Master of Science in Engineering

Challenges in Spatial Communication Using Deictic Gesture for Human-Robot Collaboration in Construction

February, 2022

Department of Architecture & Architectural Engineering

The Graduate School

Seoul National University

Sungboo Yoon

Challenges in Spatial Communication Using Deictic Gesture for Human-Robot Collaboration in Construction

by

Sungboo Yoon

A thesis submitted in partial fulfillment

of the requirements for the degree of

Master of Science in Engineering

Seoul National University

2022

Challenges in Spatial Communication Using Deictic Gesture for Human-Robot Collaboration in Construction

December, 2021

Approved by Dissertation Committee:

Seokho Chi

Moonseo Park

Changbum Ahn

Challenges in Spatial Communication Using Deictic Gesture for Human-Robot Collaboration in Construction

지도교수 박 문 서

이 논문을 공학석사 학위논문으로 제출함

2022년 2월

서울대학교 대학원

건 축 학 과

윤 성 부

윤성부의 공학석사 학위논문을 인준함

2021年 12月

위 원 장	지 석 호	(인)
부 위 원 장	박 문 서	(인)
위 원	안 창 범	(인)

Abstract

Challenges in Spatial Communication Using Deictic Gesture for Human-Robot Collaboration in Construction

Sungboo Yoon
Department of Architecture
The Graduate School
Seoul National University

Despite the recent advances in AI-based perceptual and manipulative abilities in robotics, construction robots are often interfered by other operations when deployed onsite due to the complex work sequences and cluttered working environments. Thus, in order to frequently adapt and change their work flows and methods in these situations, the robots require humans' improvisations. Along with the emerging needs for natural human-robot interaction, previous works have shown feasibility of the gesture-based human-robot interface for exchanging the spatial information, which is fundamental factors for in-situ improvisations. However, two critical gaps in evaluation exists: (1) Even though construction tasks are mostly performed in a large-scale, unstructured, and cluttered environment, previous works limited the application of the gesture-based interface for only for short-distance applications; (2) previous studies mainly focused on the evaluation of the robot performance with-

out considering various positional relationships between the robot and the human pointer. To this end, this study aims to identify the challenges associated with the deictic gesture-based target panel referencing in a large-scale environment to evaluate the potential of the deictic gestures in panel installation tasks in construction. We selectively overviewed the spatial referring abilities of both human and robot through two experiments of panel pointing with different level of precision. In the first experiment, the results presented a significant performance drop in the ceiling and angled targets, while the overall recognition accuracy was acceptable (0.864). In the second experiment, the recognition accuracy of the robot dropped by 30.18% compared to the first experiment. The results of the second experiment presented that humans tend to have consistency in the spatial referring abilities, while the robots showed statistically significant difference pertaining to the pointer-robot distances. The results also showed the performance enhancement of the target deviation by 73.49% through pointing calibration. This paper contributes to the body of knowledge by evaluating the deictic gesture-based spatial referring ability of the robot in a large-scale human-robot collaborative environment, furthering the application of the deictic gestures and robotics to panel installation work.

Keyword : Spatial Communication; Human-Robot Collaboration; Deictic Gestures

Student Number : 2020-26076

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1. Motivation

Robotic technologies are envisioned as a promising alternative for construction job sites which constantly suffer from stagnant productivity and shortage of skilled workers [1]. The McKinsey Global Institute (MGI) survey revealed that construction productivity has fallen by half since the 1960s [2]. The construction industry has also experienced the loss of 600,000 skilled workers during the2008 recession. Recent global pandemic, Coronavirus disease 2019 (Covid-19) has exacerbated the exodus, according to the report from Forbes Technology Council [3]. Meanwhile, recent advances in artificial intelligence (AI) and AI-based perceptual and manipulative abilities in robotics have led to an un-precedented increase in the robots' performance enough to perform on-site construction tasks [4,5]. Examples include semi-automated brick-laying robots [6], rebar-tying robots [7], site layout robots [8,9], and 3d printing robots [10,11]. These robots were first introduced in structured and repetitive construction tasks, benefitting from their high precision in conducting tasks and detecting minor deviations [12].

Nevertheless, such robots have been limitedly employed in real con-struction sites due to the unstructured and dynamic nature of con-struction environments [13]. When deployed onsite, construction robots

are often interfered with by other operations, workers, and equipment due to the complex work sequences and cluttered working environments. Thus, in order to frequently adapt and change their work flows and methods in these situations, the robots require humans' improvisations (adaptive decisions based on perceptual understanding and previous work experiences), making the communication between human and robot inevitable [12]. In this context, although some construction tasks have possibilities to be performed autonomously by robots, they still require human involvement to deliver in-situ improvisations.

However, current methods for human-robot communication in construction have several challenges. One of the most common methods is robot teleoperation using joysticks or control pads. However, it suffers from time delays in actuation and reduction of accuracy [14]. The communication method which makes use of BIM models has been widely used as well [6]. In this method, the user manually models the BIM model based on a predetermined work plan and transmits the model to the robot. Despite the enhanced accuracy of this method, it faces some challenges in practice due to the predefined model, because in order to handle the in-situ variations or uncertainties, the user has to manually update the BIM model during the operation. Scan-to-BIM technique has been applied recently to reduce the effort of manual BIM modeling. This technique synchronizes up-to-date data to the project cloud through 3D scanning. However, as-built BIM modeling through Scan-to-BIM technique is computationally expensive and requires the workers to constantly update new improvisations during the data synchronization, mak-

ing it unrealistic for real-time performance.

1.2. Problem Statement

Along with the emerging needs for natural human-robot interaction, previous works have shown feasibility of the gesture-based human-robot interface for exchanging the spatial information, which are fundamental factors for in-situ improvisations (Table 1-1). Previous applications include controlling a robot to change its position [14], referring to a target object [15-17] or area [18], and indicating target point on a wall surface [19-22]. Compared with other forms of interaction, human gestures are the most ideal way because they can express rich semantics, are easy to identify [23]. They are especially beneficial for construction applications because they are more natural than the other methods (i.e., wearable method) and do not require additional devices such as control pads.

However, deictic gestures are known to be inherently imprecise and ambiguous for both humans and robots. It is especially challenging for a robot to interpret the exact referent in an unstructured construction environment. Even though construction tasks are mostly performed in a large-scale, unstructured, and cluttered environment, previous works limited the application of the gesture-based interface for spatial information exchange only for short-distance applications such as tables [24] and head-up displays (HUDs) [25], where the distance to the targets are less than 2m and for known objects where one or more features (e.g., colors, shapes) of the objects are predefined.

Moreover, the deictic gestures are also known to be challenging for a human observer to interpret the exact referent of the human pointer [20,26]. While previous studies mainly focused on the evaluation of the robot performance at a fixed position, a significant knowledge gap remains on the relative spatial referring ability with various positional relationships between the robot and the human pointing subject.

Table 1-1. Summary of Previous Studies Using Deictic Gestures for Spatial Referencing.

Author, Year	Input feature (Method)	Application (Task type)	Layout (Distance)
Jirak et al., 2021	[2D Vision] Deictic gesture, fingertip (contour)	Collaborative task with the NICO robot at fixed position. 95 object configurations	3 Objects (0.4m~1m)
Mayer et al., 2020	[3D Vision] Deictic gesture (OptiTrack motion capture)	Multidirectional task in CVE. Operator in the middle of the experimental environment at fixed position	16x5 grid spanning the whole 360º rotation (0.46~1.22cm)
Jevtić et al., 2019	[3D Vision] Deictic gesture, speech	Collaborative dressing task at fixed position	4 shoes (~1m)
Williams et al., 2019	[3D Vision] Deictic gesture, speech	Object selection task with allocentric mixed reality at fixed position	19 Objects (~2m)
Obo et al., 2018	[3D Vision] Deictic gesture, verbal cues (Kinect internal software)	3 object selection tasks at 4 different locations at fixed position	4 targets (~2m)
Brand et al., 2016	[3D Vision] Deictic gesture (Leap Motion)	HUD pointing task at fixed position. 15 iterations	HUD: 24.5x5.2cm 3 and 4 segments (1m)

1.3. Research Objectives

To this end, this study evaluates the performance of the spatial communication interface using deictic gestures for human-robot collaboration in an experimental task simulating a construction environment. The latest deictic gesture-based spatial communication method was adopted, which utilizes human pose estimation with deep learning to detect the human body joints. We selected two common workspaces in construction, wall and ceiling, and conducted multiple panel pointing tasks with two different levels of precision. Finally, we evaluated the accuracy of the human-robot interface when estimating targets referenced by deictic gestures and identified some challenges and potential of the deictic gestures in spatial communication for panel installation in construction.

Chapter 2. Background

2.1. Deictic Gesture-Based Spatial Referencing

Deictic gestures are often referred to as “pointing gestures”, typically performed by extending the arm and the index finger [19,27]. In general, people often use pointing gestures to deliver spatial information to others. In other words, deictic gestures are fundamental to direct others’attention to objects and help develop a mutual understanding of objects in space [28].

Deictic gesture-based spatial referencing has been explored substantially in previous works for developing and evaluating various spatial referencing models according to task requirements. This large body of work shares the same purpose: to solve the problem of interpreting deictic gestures in order to map the referent in the environment that the user wants to indicate [29]. Tölgyessy et al. [30] presented a spatial referencing method navigating a mobile robot to an endpoint marker on the ground floor defined by a pointing gesture of a human operator. The suggested method shows the precise positioning of all the entities included in the interaction in 3D space. Jevtić et al. [20] evaluated humans’ referencing accuracy when interpreting deictic gestures for pointing the targets positioned horizontally on the wall. However, they only measured the performance in a collaborative virtual environment (CVE).

While previous works showed acceptable performance of the deictic gesture-based spatial referencing for short-distance applications, limited applications in large-scale environments need to be further evaluated.

2.2. Deictic Gesture Recognition Methods

Deictic gesture-based spatial referencing aims to exchange accurate spatial information through deictic gestures. Therefore, deictic gesture recognition has a considerable impact on the final referencing results.

Two main approaches for deictic gesture recognition have been proposed in the literature. One is a wearable sensor-based approach. This approach attempts to recognize deictic gesture by analyzing the electrical muscle stimulation (EMS) from electromyography (EMG) generated during the muscle activity [31,32], the change in measures from inertial measurement units (IMUs) [29,33], and the posture and motion data from data gloves [34]. However, although wearable sensors have the benefit of direct acquisition of the spatial posture of the pointing arm, they often require connection to a data acquisition (DAQ) device, thus restricting the applicability of this method outside of a controlled environment [35,36].

