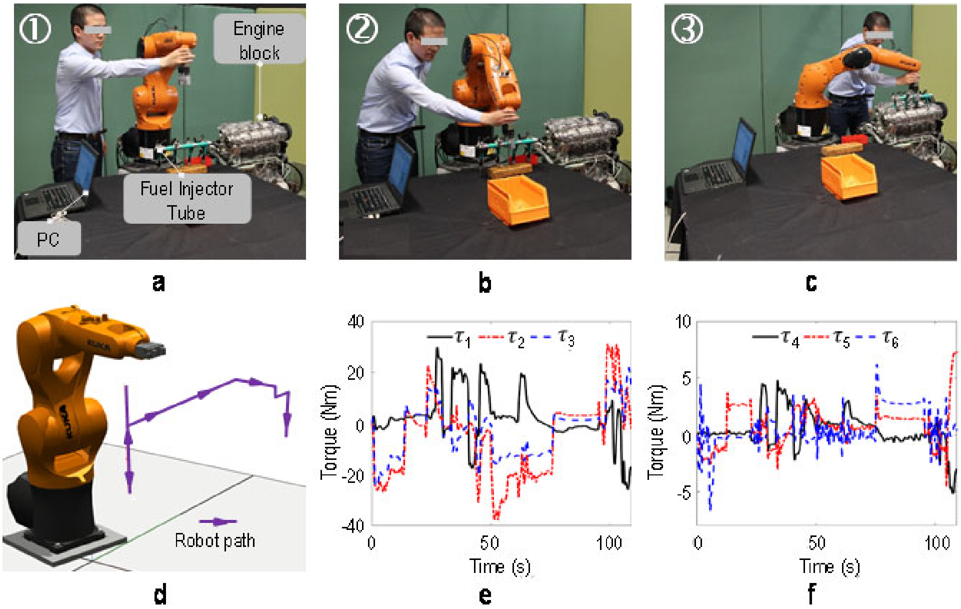
Source: Adapted and modified from Liu et al. (2021c)

To translate the contact force into precise robot movement, the Cartesian parts of the contact force are isolated to establish the robot's direction of movement. The admittance controller then adjusts the Cartesian coordinates of the robot end-effector, including its position and Euler angles. These state variables describe the Cartesian parts of the contact force and dictate the robot's motion along a designated axis. Regarding rotational manipulation, Euler angles signify the robot's orientation. The state variables governing rotational motion along the X-axis are defined, with similar mechanisms applicable to rotations about the Y and Z axes. Additionally, the first-order derivative of the Euler angle with respect to time determines the angular velocity of the robot's end-effector.

The assembly task involves inserting a fuel injection tube into four holes of a car engine's cylinder head, guided by a human operator, as shown in Figure 6(a). The assembly process begins by applying force to the robot, as shown in Figure 6(a). The contact force is then detected and converted into the robot's relative position and velocity using an admittance control policy, which guides the robot to move towards the fuel injector tube from its original configuration. Figure 6(b) shows that the robot gripper is fitted to grasp the fuel injector tube. The human operator then controls the robot to move to the engine block by applying force to the robot's wrist. Finally, the robot is controlled to adjust its orientation and then finish the insertion operation of the fuel injector tube, as shown in Figure 6(c). Figure 6(d) shows the recorded robot path indicated by purple lines when finishing the assembly task, and the arrows show the direction of the robot's motion.

The external torques calculated for Joints 1–3 and 4–6 are illustrated in Figure 6(e) and Figure 6(f), respectively, with an accuracy of 2 Nm, ensuring swift responses to human interactions. Subsequently, the adaptive admittance controller translates the detected contact force into precise motion control of the robot in Cartesian space. Notably, the torque variations and magnitudes observed in Joints 1–3 [Figure 6(e)] exceed those in Joints 4–6 [Figure 6(f)], attributed to Joints 1–3 primarily governing motion and orientation in Cartesian space, while Joints 4–6 focus on rotational aspects. Higher admittance parameters are employed in scenarios demanding precise motion control, such as inserting the cylinder head into the engine block. Conversely, lower values are chosen for swift movements. This approach facilitates closed-loop motion control, ensuring high accuracy by utilising the contact force applied by the human operator as input. Consequently, the human operator can dynamically influence robot motion by adjusting the direction and magnitude of the applied contact force, showcasing the operator's flexibility and adaptability in HRCA. Experimental results in Figures 6(e) and 6(f) demonstrate enhanced control performance, achieving robust compliance control capable of responding to minute external forces and torques with high precision.

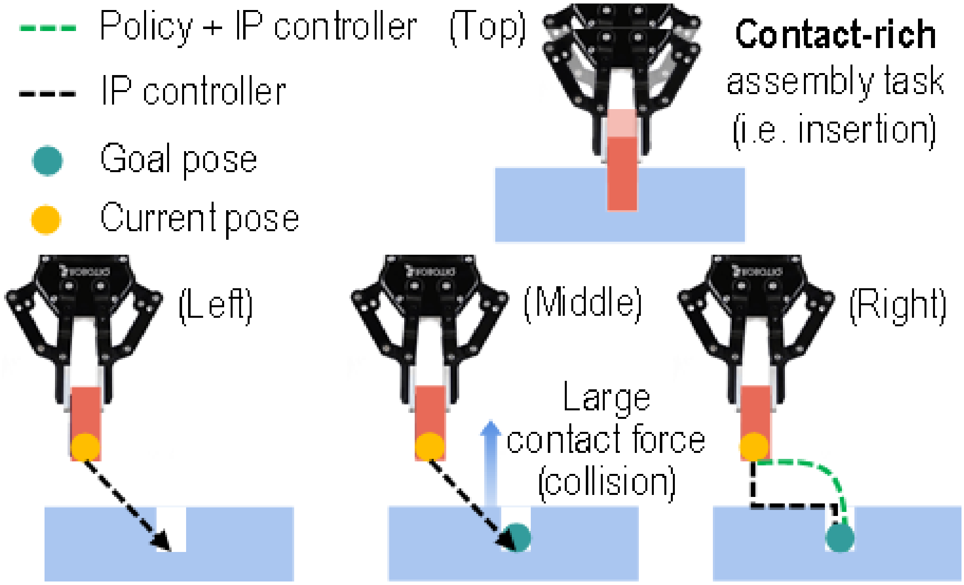
Figure 6 Experimental results, (a) experimental setup and control step 1 (b) control step 2 (c) control step 3 (d) admittance control-based robot path generated by contact force (e) contact force estimate for robot Joints 1–3 (f) contact force estimate for robot Joints 4–6 (see online version for colours)



Source: Adapted and modified from Liu et al. (2021a)

More recently, contact-rich robot manipulation has garnered significant attention, referring to robotic tasks that involve sustained or complex physical interactions between the robot, object, and environment, as illustrated in Figure 7 (Liu and Wang, 2025). Unlike simple pick-and-place actions, these tasks – such as insertion, assembly, or sliding – require the robot to precisely control force, torque, and contact dynamics throughout the manipulation process. Therefore, accurate force and torque observation is critical to the success of contact-rich manipulation tasks. Many research efforts on contact-rich manipulation conduct applications and demonstrations on collaborative robots with precise force feedback (Elguea-Aguinaco et al., 2023). For example, the use of reinforcement learning (RL) and contact force feedback in contact-rich manipulation tasks for collaborative and precision assembly was investigated (Luo et al., 2024; Liu and Wang, 2025).

Figure 7 Contact-rich manipulation task supported by RL and force feedback (see online version for colours)



Source: Adopted from Liu and Wang (2025)

3 Natural language-driven HRIs

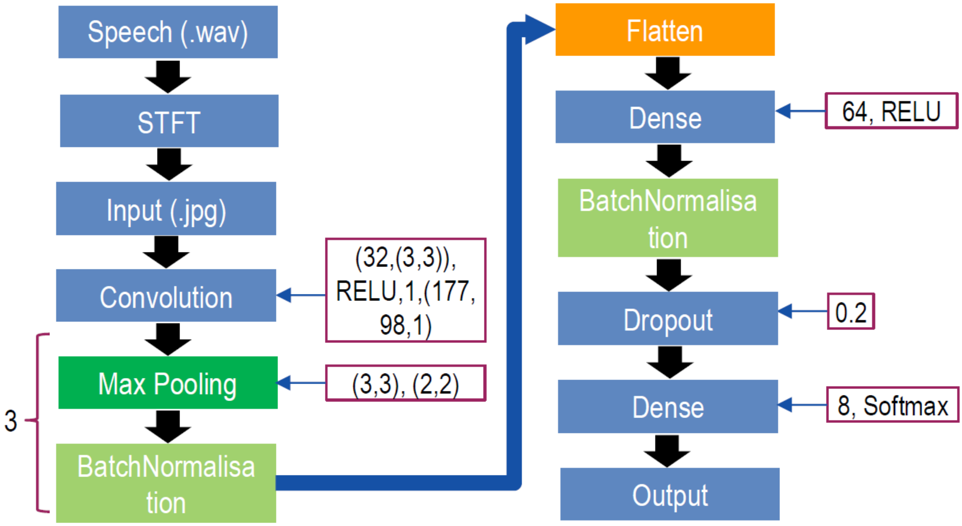
This section presents a verbal and gesture-based methodology for HRC in assembly tasks. It commences with the elucidation of signal gathering and the translation of both voice and gesture commands, followed by converting these commands into valid instructions for the robot.

3.1 Deep CNN-based voice recognition

Human voice commands are employed to oversee robot operations within an assembly environment. Given the tasks involving gripper control in robotic assembly, extensive programming is typically required, particularly when adapting to changes in the assembly environment. Hand gestures work as a nonverbal and context-dependent means of communication, enabling seamless collaboration between humans and robots, facilitating gripper control. This integration enhances flexibility and efficiency during collaborative assembly processes.

A deep CNN-based voice recognition scheme is illustrated in Figure 8. Initially, the audio data undergoes conversion into 2D spectrograms in the frequency and time domains via the short-time Fourier transformation (Liu et al., 2020). Subsequently, these generated spectrograms serve as inputs for the deep CNN model. The CNN architecture comprises convolution, max-pooling, dense, and flattened operations, culminating in a fully connected layer that outputs the final results. Within the convolution layer, a 3 × 3 image matrix filter slides over the input map to produce a feature map. This configuration utilises 32 filters with a stride length of 1 and an input shape of 177 × 98 × 1. Max-pooling operations with a 3 × 3 feature map and 2 × 2 strides then compute the maximum value of each local neighbourhood. The fully connected layer facilitates the classification of input images using a softmax function with six classes (‘*left*’, ‘*right*’, ‘*up*’, ‘*down*’, ‘*forward*’, and ‘*backwards*’). A dropout parameter value of 0.2 is utilised. The training model employs a categorical cross-entropy-based loss function, utilising a training dataset for speech recognition curated by the AIY and TensorFlow teams (Warden, 2018). This dataset comprises 18,440 audio files corresponding to the six selected voice commands. Finally, the classified results of the voice dataset are generated, achieving a classification accuracy of 92.47% on the test dataset.

Figure 8 Deep CNN-based voice recognition for robot motion control (see online version for colours)



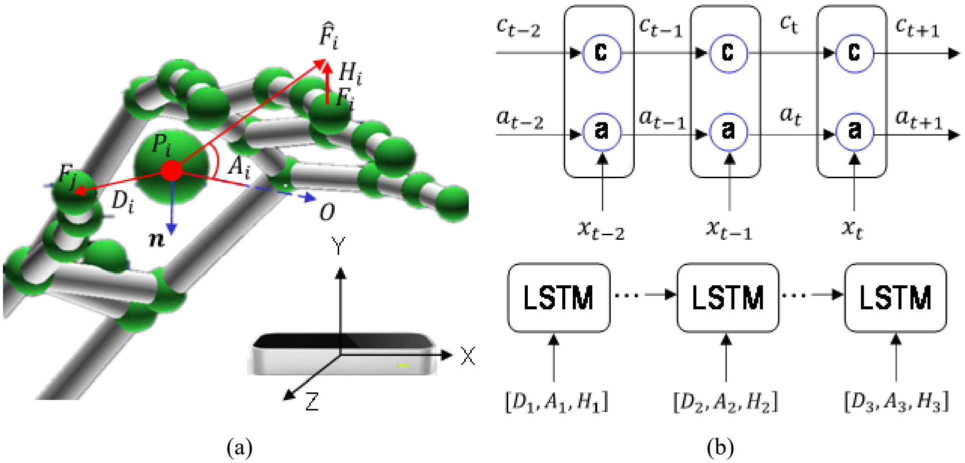
Source: Adapted and modified from Liu et al. (2021a)

Hand gestures, a nonverbal communication channel, are utilised alongside speech commands for robot control. Gesture-based manipulation of the robot gripper enhances the efficiency of robotic assembly operations. This study utilises the Leap Motion sensor to capture hand gestures and motion data, enabling the real-time representation of human hands. In the Leap Motion sensor coordinate system, depicted in Figure 9(a), hand gestures are characterised by three components in 3D space: the position of the fingertip, the palm centre, and the hand orientation. Hand orientation is determined by components parallel to the finger and perpendicular to the palm plane. Subsequently, a set of feature vectors, comprising fingertip distance, angle, and height, is extracted from sensor data.

These feature vectors serve as input for an long-short-term memory (LSTM)-based training model, facilitating gesture classification.

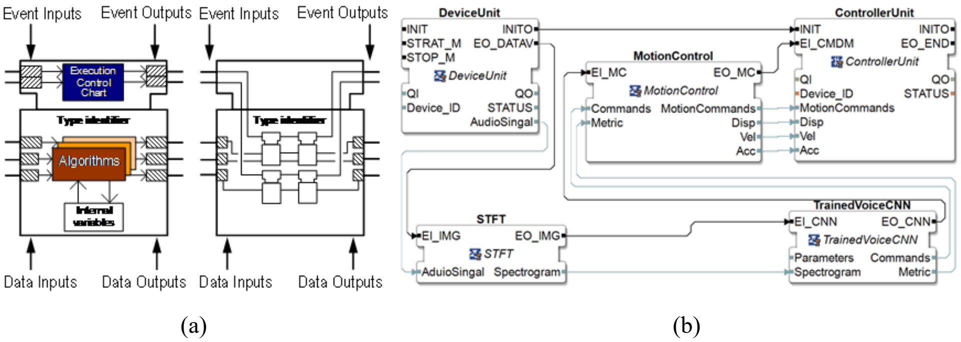
A dataset comprising 1,200 hand gesture samples is utilised for training and recognition. The efficacy of LSTM in handling time-based sequences significantly enhances the processing of hand gestures, as demonstrated in Figure 9(b). The LSTM model effectively retains and updates long-term information by adjusting its gates. In the LSTM model, the input and output sizes of the LSTM unit, with a stride length of 1, are 30 and 32, respectively, followed by the outputs of a vector with six dimensions. The accuracy achieved on the testing dataset is 95.83%.

