A Deictic Gesture-Based Human-Robot Interface for In Situ Task Specification in Construction

Sungboo Yoon[1]; Jinsik Park[2]; Moonseo Park, Ph.D.[3]; and Changbum R. Ahn, Ph.D.[4]

[1]Ph.D. Student, Department of Architecture and Architectural Engineering, Seoul National University, Seoul, Republic of Korea. Email: yoonsb24@snu.ac.kr
[2]M.S. Student, Department of Architecture and Architectural Engineering, Seoul National University, Seoul, Republic of Korea. Email: wlstlr125@snu.ac.kr
[3]Professor, Department of Architecture and Architectural Engineering, Seoul National University, Seoul, Republic of Korea. Email: mspark@snu.ac.kr
[4]Associate Professor, Department of Architecture and Architectural Engineering, Institute of Construction and Environmental Engineering, Seoul National University, Seoul, Republic of Korea (corresponding author). Email: cbahn@snu.ac.kr

ABSTRACT

Despite the potential of robotic systems for automating the construction industry, the role of human operators remains essential for the success of these systems in complex and dynamic environments. However, current human-robot interfaces are often limited to low-level interactions that require constant micromanaging of robot movements. To address this limitation, this study proposes a deictic gesture-based interface that enables high-level task specification for construction robots. To evaluate the user experience and task performance of the proposed interface, we conducted a laboratory experiment with six human subjects who interacted with the robot to make openings in drywall panels. The results show that the proposed interface significantly reduced mental demand and effort levels among the participants compared to the conventional joystick interface. Moreover, task performance using the proposed interface was comparable in accuracy and efficiency to that achieved with the joystick interface. These findings highlight the potential of the proposed deictic gesture-based interface to facilitate intuitive human-robot interaction and precise operation of construction robots, particularly in situations where as-planned building models are not readily available.

INTRODUCTION

Despite the growing expectations of advanced robotic capabilities for performing complex tasks in construction, challenges remain in integrating robotic systems into construction sites. Specifically, the current capabilities of construction robots are limited to performing simple and repetitive tasks, which often require significant input from human operators to control the robots (Liang Ci-Jyun et al. 2021). Despite the potential of robotic systems to automate the construction industry, the role of human operators remains crucial in ensuring the success of these systems in dynamic and complex construction sites (Wang et al. 2021).

In such complex environments, high-level human input plays a critical role. Human decisions about task-relevant information can help to improve robot's task performance in uncertain or ambiguous environments (Stoddard et al. 2022). For example, in the task of drywall cutting, a task specification could include the precise positions and/or angles needed for the robot to make accurate and efficient cuts (Feng et al. 2013). Such high-level specifications can help the robot to perform tasks more effectively with a high degree of autonomy, without requiring constant human intervention (Stoddard et al. 2022). Traditional methods such as control pads or buttons have proven to be difficult to achieve this level of human input (Losey et al. 2022).

Several approaches for developing intuitive human-robot interfaces that can deliver human input to construction robots are present in the literature. Interaction techniques that are easy to learn and use can significantly facilitate communication between end-users and technology (Han et al. 2020). Intuitive human-robot interfaces allow users to use body gestures (Wang et al. 2023; Wang and Zhu 2021), haptic technology (Zhu et al. 2021; Zhu Qi et al. 2022), and brain activities (Liu et al. 2021a; b) to communicate with the robot, reducing the cognitive load on the operator and increasing the efficiency of the system (Villani et al. 2018). However, these interfaces are typically limited to interactions with low-levels of human input, as they often rely on direct control strategies (Chen et al. 2020), such as 'move up' or 'stop', which are not capable of handling the technical complexity of providing high-level input for construction tasks. Such high-level input, such as choosing a workpiece or specifying precise positions and angles for cutting or assembly, is often difficult to interpret accurately and thus requires additional intelligence for interfaces (Chen et al. 2020).