Meanwhile, recent advances in computer vision technologies have brought vision-based approaches to mainstream deictic gesture recognition. Vision-based deictic gesture recognition does not require users any additional devices and only employs their pointing arms within the camera angle. Earlier approaches detected gestures through the visual features (i.e., skin-color blobs) collected from monocular cameras (e.g., RGB or infrared camera) and binocular cameras [37]. Recent works on vision-based approaches have focused on the implementation

of RGB-D cameras. Owing to the ability to augment the RGB image with depth information, RGB-D cameras are frequently being adopted in vision-based approaches.

In a vision-based approach, the deictic gesture is defined based on the relationships among the body joints. Three main models for estimating the pointing direction were developed, as illustrated in Figure 2-1 [30,38]:

- *Elbow-wrist ray model* assumes that the pointing direction is defined by a vector connecting the elbow and the wrist (hand) of the pointing arm.
- *Head-wrist ray model* assumes that the pointing direction is defined by a vector connecting the head and the wrist (hand) of the pointing arm.
- *Shoulder-wrist ray model* assumes that the pointing direction is defined by a vector connecting the shoulder and the wrist (hand) of the pointing arm.

This work evaluates the performance of the spatial referencing method using a shoulder-wrist ray model, because the elbow-wrist ray model gives lower accuracy in large-scale environments and the head-wrist ray model has potential problems associated with the occlusion in pose estimation (i.e., safety helmets) [19].

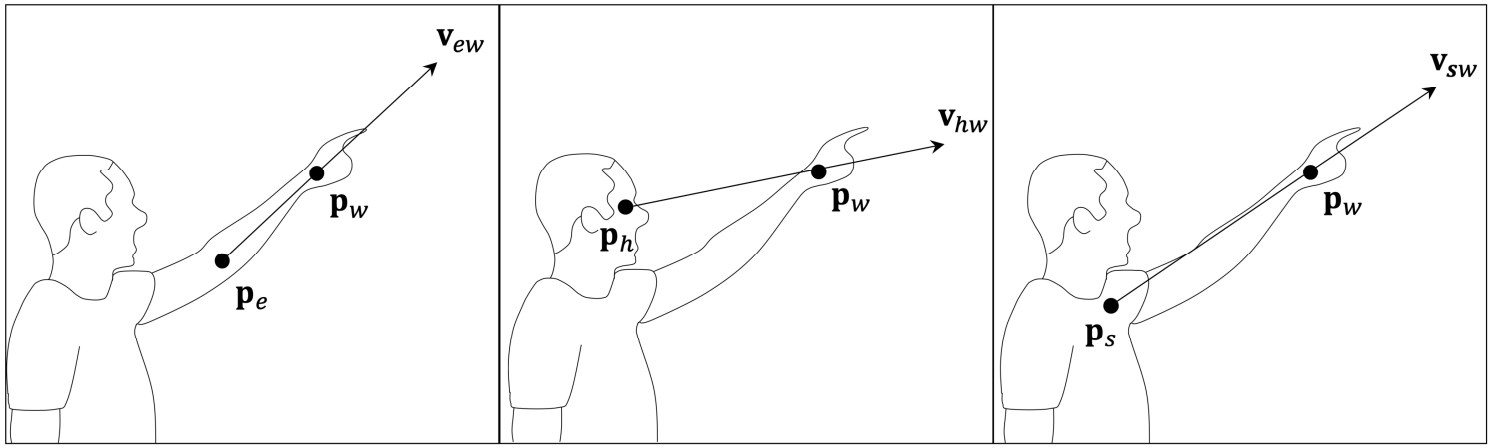

Figure 2-1. Illustration of the three models for the pointing direction estimation.

Chapter 3. Methodology

This chapter proposes the deictic gesture-based spatial referencing method. To evaluate the spatial referring performance of the collaborative robot, this work adopts current deictic gesture-based method which utilizes deep learning-based human pose estimation, as illustrated in Figure 3-2. The detection of the deictic gesture is performed based on the 3D human skeletal data extracted from the RGB and depth images. To estimate the human skeletal data, this work employs OpenPose [39] library, a real-time human pose estimation system. The library (BODY-25 model) detects 25 human body joints from each RGB image frame in 2-D coordinates. The 2-D pose coordinates are then projected to corresponding 3-D pose coordinates using the depth and camera calibration information [40]. After the 3-D pose coordinates are transformed into world coordinates to map the 3-D position in world space, pointing gesture recognition is performed (Section 3.1). If the pointing gesture is recognized, the pointing position is estimated with the ray-plane intersection. Finally, the system estimates the pointing target panel with the estimated pointing position (Section 3.2). The whole process of the system is implemented through multiple nodes and integrated with the Robot Operating System (ROS) for online operation, as illustrated in Figure 3-1. Moreover, this work proposes a pointing

calibration method to enhance the accuracy of spatial referencing, which is covered in Section 3.3.

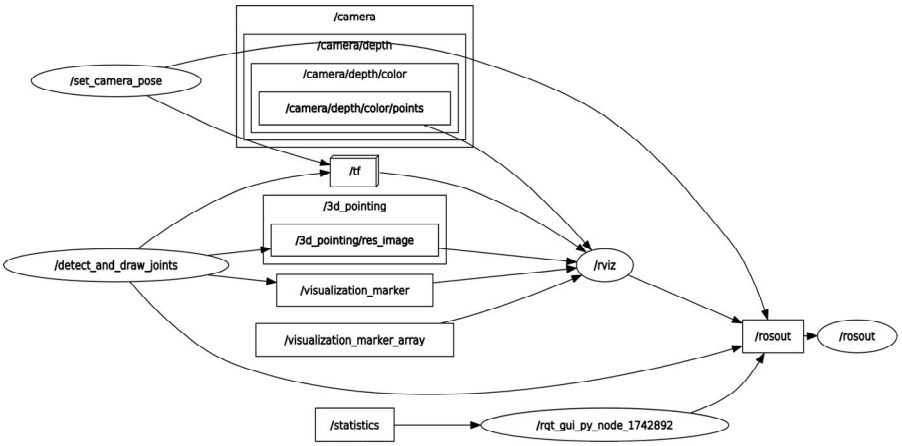

Figure 3-1. ROS node graph of the system.

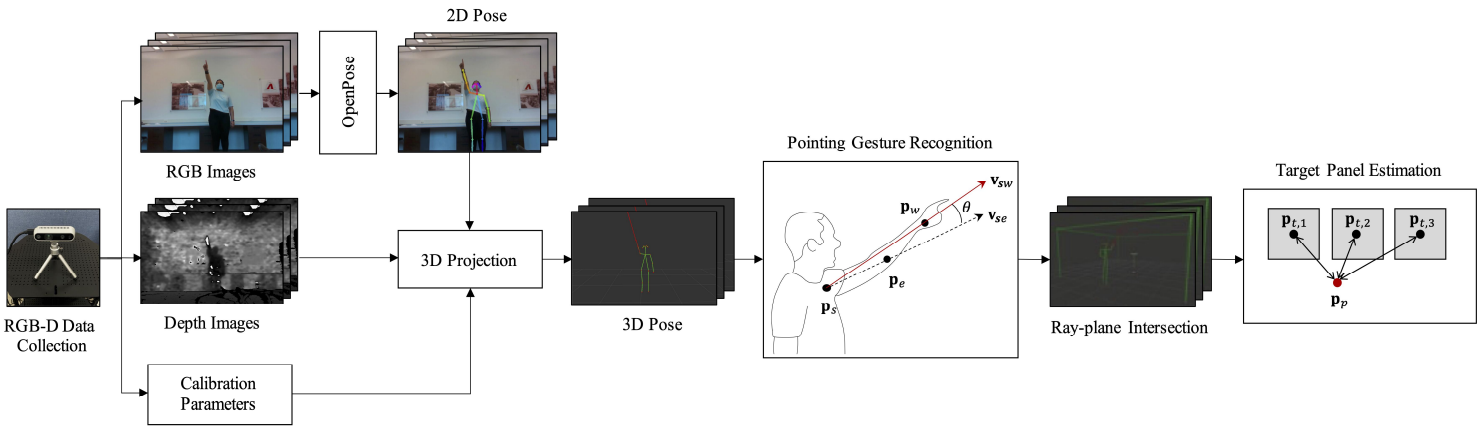

Figure 3-2. Process diagram of the deictic gesture-based spatial referencing method.

3.1. Pointing Gesture Recognition

As mentioned in Section 2, this work uses a shoulder-wrist ray model for estimating the pointing direction. Therefore, this work focuses on the position of the shoulder (1), elbow (2), and the wrist (3) joints.

$$\mathbf{p}_s = (x_s, y_s, z_s) \tag{1}$$

$$\mathbf{p}_e = (x_e, y_e, z_e) \tag{2}$$

$$\mathbf{p}_w = (x_w, y_w, z_w) \tag{3}$$

We use wrist position instead of fingers, considering the computation efficiency for further on-site applications. Given the position of the three body joints, the elbow joint angle θ is defined by:

$$\cos\theta = \frac{\mathbf{v}_{se} \cdot \mathbf{v}_{sw}}{|\mathbf{v}_{se}||\mathbf{v}_{sw}|} \tag{4}$$

$$\mathbf{v}_{se} = \mathbf{p}_e - \mathbf{p}_s \tag{5}$$

$$\mathbf{v}_{sw} = \mathbf{p}_w - \mathbf{p}_s \tag{6}$$

where $\mathbf{v}_{se}$ is the vector from the shoulder to the elbow joint (5) and $\mathbf{v}_{sw}$ is the vector from the shoulder to the wrist joint (6). If θ is below a predefined angle, the system assumes that the person is stretching their arm for pointing.

Meanwhile, before pointing target estimation, the first and fourth quartile of the frames that were initially detected as “pointing” were excluded to have a buffer for the stabilization of the pointing motion [22], as illustrated in Figure 3-3.