Figure 9 (a) Leap Motion-based hand gesture model (b) LSTM-based gesture recognition (see online version for colours)



Source: Adapted and modified from Liu et al. (2021a)

Figure 10 An example of, (a) a CFB (b) a FB network for voice recognition (see online version for colours)



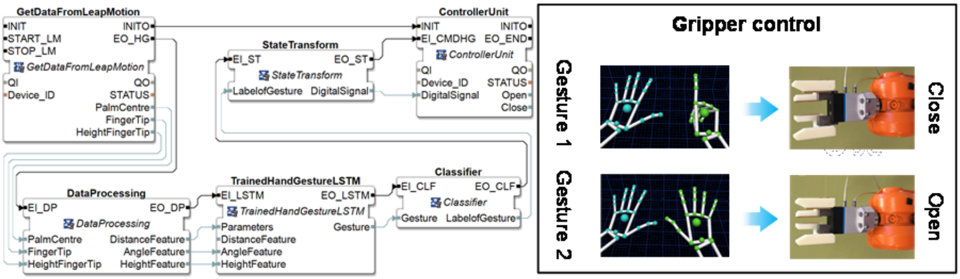
Source: Adapted and modified from Liu et al. (2021a)

Hand gestures and voice commands are translated into true robot commands to enable robot control in assembly tasks. FBs are commonly employed in robot control systems to facilitate this translation process. Figure 10(a) illustrates a composite FB (CFB), featuring data/event input/output channels and internal behaviour. Within this structure, the element known as the execution control chart (ECC) functions as a finite state machine, facilitating state transitions, algorithm execution, and control outputs.

FB networks, comprising a series of basic FBs (BFBs), are defined as leveraging voice and hand gestures for robot control. As demonstrated in Figure 10(b), an FB network dedicated to voice-based robot control is established, enabling the conversion of voice commands into valid robot control instructions. This network facilitates robot movement along specified directions through a predefined set of labelled voice commands.

The workflow of the FB network begins with the collection of audio signals through the FB (SIFB) DeviceUnit service interface. Subsequently, the audio data is processed into frequency spectrograms using the *short-time Fourier transform* (*STFT*), serving as inputs for the training model. Upon receiving the input event EI CNN, the *TrainedVoiceCNN*, equipped with the trained model, is triggered to classify commands. The detailed training parameters and algorithms were previously discussed, and the results of the classified voice commands were output via the commands channel. The *MotionControl* BFB translates human instructions into robot motion control commands to govern robot movement. Considering motion within the Cartesian space, six speech commands are utilised: '*forward*', '*left*', and '*up*' dictate motion along the positive X, Y, and Z axes, respectively, while '*backward*', '*right*', and '*down*' correspond to motion in the opposite directions. The motion control commands and parameters, including displacement (*Disp*), velocity (*Vel*), and acceleration (*Acc*), are transmitted as data output variables through *MotionCommands* to an SIFB (*ControllerUnit*) for command execution.

Figure 11 (a) FB network of hand gesture-based gripper control (b) Gestures used for robot gripper control (see online version for colours)



(a) (b)

Source: Adapted and modified from Liu et al. (2020, 2021a)

As shown in Figure 11(a), an FB network dedicated to hand gesture-based gripper control is depicted. FBs such as *GetDataFromLeapMotion* serve as intermediary communication channels, facilitating the retrieval of hand gesture data from the Leap Motion sensor. The gathered data undergoes processing through the BFB *DataProcessing*, where feature vectors corresponding to hand gestures are extracted. Specifically, the *DistanceFeature*, *AngleFeature*, and *HeightFeature* modules yield output representing respective aspects of the hand gestures.

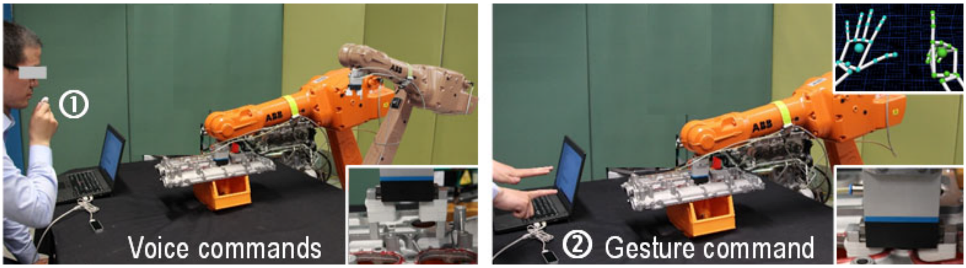
These extracted features are inputs for the trained LSTM model embedded within the *TrainedHandGestureLSTM* BFB. Comprehensive details regarding the algorithms and training parameters are elaborated above. Upon receiving the input event, the *TrainedHandGestureLSTM* BFB is activated to classify the hand gestures and produce corresponding gesture types as output.

Subsequently, the *Classifier* BFB transforms the identified gestures into predefined labels, which are emitted as output via the *LabelofGesture* channel. Since gripper control involves digital signals, the *StateTransform* BFB generates digital output signals. These signals, representing the commands to open and close the grippers, are defined as the output of the *Open* and *Close* channels within the *ControlUnit* BFB, respectively, and are ultimately executed by the robot controller.

3.2 Experiment

Figure 12 illustrates the experimental configuration for voice and gesture-based robot control. In Figure 12(a), the experiment begins with voice command-controlled robot motion. A series of voice commands, including '*left*', '*right*', '*up*', '*down*', '*forward*', and '*backwards*', are issued to control the robot's movement. Following activation by the start FB to commence assembly as shown in Figure 11(b), these voice commands are transformed into spectrograms. Robot motion is initiated by executing the '*forward*' voice command. The *MotionControl* BFB is activated to generate a set of motion control commands upon detecting the input event, indicating the identified command. Each command comprises a label corresponding to the voice command, along with the relative displacement (in mm) and TCP (tool point centre) linear velocity (in mm/s) parameters for the robot. The robot controller executes these robot control commands for assembly operations through the controller *Unit* (SIFB).

Figure 12 (a) Experiential setup (b) Results of voice-gesture based on robot control in assembly (see online version for colours)



Source: adapted and modified from Liu et al. (2021a)

Figure 12(b) displays the outcomes of hand gesture-based gripper control. Here, the human operator executes hand gestures above the Leap Motion sensor, and the collected raw gesture data undergoes feature extraction through the SIFB. The resulting feature vectors, represented as $V = [D, A, H]$, are input into the *TrainedHandGestureLSTM* BFB for the classification and identification of hand gestures, triggered by the event, as shown in Figure 10(a). The identified hand gestures are emitted as output via the *Gesture* channel, which is then translated into predefined gesture labels.

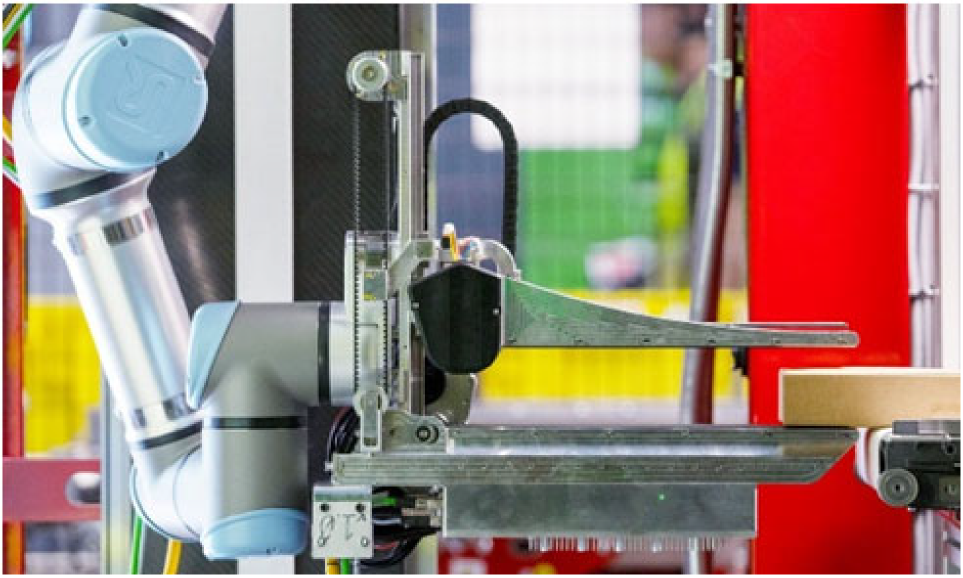
As depicted in the top and bottom insets of Figure 12, the hand gestures are labelled as '*Gesture 1*' and '*Gesture 2*', corresponding to closing and opening the gripper, respectively. These gesture labels are converted into a state variable using the

StateTransform BFB, generating the digital output signals for gripper control. The results of gripper control are displayed on the right side of the subfigure.

In recent years, human touch gestures have been used as intuitive communicative signs for effective physical interactions between humans and robots. For example, five representative types of physical touches with robot links were defined and trained using DL algorithms for intuitive interactions (Jung et al., 2025). In addition, the semantic-pose to motion model translates dynamic hand gestures by analysing their semantic meaning and spatial configuration. It enables real-time control of complex robotic systems, such as quadruped robots equipped with robotic arms, by translating 3D hand skeleton data into precise mechanical actions (Delmas et al., 2025; Xie et al., 2025). By incorporating spatial and temporal components, this DL approach facilitates more intuitive non-verbal communication between humans and robots (Roy et al., 2024).

Furthermore, integrating LLMs with gesture recognition systems has led to the development of GestLLM (Kobzarev et al., 2025). It leverages advanced feature extraction techniques to interpret a wide array of hand gestures, including those not commonly found in traditional datasets, thereby enhancing the inclusivity and flexibility of HRIs. More recently, tactile technologies demonstrate promising capabilities in improving interactive robot manipulation. Specifically, a notable recent application of tactile technology in robotic manipulation is Amazon's deployment of the Vulcan robot in its fulfilment centres (Park et al., 2025), as shown in Figure 13. Vulcan is equipped with advanced touch-sensing capabilities that enable it to identify and retrieve specific products from storage shelves, and it also provides enormous potential for reshaping conventional robot interactions.

Figure 13 Vulcan: Amazon's first robot with a sense of touch (see online version for colours)

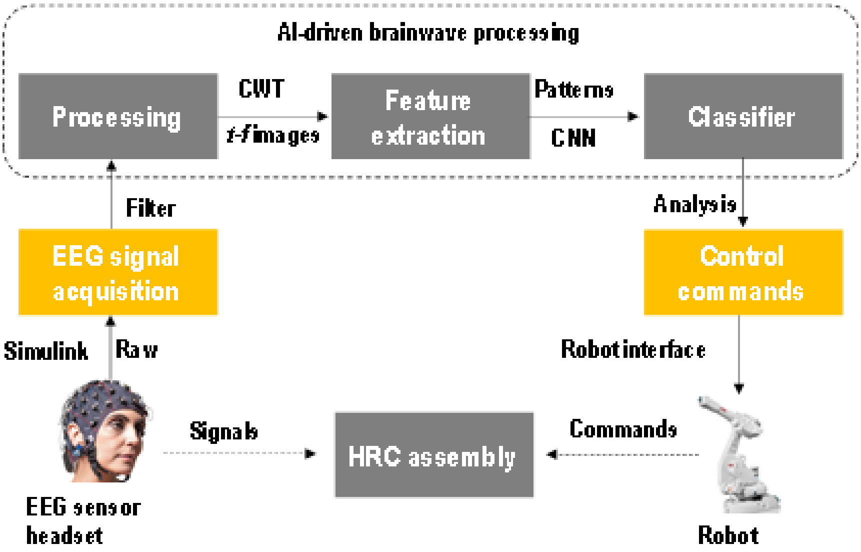


Source: Adapted from Park et al. (2025)

4 Leveraging neural signals for robot control

This section introduces a novel control approach to HRCA, facilitated by FBs and guided by brainwaves, illustrated in Figure 14. Within this context, robot control engaged in collaborative assembly is facilitated by FBs, utilising brainwaves as inputs for macro control commands. This section explores a stimulus-free EEG signal acquisition method that mitigates the limitations of stimulus-based EEG approaches. Subsequently, an FB-based HRCA strategy is introduced. Following this, a DL-driven classification system of the brainwaves is developed to achieve robot control with high accuracy and low latency. This system employs a wavelet transform for feature extraction, generating a feature matrix used as input. A robust translation of brainwave commands into robot control commands is then implemented to enhance collaborative assembly efforts.

Figure 14 Concept and pipeline of brain robotics in assembly (see online version for colours)

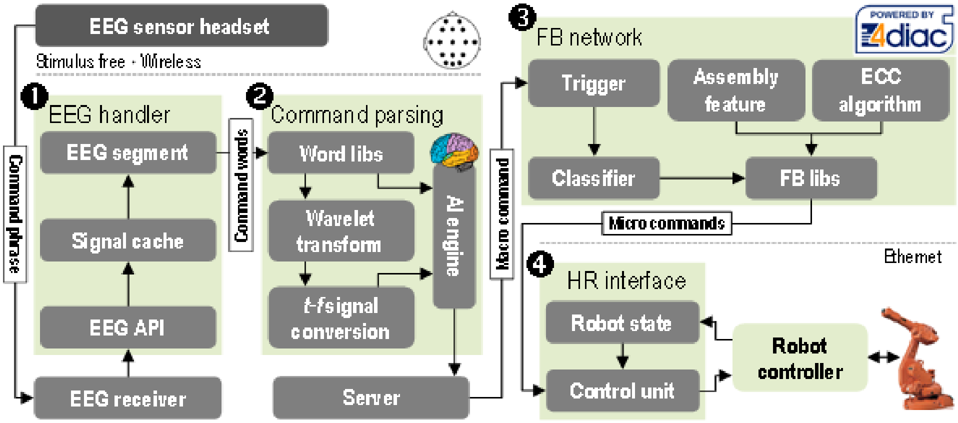


Source: Adapted and modified from Liu et al. (2021e)

As shown in Figure 14, the EEG signal processing workflow in HRCA commences with the collection of EEG signals from a sensor headset. Subsequently, the signals undergo processing through filters and artefact removal modules. After eliminating noise, the signals maintain consistent amplitude over time, yet the specific brainwave frequencies remain unidentifiable. A continuous wavelet transform (CWT) is applied to uncover frequency characteristics, generating time-frequency (*t-f*) visual representations. These *t-f* images are then processed by a CNN model to detect distinct brainwave patterns. A dedicated evaluation framework is introduced to measure the accuracy of the classification model. Based on this, a method is implemented within FBs to translate CNN outputs into robot control commands. They are finally sent to the robot controller through an interfacing system for execution.

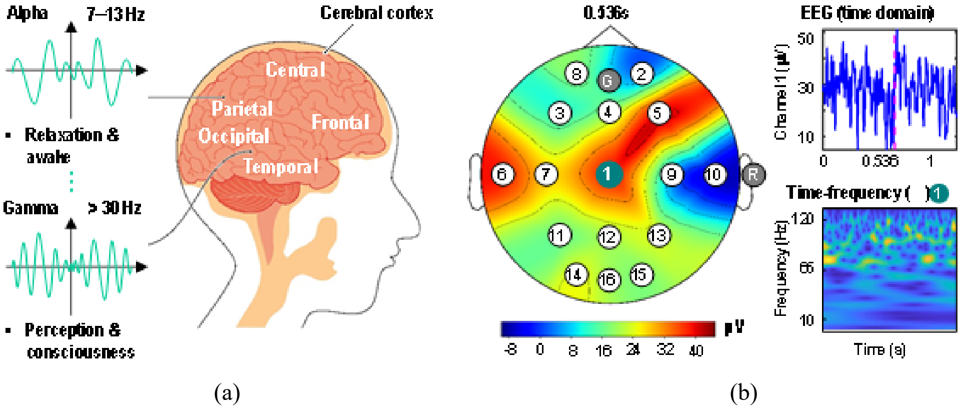
The architecture of the built system comprises four modules designed for HRCA, as depicted in Figure 15. Brainwaves serve as control input commands to the FBs, providing micro commands for adaptive robot control and collaborative assembly. Module 1 starts with brainwave collection from an EEG sensor headset and segments the brainwaves of a command phrase, removing background noise and isolating command words. Module 2 utilises the wavelet transform to convert brainwaves into feature inputs for a CNN, enabling high-accuracy command classification. Module 3 constitutes an FB network embedded with control algorithms activated by the classified commands. AF-based FBs are defined within this module and employed for collaborative assembly, converting brainwaves into robot control commands sent to module 4 for execution via a humanrobot interface.