To this end, this study proposes a deictic gesture-based interface that can handle in situ task specification as high-level inputs for construction robots (Figure 1). Deictic gestures, which are commonly used to convey spatial information in complex environments (Alibali 2005), are used as input for the interface, which combines laser pointing as fine-tuning techniques to accurately estimate task locations in 3D workspaces and perform corresponding motion planning. To evaluate the performance of our proposed interface, we conducted a laboratory experiment in which six human subjects interacted with the robot to make openings in drywall panels for installing steel vents using deictic gesture-based interaction method. The findings of this study highlight the opportunity of using our deictic gesture-based interface to facilitate intuitive human-robot interaction and precise operation of construction robots.

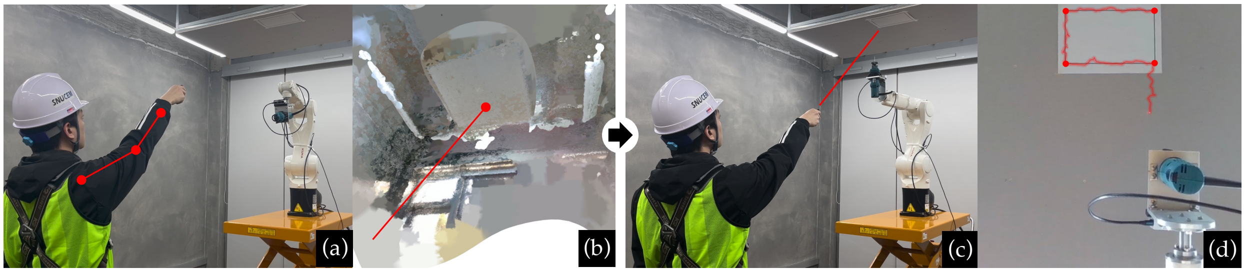

Figure 1. Task specification sequence using our gesture-based human-robot interface. (a-b): The human operator uses deictic gestures to indicate the target location; (c-d): After the robot moves to the detection pose, laser pointing is used to fine-tune the target location.

METHODOLOGY

An overview of the workflow for the proposed deictic gesture-based human-robot interface is shown in Figure 2. The proposed interface comprises of three primary components, namely environment mapping, pointing detection, and laser point detection, as well as a motion planner. The entire system is integrated with the Robot Operating System (ROS), where each component is implemented as an individual ROS node. All RGB-D data is collected in real-time through the

RGB-D camera mounted on the robot end effector, which is initially oriented towards the human operator.

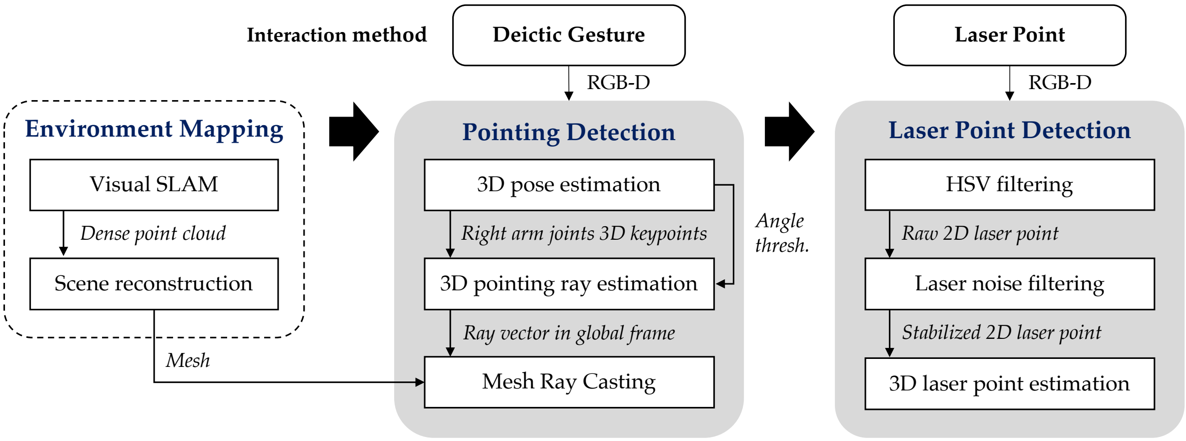

Figure 2. Workflow of the deictic gesture-based human-robot interface.