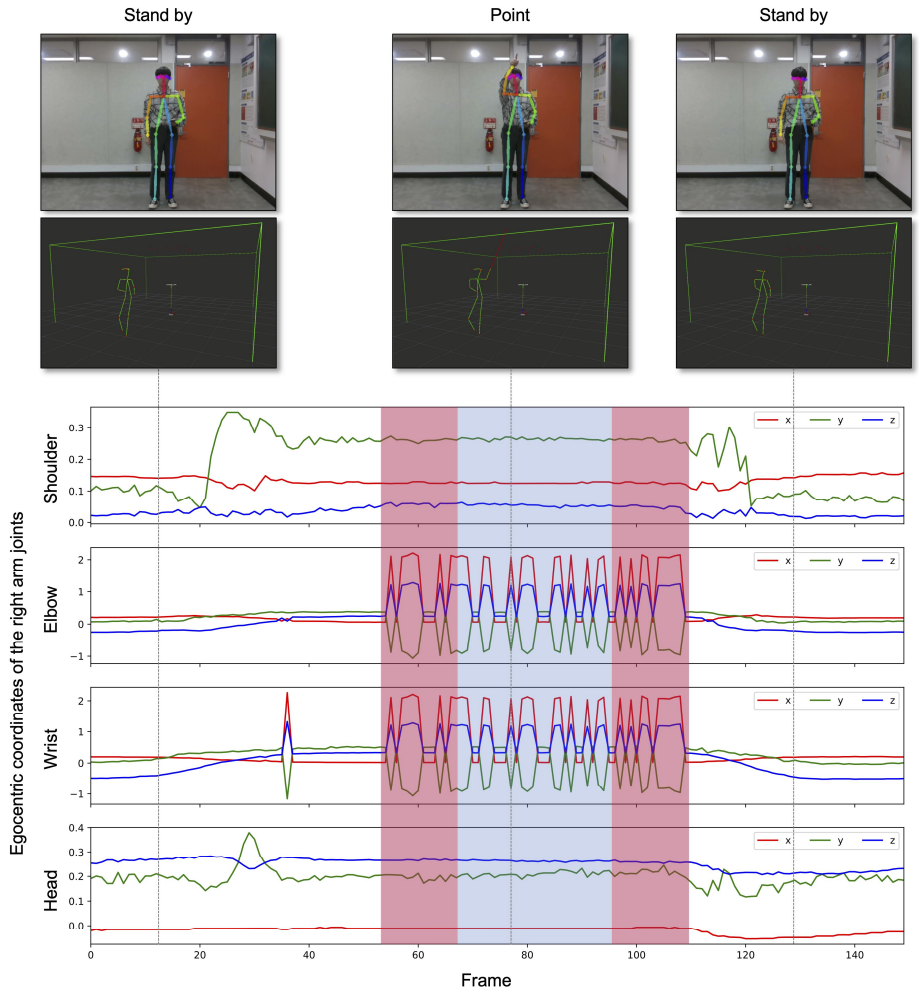

Figure 3-3. Egocentric (human body relative) 3-D coordinates of the head and the three right arm joints (shoulder, elbow, and wrist). Red shaded area represents excluded frames and the blue shaded area represents the candidate frames used for pointing target estimation.

3.2. Pointing Target Estimation

We first estimate the pointing target position with the stabilized candidate frames in order to predict the target panel. As mentioned in Section 3.1, the pointing direction is defined by a straight line starting from the shoulder to the wrist joint:

$$\mathbf{s} = \mathbf{p}_s + \lambda(\mathbf{p}_w - \mathbf{p}_s), \ \lambda \in \mathbb{R} \tag{7}$$

For the targets on a ceiling, the targets are parallel to the floor at a constant height of the ceiling h. Therefore, the pointing target position $\mathbf{p}_p$ is calculated as follows:

$$\mathbf{p}_p = (x_p, y_p, z_p) \tag{8}$$

$$x_p = x_s + \frac{h - z_s}{z_w - z_w}(x_w - x_s) \tag{9}$$

$$y_p = y_s + \frac{h - z_s}{z_w - z_w}(y_w - y_s) \tag{10}$$

$$z_p = h \tag{11}$$

Meanwhile, for the targets on a wall, the targets are parallel to the wall at a constant distance d. Thus, akin to ceiling, the pointing target position $\mathbf{p}_p$ in this case is computed as:

$$x_p = d \tag{12}$$

$$y_p = y_s + \frac{h - x_s}{x_w - x_w}(y_w - y_s) \tag{13}$$

$$z_p = z_s + \frac{h - x_s}{x_w - x_w}(z_w - z_s) \tag{14}$$

The target panel is then predicted using the pointing target position.

Let $\mathbf{p}_{t,i}$ be the center point of the target panel index $i \in \mathbb{R}\{1,2,3,\cdots,n\}$, where $n \in \mathbb{N}$. A panel with the closest Euclidean distance from the center point is selected as a target panel i_p.

$$i_p = \underset{i}{\operatorname{argmin}}(|\mathbf{p}_{\hat{p}} - \mathbf{p}_{t,i}|) \tag{15}$$

3.3. Pointing Calibration

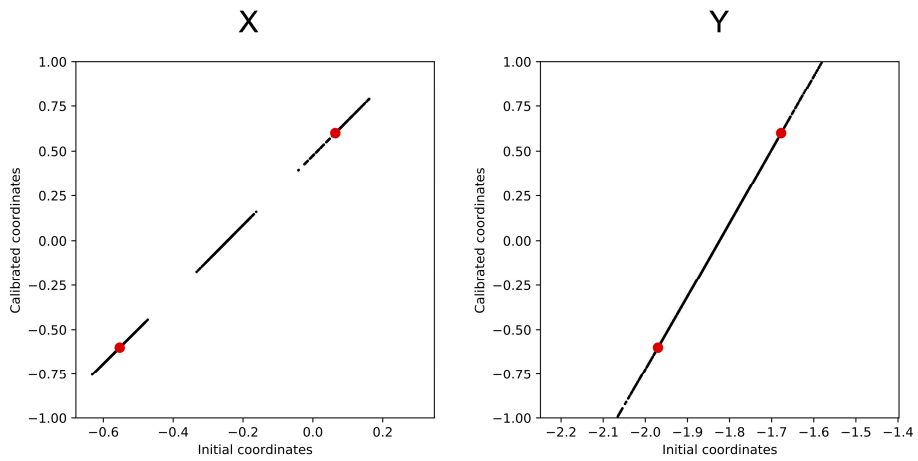

Figure 3-4. Pointing calibration through linear regression. Red dots represent the external points and the black dots represent the collected data points.

This work performs pointing calibration before the main experiments. Pointing gestures vary in motion and accuracy depending upon the subjects. Therefore, the estimated points should be calibrated to compensate for the variations in the subjects' pointing gestures [17].

In the calibration phase, the subjects were asked to point at the external panels. The system estimates the pointing position of the external panels through pointing gesture recognition and pointing target estimation. The mean values of each x and y coordinate are then calculated using the pointing positions. Finally, the parameters of the linear fitting function are computed with the initial mean coordinates.

In the experiment, the linear regression model was applied for collected and estimated data points. Figure 3-4 shows the initial and calibrated coordinates with the external target points.

Chapter 4. Low-Level Panel Pointing Tasks

In Chapter 4 and 5, two experiments are presented to evaluate the panel referring abilities of the robot and human in a large-scale environment.

In Chapter 4 (**E1**), this work evaluates the panel referring ability of the robot with panel pointing tasks which require low-level precision. In other words, this work performs pointing experiments based on large panel targets that are located one-way on the wall and ceiling surface, which are the most basic and common workspaces in construction.

In Chapter 5 (**E2**), this work evaluates the panel referring abilities of both the human and robot observers in the human-human interaction and human-robot interaction respectively, with panel pointing tasks which require high-level precision. In this case, this work performs pointing experiments based on small panel targets that are located two-way on the ceiling surface. We identified the challenges related to the deictic gesture-based spatial referencing in a large-scale environment and with various positional relationships of the human and robot.

4.1. Experimental Setup

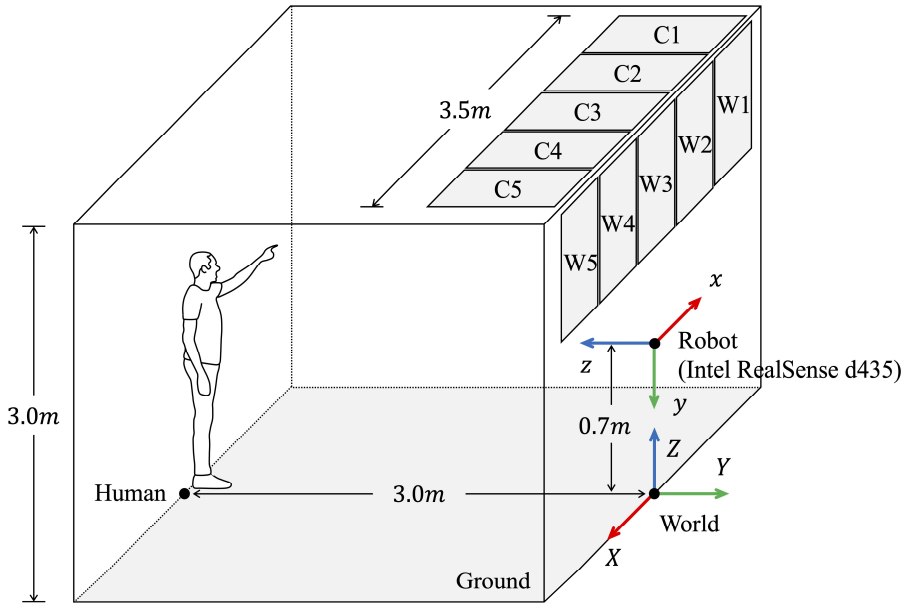

Figure 4-1. Illustration of the experimental environment for the first experiment (**E1**).

The experimental setup for the first experiment (**E1**) is depicted in Figure 4-1. We use Intel RealSense™ Depth Camera D435 for data collection. The operational and module specifications of the camera are shown in Table 4-1. The RGB and depth images are simultaneously captured at a frame rate of up to 30 fps and with an image resolution of 640 x 480 pixels and 840 x 480 pixels, respectively. It is installed at the position of the robot facing participants, at the height of 0.7m and the pointing subject is located at the position of the human, 3.0m away from the camera, as shown in Figure 4-1.

Five target panels with an equal size of 0.7 x 0.7m are located side

by side on both ceiling and wall.