Figure 15 System design of brainwave-driven HRCA (see online version for colours)



Source: Adapted and modified from Liu et al. (2021e)

Figure 16 (a) Structure and waves of the brain (b) Visualisation of brainwaves in terms of time, time-frequency domain and topographic map (see online version for colours)

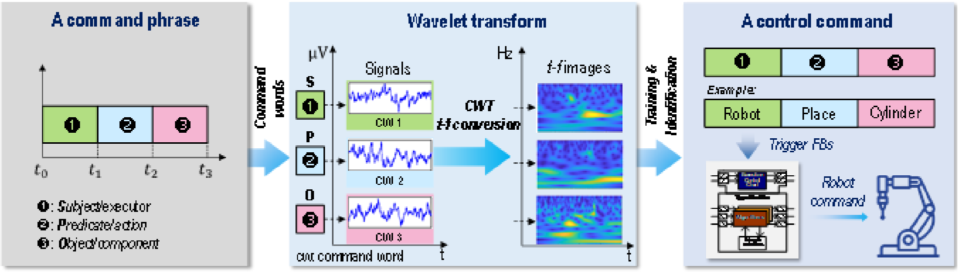


Source: Adapted and modified from Wang et al. (2021)

The activity of neurons, EEG power topography, and brain structure activation are depicted in Figure 16. Figure 16(a) shows the brain's structure, functions, and waves. The left-side sub-figure of Figure 16(b) represents a topographic map at 0.536 s with 16 channels and a reference channel (R). Blue and red areas indicate fewer neurons firing, and the highest neuron activity, respectively. The right-side sub-figure (top) illustrates the time-domain brainwaves for Channel 1, and the bottom displays the *t-f* image of the signals. The *t-f* features of the brain signals indicated by bright yellow spots in the image can be observed.

During EEG signal recording, an operator issues a command phrase consisting of a subject, predicate, and object, such as ‘robot place block’. Each command word in the phrase denotes an executor, an action, and a part to be acted upon, respectively, as illustrated in the ‘A command phrase’ of Figure 17. The invoked electrical potential of the cerebral cortex is measured as EEG signals. Simultaneously, the human operator maintains a rhythmic thought pattern to issue command words. Specifically, the subject, predicate, and object are contemplated for one second each, with a one-second pause in between. Subsequently, the EEG signals undergo filtering using online band-pass and notch filters with a frequency range of 0.1–100 Hz and 48–52 Hz, respectively. The former captures signals within the desired frequency range, and the latter eliminates electrical noise.

Figure 17 A pipeline of brainwave to robot control commands (see online version for colours)



Following an extensive investigation into feature extraction and performance evaluation, as discussed in Liu et al. (2024a), the capability of wavelet transform to unveil characteristics inherent in non-stationary signals has proven effective in extracting signal features related to brainwave commands. The wavelet transform of a signal $x(t)$ is formulated as

$$wt(\tau, s, x, \Psi) = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} x(t) \Psi^* \left(\frac{t - \tau}{s} \right) dt : s \in \mathbb{R}^+, \tau \in \mathbb{R} \tag{7}$$

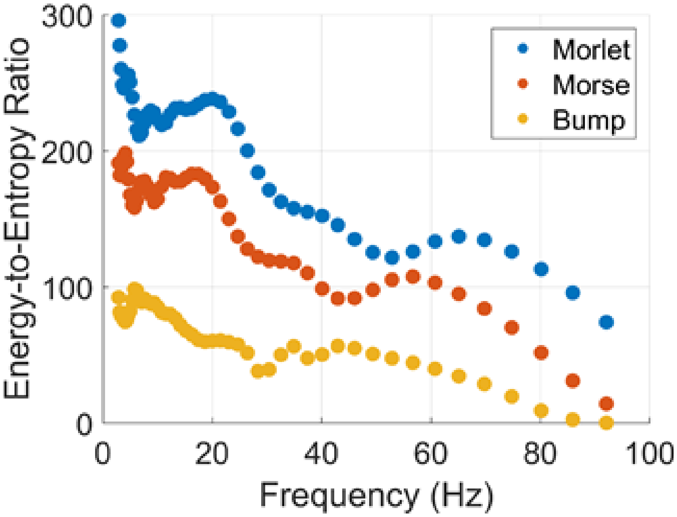
where wt is the wavelet coefficient. s and τ are the scaling and shifting parameters, respectively. Ψ and Ψ^* denote the base wavelet and its complex conjugate, respectively. The wavelet transform can extract signal features over the entire signal spectrum by adjusting s and τ. A set of commonly used base wavelets are adopted in the wavelet transform for the feature extraction, and they are the base wavelets of ‘*B-Spline*’, ‘*Bump*’, ‘*Gaussian*’, ‘*Harmonic*’, ‘*Morlet*’, ‘*Morse*’, and ‘*Shannon*’.

The energy-to-Shannon entropy ratio of the EEG signals serves as a criterion for selecting the most suitable base wavelet. The energy-to-entropy ratio (R) is described as

$$R(s) = \sum_{i=1}^{N} |wt(\tau, s)|^2 \Big/ \left(-\sum_{i=1}^{N} p_i / \log_2 p_i \right) \tag{8}$$

The values of R represent the amount of relevant information within the brain signals, and the appropriate selection of base wavelets can be determined by maximising R across a set of candidates. Using 12,800 EEG signals from 16 channels and involving nine command words, the respective mean values $\overline{R}$ of R are calculated. The mean values for the '*B-Spline*', '*Bump*', '*Gaussian*', '*Harmonic*', '*Morse*', '*Shannon*', and '*Morlet*' base wavelets are 80, 66, 73, 85, 130, 149, and 172, respectively. Figure 18 illustrates the energy-to-entropy ratio of Morlet, Morse, and Bump wavelets across frequencies. Consequently, the Morlet base wavelet is selected due to its highest R value among the candidates.

Figure 18 Energy-to-entropy ratio of the EEG signals for different wavelets (see online version for colours)



The effectiveness of CNNs in discerning patterns within images has rendered them adept at classifying *t-f* images related to robot commands. In the current study, three criteria guide the selection of a CNN:

1 classification accuracy

2 training time

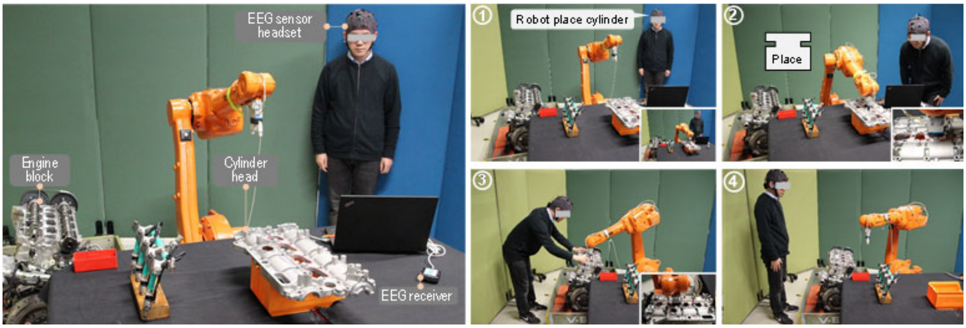
3 computational efficiency.

To meet these criteria, a comparative analysis is conducted on 14 pre-trained CNNs detailed in Canziani et al. (2016). The original training of these networks focused on classifying natural images. To adapt them for the current task, they are fine-tuned using time-frequency representations associated with robotic control signals. Consequently, depending on their convolutional or fully connected nature, the selected networks (14) are trained using EEG signals, either as images or vectorised representations. Ultimately, VGG16 (Simonyan and Zisserman, 2014) is selected due to its high accuracy (97% across all sensor data) and efficiency (convergence in 150 epochs) in classifying EEG signals. Convergence is less than a 0.5% improvement in training accuracy over 10 consecutive epochs.

The FB network is employed here to translate EEG signals into robot control commands, as depicted in ‘a robot command’ in Figure 15. All FBs operate within an IEC 61499 runtime environment (4Diac FORTE). When triggered by an EEG event, a command phrase is parsed into command words, including subjects, predicates, and objects. Specifically, these are ‘robot’, ‘place’, and ‘cylinder’, respectively, with their respective values assigned to the output variables of the robot, actions, and parts. The predicates’ content correlates with predefined AF and serves as a trigger event for AFbased FBs in assembly operations. Consequently, the ‘*Place*’ output events activate the AF-based FBs of the place AF FB. Here, assembly reference and constraints, including part information and constraints, function as the control input for assembly and motion planning. Ultimately, robot control commands, comprising a series of gripper and motion instructions, are dispatched for task execution.

In this section, a series of assembly scenarios are devised to assess the efficacy of the HRCA system developed. Figure 19(a) illustrates the experimental arrangement, which links an industrial robot controller to a receiver for EEG signal transmission via Ethernet and USB connections. The assembly tasks involve positioning a cylinder head onto an engine block.

Figure 19 (a) Experimental setup (b) Examples of brainwave-driven assembly (see online version for colours)



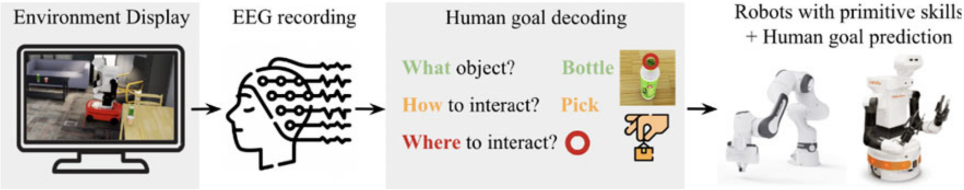
(a) (b)

Source: Adapted and modified from Wang et al. (2021)

As depicted in Figure 19, the assembly procedure comprises four sequential control steps. Step 1 initiates with a brainwave command, ‘robot place cylinder’. The SIFB (SEVER) facilitates the transmission of recorded EEG signals to the classifier, segmenting the command phrases into subject, predicate, and object. The resultant feature matrix serves as input for the trained model, yielding classified results. These context-based commands – ‘robot’, ‘place’, and ‘cylinder’ – activate predefined FBs via their predicate components. The subject and object values, containing executor (robot) and part (cylinder) information, dictate the cylinder’s position, the engine block’s location, and the cylinder’s grasping point in a robotic coordinate system. Step 2 is triggered by the ‘*Place*’ event, engaging the ‘*Place_AF*’ FB. Within this FB, gripper and motion control algorithms generate valid control commands, executed on the robot controller via the *CLIENT* SIFB. Following these commands, the robot manoeuvres to and securely grasps the cylinder head with millimetre precision, as depicted in the inset. Step 3 controls the robot to place the cylinder head on the engine block, with the human operator working in tandem to fine-tune the cylinder's position and assess assembly quality, as illustrated in the inset. Finally, the human operator secures the components upon the robot's return to its initial position in Step 4.

In addition to a comprehensive review of brain-robot interaction systems using EEG signals (Zhang et al., 2024), Liu et al. (2024a) also provide a thorough review of the past and current status of EEG signal use for robot control. It describes wide applications of EEG-controlled robotics, including industrial settings, mobile assistants, exoskeleton robotics, and human intent detection. More recently, numerous research efforts on EEG-based robotic applications have been actively explored, including the use of EEG signals for robot teleoperation (Zhang et al., 2025b), mobile robot obstacle avoidance using EEG signals (Omer et al., 2025), and multi-brain to multi-robot interactions (Ouyang et al., 2024). As shown in Figure 20, neural signal operated intelligent robot (NOIR) is a brain-robot interface system developed by researchers at Stanford University (Zhang et al., 2023b). It enables people to direct robots to perform everyday tasks using brain signals, such as moving objects, cleaning countertops, playing tic-tac-toe, petting a robot dog, and even cooking a simple meal. The system comprises a modular EEG signal decoding pipeline and a module linking the signals with a set of robotic skills. The robots can learn to predict human intended goals, reducing the human effort required for decoding.

Figure 20 A NOIR system to perform everyday activities through brain signals (see online version for colours)



Source: Adapted from Zhang et al. (2023b)

As pointed out in Li et al. (2025) and Liu et al. (2024a), EEG-based control systems hold promise for human-machine interaction. However, their readiness for deployment in real-world industrial environments remains limited. EEG signals are highly susceptible to noise from electrical interference and user movement, which can significantly degrade signal quality in high-paced, noisy settings. Additionally, the requirement for precise electrode placement, often involving wet sensors, poses practical challenges for routine use outside controlled laboratory conditions. Real-time processing constraints, user fatigue, and inter-individual variability in EEG patterns further complicate reliable implementation. Thus, while the technology advances, substantial work remains to improve robustness, usability, and scalability before EEG-based control can be reliably deployed in industrial contexts.

5 DT-based HRCA

This section commences with a discussion on the concept of DTs, elucidating the implications pertinent to HRCA environments. Subsequently, a comprehensive review of

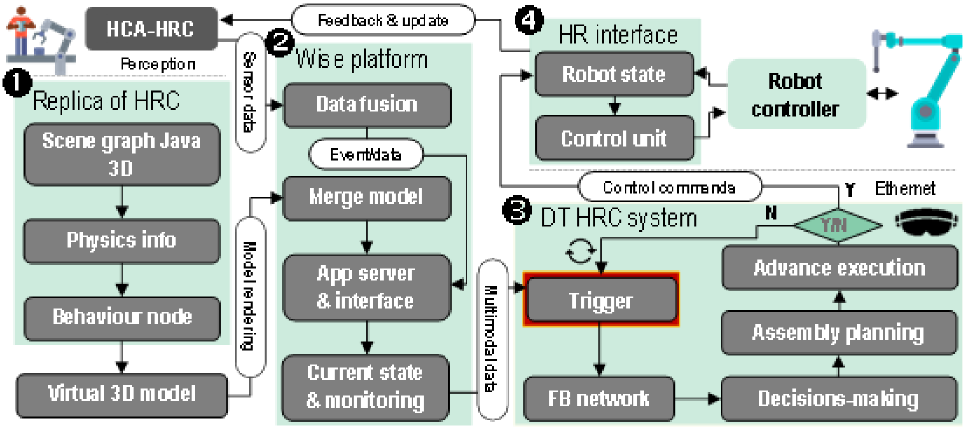
DT-based HRCA is presented, encompassing aspects such as system architecture, AI, and HRI. This analysis establishes the groundwork for investigating prospective developments and identifying emerging trends within the HRCA domain.

5.1 DT in HRCA

Grieves (2014) initially proposed the concept of DT in his 2003 lecture on product lifecycle management. The DT consists of three core elements: the tangible object, its corresponding digital model, and a two-way data flow that continuously synchronises information between them. Over the subsequent two decades, the DT concept has undergone a process of continuous evolution and expansion. For instance, this framework was extended to a 5D model of the DT, which encompasses physical and virtual components, data, and services (Tao et al., 2018). In contemporary manufacturing, which aims to advance highly automated processes, the DT is anticipated to assume a critical role, particularly in HRC (Ramasubramanian et al., 2022). The assembly task in HRC poses a common and complex problem within modern manufacturing settings. Compared to fully automated robotic systems, HRC systems are more adept at adapting to fluctuating production demands and managing the assembly tasks that require customisation.