Environment mapping. Before receiving task specifications from the human operator, the robot conducts a preliminary exploration of its surrounding unknown environment autonomously to build a global 3D map. First, the robot end effector traverses its surroundings along a predetermined path to collect the 3D point cloud data of the environment. For point cloud data collection, we employ the RTAB-Map (Real-Time Appearance-Based Mapping) (Labbe and Michaud 2014), a well-known visual Simultaneous Localization and Mapping (vSLAM) algorithm. Next, the triangle meshes are created from the dense point cloud using Poission surface reconstruction method (Kazhdan et al. 2006). This pre-collected mesh model is then saved as a global 3D map to allow the human operator to command target location on the surface of the environment (Figure 1 (b)).

Pointing detection. Given the task specifications through the human operator's deictic gesture, the target location within the global 3D map is estimated. We obtain the 3D positions of the shoulder, elbow, and wrist joints through 3D human pose estimation method (Zimmermann et al. 2018) (Figure 1 (a)). If a deictic gesture is detected based on the predefined elbow joint angle threshold (Yoon et al. 2021), the deictic gesture ray vector is defined as the vector from the shoulder to the wrist joint. Next, assuming it is known that the gesture ray intersects with a triangle from mesh with index i, the intersection point P can be estimated as follows:

$$P = \mathbf{o} + \frac{\mathbf{n}_\mathrm{i} \cdot (\mathbf{v}_0 - \mathbf{o})}{\mathbf{n}_\mathrm{i} \cdot \mathbf{g}} \cdot \mathbf{g} \tag{1}$$

where $\mathbf{o}$ is the origin of the gesture ray, and $\mathbf{g}$ is the gesture ray vector, $\mathbf{n}_\mathrm{i}$ is the normal vector of the intersection triangle with index i, and $\mathbf{v}_0$ is one of the vertices of the intersection triangle. The intersection point P, represented in the world coordinate system, is regarded as the target location and used to direct the robot towards the detection pose.

Laser point detection. After the robot moves to the detection pose (Figure 1 (c)), the laser pointing technique is utilized to fine-tune the target location (Figure 1 (d)). The detection of laser points involves computing the image moment of the HSV image, followed by the conversion of the 2D laser point coordinates to 3D coordinates using the depth image and camera intrinsic. First, the input RGB image I_{RGB} is transformed into the HSV image I_{HSV}. To segment the red spots from the HSV image, the hue layer is limited to a range of [165 – 179], while the saturation layer is limited to [10 – 255], and the value layer to [200 – 255]. Furthermore, to distinguish the actual laser spot, a threshold is applied to the values of the segmented areas. The center of mass is then computed as follows:

$$M_{ij} = \sum_{x}\sum_{y} x^{i} y^{j} I_{HSV}(x, y) \tag{2}$$

$$x_c = \left\lfloor \frac{M_{10}}{M_{00}} \right\rfloor, y_c = \left\lfloor \frac{M_{01}}{M_{00}} \right\rfloor \tag{3}$$

where $I_{HSV}(x, y)$ is the thresholded image and M_{ij} is the $(i + j)^{th}$ order image moment of $I_{HSV}(x, y)$. The mass center yields the 2D laser point coordinates (x_c, y_c) in the image plane. To minimize the jitter and lag of the raw 2D laser point, One-Euro filter (Casiez et al. 2012) is applied. Next, 2D laser point coordinates are transformed into 3D space using the inverse projection for the pinhole camera model (Sprute et al. 2019):

$$X_c = \frac{x_c - c_x}{f_x} d, Y_c = \frac{y_c - c_y}{f_y} d, Z_c = d \tag{4}$$