Table 4-1. Specifications of Intel RealSense™ Depth Camera D435.

Specification	Value
Dimensions (L x D x H)	90 mm x 25 mm x 25 mm
Operating range (Min-Max)	~.11m - 10m
Depth accuracy	< 2% at 2m
Depth resolution and frame rate	Up to 1280 x 720 / 90 fps
Depth Field of View (FOV)	87° × 58°
RGB resolution and frame rate	Up to 1920 x 1080 / 30 fps
RGB Field of View (FOV)	69° × 42°
Depth technology	Stereoscopic

4.2. Procedure

Four participants (two males and two females) were recruited to perform the pointing tasks (Figure 4-2). Each participant performed two experiments, 15 iterations for each experiment. A single iteration consists of 10 pointing trials: five ceiling panels (from C1 to C5) and five wall panels (from W1 to W5), in sequential order. In sum, 2 x 15 x 10 = 300 trials for each participant were obtained.

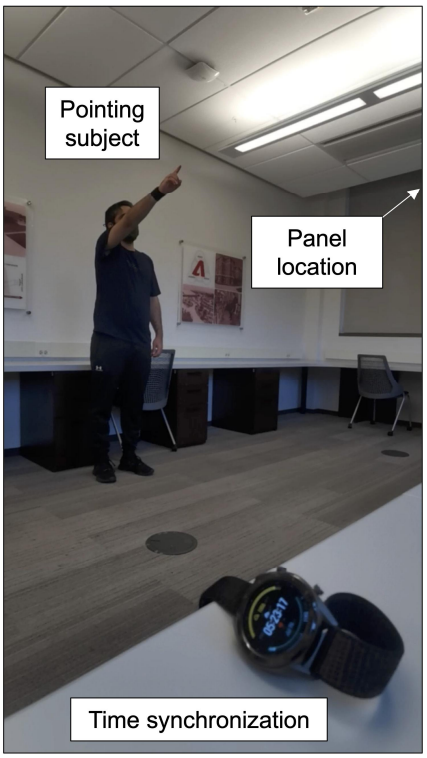

Figure 4-2. Picture of the Subject 1 conducting the first experiment (**E1**).

4.3. Evaluation

We use the following metrics to evaluate the spatial referring ability of the human and robot. We use these two metrics for both experiments (**E1** and **E2**)

Recognition accuracy. Recognition accuracy is calculated by F1 score. F1 score is defined by:

$$F1score = \frac{2 \times Precision \times Recall}{\mathrm{Pr}ecision + Recall} \tag{16}$$

$$where,\ \mathrm{Pr}ecision = \frac{TP}{TP + FP} \tag{17}$$

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

For each pointed position, this work consider it as:

True positive (TP) if it is classified as a correct target panel

False negative (FN) if it is classified as other target panels

False positive (FP) if the subject is pointing at other target panels

True negative (TN) if the subject is pointing at other target panels but classified correctly.

Deviation from target. Deviation from target ϵ is calculated by Euclidean distance between the estimated pointing position $\mathbf{p}_p$ and the center point of the target panel $\mathbf{p}_t$.

$$\epsilon = |\mathbf{p}_p - \mathbf{p}_t| \tag{19}$$

4.4. Results

A total of 1,200 pointing trials of four pointing subjects were evaluated in an offline setting in order to validate the performance of the spatial referencing method. The main results are shown in Table 4-2.

Table 4-2. F1 score and deviation from target (Mean ± SD).

Panel ID	F1 Score		Deviation from target [m]	
	Ceiling	Wall	Ceiling	Wall
C1	0.849	0.948	1.118 ± 0.243	0.561 ± 0.204
C2	0.826	0.802	1.159 ± 0.197	0.325 ± 0.134
C3	0.837	0.946	1.121 ± 0.216	0.282 ± 0.196
C4	0.756	0.861	1.090 ± 0.319	0.352 ± 0.145
C5	0.808	0.921	1.135 ± 0.316	0.530 ± 0.180
Avg.	0.815	0.896	1.125 ± 0.263	0.410 ± 0.174

4.4.1 Recognition Accuracy

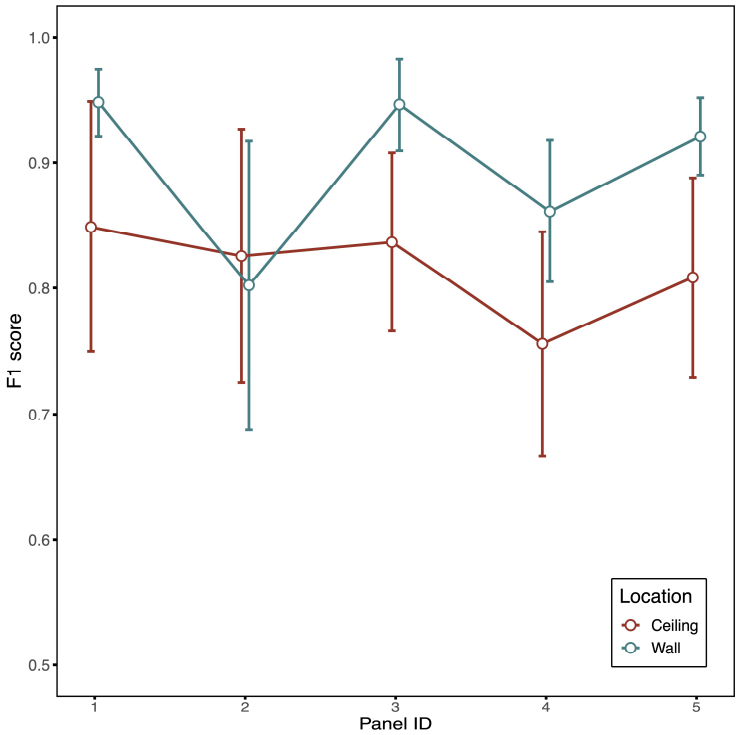

Figure 4-3. F1 score of the ceiling and wall pointing tasks.

The F1 score of the predicted target panels by robot is shown in Figure 4-3. We found that the ceiling pointing task shows a lower recognition accuracy. The mean F1 score of the ceiling pointing task was 0.815, while the F1 score of the wall pointing task was 0.896.

Higher error in the estimation of the pointing position increases the probability of inferring the wrong panels located nearby, which in turn lowers the recognition accuracy. This tendency is especially salient in the panels near the center.

4.4.2 Deviation from Target

The experimental results of deviation from target is shown in Figure 4-4. It was observed that on average, the ceiling pointing task yielded a higher mean deviation from target compared to the wall pointing task (Table 4-3). In addition, the mean deviation from target was higher when pointing at the side panels (C1/C5 and W1/W5) than the panels near the center (C2-C4 and W2-W4). This phenomenon will be expanded up in Section 4.5.

Meanwhile, a similar tendency between all four participants was observed as illustrated in Figure 4-5. Among the subjects, Subject 3 reached the highest mean target deviation for the ceiling pointing task with 0.482m of result gap between the lowest, Subject 1. For the wall, Subject 2 showed the highest distance error with 0.207m of result gap between the lowest, Subject 1.

Table 4-3. Mean deviation from target by workspaces and locations. The differences are statistically significant with both $p < 0.001$.

		Deviation from target [m]	**p-value**
Workspace	Ceiling	1.125 ± 0.263	$p < 2.2e\text{-}16$
	Wall	**0.410 ± 0.174**	
Angle	Side	0.836 ± 0.241	$p < 0.001$
	Center	**0.721 ± 0.210**	

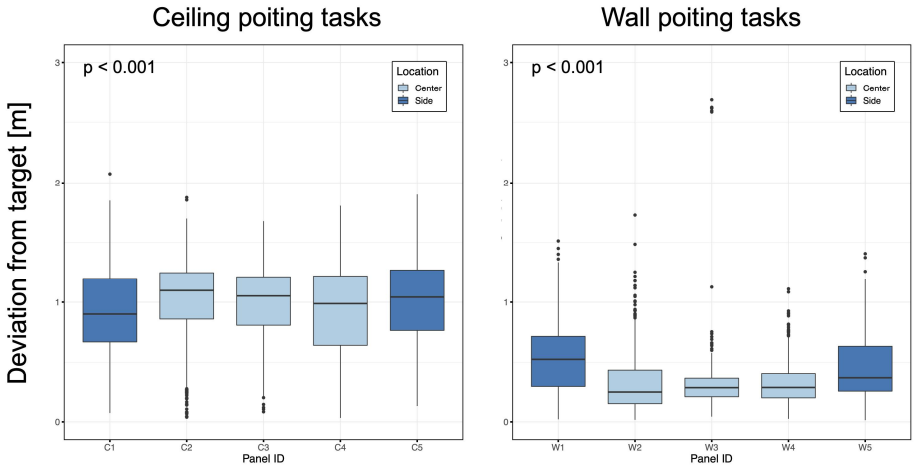

Figure 4-4. Deviation from target of the ceiling and wall pointing tasks. C1-C5 and W1-W5 refers to target ceiling and wall panels from left to right, respectively. Shaded areas represent the side panels of each surface and the rest represent the center panels of each surface.

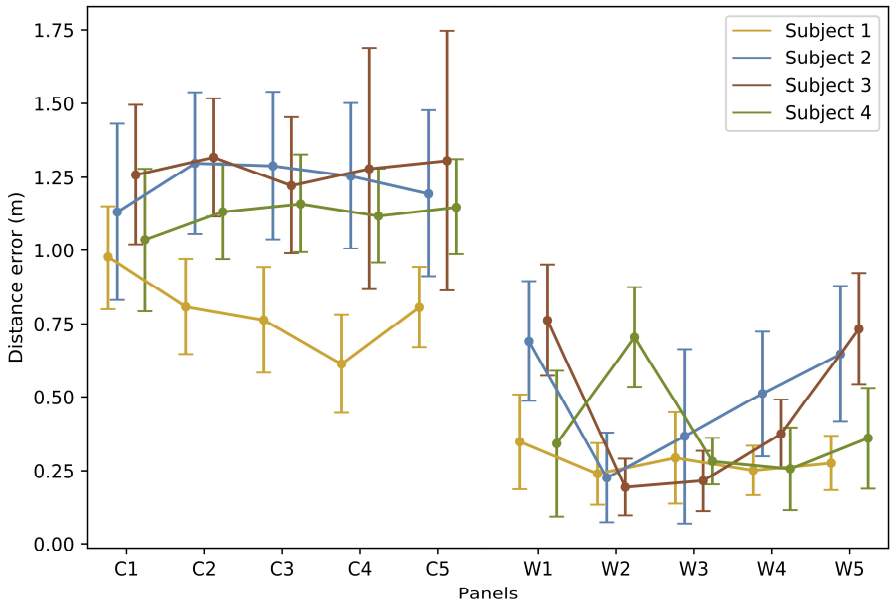

Figure 4-5. Deviation from target by the pointing subject.