Nevertheless, the simultaneous execution of intricate assembly operations for humans and robots within a shared operational area, without rigid physical boundaries, necessitates implementing advanced cognitive capabilities, interactive engagement, safety measures, and adaptability. One method involves creating detailed virtual simulation environments that can accurately capture the characteristics of the real-world system (Kritzinger et al., 2018). Consequently, the DT, renowned for its capacity to accurately mirror a physical system in real-time, is becoming increasingly prevalent in HRCA scenarios (Liu et al., 2025), as shown in Figure 21.

Figure 21 System design of the DT-driven HRCA (see online version for colours)

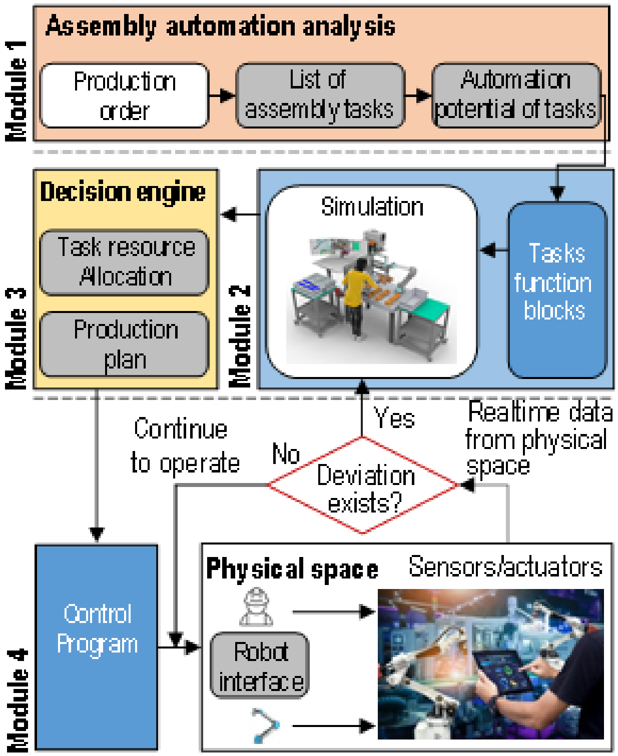


Source: Adapted from Liu et al. (2022b)

5.2 System architecture

The architectural designs of DT-based HRCA systems exhibit considerable diversity, especially in Industry 4.0. These architectures are characterised by integrating the real-time monitoring capabilities of DTs with the efficiency of robots, all of which are supported by human operators. Consequently, the DT-based HRCA system architecture design generally considers hardware and software configuration, HRIs and data management. The physical hardware configuration initially encompasses the robotic components and associated sensors. Another critical aspect is the software configuration, which is geared towards developing high-fidelity DT models for seamless integration with the physical system. Furthermore, it is imperative to ensure bidirectional communication between the robotic system and its DT.

Figure 22 An event-driven simulation-based DT of the HRCA system (see online version for colours)



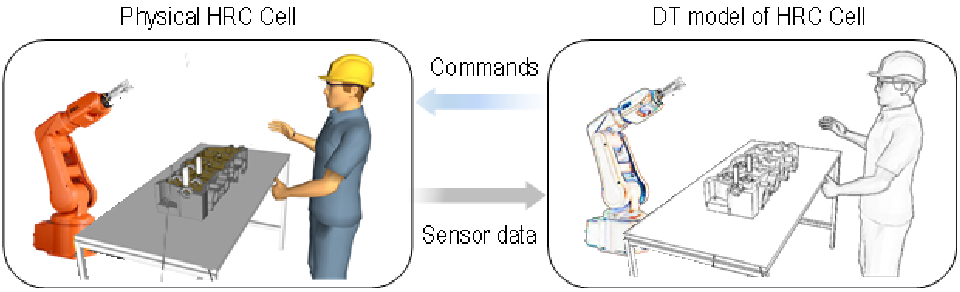
Source: Adapted and modified from Bilberg and Malik (2019)

In contrast, HRIs facilitate safe, efficient, and ergonomic operations through intuitive interfaces that promote a smooth collaborative environment. Data management strategies are also paramount for successfully operating a DT-based HRCA system. It is essential to maintain an adequate data flow between the physical and virtual components of the system to preserve its integrity and performance. This process involves not only the collection and analysis of real-time data to enhance system insight and intelligence but also the protection of data security and privacy.

The modularity of these architectures stems from the inherent complexities of modern manufacturing tasks. The modular architecture of the DT-driven HRCA, designed to facilitate human-centric assembly (HCA) (Liu et al., 2022b), is illustrated in Figure 18. This system comprises four modules: a replica of HRC, a platform for data fusion and state monitoring, a DT system, and a human-robot interface. Such a configuration promotes the seamless integration of data and interaction, enabling both humans and robots to perform collaboratively with high efficiency and adaptability in real-time environments. As another example, an event-driven, simulation-based decision-making system was proposed and structured into four modules, as shown in Figure 22: assembly analysis, simulation, decision, and control (Bilberg and Malik, 2019). Additionally, a DT approach designed to enhance the reconfiguration of HRCA lines was presented (Kousi et al., 2021), with a primary emphasis placed on the flexibility of assembly systems.

Furthermore, certain architectural frameworks are designed to address specific products. A cyber-physical system (CPS) framework for assembling the interior of an aircraft fuselage was proposed by Franceschi et al. The framework comprises three principal components: a physical system featuring a robot and a human operator, a virtual system equipped with a DT and a Dashboard for real-time status updates, and a Mainframe for task and data management, enabling flexible and modular assembly. A DT-based HRC framework in the construction field, which incorporates design, learning by demonstration, and robot control modules, was investigated to address the assembly of a wooden structure, an interlocking wooden joint (Kramberger et al., 2022). As shown in Figure 23, a digital representation of a physical HRC workcell for assembly tasks was presented to build a DT-driven collaborative system, promoting flexibility and adaptability through real-time monitoring and operations (Liu et al., 2025).

Figure 23 Digital representation of an HRCA system (see online version for colours)



Source: Adapted and modified from Liu et al. (2025)

5.3 *Artificial intelligence*

The concept of HRCA capitalises on the convergence of human cognitive skills and robotic precision to enhance manufacturing procedures. Recent AI advancements, particularly DT deployment, have significantly transformed HRCA systems. These advancements are characterised by AI technologies, including machine learning (ML), DL, and RL. They play a critical role in refining the capabilities of DT in HRCA systems. ML technology enables DTs to learn from extensive operational data. DL, a subset of ML, is recognised for its proficiency in managing complex data structures for detecting patterns and anomalies within the manufacturing process (Tuli et al., 2021; Zhang et al., 2023a). Robots are empowered by RL to make decisions based on real-time feedback, optimising their actions and improving their adaptability to new and evolving environments (Lv et al., 2021). An AI-driven HRCA system that simulates the movement of robots as well as humans through DT and utilises a CNN to detect the positions of assembly parts and human workers was proposed (Dimitropoulos et al., 2021). Ergonomics is prioritised with this system, thereby ensuring a safer and more efficient collaboration. Additionally, a heuristic search algorithm is employed to dynamically adjust the robot's behaviour in response to human actions.

Uncertainty in human behaviour is more prevalent than robotic motion due to humans' greater autonomy and spontaneity (Zheng et al., 2023). Therefore, it is crucial to recognise and predict human actions during assembly tasks. Building on this premise, a hybrid AI model that integrates a spatial-temporal GCN with a one-dimensional CNN was developed (Gao et al., 2023). This model employs a combination of iterative closest point and PointNet algorithms, alongside a pixel-wise voting network, to accurately identify the types of parts and their poses (Zhang et al., 2023a). Such advancements significantly reduce assembly failures, improve safety, and minimise the risk of human errors in HRCA setups. Moreover, a DL-based strategy to overcome visual occlusions in HRCA environments was proposed (Zhang et al., 2024). This occlusion-robust mesh recovery algorithm aids in reconstructing occluded human bodies, thereby enabling the precise creation of a human DT and the planning of robot trajectories that seamlessly adapt to human movements without compromising safety or efficiency.

Compared to manual assembly or pre-programmed assembly methods, one notable advantage of DT-based HRCA is the use of RL models. These models optimise the assembly system by providing optimal action sequences and improving its learning capabilities (Lv et al., 2021). A deep deterministic policy gradient (DDPG) technique was utilised in Sun et al. (2022) to adaptively optimise robot motion paths, explicitly targeting the assembly of complex products. Furthermore, a double-DDPG model was introduced, which demonstrably enhances the efficiency of task allocation and reduces the workload associated with manual operations in HRCA systems (Lv et al., 2021).

5.4 Interaction technology

In DT-based HRCA scenarios, HRI is a pivotal process that significantly enhances efficiency and safety. The integration of HRI technologies, such as mixed reality (MR), augmented reality (AR), and virtual reality (VR), into DT-based HRCA systems represents a significant evolution in manufacturing, placing a primary focus on designing processes around human needs and interactions. As these technologies evolve and mature, their role in enabling seamless and efficient HRI is set to grow, paving the way for more innovative and adaptable manufacturing solutions. VR utilises computer technology and software to create an interactive 3D virtual environment. In HRI, the DT provides an accurate digital model, facilitating complete process simulation before the actual assembly. Throughout the assembly process, the data analytical capabilities of the DT model are utilised to adjust the robot's behaviour and strategy in real-time. VR intuitively enables operators to interact with these models, offering a comprehensive immersive experience. Integrating VR and DT in HRCA establishes a seamless and efficient platform for collaboration between operators and robots, significantly enhancing safety, production efficiency, and quality. A VR-based framework designed for the immersive design and simulation of HRCA environments was developed in Malik et al. (2020).

AR superimposes virtual objects and digital information onto the real world, seamlessly blending digital content with the physical environment. Unlike VR, which constructs an entirely new, immersive environment, AR augments the user's perception of existing surroundings by integrating interactive digital elements such as images, text, and sounds directly into their field of vision. The user's experience is enriched by blending real and virtual worlds, allowing them to interact with both simultaneously, thereby offering an augmented perception of reality. A methodology combining AR and DT was developed to improve the assembly tasks of HRC scenarios (Blaga and Tamas, 2018). This innovative approach leverages the strengths of both AR and DT: AR provides immediate, intuitive interaction with digital enhancements, while DT offers a precise, dynamic replica of physical assets. Consequently, it effectively bridges the gap between digital planning and physical execution. Additionally, a human-cyber-physical assembly system that merges DT and AR was introduced to minimise human errors and enhance safety (Zhang et al., 2023a). This synergistic technology is poised to transform traditional manufacturing processes by increasing accuracy and reliability in dynamic environments.

MR refers to the convergence of the real and virtual worlds to form a new visual environment, where physical objects coexist and interact in real-time. MR represents a synthesis of AR and VR technologies. Although research on the integration of MR and DT in HRCA is currently limited, the capabilities and potential of MR within this field are well-established (David et al., 2023; Li et al., 2024). This emerging technology has the potential to transform the field by enabling more intricate interactions and comprehensive simulations, which could lead to breakthroughs in design, training, and operations across various industries.

6 Skeleton-gesture driven HRC

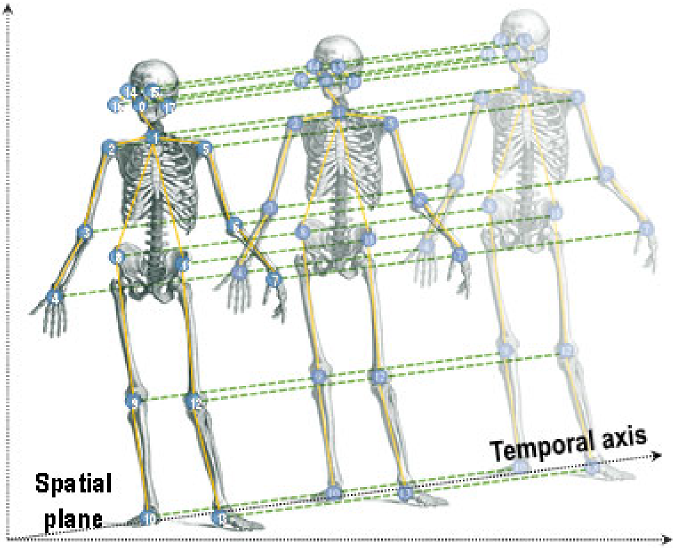
Human gestures delineated through skeletal gestures constitute a pivotal modality within the domain of HRC. This section commences with an introduction to the human skeleton as a fundamental representation of human factors, emphasising motion processing. Subsequently, we comprehensively review skeleton-gesture-driven HRC, exploring its applications and implications in this field.

6.1 Human skeleton as a human factor representation

The dynamics of the human skeleton encode critical information pertinent to human motion, rendering it one of the most frequently utilised modalities for the analysis of human actions and motion studies. This emphasis on skeletal dynamics is predicated on their abilities to provide a structured and measurable framework for quantitatively assessing and interpreting human movement (Liu et al., 2016; Du et al., 2015).

The human skeleton can be effectively represented as a time series of human joint locations, encapsulated in 2D or 3D coordinates. Subsequent studies on human factors, such as human action recognition and motion estimation, are facilitated by analysing the patterns embedded within these individual frames or time series. A topical human skeleton can be seen in Figure 24. Skeleton can be collected via various approaches, such as wearable motion capture devices (Zhou et al., 2023), depth sensors (Zhang et al., 2020a), and daily cameras with algorithmic estimators (Cao et al., 2021). Previous methodologies employed in human action recognition predominantly utilised joint coordinates at discrete time steps to construct feature vectors. These vectors were then subjected to temporal processing to elucidate human movement patterns over time. This approach, while foundational, typically involved straightforward aggregation of spatial data per time point, followed by sequential analysis to interpret the temporal dynamics inherent in human actions. After this initial phase, the representation of the human skeleton was enhanced through integration with more sophisticated data structures such as graphs, which foreground the connections among joints. This advancement enabled the concurrent processing of spatial and temporal information using deep neural networks (Yan et al., 2018).

Figure 24 An illustration of the human skeleton (see online version for colours)



Source: Adapted and modified from Liu et al. (2023)

Compared to other modalities, such as images and videos, human skeletons typically require lower data volumes, which offer distinct advantages for transmitting human motion data across communication networks (Liu et al., 2023). However, human skeletons can sometimes be too information-dense to be used in isolation. Consequently, RGB and depth data frequently serve as supplementary inputs to enhance the utility of skeleton data.

On the other hand, human skeletons analysed independently of RGB data are often recognised for their superior individual generalisation capabilities. This is because the colour information about backgrounds and clothing, which is non-essential for analysing individual movements, is absent, reducing the potential for extraneous variables to influence the generalisation process. This delineation highlights the nuanced trade-offs between data complexity and processing efficiency in human motion analysis. For instance, a deep neural model often processes a higher volume of input data with more parameters.