where d is the depth value of the corresponding pixel in the depth image, f_x and f_y are the focal lengths of the camera, c_x and c_y are the principle coordinates of the camera, X_c, Y_c, and Z_c are the 3D laser point coordinates in the camera coordinate system. Finally, the 3D coordinates of the laser point from the camera coordinate system $L_c = (X_c, Y_c, Z_c)$ are transformed to the world coordinate system $L_w = (X_w, Y_w, Z_w)$, where $L_w = [R|t]L_c$, R and t are the rotation matrix and translation vector, respectively, obtained from the camera extrinsic. The intrinsic and extrinsic camera parameters are acquired through a calibration process.

Motion planning. Given the target location using gesture and laser pointer, the robotic system performs motion planning. To compute the desired end effector pose, we assume that the mobile base is set at the desired position and that if the estimated 3D coordinates of the specified location are within the robot's reachable space, the target is located on the ceiling, parallel to the ground. Moreover, we set an offset of 30 cm in the direction of the normal vector from the target to avoid collision with the workpiece. For motion planning, we utilize the Stochastic Trajectory Optimization for Motion Planning (STOMP) algorithm, owing to its demonstrated ability to generate smooth trajectories in real-time (Choi et al. 2022).

EXPERIMENTS

To evaluate the performance of the deictic gesture-based human-robot interface, we designed a drywall cutting experiment in an environment with unfinished concrete walls. The hardware configuration consisted of a mobile base (table lift), a KUKA KR 6 R 900 6-DoF manipulator, and an end-effector equipped with a Makita 3706 drywall cut-out tool and an Intel RealSense D435 RGB-D camera. The experimental task involved performing a drywall cutting to make openings for a 10" w X 6" h steel air grill. The objective of human-robot interaction was to deliver target locations to the robot, as shown in Figure 1. This involved determining the four corner points of the opening and delivering the locations using human-robot interface. There were no experimental instructions on which order to deliver the four locations. For each corner point, participants were required to notify an experimenter once they had moved the robot to the desired pose, thus allowing the experimenter to record the position of the tool center point. The experiment was designed as a within-subjects experiment, varying the interaction methods: our proposed deictic gesture-based interface and the joystick interface. The joystick interface was chosen as the baseline modality for comparison, given its conventional acceptance in construction robotics. In the case of the joystick interface, participants were interfaced with an Xbox 360 controller, which mapped the three translational and three rotational dimensions of the end effector into two analog sticks and buttons.

We recruited six participants (five male and one female) for the experiment. The participants were all right-handed and aged 25.3 on average (SD = 2.45). Each participant was asked to perform two task trials, randomized across two interaction methods. To measure the performance of our interface, we measured task effectiveness using the pairwise distance error for each corner point and task efficiency using the time needed for a user to complete a task instruction. Additionally, to obtain a subjective assessment of the user experience, we conducted a post-experiment survey, wherein the system usability was evaluated using the System Usability Scale (SUS) and cognitive load was evaluated using the adapted NASA Task Load Index (NASA-TLX) questionnaire.

RESULTS

The results of task performance and usability are presented in Figure 3. Overall, our interface was able to accurately estimate the task locations, complete the task efficiently, and provide acceptable usability. Participants, on average, completed the tasks with a maximum distance error of 13 mm and within 200 seconds. The results suggest that, although certain tendencies were discerned, no statistically significant difference was found between the two interaction methods with respect to task effectiveness, efficiency, and system usability. Moreover, our interface has the potential to adapt to individual differences and mitigate the effects on task performance, resulting in a more robust and reliable performance across users. It was revealed that there were no significant differences between the six subjects with regards to the measures of task performance (distance error: $F(1,10) = 0.254, p = 0.625, \eta_{\mathrm{p}}^2 = 0.025$; completion time: $F(1,10) = 0.61, p = 0.453, \eta_{\mathrm{p}}^2 = 0.057$).