Chapter 5. High-Level Panel Pointing Tasks

5.1. Experimental Setup

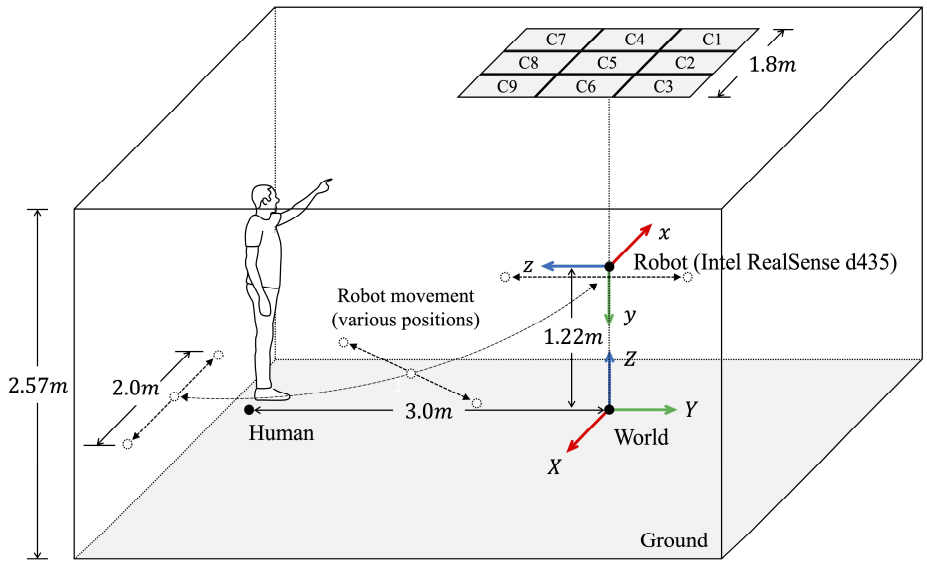

Figure 5-1. Illustration of the experimental environment for the second experiment (**E2**).

The experimental setup for the second experiment (**E2**) is depicted in Figure 5-1. Nine target panels with an equal size of 0.6 x 0.6m are installed on the ceiling in a 3 x 3 array, starting from the far left position from a pointier. The pointer is located 3.0m away from the origin of the world coordinate system, which is located at the projected position of the center point of the C5 panel.

As with the first experiment, this work uses Intel RealSense™ Depth Camera D435 to collect RGB and depth images. The camera is initially

located at the origin, at a height of 1.22m facing participants. However, in the second experiment, the robot, as well as human, moves around the subject to observe and estimate the pointing position with various distances (2, 3, and 4m) and angles (0°, 45°, and 90°), resulting in nine combinations of location (Figure 5-2). In sum, pointing data from nine locations observed by both human and robot was collected.

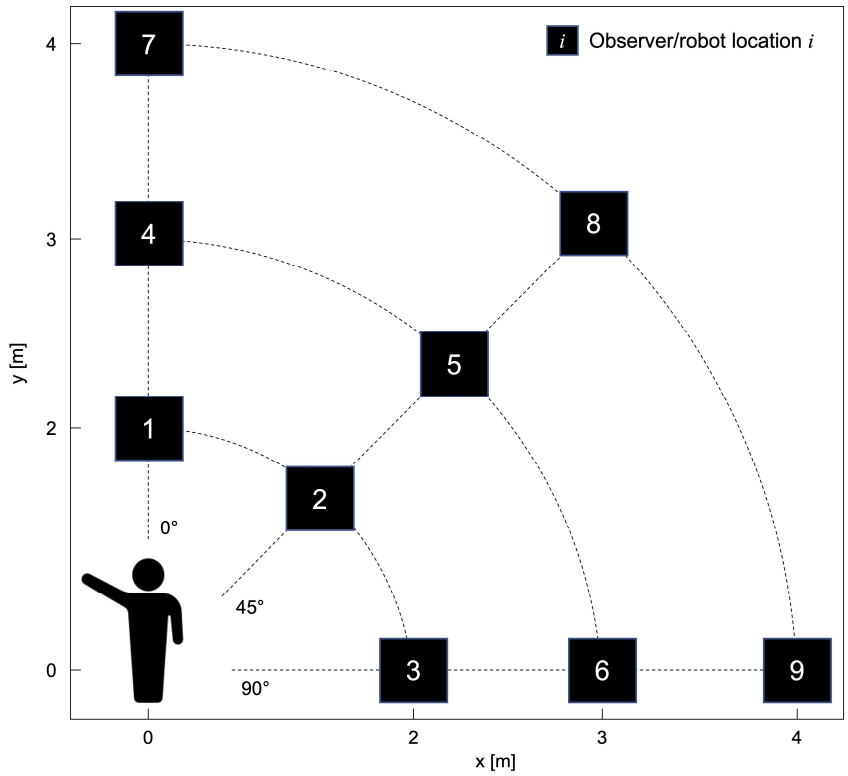

Figure 5-2. Illustration of the egocentric distances and angles of the nine observation positions for human and robot.

5.2. Procedure

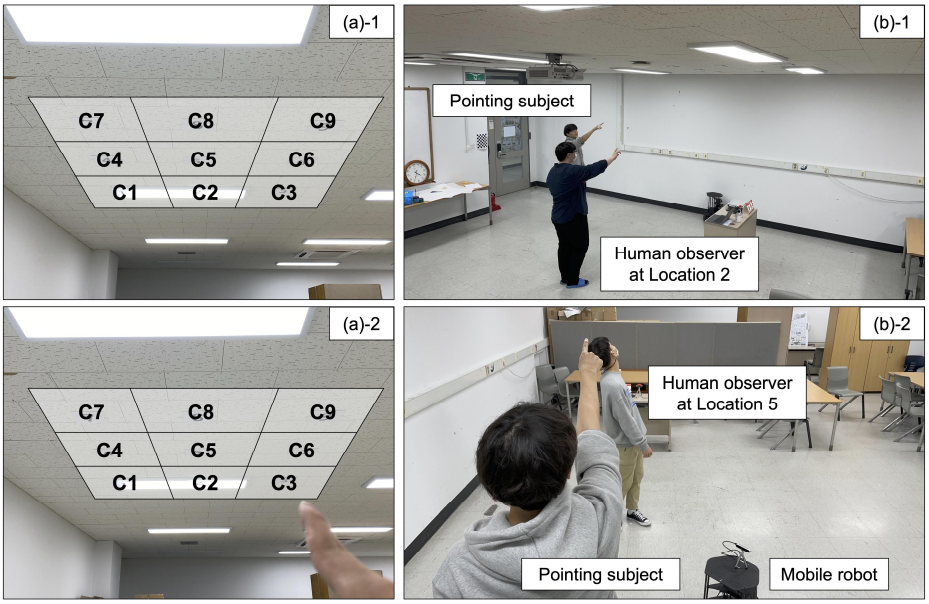

Figure 5-3. (a) First person view of the C3 panel pointing trial. (b) Picture of the pointing subject, human observer at location 2 and 5, and the mobile robot conducting the second experiment (**E2**).

11 participants (seven males and four females) were recruited to participate in the second experiment. The average age of the participants were 27.2 years (SD=3.1), average height was 171.9cm (SD=9.6), and the average weight was 67.0kg (SD=14.7).

One of the participants (male) was asked to perform the pointing task and the robot collected data with a camera at nine observation positions. For each position, the pointing subject performed 5 iterations for each panel. A single iteration consists of nine pointing trials (ceiling panels C1 to C9). Moreover, the 5 x 9 = 45 pointing trials were ordered randomly rather than in a sequential order, considering the human

predictions. In sum, 9 x 5 x 9 = 405 trials of pointing data for robot in a single experiment was obtained.

Meanwhile, the rest of the 10 participants (six males and four females) were asked to predict which panel the subject is pointing at, also at nine observation positions.

A single experiment took approximately an hour per participant, including the orientation.

5.3. Results

5.3.1. Recognition Accuracy

The recognition accuracy results showed that the performance of both human and robot were comparable on average. In addition, while various observation positions showed a considerable impact on the performance of the robot, especially the pointer-robot distance, human observers showed a consistent performance, regardless of the observation positions.

The average recognition accuracy of all panels and observation positions is shown in Table 4-5. Human observers predicted target panels at an F1 score of M=0.567 (SD=0.097), precision of M=0.562 (SD=0.124), and recall of M=0.599 (SD=0.148). Although the robot's recognition accuracy was higher in precision (M=0.590, SD=0.280), the recall (M=0.591, SD=0.278) and F1 score (M=0.569, SD=0.251) did not present a significant difference in values. A sample of target recognition cases is shown in Figure 5-4.

Table 5-1. Evaluation results (Mean ± SD): Recognition accuracy (F1 score, precision, and recall).

	Human	**Robot**
F1 score	0.567 ± 0.097	**0.569 ± 0.251**
Precision	0.562 ± 0.124	**0.590 ± 0.280**
Recall	**0.599 ± 0.148**	0.591 ± 0.278

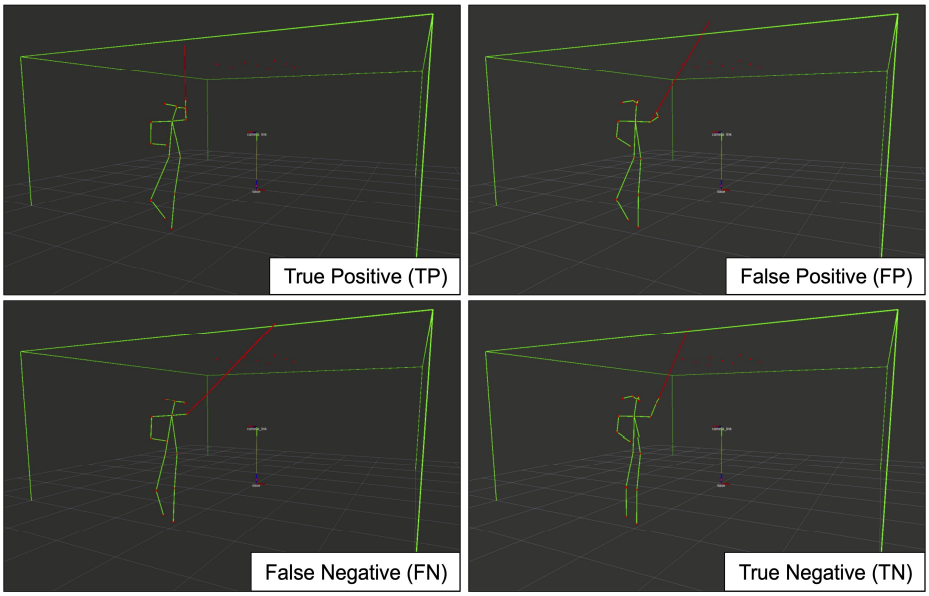

Figure 5-4. Sample target panel recognition cases.