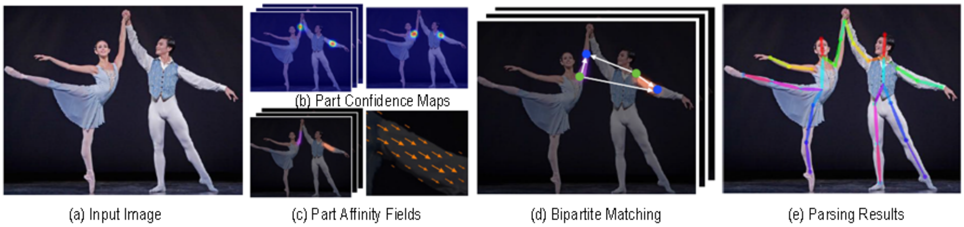
6.2 Sensing techniques for human skeletal representation

Various sensors and algorithms facilitate the tracking and estimation of human skeletons, enabling precise motion capture, activity recognition, and interaction across various applications. Sensing techniques for representing human skeletons can be categorised as follows:

- *Optical motion capture systems:* This motion capture system uses multiple cameras or receivers to track the position of reflective markers or LEDs on a person's body. In industry, this method is often used in animation, film production, and biomechanics. For instance, OptiTrack (Nagymáté and Kiss, 1970) is such a system widely used in human movement analysis.
- *Inertial measurement units (IMUs) are on the human body:* IMUs utilise accelerometers, gyroscopes, and sometimes magnetometers to estimate joint angles and body orientation. This kind of technology has been used in research of human functional activities (Cudejko et al., 2022), as well as HRC (Zhou et al., 2023, 2024).
- *Depth cameras:* Such cameras use stereo imaging with multiple cameras or special lighting, such as infrared or structured light. For instance, sensors such as Intel RealSense (Keselman et al., 2017) or Microsoft Kinect (Zhang, 2012) utilise structured light or time-of-flight techniques to capture depth information and reconstruct 3D models of the environment and humans. Depth cameras, such as Kinect, can generate human skeletons directly for HRC.
- *Computer vision and ML-driven algorithmic estimators:* Such algorithms can process images or video from standard cameras to detect and track human joints using methods like pose estimation, for example, the OpenPose motion estimator (Cao et al., 2021). Figure 25 illustrates the OpenPose pipeline.

The techniques above can be further divided into natural and non-natural human-robot interfaces. Natural interfaces do not require the human operator to wear any devices, reducing their physical burden. Conversely, though potentially uncomfortable in practical applications, non-natural interfaces can capture human motion with greater precision.

Figure 25 An illustration of the human skeleton generated by the OpenPose pipeline (see online version for colours)



Source: Adapted and modified from Cao et al. (2021)

Specifically, optical tracking tags and IMUs require attachment to the human body before use. Although these devices can be integrated into gloves, helmets, or shoes or attached to clothing, wearing them inevitably results in discomfort. Additionally, optical tracking requires the installation of receiving cameras or antennas, often mounted on ceilings, prior to use. For IMU-based tracking systems, powering the IMUs attached to the human body presents a practical challenge. Issues like uncomfortable wiring connections and limited battery life in wireless setups are common. Moreover, such systems tend to be costly due to their numerous components.

Conversely, camera-based sensing technologies provide a more natural human-robot interface, enabling individuals to interact with robots in a manner similar to interacting with other humans in daily activities. Additionally, applying computer vision and ML driven algorithms to standard, cost-effective cameras enhances their popularity in practical applications, providing a more accessible and user-friendly approach to HRC.

6.3 Skeleton processing for HRC

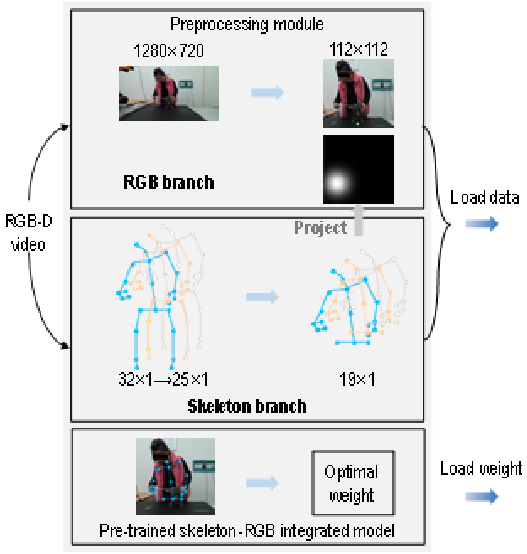
Skeleton-gesture-driven HRC represents a field of research that utilises sensors to capture the human skeletal structure in real-time and subsequently modulates the robotic system to align with these inputs, thereby establishing a cooperative framework wherein humans and robots synergistically execute shared tasks. Conceptually, these modifications may encompass the control and planning mechanisms within the robotic system, which are manifested at both the motion and task levels in robotics (Liu et al., 2022c).

The application at the task level represents the most extensively examined area where the processing of the human skeleton is utilised. This prominence stems from the real-time perception of human status being a prerequisite for a more agile and flexible HRC. By achieving this, robots can frequently re-plan tasks to accommodate the human collaborator within a shared space (Wang et al., 2019). For instance, Zhang et al. (2022b) used an STGCN network and a YOLOX network to recognise the assembly intention from human operators. In this work, the human skeleton is used to represent the assembly motion of a human. The human skeleton data is transmitted into STGCN for processing. Yasar and Iqbal (2021) used a gated recurrent unit (GRU) network to process the human skeleton data in HRC scenarios. Zhang et al. (2020b) used an RNN for human motion prediction in HRC. Discrete human motion states were first designed, and then the human skeleton was used as input for the RNN, followed by results reflecting the next states. Zhang et al. (2022a) embodied LSTM networks concerning motion representation in HRC. In this work, the collaborative task was represented by the graph. The recognition of human motion updates the graph. Li et al. (2022) combined the human skeleton with RGB data, and recognised five different motions in HRC.

The literature reviewed shows that recent research on human skeleton processing for HRC has predominantly utilised conventional DL models that incorporate time series inputs, such as RNNs, LSTMs, and GRUs. Subsequently, emphasis has shifted towards topology information. Notable studies have demonstrated the effective use of graph networks, given that the human skeleton can be accurately represented as a graph. However, the majority of studies have overlooked several issues that are both significant and practical. For instance, much of the research has concentrated solely on human motions that are readily distinguishable from one another, indicating that these motions or actions are inherently distinct. Conversely, similar motions and their corresponding manual procedures in HRC tasks, such as assembly and disassembly, have not been thoroughly investigated. Actions like raising arms, showing hands, and crossing hands, significant elements of human body language, are seldom observed in practical HRCs, particularly in industrial settings.

Furthermore, many published studies have analysed the human skeleton without accounting for noise or occlusions. For example, occlusions of the human skeleton are pretty common during interactions between humans and robots, where parts of the human skeleton may be obscured, or some joints may not be accurately captured. Despite the prevalence and severity of these challenges, they are frequently disregarded in the literature. To this end, Chen et al. (2020) investigated the repetitive assembly actions using object detection and skeleton-based pose estimation. However, this work was not deeply associated with robotic applications. Zhang et al. (2024) studied similar human action in HRCA with skeleton and RGB integrated input, shown in Figure 26. However, HRIs were not well-considered. Recently, Liu et al. (2023) studied similar repetitive manual procedures and human-robot interferences combined. In their work, offline manual procedure recognition was realised, and online recognition and robotic procedure generation were also studied.

Figure 26 An illustration of online human skeleton processing (see online version for colours)



```
Online human action prediction
Given: RGB-D video stream
Preprocessing module I
Pre-trained model H
Procedure: Initialize video sequence E;
Initialize obtaining frames J;
Initialize the input data L in the storage sequence S;
Initialize frame count O;
Given a limited number of frames Q;
while start device do:
    Start capturing the video stream;
    if frame count O % Q = = 0
        Save data of J frames from E to S;
        Feed input data L (RGB images and
        skeleton sequences) from S to I;
        Feed processed data including mask to H;
        Output prediction;
        Delete 0.8L in the storage sequence S;
        frame count O − 0.2O
    end if
    frame count += 1;
end while
```

Source: Adapted and modified from Zhang et al. (2024)

From the literature, it is apparent that the human skeleton serves as an ideal representation of human motion in HRC. It effectively captures the position and gesture of human motion at any given timestamp. It can also convey semantic information, such as the specifics of ongoing or completed assembly or disassembly procedures. When viewed from a broader perspective, the human skeleton data along the temporal axis can be structured as a spatial-temporal graph. This data configuration has facilitated the development of spatial-temporal integrated processing approaches, such as deep graph-based neural models. These models leverage the rich, layered information embedded within the spatial-temporal graph to enhance the model performance (Duan et al., 2022). It should also be noted that, while the human skeleton representation has a lower data volume, which improves data transmission efficiency, it can be augmented with additional modalities, such as RGB images and depth point clouds. This capability allows for an in-depth understanding of human motion by integrating diverse data types that provide complementary perspectives and depth to the skeletal data.

However, some challenges remain insufficiently addressed in the literature. For instance, few studies have considered factors such as similarity, repetition, and human-robot interference, despite their prevalence in practical applications. Neglecting these challenges impedes the transition of laboratory-based research to real-world environments. Additionally, many studies have focused solely on offline recognition, deferring the exploration of real-time, online applications to future research. Furthermore, integrating human skeleton data with robotic task planning, such as generating procedural tasks for robots in collaborative settings, is only beginning to be explored. These issues underscore the need for continued research, which we will discuss in the sections on future directions and open problems.

7 Future directions

Future research directions are closely aligned with the specific limitations and unresolved challenges discussed in previous sections. By systematically reflecting on the technical constraints, usability issues, and deployment barriers identified, these proposed directions aim to provide a roadmap for advancing the field. Rather than offering abstract suggestions, each direction is grounded in a concrete gap or bottleneck observed in current methodologies, ensuring relevance and coherence with the state of the art.

7.1 LLMs for natural HRIs

The progression of LLMs has showcased remarkable proficiency in language comprehension, reasoning, and generation tasks, effectively bridging longstanding gaps in natural language processing and understanding. These models, particularly those trained on multimodal data, have shown strong generalisation capabilities across diverse tasks with minimal or no fine-tuning. This makes them well-suited for integration into HCA, where natural, intuitive interactions between humans and robots are critical. Within an assembly environment, operators are often required to convey intentions, provide feedback, or issue instructions – all of which can be naturally expressed through language. LLMs offer a compelling opportunity to serve as the central interpretive interface between human operators and robotic systems. The integration of LLMs into HCA systems enables more seamless and context-aware interactions. Consider GPT-4, a state-of-the-art multimodal model capable of processing textual and visual inputs and generating highly contextualised outputs. For example, an operator could upload an image of a partially assembled component and ask the system, in natural language, how to proceed. The model could then interpret the visual scene, reason over the current assembly stage, and generate step-by-step instructions tailored to the specific task. These instructions could then be translated into executable robot commands via downstream modules for perception and control, enabling fluid HRC without requiring specialised programming.

Unlike traditional communication channels such as GUI-based programming or predefined command sets, LLMs offer a unified interface that supports flexible input modalities, including text, speech, gestures, and gaze. This flexibility is crucial in dynamic manufacturing environments where operators prefer or require different interaction modes. Moreover, thanks to their extensive pre-training on internet-scale datasets – including technical manuals, instructional videos, and synthetic data – LLMs can generalise to new tasks and adapt to various assembly scenarios with minimal supervision. Further improvements can be realised by enriching their contextual awareness by integrating LLMs with real-time multimodal sensor streams (e.g., RGB-D cameras, wearable devices, haptic sensors). Fine-tuning or instruction tuning using domain-specific assembly data can enhance their relevance and safety. With continued advances in grounding language to perception and action, LLMs stand to become a foundational component of next-generation HRC systems, enabling robots to understand human commands and proactively assist, adapt, and collaborate in increasingly complex and dynamic environments.

7.2 Autonomous robots assisting humans in tasks

In the context of on-demand robot-human assistance, the next generation of collaborative robotic systems must recognise when and where assistance is needed and execute such tasks autonomously and effectively. This vision requires robots to exhibit a degree of cognitive and motor intelligence that mirrors human flexibility and adaptability. Specifically, robots must interpret human intentions, predict real needs, and respond with context-appropriate actions. Intention detection, behaviour understanding, and task inference are critical components of this high-level decision-making process, enabling robots to support human operators proactively during complex assembly procedures.

Traditional industrial robots are often constrained by rigid programming and limited environmental awareness, making them unsuitable for dynamic and unstructured human-centric environments (Liu et al., 2024c; Wang et al., 2025c). To overcome these limitations, autonomous mobile robotic platforms equipped with integrated perception, navigation, and manipulation capabilities are emerging as viable solutions. These robots can serve as intelligent extensions of human capability – navigating workspaces, fetching tools or parts, and adjusting their actions based on evolving task contexts. Multimodal communication plays a central role in facilitating this collaboration. Humans can issue commands through various forms, including speech, text, gesture, and visual demonstrations. These instructions are interpreted through learning-based instruction reasoning systems that map high-level human intent to executable robot actions. For example, a spoken command such as ‘bring me the torque wrench next to the blue valve’ can be grounded through object detection, spatial reasoning, and semantic understanding. This enables the robot to identify the correct object and deliver it to the operator with minimal supervision. Advanced vision systems further enhance the robot’s autonomy by allowing visual reconstruction of the workspace, object pose estimation, and consistency tracking over time. This empowers robots to recognise partially completed assemblies, identify missing components, and suggest next steps or corrections. By understanding both the nature and the spatial configuration of tasks, autonomous robots can become proactive collaborators, anticipating needs, reducing cognitive load on human workers, and ultimately improving efficiency, safety, and flexibility in HRCA settings.

7.3 Robotics foundation models for general-purpose manipulation

The emergence of GPT marked a paradigm shift in AI, demonstrating that large-scale models pre-trained on diverse internet-scale datasets can generalise across tasks with remarkable effectiveness. Initially designed for language, these foundation models have since been extended to vision, audio, and multimodal domains, enabling new forms of interaction, reasoning, and generalisation. Their success has inspired similar developments in robotics, where integrating large-scale learning with physical embodiment opens new frontiers for general-purpose manipulation. Robotics foundation models aim to replicate the success of GPT-like architectures by training on massive, heterogeneous datasets composed of both simulated and real-world robot interactions. These datasets may include visual observations, action trajectories, proprioceptive feedback, and multimodal instructions (e.g., language, demonstrations). Rather than being narrowly tailored to a single task or robot, these models are designed to generalise across functions, environments, and hardware platforms. This flexibility is crucial in human-centric domains such as manufacturing and assembly, where task variability and environment complexity demand adaptable and reusable robotic intelligence.

The development of robotics foundation models is accelerating, supported by increasing access to large-scale robot learning datasets collected across distributed labs and industrial deployments. RT-1 (Brohan et al., 2022), RT-2 (Brohan et al., n.d.), and OpenVLA (Kim et al., 2024b) demonstrate early examples of scalable frameworks for training unified policies across hundreds of manipulation tasks and diverse robot morphologies. These models learn to manipulate objects, use tools, and navigate in mobile environments, often guided by high-level language commands or visual goals. Encoding semantic understanding and physical interaction patterns offers a pathway to zero-shot generalisation, where a robot can perform a novel task based on prior experience and contextual cues without additional training. Integrating foundation models into robotic systems promises to drastically reduce the time and cost associated with programming robots for new tasks. When combined with real-time sensory feedback and fine-tuning on task-specific demonstrations, these models can serve as general-purpose agents capable of assisting in various applications, from precision manufacturing to home assistance, pushing robotics closer to truly versatile and intelligent autonomy.

7.4 RL-assisted DT-based HRCA

The integration of RL into DT-based HRCA represents a promising frontier for next generation intelligent manufacturing systems. RL offers a robust framework for robots to learn optimal policies through trial and error, guided by feedback from the environment. When coupled with DTs – virtual replicas of physical systems that mirror real-time operations – RL can enable robots to adapt dynamically to evolving conditions, optimise task performance, and enhance safety in collaborative scenarios. In highly variable manufacturing settings where customisation, flexibility, and responsiveness are key, RL assisted DT systems can simulate and evaluate numerous strategies before applying them in the physical world. This reduces risk, accelerates learning, and minimises costly errors on the factory floor. The DT serves as a real-time mirror and a sandbox environment where policies can be pre-trained or fine-tuned. This is especially beneficial for tasks that involve close collaboration with human operators, where safety, precision, and contextual understanding are critical.