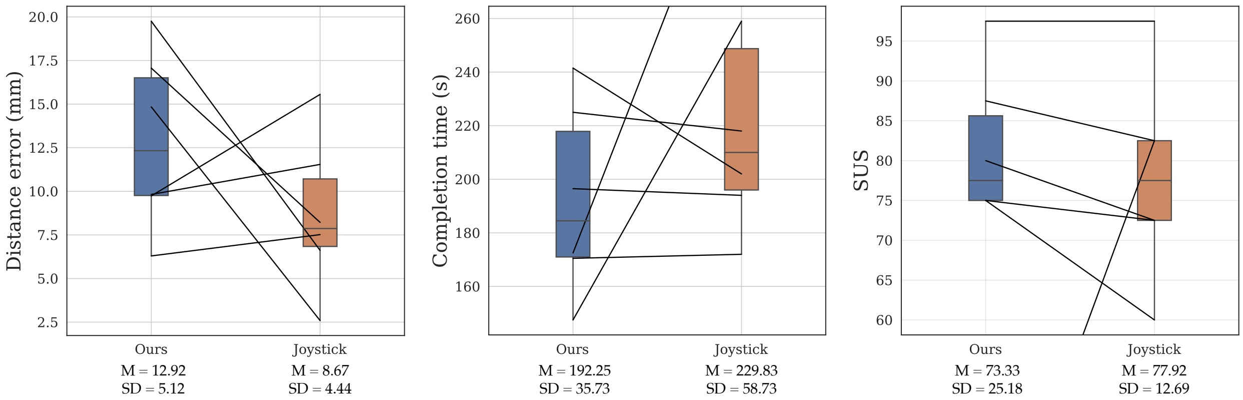

Figure 3. Results of distance error, completion time, and SUS by interaction method.

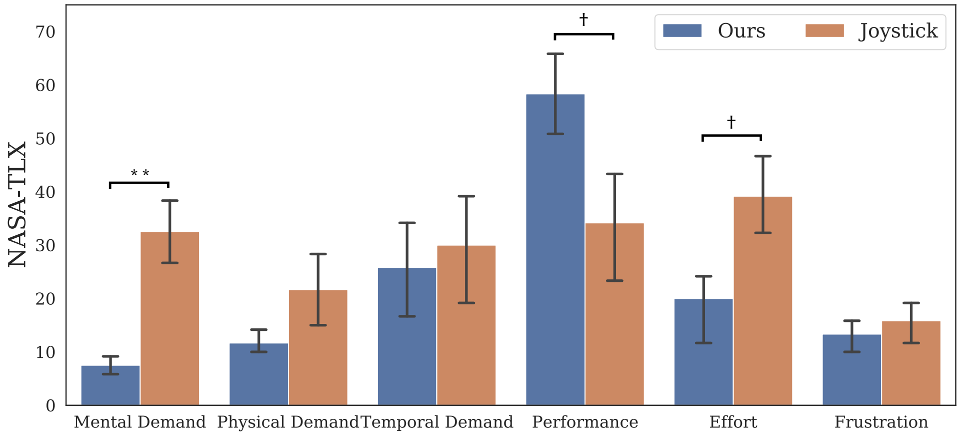

Figure 4. NASA-TLX scores by interaction method. **p<0.01; †p<0.1.

The results of NASA-TLX questionnaires are shown in Figure 4. The results indicate that on average, our interface demonstrated better scores in the dimensions of mental demand, physical demand, temporal demand, effort, and frustration. Conversely, joystick recorded higher scores in performance. Moreover, our interface showed a significant difference in mental demand ($t(5) = -4.038, p = 0.0099$) and a marginal difference in effort levels compared to joystick ($t(5) = -2.253, p = 0.07$).