Figure 5-5 shows the effect of pointer-observer angle on the average recognition accuracy of nine panels. This work conducted one-way analysis of variance (ANOVA) to determine whether the values of the F1 score, precision, and recall is different for the three pointer-observer angles. The analysis showed no significant difference on the recognition accuracy at all distances (2, 3, and 4m) for both human ($p > 0.102$) and robot ($p > 0.069$). Therefore, the angle of the pointing subject and the observer has no significant effect on the recognition accuracy for both human and robot observers. Moreover, the results showed that the recognition accuracy of the robot remains higher than the recognition accuracy of the human observer until the distance of 3m, but the performance degrades, thus resulting in lower recognition accuracy than human.

Meanwhile, the effect of pointer-observer distance on the average recognition accuracy of nine panels is shown in Figure 5-6. The one-way ANOVA was conducted to determine whether the values of the F1 score, precision, and recall is different for the three pointer-observer distances. The analysis showed a significant effect of pointer-observer distance on the recognition accuracy for robot ($p < 0.05$). The performance drop was much worse in the angle of 0 and 45 degrees. However, in the case of the human observers, the analysis showed no significant difference on the recognition accuracy ($p > 0.088$) at all angles (0°, 45°, and 90°). Thus, the effect of distance of the pointing subject and the observer on the recognition accuracy is only significant for the robot.

Table 5-2 and Figure 5-7 show the human recognition accuracy of the nine panels and confusion matrix, respectively. The results show that the F1 score reaches its highest for panel C8 (0.657) and higher scores for the two panels located side by side (C7=0.629, C9=0.643) than the rest of the panels. The results indicate that the performance of the human observer is enhanced when the target panel is installed at a close distance to the pointing subject. However, there was no significant performance gap between the panels in the same row.

Table 5-3 and Figure 5-8 show the robot recognition accuracy of the nine panels and confusion matrix, respectively. The results show that the F1 score reaches its highest for panel C7 (0.725). In the case of the robot, the performance gap between the highest F1 score and the lowest F1 score is bigger than the F1 score of human, while the per-

formance is mostly enhanced when the target panel is located at the four corners.

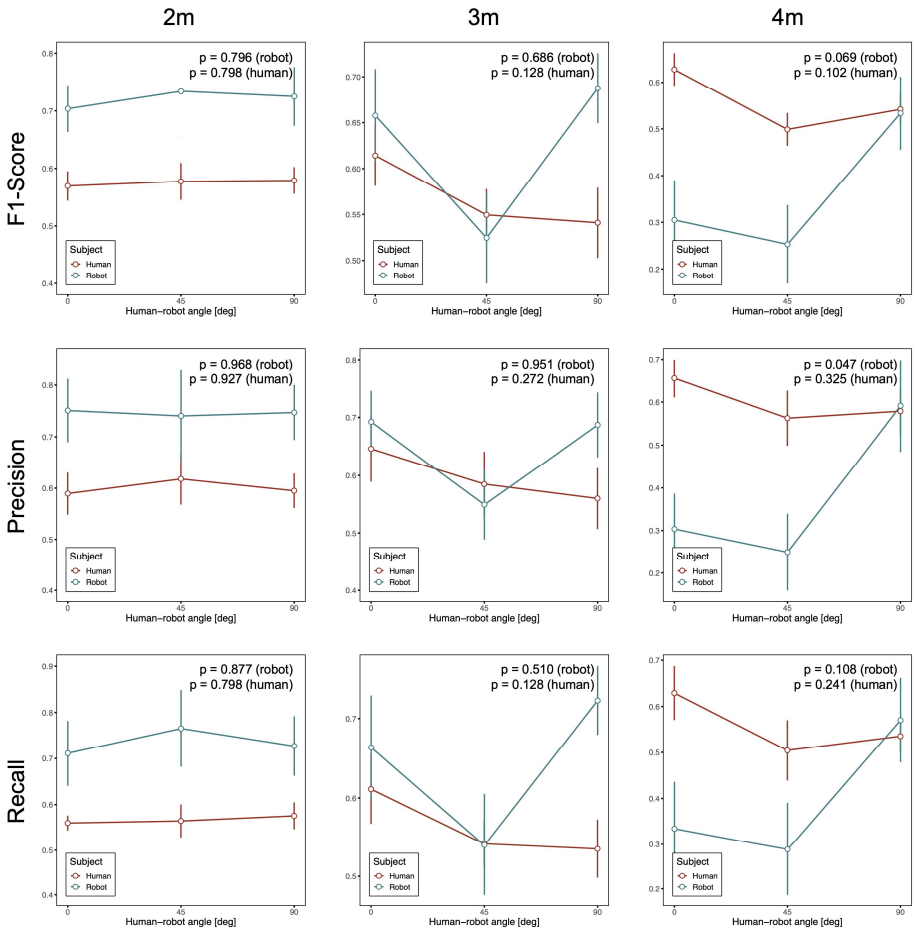

Figure 5-5. Effect of pointer-observer angle on the F1 score, precision, and recall of the predictions by the human observer (red) and the robot (green).

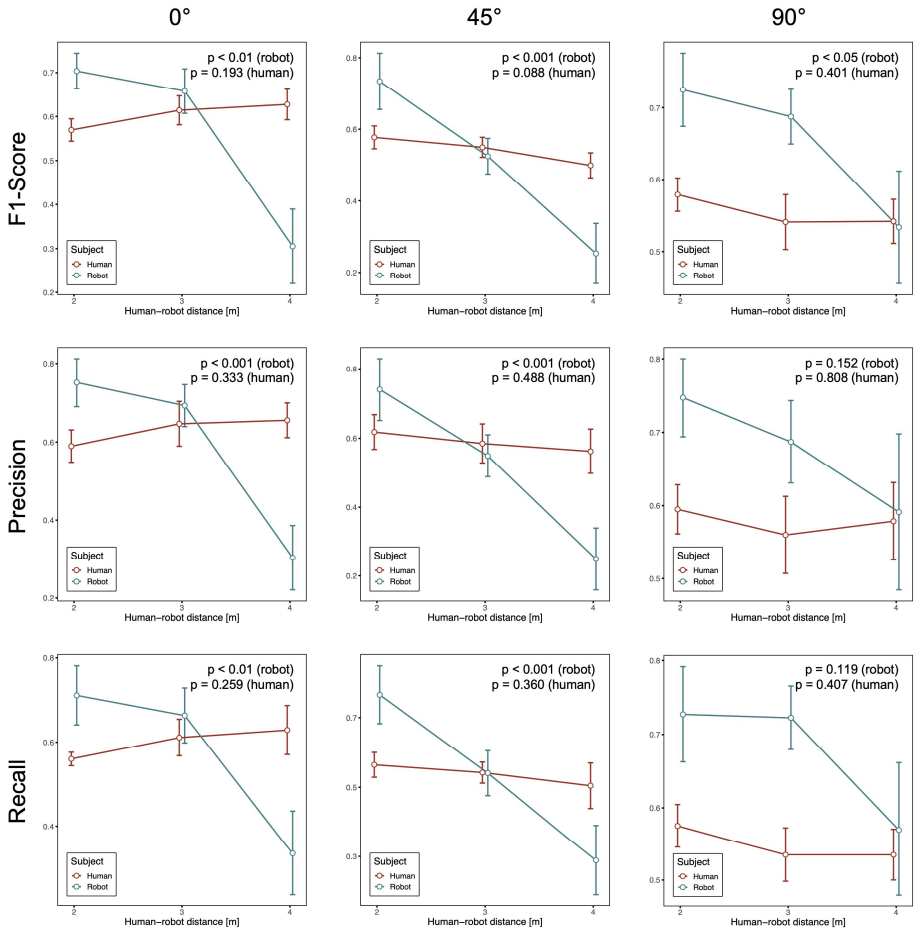

Figure 5-6. Effect of pointer-observer distance on the F1 score, precision, and recall of the predictions by the human observer (red) and the robot (green).

Table 5-2. Human Recognition accuracy (F1 score, precision, and recall) of nine panels.

Panels	F1 score	Precision	Recall
C1	0.593	0.590	0.596
C2	0.602	0.571	0.638
C3	0.596	0.558	**0.640**
C4	0.482	0.446	0.524
C5	0.453	0.435	0.471
C6	0.472	0.463	0.482
C7	0.629	0.674	0.589
C8	**0.657**	0.739	0.591
C9	0.643	**0.826**	0.527

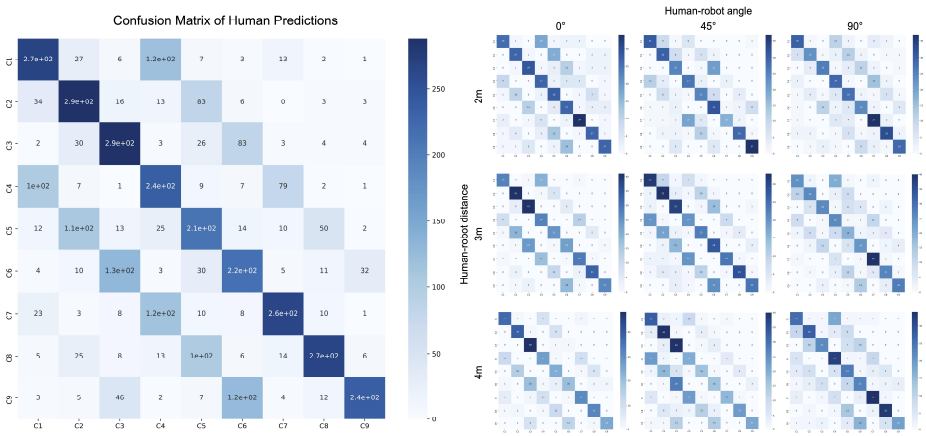

Figure 5-7. Confusion matrix of human predictions on nine target panels: average (left) and by nine observation positions (right).