By integrating continuous real-time sensor data from the physical workspace, RL agents operating within the DT environment can make informed decisions based on accurate, up-to-date representations of the assembly process. This enables proactive adjustments to robot behaviour in response to human motion, changes in component availability, or unexpected disturbances. Such capabilities enhance robotic perception, mobility, and adaptability in shared workspaces. RL also plays a key role in promoting coordinated behaviour among heterogeneous agents in multi-robot systems (Wang et al., 2025a). Robots can learn to cooperate, divide labour, and sequence actions efficiently through multi-agent RL, improving throughput and reducing operator burden. Future RL assisted DT-based HRCA systems are expected to become increasingly autonomous, intelligent, and context-aware. They will support more natural forms of HRI, enable rapid task reconfiguration, and scale across different production environments with minimal human intervention. This RL, DT, and HRC convergence holds transformative potential for Industry 5.0, where human creativity is augmented by flexible, learning-enabled robotic collaborators operating in digitally orchestrated ecosystems.

7.5 DT and MR-enabled HRCA

The DT and MR technologies offer transformative potential for enhancing HRCA in smart manufacturing. Combining DTs' real-time, data-driven capabilities with the immersive, interactive interfaces of MR, including both VR and AR, enables human operators to engage with digital representations of assembly tasks in a physically grounded context. One of the key benefits of DT-MR integration lies in the ability to visualise and interact with virtual models of robotic systems and assembly components directly within the physical workspace. Operators can observe projected robotic motions, assess task feasibility, and simulate assembly sequences before physical execution. This level of transparency supports intuitive task planning, helps identify potential spatial or temporal conflicts early in the design phase, and enhances situational awareness during collaboration. As a result, errors, downtime, and safety risks can be significantly reduced. Moreover, MR interfaces empower users to manipulate and configure DT models through natural gestures, voice commands, or handheld controllers, fostering a more user-friendly and accessible interaction paradigm. Engineers and technicians can explore component relationships, inspect machine internals, and train on complex procedures without requiring direct access to physical systems, minimising production interruptions and lowering training costs.

In practice, this integration enables a continuous feedback loop, where the DT updates in real-time based on sensor data from the physical environment. At the same time, MR interfaces display this evolving state to users, allowing for immediate adjustments and collaborative problem-solving. This duality of real and virtual enhances planning and execution, enabling dynamic reconfiguration of tasks and rapid adaptation to unexpected events or changes in production requirements. Looking ahead, the widespread application of DT and MR in HRCA is anticipated to become a cornerstone of Industry 5.0, where the synergy between human intuition and robotic precision is maximised. As hardware becomes more lightweight and software more interoperable, these technologies are essential to bridge the cognitive and physical gap between robots and humans, enabling a new era of intelligent, immersive, and efficient collaboration on the factory floor.

7.6 Advanced ML-driven algorithmic estimator for human skeleton

Recent progress in ML has enabled the creation of skeleton estimation algorithms that are far more user-friendly and widely available, cost-effective, and user-friendly than traditional motion capture systems. These estimators, powered by DL models such as CNNs, GNNs, and transformer-based architectures, can infer full-body human poses from RGB videos or depth camera inputs, making them suitable for deployment in portable, real-time, and even embedded environments. As a result, they are particularly attractive for use in HRCA, where flexibility, affordability, and ease of integration are critical. Despite these advantages, current learning-based skeleton estimators still face notable limitations in accuracy, robustness, and temporal stability, particularly when compared to high-accuracy motion capture systems or inertial measurement unit (IMU)-based solutions. Estimation errors can accumulate in dynamic scenes, under occlusions, or during fast human movements. These challenges can hinder the reliable understanding of motion and intention prediction in collaborative robotic systems, where safety and precision are paramount.

There is substantial room for improvement by incorporating domain knowledge from HRI contexts into model design and training. For example, task-specific priors, physical constraints of human movement, and collaborative context cues can be embedded into the estimation pipeline to enhance robustness and reduce ambiguity. Furthermore, integrating multimodal data, such as combining 2D video with sparse depth input or wearable inertial data, can significantly improve estimator performance while maintaining cost efficiency. Developing advanced algorithmic skeleton estimators tailored for HRCA remains a vibrant and open research direction. Future estimators could leverage large-scale datasets collected from collaborative assembly scenarios to better model interaction dynamics. Additionally, incorporating real-time feedback from robots and DTs could enable adaptive estimation strategies that self-correct based on environmental and task constraints. By closing the performance gap between lightweight estimators and high-fidelity tracking systems, these ML-driven methods can unlock scalable, real-world deployment of intelligent, perception-aware robotic systems capable of understanding, predicting, and responding to human behaviour with greater accuracy and reliability.

7.7 Robust online procedure recognition with human skeleton

Robust and real-time recognition of manual procedures is critical to effective HRC, particularly in dynamic and safety-sensitive environments such as manufacturing and assembly. Leveraging human skeletal data extracted from vision-based or sensor-driven pose estimation systems provides a promising avenue for understanding human actions at a fine-grained level. However, online procedure recognition – where systems must make decisions continuously and adaptively during task execution – presents several technical challenges that remain insufficiently addressed. One major challenge lies in optimising the temporal representation of skeletal data. Sliding window techniques are often used to segment continuous streams of skeleton data for processing, but selecting an appropriate window size is non-trivial. A window that is too short may fail to capture the semantic structure of complex actions, leading to frequent false positives or misclassifications. Conversely, overly long windows may introduce latency and blur transitions between procedures, reducing system responsiveness. Adaptive windowing mechanisms that dynamically adjust based on motion intensity or contextual cues could significantly improve segmentation fidelity.

Another key issue is the continuity of recognition results. In real-world collaborative tasks, transitions between procedures, such as reaching, grasping, assembling, or handing over tools, often occur fluidly and without clear boundaries. Existing recognition systems may suffer from jitter or abrupt switching between predicted classes, undermining the interpretability and stability of the recognition pipeline. Incorporating transition smoothing strategies, confidence-based filtering, and temporal regularisation techniques can enhance the coherence of online outputs. Furthermore, reducing short-term misrecognitions is essential for building trust in robotic systems. Even brief errors in procedure classification can lead to unsafe or inappropriate robot behaviour. Future work should explore hybrid models that combine skeleton-based cues with additional modalities, such as audio, eye gaze, or object state recognition, to cross-validate predictions and enhance robustness.

In summary, online procedure recognition using human skeleton data remains an active and open research frontier. Improving temporal modelling, transition handling, and multi-modal fusion are key directions that will enable more accurate, interpretable, and reliable real-time recognition systems, laying the groundwork for fluid, responsive, and intelligent HRC.

7.8 Robotic task planning with online human skeleton processing

The ultimate objective of human motion recognition in collaborative settings is not simply to classify or label actions, but to enable intelligent, context-aware robotic behaviour that supports and complements human partners. In this regard, integrating online human skeleton processing into robotic task planning represents a critical yet underdeveloped research direction. By leveraging real-time skeletal data to infer human intentions, goals, and actions, robotic systems can proactively adapt their plans to ensure synchronised, safe, and effective collaboration. While recent advances in human pose estimation and action recognition have improved the reliability of skeletal tracking, translating this data into actionable robotic plans remains a complex challenge. One of the key gaps lies in the absence of unified frameworks that bridge perception and planning. For example, joint task representations – formal models that encode human and robot actions, their temporal dependencies, and shared objectives – are still in their early stages of development. Without such representations, robots lack the structure necessary to interpret human movements concerning evolving task contexts and to reason about their contributions accordingly.

Furthermore, planning methodologies that account for real-time human motion are often computationally intensive or rigid, limiting their scalability and responsiveness. Future work should explore lightweight, adaptive planning algorithms that can operate under uncertainty, incorporate probabilistic forecasts of human motion, and adjust plans incrementally as new skeletal data becomes available. Integration with DT environments can also support predictive simulations and dynamic re-planning, offering a virtual testing ground for collaborative strategies before they are physically executed. The fusion of skeleton-based observations with other data streams, such as speech, object tracking, or gaze estimation, can further enrich the robot's understanding of human intent. This multimodal integration is crucial for developing context-aware planners that can provide nuanced responses. Finally, embedding online human skeleton processing into robotic task planning remains a frontier with vast potential. Addressing challenges related to joint task representation, adaptive planning, and multimodal fusion will be key to unlocking more natural, responsive, and efficient HRC across various industrial and service-oriented applications.

8 Conclusions

This study presents a comprehensive overview of multimodal HRC in manufacturing. The multimodal communication channels composed of voice, gesture, haptics, and brainwaves were investigated and adopted for natural HRIs as well as action and assembly control. Specifically, it is composed of voice instruction-controlled robot motion in Cartesian space, hand gesture instructions for robot gripper control, and brainwave-driven HRCA, especially in noisy environments with unreliable voice recognition or when operators are occupied with other tasks and unable to make gestures. In parallel, a sensorless haptic control approach to HRCA enables human operators to haptically control industrial robots without requiring additional sensors during collaborative assembly. This offers multimodal support to the HRCA as an alternative to contactless commands. Meanwhile, a DT model of physical HRC workcells was developed to facilitate collaborative assembly. To better reveal and predict human behaviours in assembly, the use of the human body skeleton to facilitate efficient HRC was also investigated. Along with the advancement of HRC, future directions are highlighted.

Acknowledgements

The authors acknowledge support from the Swedish Research Council under Award 2023-00493, the Excellence in Production Research (XPRES), and the National Academic Infrastructure for Super Computing in Sweden (NAISS) under Award 2024/5-164 and 2025/22-791. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

References

Adamson, G., Wang, L. and Moore, P. (2017) ‘Feature-based control and information framework for adaptive and distributed manufacturing in cyber physical systems’, *Journal of Manufacturing Systems*, Vol. 43, pp.305–315, https://doi.org/10.1016/j.jmsy.2016.12.003.

Al-Bender, F., Lampaert, V. and Swevers, J. (2005) ‘The generalized Maxwell-slip model: a novel model for friction simulation and compensation’, *IEEE Transactions on Automatic Control*, Vol. 50, No. 11, pp.1883–1887, https://doi.org/10.1109/TAC.2005.858676.

Al-Saegh, A., Dawwd, S.A. and Abdul-Jabbar, J.M. (2021) ‘Deep learning for motor imagery EEG-based classification: a review’, *Biomedical Signal Processing and Control*, Vol. 63, p.102172, https://doi.org/10.1016/j.bspc.2020.102172.

Bilberg, A. and Malik, A.A. (2019) ‘Digital twin driven human–robot collaborative assembly’, *CIRP Annals*, Vol. 68, No. 1, pp.499–502, https://doi.org/10.1016/j.cirp.2019.04.011.

Blaga, A. and Tamas, L. (2018) 'Augmented reality for digital manufacturing', *2018 26th Mediterranean Conference on Control and Automation (MED)*, pp.173–178, https://doi.org/10.1109/MED.2018.8443028.

Breque, M., De Nul, L. and Petridis, A. (2021) *Industry 5.0: Towards a Sustainable, Human Centric and Resilient European Industry*, Publications Office [online] https://data.europa.eu/doi/10.2777/308407 (accessed 10 October 2025).

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., Florence, P., Fu, C., Arenas, M.G., Gopalakrishnan, K., Han, K., Hausman, K., Herzog, A., Hsu, J., Ichter, B. and Zitkovich, B. (n.d.) *RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control*, https://doi.org/10.48550/arXiv.2307.15818.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jackson, T., Jesmonth, S., Joshi, N.J., Julian, R., Kalashnikov, D., Kuang, Y. and Zitkovich, B. (2022) *RT-1: Robotics Transformer for Real-World Control at Scale (Version 2)*, arXiv, https://doi.org/10.48550/ARXIV.2212.06817.

Canziani, A., Paszke, A. and Culurciello, E. (2016) *An Analysis of Deep Neural Network Models for Practical Applications (Version 4)*, arXiv, https://doi.org/10.48550/ARXIV.1605.07678.

Cao, Z., Hidalgo, G., Simon, T., Wei, S-E. and Sheikh, Y. (2021) 'OpenPose: realtime multi-person 2D pose estimation using part affinity fields', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 43, No. 1, pp.172–186, https://doi.org/10.1109/TPAMI.2019.2929257.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P.S., Yang, Q. and Xie, X. (2024) 'A survey on evaluation of large language models', *ACM Transactions on Intelligent Systems and Technology*, Vol. 15, No. 3, pp.1–45, https://doi.org/10.1145/3641289.

Chen, C., Wang, T., Li, D. and Hong, J. (2020) 'Repetitive assembly action recognition based on object detection and pose estimation', *Journal of Manufacturing Systems*, Vol. 55, pp.325–333, https://doi.org/10.1016/j.jmsy.2020.04.018.

Cheng, G., Ehrlich, S.K., Lebedev, M. and Nicolelis, M.A.L. (2020) 'Neuroengineering challenges of fusing robotics and neuroscience', *Science Robotics*, Vol. 5, No. 49, p.eabd1911, https://doi.org/10.1126/scirobotics.abd1911.

Craik, A., He, Y. and Contreras-Vidal, J.L. (2019) 'Deep learning for electroencephalogram (EEG) classification tasks: a review', *Journal of Neural Engineering*, Vol. 16, No. 3, p.31001, https://doi.org/10.1088/1741-2552/ab0ab5.

Cudejko, T., Button, K. and Al-Amri, M. (2022) 'Validity and reliability of accelerations and orientations measured using wearable sensors during functional activities', *Scientific Reports*, Vol. 12, No. 1, p.14619, https://doi.org/10.1038/s41598-022-18845-x.

David, J., Coatanéa, E. and Lobov, A. (2023) 'Deploying OWL ontologies for semantic mediation of mixed reality interactions for human-robot collaborative assembly', *Journal of Manufacturing Systems*, Vol. 70, pp.359–381, https://doi.org/10.1016/j.jmsy.2023.07.013.

Delmas, G., Weinzaepfel, P., Moreno-Noguer, F. and Rogez, G. (2025) 'PoseEmbroider: towards a 3D, visual, semantic-aware human pose representation', in Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T. and Varol, G. (Eds.): *Computer Vision – ECCV 2024*, Springer Nature, Switzerland, Vol. 15129, pp.55–73, https://doi.org/10.1007/978-3-031-73209-6_4.

Dimitropoulos, N., Togias, T., Zacharaki, N., Michalos, G. and Makris, S. (2021) 'Seamless human-robot collaborative assembly using artificial intelligence and wearable devices', *Applied Sciences*, Vol. 11, No. 12, p.5699, Switzerland, https://doi.org/10.3390/app11125699.

Djemal, R. and Ko, W. (2020) 'Comprehensive review on brain-controlled mobile robots and robotic arms based on electroencephalography signals', *Intelligent Service Robotics*, Vol. 13, No. 4, pp.539–563, https://doi.org/10.1007/s11370-020-00328-5.

Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W. and Chebotar, Y. (2023) 'Palm-e: an embodied multimodal language model'.

Du, Y., Wang, W. and Wang, L. (2015) 'Hierarchical recurrent neural network for skeleton based action recognition', *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1110–1118, https://doi.org/10.1109/CVPR.2015.7298714.

Duan, H., Wang, J., Chen, K. and Lin, D. (2022) 'PYSKL: towards good practices for skeleton action recognition', *Proceedings of the 30th ACM International Conference on Multimedia*, pp.7351–7354, https://doi.org/10.1145/3503161.3548546.