CONCLUSION

In this study, we developed and tested a novel interface that can handle deictic gesture-based task specification for giving four drywall cutting locations to a construction robot. The developed interface is capable of motion planning and execution when provided with task locations, allowing human operators to focus on higher-level aspects of the task, rather than micromanaging the robot's every move. To the best of our knowledge, deictic gesture-based human-robot interaction methods have not been widely applied in construction robotics. Although deictic gesture-based interaction methods have primarily been used in other domains, such as manufacturing (Neto et al. 2019) and logistics (Guzzi et al. 2022), adapting such techniques to the construction domain presents several challenges, including differences in the physical environment and level of precision. Despite these challenges, the results indicate that the implementation of deictic gestures to human-robot interface led to a statistically significant reduction in mental demand and effort levels among the participant, compared to the conventional joystick interface. Moreover, our deictic gesture-based interface demonstrated task performance comparable in accuracy and efficiency to that achieved using joystick. These investigations indicate an opportunity to enhance intuitive and accurate operation of the semi-autonomous construction robots, in situations where communication about human decisions on task specifications are required.

The proposed deictic gesture-based human-robot interface has limitations, as it was evaluated with only a single human operator. In order to enhance the applicability of the proposed interface in the real-world construction sites with multiple workers, future improvements can be achieved through the integration of learning-based algorithms capable of identifying the appropriate timing for initiating interactions and accurately discerning the worker providing task specifications.

ACKNOWLEDGEMENTS

This study was supported by the New Faculty Startup Fund and the Institute of Construction and Environmental Engineering (ICEE) at Seoul National University (SNU). Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of ICEE or SNU.

REFERENCES

Alibali, M. W. 2005. "Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information." *Spatial Cognition and Computation*. Lawrence Erlbaum Associates, Inc.

Casiez, G., N. Roussel, and D. Vogel. 2012. "1 € filter: a simple speed-based low-pass filter for noisy input in interactive systems." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, 2527–2530. New York, NY, USA: Association for Computing Machinery.

Chen, X., X. Huang, Y. Wang, and X. Gao. 2020. "Combination of Augmented Reality Based Brain- Computer Interface and Computer Vision for High-Level Control of a Robotic Arm." *IEEE Trans. Neural Syst. Rehabil. Eng.*, 28 (12): 3140–3147. https://doi.org/10.1109/TNSRE.2020.3038209.

Choi, A., M. K. Jawed, and J. Joo. 2022. "Preemptive Motion Planning for Human-to-Robot Indirect Placement Handovers." *arXiv [cs.RO]*.

Feng, C., N. Fredricks, and V. R. Kamat. 2013. "Human-robot integration for pose estimation and semi-autonomous navigation on unstructured construction sites." *Proceedings of the 30th International Symposium on Automation and Robotics in Construction and Mining (ISARC 2013): Building the Future in Automation and Robotics*. International Association for Automation and Robotics in Construction (IAARC).

Guzzi, J., G. Abbate, A. Paolillo, and A. Giusti. 2022. "Interacting with a Conveyor Belt in Virtual Reality using Pointing Gestures." *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '22, 1194–1195. IEEE Press.

Han, J., G. Ajaykumar, Z. Li, and C.-M. Huang. 2020. "Structuring Human-Robot Interactions via Interaction Conventions." *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 341–348. ieeexplore.ieee.org.

Kazhdan, M., M. Bolitho, and H. Hoppe. 2006. "Poisson surface reconstruction." *Proceedings of the fourth Eurographics symposium on Geometry processing*. cse.iitd.ac.in.

Labbe, M., and F. Michaud. 2014. "Online global loop closure detection for large-scale multi-session graph-based SLAM." *Rep. U. S.* ieeexplore.ieee.org.

Liang Ci-Jyun, Wang Xi, Kamat Vineet R., and Menassa Carol C. 2021. "Human–Robot Collaboration in Construction: Classification and Research Trends." *J. Constr. Eng. Manage.*, 147 (10): 03121006. American Society of Civil Engineers. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002154.