Table 5-3. Robot Recognition accuracy (F1 score, precision, and recall) of nine panels.

Panels	F1 score	Precision	Recall
C1	0.632	0.699	0.605
C2	0.454	0.531	0.428
C3	0.711	0.685	0.760
C4	0.416	0.448	0.405
C5	0.554	0.538	0.601
C6	0.450	0.451	0.486
C7	**0.725**	0.615	**0.919**
C8	0.556	0.574	0.580
C9	0.627	**0.769**	0.539

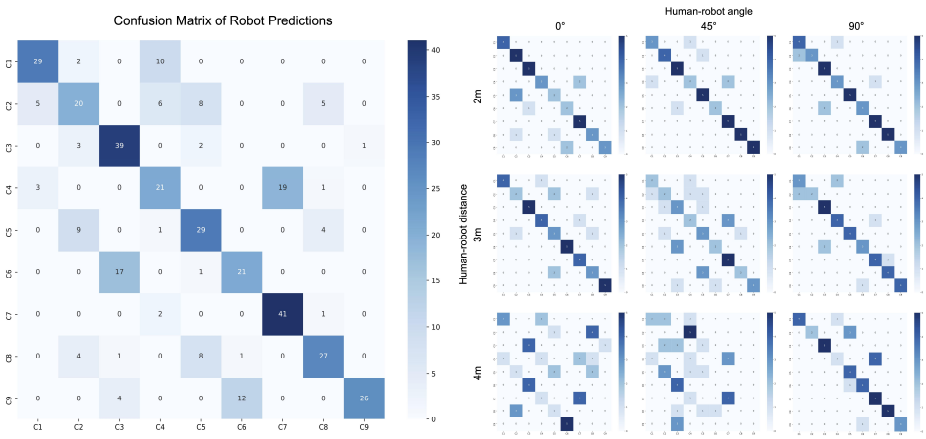

Figure 5-8. Confusion matrix of robot predictions on nine target panels: average (left) and by nine observation positions (right).

5.3.2. Deviation from Target

Table 5-4. Mean deviation from target (Mean ± SD).

Metrics	Values
Deviation from target [m]	0.449 ± 0.505
Horizontal component [m]	0.151 ± 0.156
Vertical component [m]	0.393 ± 0.504

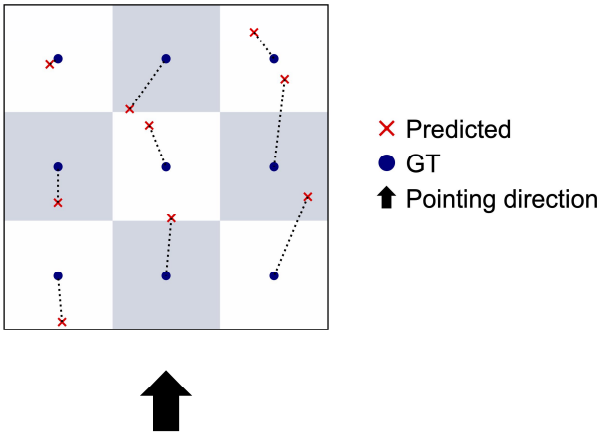

Figure 5-9. Estimated target positions by the robot and the deviation from target. Red markers represent the predicted points and the Blue dots represent the ground truth (center points of the each panel).

Figure 5-9 shows the average estimated target positions and their deviations from target of nine observation positions. As shown in Table 5-4, the robot estimated target positions with the mean deviation of M=0.449m (SD=0.505). The vertical component of the deviation was dominant (M=0.393m, SD=0.504) compared to the horizontal component (M=0.151m, SD=0.156).

Figure 5-10 shows the effect of the pointer-robot angle on the average deviation from target of nine target panels. This work conducted

one-way ANOVA to determine whether the target deviation is different for the three pointer-robot angles. The analysis showed no significant difference on the target deviation for 2m of distance ($p=0.498$) and 3m of distance ($p=0.815$). While the deviation increases as the robot moves away from the pointing subject, at 4m of distance, the deviation shows a drop in the 90° of pointer-robot angle, as well as the significant p-value ($p<0.001$).

In the same context, the one-way ANOVA results of the effect of pointer-robot distance on target deviation presented no significant difference on the target deviation at the pointer-robot angle of 90° ($p=0.824$), as shown in Figure 5-11. Meanwhile, target deviation showed an increasing trend at the pointer-robot angle of 0° and 45°, with the significant p-values ($p<0.001$). Moreover, the results presented that the vertical component dominates in the target deviation at the pointer-robot angle of 0° and both vertical and horizontal component had an impact on the target deviation at the pointer-robot angle of 45°.

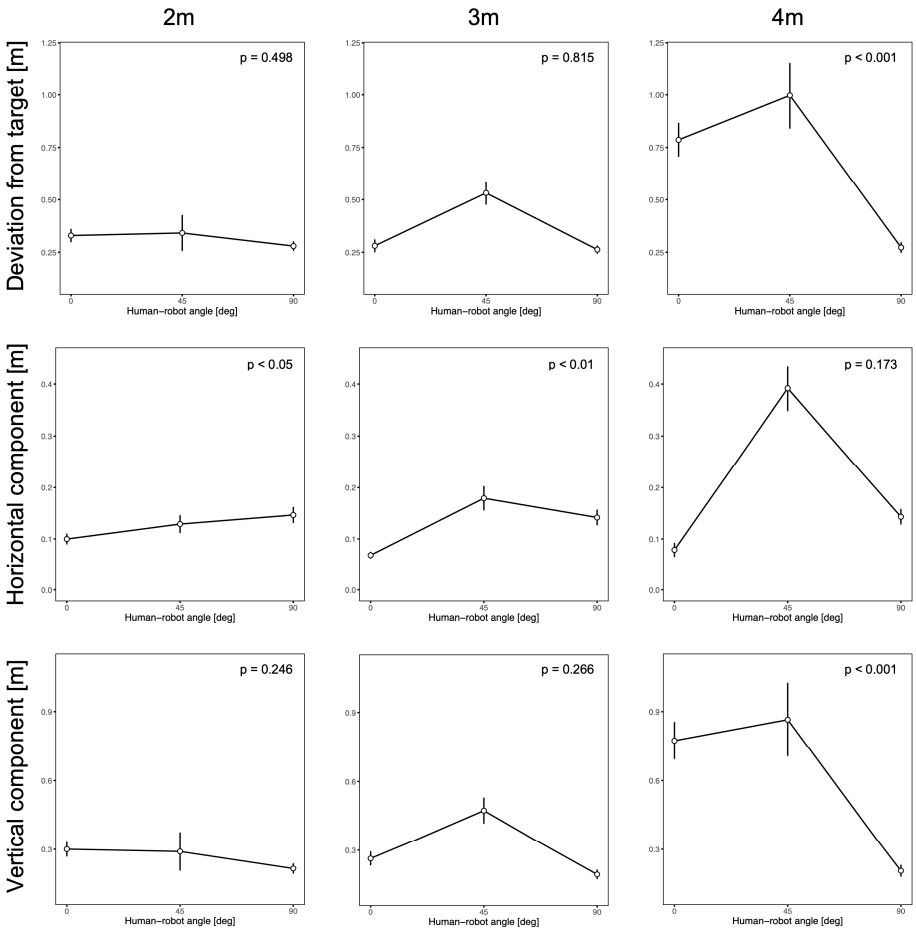

Figure 5-10. Effect of pointer-robot angle on the deviation from target and its horizontal and vertical component.

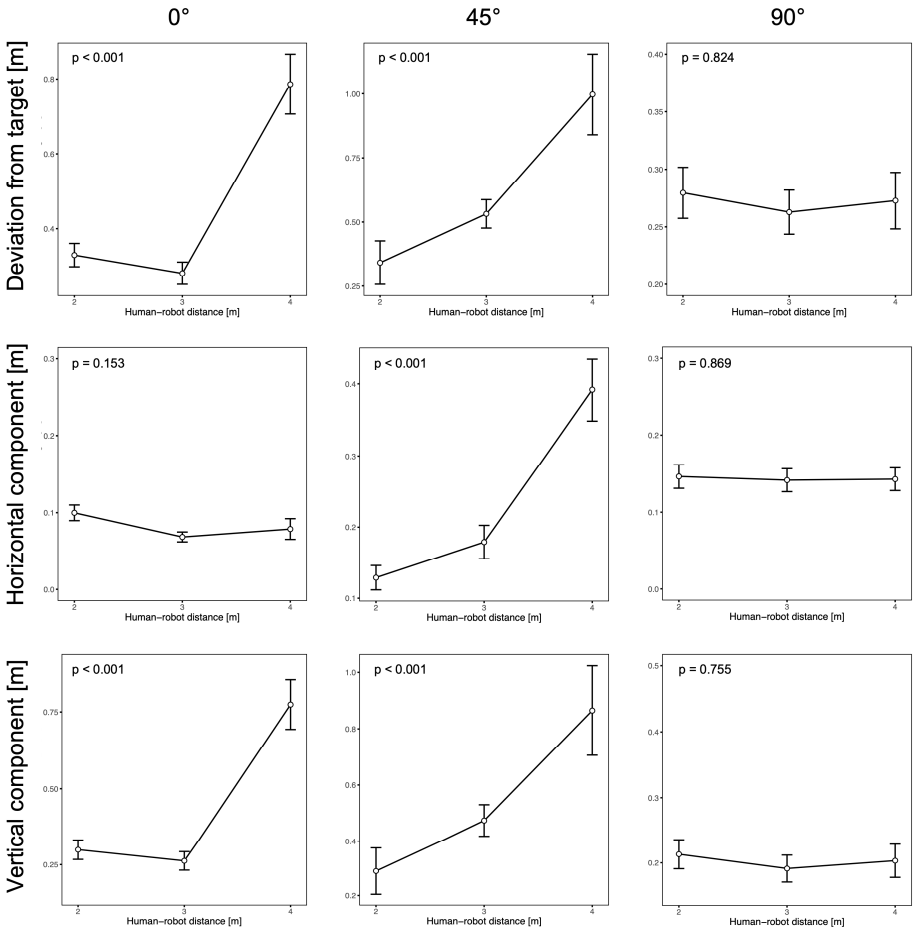

Figure 5-11. Effect of pointer-robot distance on the deviation from target and its horizontal and vertical component.