Elguea-Aguinaco, Í., Serrano-Muñoz, A., Chrysostomou, D., Inziarte-Hidalgo, I., Bøgh, S. and Arana-Arexolaleiba, N. (2023) 'A review on reinforcement learning for contact-rich robotic manipulation tasks', *Robotics and Computer-Integrated Manufacturing*, Vol. 81, p.102517, https://doi.org/10.1016/j.rcim.2022.102517.

Ferdaus, M.M., Abdelguerfi, M., Ioup, E., Niles, K.N., Pathak, K. and Sloan, S. (2024) *Towards Trustworthy AI: A Review of Ethical and Robust Large Language Models (Version 1)*, arXiv, https://doi.org/10.48550/ARXIV.2407.13934.

Gao, F., Xia, L., Zhang, J., Liu, S., Wang, L. and Gao, R.X. (2024) 'Integrating large language model for natural language-based instruction toward robust human-robot collaboration', *Procedia CIRP*.

Gao, Z., Yang, R., Zhao, K., Yu, W., Liu, Z. and Liu, L. (2023) 'Hybrid convolutional neural network approaches for recognizing collaborative actions in human-robot assembly tasks', *Sustainability*, Vol. 16, No. 1, p.139, https://doi.org/10.3390/su16010139.

Grieves, M. (2014) *Digital Twin: Manufacturing Excellence through Virtual Factory Replication*, White Paper.

Gustavsson, P., Syberfeldt, A., Brewster, R. and Wang, L. (2017) 'Human-robot collaboration demonstrator combining speech recognition and haptic control', *Procedia CIRP*, Vol. 63, pp.396–401, https://doi.org/10.1016/j.procir.2017.03.126.

He, B., Yuan, H., Meng, J. and Gao, S. (2020) 'Brain-computer interfaces', in He, B. (Ed.): *Neural Engineering*, pp.131–183, Springer International Publishing, https://doi.org/10.1007/978-3-030-43395-6_4.

He, C., Chen, Y-Y., Phang, C-R., Stevenson, C., Chen, I-P., Jung, T-P. and Ko, L-W. (2023) 'Diversity and suitability of the state-of-the-art wearable and wireless EEG systems review', *IEEE Journal of Biomedical and Health Informatics*, Vol. 27, No. 8, pp.3830–3843, https://doi.org/10.1109/JBHI.2023.3239053.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B. and Liu, T. (2025) 'A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions', *ACM Transactions on Information Systems*, Vol. 43, No. 2, pp.1–55, https://doi.org/10.1145/3703155.

Iba, S., Paredis, C.J.J. and Khosla, P.K. (2005) 'Interactive multimodal robot programming', *The International Journal of Robotics Research*, Vol. 24, No. 1, pp.83–104, https://doi.org/10.1177/0278364904049250.

Islam, M.J., Ho, M. and Sattar, J. (2019) 'Understanding human motion and gestures for underwater human-robot collaboration', *Journal of Field Robotics*, Vol. 36, No. 5, pp.851–873, https://doi.org/10.1002/rob.21837.

Jung, D., Gu, C., Park, J. and Cheong, J. (2025) 'Touch gesture recognition-based physical human-robot interaction for collaborative tasks', *IEEE Transactions on Cognitive and Developmental Systems*, Vol. 17, No. 2, pp.421–435, https://doi.org/10.1109/TCDS.2024.3466553.

Kardos, C., Kovács, A. and Váncza, J. (2017) 'Decomposition approach to optimal feature-based assembly planning', *CIRP Annals*, Vol. 66, No. 1, pp.417–420, https://doi.org/10.1016/j.cirp.2017.04.002.

Kawala-Sterniuk, A., Browarska, N., Al-Bakri, A., Pelc, M., Zygarlicki, J., Sidikova, M., Martinek, R. and Gorzelanczyk, E.J. (2021) 'Summary of over fifty years with brain-computer interfaces – a review', *Brain Sciences*, Vol. 11, No. 1, p.43, https://doi.org/10.3390/brainsci11010043.

Keemink, A.Q., Van Der Kooij, H. and Stienen, A.H. (2018) 'Admittance control for physical human-robot interaction', *The International Journal of Robotics Research*, Vol. 37, No. 11, pp.1421–1444, https://doi.org/10.1177/0278364918768950.

Keselman, L., Iselin Woodfill, J., Grunnet-Jepsen, A. and Bhowmik, A. (2017) 'Intel realsense stereoscopic depth cameras', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp.1–10.

Khalil, W. and Dombre, E. (2002) *Modeling, Identification and Control of Robots*, Elsevier, https://doi.org/10.1016/B978-1-903996-66-9.X5000-3.

Kim, C.Y., Lee, C.P. and Mutlu, B. (2024a) 'Understanding large-language model (LLM)-powered human-robot interaction', *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pp.371–380, https://doi.org/10.1145/3610977.3634966.

Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., Vuong, Q., Kollar, T., Burchfiel, B., Tedrake, R., Sadigh, D., Levine, S., Liang, P. and Finn, C. (2024b) *OpenVLA: An Open-Source Vision-Language-Action Model (Version 3)*, arXiv, https://doi.org/10.48550/ARXIV.2406.09246.

Kobzarev, O., Lykov, A. and Tsetserukou, D. (2025) *GestLLM: Advanced Hand Gesture Interpretation via Large Language Models for Human-Robot Interaction (Version 2)*, arXiv, https://doi.org/10.48550/ARXIV.2501.07295.

Kokkalis, K., Michalos, G., Aivaliotis, P. and Makris, S. (2018) 'An approach for implementing power and force limiting in sensorless industrial robots', *Procedia CIRP*, Vol. 76, pp.138–143, https://doi.org/10.1016/j.procir.2018.01.028.

Kousi, N., Gkournelos, C., Aivaliotis, S., Lotsaris, K., Bavelos, A.C., Baris, P., Michalos, G. and Makris, S. (2021) 'Digital twin for designing and reconfiguring human-robot collaborative assembly lines', *Applied Sciences*, Vol. 11, No. 10, p.4620, https://doi.org/10.3390/app11104620.

Kragic, D., Gustafson, J., Karaoguz, H., Jensfelt, P. and Krug, R. (2018) 'Interactive, collaborative robots: challenges and opportunities', *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp.18–25, https://doi.org/10.24963/ijcai.2018/3.

Kramberger, A., Kunic, A., Iturrate, I., Sloth, C., Naboni, R. and Schlette, C. (2022) 'Robotic assembly of timber structures in a human-robot collaboration setup', *Frontiers in Robotics and AI*, Vol. 8, p.768038, https://doi.org/10.3389/frobt.2021.768038.

Kritzinger, W., Karner, M., Traar, G., Henjes, J. and Sihn, W. (2018) 'Digital twin in manufacturing: a categorical literature review and classification', *IFAC-PapersOnLine*, Vol. 51, No. 11, pp.1016–1022, https://doi.org/10.1016/j.ifacol.2018.08.474.

Krüger, J., Lien, T.K. and Verl, A. (2009) 'Cooperation of human and machines in assembly lines', *CIRP Annals*, Vol. 58, No. 2, pp.628–646, https://doi.org/10.1016/j.cirp.2009.09.009.

Kuang, J., Shen, Y., Xie, J., Luo, H., Xu, Z., Li, R., Li, Y., Cheng, X., Lin, X. and Han, Y. (2025) 'Natural language understanding and inference with MLLM in visual question answering: a survey', *ACM Computing Surveys*, Vol. 57, No. 8, pp.1–36, https://doi.org/10.1145/3711680.

Li, C., Zheng, P., Zhou, P., Yin, Y., Lee, C.K.M. and Wang, L. (2024) 'Unleashing mixed-reality capability in deep reinforcement learning-based robot motion generation towards safe human-robot collaboration', *Journal of Manufacturing Systems*, Vol. 74, pp.411–421, https://doi.org/10.1016/j.jmsy.2024.03.015.

Li, H., Li, X. and Millán, J.R.D. (2025) 'Noninvasive EEG-based intelligent mobile robots: a systematic review', *IEEE Transactions on Automation Science and Engineering*, Vol. 22, pp.6291–6315, https://doi.org/10.1109/TASE.2024.3441055.

Li, S., Zheng, P., Fan, J. and Wang, L. (2022) 'Toward proactive human–robot collaborative assembly: a multimodal transfer-learning-enabled action prediction approach', *IEEE Transactions on Industrial Electronics*, Vol. 69, No. 8, pp.8579–8588, https://doi.org/10.1109/TIE.2021.3105977.

Li, S., Zheng, P., Liu, S., Wang, Z., Wang, X.V., Zheng, L. and Wang, L. (2023a) 'Proactive human–robot collaboration: mutual-cognitive, predictable, and self-organising perspectives', *Robotics and Computer-Integrated Manufacturing*, Vol. 81, p.102510, https://doi.org/10.1016/j.rcim.2022.102510.

Li, X., Zhang, M., Geng, Y., Geng, H., Long, Y., Shen, Y., Zhang, R., Liu, J. and Dong, H. (2023b) *ManipLLM: Embodied Multimodal Large Language Model for Object-Centric Robotic Manipulation (Version 1)*, arXiv, https://doi.org/10.48550/ARXIV.2312.16217.

Lin, K., Li, Y., Sun, J., Zhou, D. and Zhang, Q. (2020) 'Multi-sensor fusion for body sensor network in medical human-robot interaction scenario', *Information Fusion*, Vol. 57, pp.15–26, https://doi.org/10.1016/j.inffus.2019.11.001.

Liu, H. and Wang, L. (2018) 'Gesture recognition for human-robot collaboration: a review', *International Journal of Industrial Ergonomics*, Vol. 68, pp.355–367, https://doi.org/10.1016/j.ergon.2017.02.004.

Liu, H., Fang, T., Zhou, T., Wang, Y. and Wang, L. (2018) 'Deep learning-based multimodal control interface for human-robot collaboration', *Procedia CIRP*, Vol. 72, pp.3–8, https://doi.org/10.1016/j.procir.2018.03.224.

Liu, J., Shahroudy, A., Xu, D. and Wang, G. (2016) 'Spatio-temporal LSTM with trust gates for 3D human action recognition', in Leibe, B., Matas, J., Sebe, N. and Welling, M. (Eds.): *Computer Vision – ECCV 2016*, Springer International Publishing, Vol. 9907, pp.816–833, https://doi.org/10.1007/978-3-319-46487-9_50.

Liu, S. and Wang, L. (2025) 'Vision intelligence-conditioned reinforcement learning for precision assembly', *CIRP Annals*, p.S0007850625000642, https://doi.org/10.1016/j.cirp.2025.04.016.

Liu, S., Guo, D., Liu, Z., Wang, T., Qin, Q., Wang, X.V. and Wang, L. (2025) 'A digital twin-enabled approach to reliable human-robot collaborative assembly', in Wang, B., Zheng, P., Wang, L. and Mourtzis, D. (Eds.): *Human-Centric Smart Manufacturing Towards Industry 5.0*, pp.281–304, Springer Nature, Switzerland, https://doi.org/10.1007/978-3-031-82170-7_12.

Liu, S., Wang, L. and Gao, R.X. (2024a) 'Cognitive neuroscience and robotics: advancements and future research directions', *Robotics and Computer-Integrated Manufacturing*, Vol. 85, p.102610, https://doi.org/10.1016/j.rcim.2023.102610.

Liu, S., Wang, L. and Wang, X.V. (2020) 'Symbiotic human-robot collaboration: multimodal control using function blocks', *Procedia CIRP*, Vol. 93, pp.1188–1193, https://doi.org/10.1016/j.procir.2020.03.022.

Liu, S., Wang, L. and Wang, X.V. (2021a) 'Function block-based multimodal control for symbiotic human-robot collaborative assembly', *Journal of Manufacturing Science and Engineering*, Vol. 143, No. 9, p.91001, https://doi.org/10.1115/1.4050187.

Liu, S., Wang, L. and Wang, X.V. (2021b) 'Sensorless force estimation for industrial robots using disturbance observer and neural learning of friction approximation', *Robotics and Computer-Integrated Manufacturing*, Vol. 71, p.102168, https://doi.org/10.1016/j.rcim.2021.102168.

Liu, S., Wang, L. and Wang, X.V. (2021c) 'Sensorless haptic control for human-robot collaborative assembly', *CIRP Journal of Manufacturing Science and Technology*, Vol. 32, pp.132–144, https://doi.org/10.1016/j.cirpj.2020.11.015.

Liu, S., Wang, L. and Wang, X.V. (2021d) 'Sensorless haptic control for physical human-robot interaction', in Wang, L., Wang, X.V., Váncza, J. and Kemény, Z. (Eds.): *Advanced Human-Robot Collaboration in Manufacturing*, pp.319–350, Springer International Publishing, https://doi.org/10.1007/978-3-030-69178-3_13.

Liu, S., Wang, L. and Wang, X.V. (2022a) 'Multimodal data-driven robot control for human-robot collaborative assembly', *Journal of Manufacturing Science and Engineering*, Vol. 144, No. 5, p.51012, https://doi.org/10.1115/1.4053806.

Liu, S., Wang, L., Wang, X.V., Cooper, C. and Gao, R.X. (2021e) 'Leveraging multimodal data for intuitive robot control towards human-robot collaborative assembly', *Procedia CIRP*, Vol. 104, pp.206–211, https://doi.org/10.1016/j.procir.2021.11.035.

Liu, S., Wang, X.V. and Wang, L. (2022b) 'Digital twin-enabled advance execution for human-robot collaborative assembly', *CIRP Annals*, Vol. 71, No. 1, pp.25–28, https://doi.org/10.1016/j.cirp.2022.03.024.

Liu, S., Zhang, J., Gao, R.X., Wang, X.V. and Wang, L. (2024b) 'Vision-language model-driven scene understanding and robotic object manipulation', *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*, pp.21–26, https://doi.org/10.1109/CASE59546.2024.10711845.

Liu, S., Zhang, J., Yi, S., Gao, R., Mourtzis, D. and Wang, L. (2024c) 'Human-centric systems in smart manufacturing', in *Manufacturing from Industry 4.0 to Industry 5.0*, pp.181–205, Elsevier, https://doi.org/10.1016/b978-0-443-13924-6.00006-5.

Liu, Z., Liu, Q., Xu, W., Wang, L. and Ji, Z. (2023) 'Adaptive real-time similar repetitive manual procedure prediction and robotic procedure generation for human-robot collaboration', *Advanced Engineering Informatics*, Vol. 58, p.102129, https://doi.org/10.1016/j.aei.2023.102129.

Liu, Z., Liu, Q., Xu, W., Wang, L. and Zhou, Z. (2022c) 'Robot learning towards smart robotic manufacturing: a review', *Robotics and Computer-Integrated Manufacturing*, Vol. 77, p.102360, https://doi.org/10.1016/j.rcim.2022.102360.

Luo, J., Xu, C., Wu, J. and Levine, S. (2024) *Precise and Dexterous Robotic Manipulation via Human-in-the-Loop Reinforcement Learning (Version 2)*, arXiv, https://doi.org/10.48550/ARXIV.2410.21845.

Lv, Q., Zhang, R., Sun, X., Lu, Y. and Bao, J. (2021) 'A digital twin-driven human-robot collaborative assembly approach in the wake of COVID-19', *Journal of Manufacturing Systems*, Vol. 60, pp.837–851, https://doi.org/10.1016/j.jmsy.2021.02.011.