Liu, Y., M. Habibnezhad, and H. Jebelli. 2021a. "Brain-computer interface for hands-free teleoperation of construction robots." *Autom. Constr.*, 123: 103523. Elsevier. https://doi.org/10.1016/j.autcon.2020.103523.

Liu, Y., M. Habibnezhad, and H. Jebelli. 2021b. “Brainwave-driven human-robot collaboration in construction.” *Autom. Constr.*, 124: 103556. Elsevier B.V. https://doi.org/10.1016/j.autcon.2021.103556.

Losey, D. P., H. J. Jeon, M. Li, K. Srinivasan, A. Mandlekar, A. Garg, J. Bohg, and D. Sadigh. 2022. “Learning latent actions to control assistive robots.” *Auton. Robots*, 46 (1): 115–147. https://doi.org/10.1007/s10514-021-10005-w.

Neto, P., M. Simão, N. Mendes, and M. Safeea. 2019. “Gesture-based human-robot interaction for human assistance in manufacturing.” *Int. J. Adv. Manuf. Technol.*, 101 (1): 119–135. https://doi.org/10.1007/s00170-018-2788-x.

Sprute, D., K. Tönnies, and M. König. 2019. “This Far, No Further: Introducing Virtual Borders to Mobile Robots Using a Laser Pointer.” *2019 Third IEEE International Conference on Robotic Computing (IRC)*, 403–408. Institute of Electrical and Electronics Engineers Inc.

Stoddard, B., M. Cravetz, T. Player, and H. Knight. 2022. “A Haptic Multimodal Interface with Abstract Controls for Semi-Autonomous Manipulation.” *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, HRI ’22, 1206–1207. IEEE Press.

Villani, V., F. Pini, F. Leali, and C. Secchi. 2018. “Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications.” *Mechatronics* , 55: 248–266. https://doi.org/10.1016/j.mechatronics.2018.02.009.

Wang, X., C.-J. Liang, C. C. Menassa, and V. R. Kamat. 2021. “Interactive and immersive process-level digital twin for collaborative human–robot construction work.” *J. Comput. Civ. Eng.*, 35 (6): 04021023. American Society of Civil Engineers (ASCE). https://doi.org/10.1061/(asce)cp.1943-5487.0000988.

Wang, X., D. Veeramani, and Z. Zhu. 2023. “Wearable Sensors-Based Hand Gesture Recognition for Human–Robot Collaboration in Construction.” *IEEE Sens. J.*, 23 (1): 495–505. https://doi.org/10.1109/JSEN.2022.3222801.

Wang, X., and Z. Zhu. 2021. “Vision–based framework for automatic interpretation of construction workers’ hand gestures.” *Autom. Constr.*, 130: 103872. Elsevier B.V. https://doi.org/10.1016/j.autcon.2021.103872.

Yoon, S., Y. Kim, C. Ahn, and M. Park. 2021. “Challenges in deictic gesture-based spatial referencing for human-robot interaction in construction.” *Proceedings of the 38th International Symposium on Automation and Robotics in Construction (ISARC)*, 491–497. Waterloo, Canada, Waterloo: International Association for Automation and Robotics in Construction (IAARC).

Zhu, Q., J. Du, Y. Shi, and P. Wei. 2021. “Neurobehavioral assessment of force feedback simulation in industrial robotic teleoperation.” *Autom. Constr.*, 126: 103674. https://doi.org/10.1016/j.autcon.2021.103674.

Zhu Qi, Zhou Tianyu, Xia Pengxiang, and Du Jing. 2022. “Robot Planning for Active Collision Avoidance in Modular Construction: Pipe Skids Example.” *J. Constr. Eng. Manage.*, 148 (10): 04022114. American Society of Civil Engineers. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002374.

Zimmermann, C., T. Welschehold, C. Dornhege, W. Burgard, and T. Brox. 2018. “3D Human Pose Estimation in RGBD Images for Robotic Task Learning.” *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 1986–1992. IEEE Press.