Makris, S., Tsarouchi, P., Surdilovic, D. and Krüger, J. (2014) 'Intuitive dual arm robot programming for assembly operations', *CIRP Annals*, Vol. 63, No. 1, pp.13–16, https://doi.org/10.1016/j.cirp.2014.03.017.

Malik, A.A., Masood, T. and Bilberg, A. (2020) 'Virtual reality in manufacturing: immersive and collaborative artificial-reality in design of human-robot workspace', *International Journal of Computer Integrated Manufacturing*, Vol. 33, No. 1, pp.22–37, https://doi.org/10.1080/0951192X.2019.1690685.

Nagymáté, G. and Kiss, R.M. (1970) 'Application of OptiTrack motion capture systems in human movement analysis', *Recent Innovations in Mechatronics*, Vol. 5, No. 1, https://doi.org/10.17667/riim.2018.1/13.

Neto, P., Norberto Pires, J. and Paulo Moreira, A. (2010) 'High-level programming and control for industrial robotics: using a hand-held accelerometer-based input device for gesture and posture recognition', *Industrial Robot: An International Journal*, Vol. 37, No. 2, pp.137–147, https://doi.org/10.1108/01439911011018911.

Ngo, V-T. and Liu, Y-C. (2024) 'Adaptive impedance and admittance controls for physical human-robot interaction with force-sensorless', in *2024 American Control Conference (ACC)*, IEEE, pp.3791–3796.

Noohi, E., Zefran, M. and Patton, J.L. (2016) 'A model for human-human collaborative object manipulation and its application to human-robot interaction', *IEEE Transactions on Robotics*, Vol. 32, No. 4, pp.880–896, https://doi.org/10.1109/TRO.2016.2572698.

Omer, K., Ferracuti, F., Freddi, A., Iarlori, S., Vella, F. and Monteriù, A. (2025) 'Real-time mobile robot obstacles detection and avoidance through EEG signals', *Brain Sciences*, Vol. 15, No. 4, p.359, https://doi.org/10.3390/brainsci15040359.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J. and Zoph, B. (2023) *GPT-4 Technical Report (Version 6)*, arXiv, https://doi.org/10.48550/ARXIV.2303.08774.

Ott, C., Mukherjee, R. and Nakamura, Y. (2010) 'Unified impedance and admittance control', *2010 IEEE International Conference on Robotics and Automation*, pp.554–561, https://doi.org/10.1109/ROBOT.2010.5509861.

Ouyang, J., Wu, M., Li, X., Deng, H., Jin, Z. and Wu, D. (2024) 'NeuroBCI: multi-brain to multi-robot interaction through EEG-adaptive neural networks and semantic communications', *IEEE Transactions on Mobile Computing*, Vol. 23, No. 12, pp.14622–14637, https://doi.org/10.1109/TMC.2024.3446829.

Park, K., Kulick, J., Melkozerov, A., Vilagrasa, R.A., Lembono, T.S., Neubauer, V., Minichev, A., Turner, K., Agrigoroaiei, O., Klink, P., Lee, J., Dwivedi, K., Eltayeb, M., Posner, I., Parness, A. and Erdogan, C. (2025) *Vulcan Pick: A Robotic System for Picking Targeted Objects from Fabric Pods*, pp.1–16 [online] https://www.amazon.science/publications/vulcan-pick-a-robotic-system-for-picking-targeted-objects-from-fabric-pods.

Pavlovic, V.I., Sharma, R. and Huang, T.S. (1997) 'Visual interpretation of hand gestures for human-computer interaction: a review', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, pp.677–695, https://doi.org/10.1109/34.598226.

Perzanowski, D., Schultz, A.C., Adams, W., Marsh, E. and Bugajska, M. (2001) 'Building a multimodal humanrobot interface', *IEEE Intelligent Systems*, Vol. 16, No. 1, pp.16–21, https://doi.org/10.1109/MIS.2001.1183338.

Ramasubramanian, A.K., Mathew, R., Kelly, M., Hargaden, V. and Papakostas, N. (2022) 'Digital twin for human-robot collaboration in manufacturing: review and outlook', *Applied Sciences*, Vol. 12, No. 10, p.4811, https://doi.org/10.3390/app12104811.

Roy, L., Croft, E.A. and Kulić, D. (2024) 'Learning to communicate functional states with nonverbal expressions for improved human-robot collaboration', *IEEE Robotics and Automation Letters*, Vol. 9, No. 6, pp.5393–5400, https://doi.org/10.1109/LRA.2024.3384037.

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T.H. and Faubert, J. (2019) 'Deep learning-based electroencephalography analysis: a systematic review', *Journal of Neural Engineering*, Vol. 16, No. 5, p.51001, https://doi.org/10.1088/1741-2552/ab260c.

Simonyan, K. and Zisserman, A. (2014) *Very Deep Convolutional Networks for Large-Scale Image Recognition (Version 6)*, arXiv, https://doi.org/10.48550/ARXIV.1409.1556.

Sun, T., Sun, J., Lian, B. and Li, Q. (2024) 'Sensorless admittance control of 6-DoF parallel robot in human-robot collaborative assembly', *Robotics and Computer-Integrated Manufacturing*, Vol. 88, p.102742, https://doi.org/10.1016/j.rcim.2024.102742.

Sun, X., Zhang, R., Liu, S., Lv, Q., Bao, J. and Li, J. (2022) 'A digital twin-driven human-robot collaborative assembly-commissioning method for complex products', *The International Journal of Advanced Manufacturing Technology*, Vol. 118, Nos. 9–10, pp.3389–3402, https://doi.org/10.1007/s00170-021-08211-y.

Swevers, J., Al-Bender, F., Ganseman, C.G. and Projogo, T. (2000) 'An integrated friction model structure with improved presliding behavior for accurate friction compensation', *IEEE Transactions on Automatic Control*, Vol. 45, No. 4, pp.675–686, https://doi.org/10.1109/9.847103.

Taesi, C., Aggogeri, F. and Pellegrini, N. (2023) 'COBOT applications – recent advances and challenges', *Robotics*, Vol. 12, No. 3, p.79, https://doi.org/10.3390/robotics12030079.

Tang, X., Li, W., Li, X., Ma, W. and Dang, X. (2020) 'Motor imagery EEG recognition based on conditional optimization empirical mode decomposition and multi-scale convolutional neural network', *Expert Systems with Applications*, Vol. 149, p.113285, https://doi.org/10.1016/j.eswa.2020.113285.

Tao, F., Zhang, M., Liu, Y. and Nee, A.Y.C. (2018) 'Digital twin driven prognostics and health management for complex equipment', *CIRP Annals*, Vol. 67, No. 1, pp.169–172, https://doi.org/10.1016/j.cirp.2018.04.055.

Tsarouchi, P., Matthaiakis, A-S., Makris, S. and Chryssolouris, G. (2017) 'On a human-robot collaboration in an assembly cell', *International Journal of Computer Integrated Manufacturing*, Vol. 30, No. 6, pp.580–589, https://doi.org/10.1080/0951192X.2016.1187297.

Tuli, T.B., Kohl, L., Chala, S.A., Manns, M. and Ansari, F. (2021) 'Knowledge-based digital twin for predicting interactions in human-robot collaboration', *2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, pp.1–8, https://doi.org/10.1109/ETFA45728.2021.9613342.

Wang, G., Zhang, C., Liu, S., Zhao, Y., Zhang, Y. and Wang, L. (2025a) 'Multi-robot collaborative manufacturing driven by digital twins: advancements, challenges, and future directions', *Journal of Manufacturing Systems*, Vol. 82, pp.333–361, https://doi.org/10.1016/j.jmsy.2025.06.014.

Wang, J., Shi, E., Hu, H., Ma, C., Liu, Y., Wang, X., Yao, Y., Liu, X., Ge, B. and Zhang, S. (2025b) 'Large language models for robotics: opportunities, challenges, and perspectives', *Journal of Automation and Intelligence*, Vol. 4, No. 1, pp.52–64, https://doi.org/10.1016/j.jai.2024.12.003.

Wang, L. (2022) 'A futuristic perspective on human-centric assembly', *Journal of Manufacturing Systems*, Vol. 62, pp.199–201, https://doi.org/10.1016/j.jmsy.2021.11.001.

Wang, L., Gao, R., Váncza, J., Krüger, J., Wang, X.V., Makris, S. and Chryssolouris, G. (2019) 'Symbiotic human-robot collaborative assembly', *CIRP Annals*, Vol. 68, No. 2, pp.701–726, https://doi.org/10.1016/j.cirp.2019.05.002.

Wang, L., Gao, R.X., Krüger, J. and Váncza, J. (2025c) 'Human-centric assembly in smart factories', *CIRP Annals*, p.S0007850625001064, https://doi.org/10.1016/j.cirp.2025.04.058.

Wang, L., Liu, S., Cooper, C., Wang, X.V. and Gao, R.X. (2021) 'Function block-based human-robot collaborative assembly driven by brainwaves', *CIRP Annals*, Vol. 70, No. 1, pp.5–8, https://doi.org/10.1016/j.cirp.2021.04.091.

Wang, L., Liu, S., Liu, H. and Wang, X.V. (2020) 'Overview of human-robot collaboration in manufacturing', *Proceedings of 5th International Conference on the Industry 4.0 Model for Advanced Manufacturing: AMP 2020*, pp.15–58.

Wang, P., Liu, H., Wang, L. and Gao, R.X. (2018) 'Deep learning-based human motion recognition for predictive context-aware human-robot collaboration', *CIRP Annals*, Vol. 67, No. 1, pp.17–20.

Warden, P. (2018) *Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition (Version 1)*, arXiv, https://doi.org/10.48550/ARXIV.1804.03209.

Xie, J., Xu, Z., Zeng, J., Gao, Y. and Hashimoto, K. (2025) 'Human-robot interaction using dynamic hand gesture for teleoperation of quadruped robots with a robotic arm', *Electronics*, Vol. 14, No. 5, p.860, https://doi.org/10.3390/electronics14050860.

Xue, T., Wang, W., Ma, J., Liu, W., Pan, Z. and Han, M. (2020) 'Progress and prospects of multimodal fusion methods in physical human-robot interaction: a review', *IEEE Sensors Journal*, Vol. 20, No. 18, pp.10355–10370, https://doi.org/10.1109/JSEN.2020.2995271.

Yan, S., Xiong, Y. and Lin, D. (2018) 'Spatial temporal graph convolutional networks for skeleton-based action recognition', *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, No. 1, https://doi.org/10.1609/aaai.v32i1.12328.

Yasar, M.S. and Iqbal, T. (2021) 'A scalable approach to predict multi-agent motion for human-robot collaboration', *IEEE Robotics and Automation Letters*, Vol. 6, No. 2, pp.1686–1693, https://doi.org/10.1109/LRA.2021.3058917.

Yi, S., Liu, S., Yang, Y., Yan, S., Guo, D., Wang, X.V. and Wang, L. (2024) 'Safety-aware human-centric collaborative assembly', *Advanced Engineering Informatics*, Vol. 60, p.102371.

Zhang, C., Zhang, Y., Liu, S. and Wang, L. (2025a) 'Transfer learning and augmented data-driven parameter prediction for robotic welding', *Robotics and Computer-Integrated Manufacturing*, Vol. 95, p.102992, https://doi.org/10.1016/j.rcim.2025.102992.

Zhang, C., Zhou, G., Ma, D., Wang, R., Xiao, J. and Zhao, D. (2023a) 'A deep learning-enabled human-cyberphysical fusion method towards human-robot collaborative assembly', *Robotics and Computer-Integrated Manufacturing*, Vol. 83, p.102571, https://doi.org/10.1016/j.rcim.2023.102571.

Zhang, J., Li, P., Zhu, T., Zhang, W-A. and Liu, S. (2020a) 'Human motion capture based on kinect and IMUs and its application to human-robot collaboration', *2020 5th International Conference on Advanced Robotics and Mechatronics (ICARM)*, pp.392–397, https://doi.org/10.1109/ICARM49381.2020.9195342.

Zhang, J., Liu, H., Chang, Q., Wang, L. and Gao, R.X. (2020b) 'Recurrent neural network for motion trajectory prediction in human-robot collaborative assembly', *CIRP Annals*, Vol. 69, No. 1, pp.9–12.

Zhang, R., Lee, S., Hwang, M., Hiranaka, A., Wang, C., Ai, W., Tan, J.J.R., Gupta, S., Hao, Y., Levine, G., Gao, R., Norcia, A., Fei-Fei, L. and Wu, J. (2023b) *NOIR: Neural Signal Operated Intelligent Robots for Everyday Activities (Version 1)*, arXiv, https://doi.org/10.48550/ARXIV.2311.01454.

Zhang, R., Lv, J., Li, J., Bao, J., Zheng, P. and Peng, T. (2022a) 'A graph-based reinforcement learning-enabled approach for adaptive human-robot collaborative assembly operations', *Journal of Manufacturing Systems*, Vol. 63, pp.491–503, https://doi.org/10.1016/j.jmsy.2022.05.006.

Zhang, W., Wang, T., Qin, C., Xu, B., Hu, H., Wang, T. and Shen, Y. (2025b) 'Vibration stimulation enhances robustness in teleoperation robot system with EEG and eye-tracking hybrid control', *Frontiers in Bioengineering and Biotechnology*, Vol. 13, p.1591316, https://doi.org/10.3389/fbioe.2025.1591316.

Zhang, Y., Ding, K., Hui, J., Liu, S., Guo, W. and Wang, L. (2024) 'Skeleton-RGB integrated highly similar human action prediction in human-robot collaborative assembly', *Robotics and Computer-Integrated Manufacturing*, Vol. 86, p.102659, https://doi.org/10.1016/j.rcim.2023.102659.

Zhang, Y., Ding, K., Hui, J., Lv, J., Zhou, X. and Zheng, P. (2022b) 'Human-object integrated assembly intention recognition for context-aware human-robot collaborative assembly', *Advanced Engineering Informatics*, Vol. 54, p.101792, https://doi.org/10.1016/j.aei.2022.101792.

Zhao, S., Liu, S., Jiang, Y., Zhao, B., Lv, Y., Zhang, J., Wang, L. and Zhong, R.Y. (2025a) 'Industrial foundation models (IFMs) for intelligent manufacturing: a systematic review', *Journal of Manufacturing Systems*, Vol. 82, pp.1–30.

Zhao, S., Zhang, G., Liu, S., Zhang, J., Bandara, H., Zhong, R.Y. and Wang, L. (2025b) 'Interpretable verification mechanism for zero-hallucination industrial large model in intelligent manufacturing', *Engineering*, DOI: https://doi.org/10.1016/j.eng.2025.08.023.

Zheng, P., Li, S., Fan, J., Li, C. and Wang, L. (2023) 'A collaborative intelligence-based approach for handling human-robot collaboration uncertainties', *CIRP Annals*, Vol. 72, No. 1, pp.1–4, https://doi.org/10.1016/j.cirp.2023.04.057.

Zhou, H., Wang, L., Pang, G., Shen, H., Wang, B., Wu, H. and Yang, G. (2024) 'Toward human motion digital twin: a motion capture system for human-centric applications', *IEEE Transactions on Automation Science and Engineering*, pp.1–12, https://doi.org/10.1109/TASE.2024.3363169.

Zhou, H., Yang, G., Wang, B., Li, X., Wang, R., Huang, X., Wu, H. and Wang, X.V. (2023) 'An attention-based deep learning approach for inertial motion recognition and estimation in human-robot collaboration', *Journal of Manufacturing Systems*, Vol. 67, pp.97–110, https://doi.org/10.1016/j.jmsy.2023.01.007.