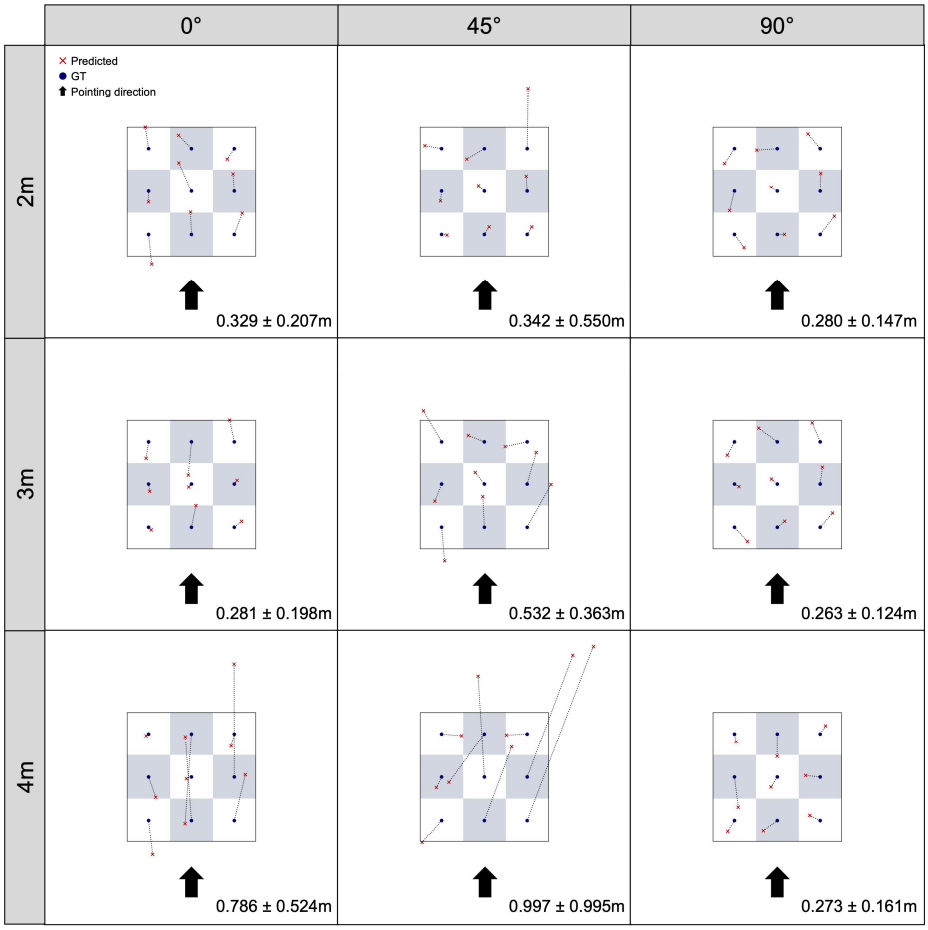

Figure 5-12. Estimated target positions by the robot and the deviation from target for nine observation positions. Red markers represent the predicted points and the Blue dots represent the ground truth (center points of the each panel).

Chapter 6. Discussion

In this chapter, the results from Chapter 4 and 5 will be leveraged for some discussions and to produce guidelines for improving the performance of the deictic gesture-based spatial referencing method for applications in collaborative construction robots.

6.1. Challenges in Low-Level Precision Tasks

Our experimental results present that the deictic gesture-based method can interpret the spatial information at the most common workspaces in construction with a mean target deviation of 0.767m and the recognition accuracy of 0.864.

The results show worse performance in the ceiling with a mean target deviation of 1.125m, which was 0.715m higher than the mean target deviation of the wall. In recognition of the panel, the mean F1 score dropped at a rate of 8.98% at the ceiling compared to the wall. These measures indicate that variation in plane causes a performance gap in the spatial referring ability regarding the recognition accuracy.

Furthermore, the mean target deviation tends to increase by 15.91% when the location of the target panel changes from the center to side. In this situation, the human mainly delivers the angle information to the robot, because the target panels share the same plane.

In general, these tendencies of the result can be explained by the pinhole projection model, as illustrated in Figure 4-6. Human eyes see the world via pinhole projection. The 3D world (on the world coordinate system) is projected onto a flat projection plane: this plane is focal length d away from the projective center along the Z_h axis (on the human coordinate system), the gaze direction [41]. Thus, a 3D point $\mathbf{p} = (x_h, y_h, z_h)$ in the human coordinate system is projected to 2D coordinates $\mathbf{p}'$ on the projection plane at a rate of d/z_h:

$$\mathbf{p}' = \frac{d}{z_h}(x_h, y_h) \tag{20}$$

Therefore, the area of the target panel is also projected to the projection plane, affecting the visible area of the panel with respect to the rate of d/z_h.

The top and side views of the experiment setup are depicted in Figure 4-7. $\mathbf{A_0}$ and $\mathbf{A_1}$ refers to the visible area of the panels projected on the projection plane, perpendicular to the gaze direction Z_0 and Z_1 (Here, this work assumes a person gazes at the center of the target panels when pointing). In both situations, $\mathbf{A_0}$ is larger than $\mathbf{A_1}$ due to the difference in between the angles θ_0 and θ_1, as well as the position of the panels. A smaller visible area hinders the subjects from pointing precisely while maintaining consistency. Thus, compared to the targets with large visible areas (wall and center panels), the performance degrades in the targets with small visible areas (ceiling and side panels). Overall, it can be noted that the visible area of the target is a crucial factor for human's ability of interpreting deictic gesture-based spatial communication for both wall and ceiling conditions. Therefore, in practice, one can expect lower performance in referencing a distanced and angled regions of interest in overhead operations (e.g., electrical wiring, plumbing, and interior finishing work). In such situations, collaborative robots need mobility for estimation of the workspace geometry through navigating themselves closer to the target.

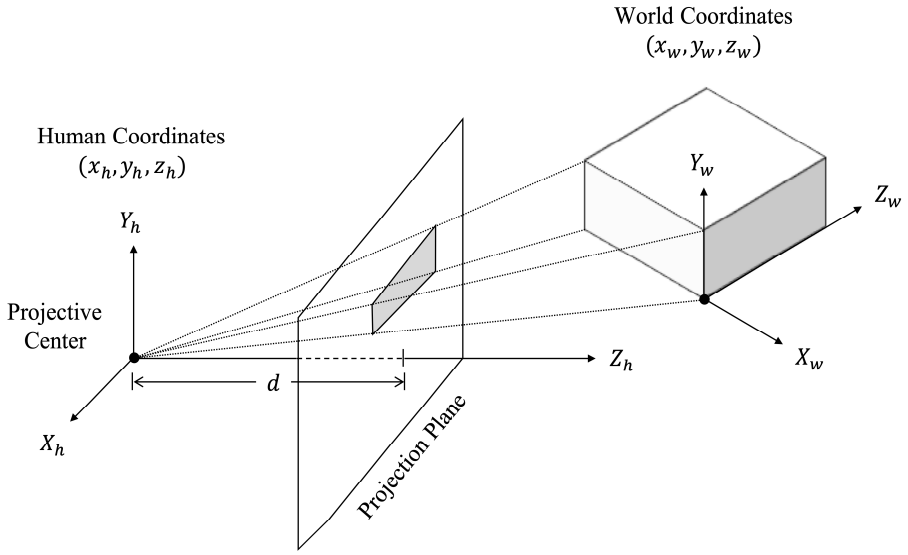

Figure 6-1. Pinhole projection model [47].

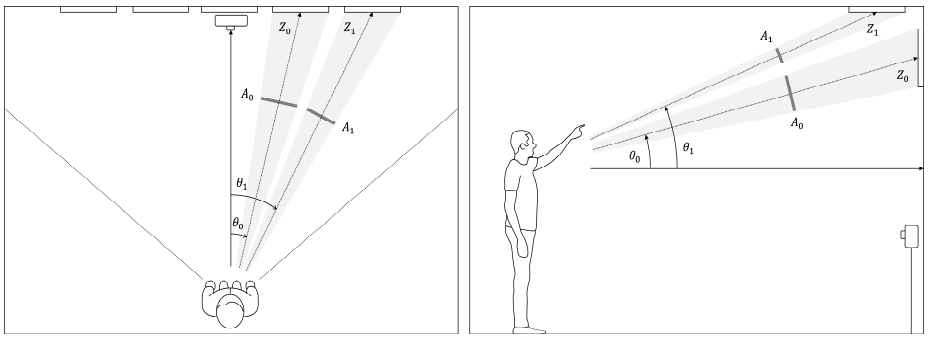

Figure 6-2. The top (left) and side (right) views of the experimental environment. The panels were spaced for a better understanding.

6.2. Challenges in High-Level Precision Tasks

The results of experiment 2 (**E2**) present that although spatial referencing through deictic gestures is intuitive and natural to humans, it may be difficult to interpret and predict targets solely on them in these high-level tasks not only for robots, but also for humans. In this section, the results of experiment 2 (**E2**) will be discussed in detail.

The results show that the deictic gesture-based method can estimate the pointing position for the nine ceiling target panels at an mean F1 score of 0.569, which is 30.18% lower than the mean F1 score of the low-level pointing tasks. Human observers showed even worse performance, with the F1 score of 0.567.

However, there were no significant variations in the performance among the observation positions in the case of humans. In other words, humans tend to have a consistent spatial referring ability, regardless of the observation positions. Robots, on the other hand, tend to show a constant decrease in the recognition accuracy when the distance of the pointing subject and the robot increases, while limited effects of the angle were observed. In particular, the accuracy significantly dropped at the distance of 4m. This is not surprising as it is specified that the ideal range of Intel RealSense™ Depth Camera D435 is 0.3m to 3m.

In addition, human observers showed lower recognition accuracy in the middle row panels (C4, C5, and C6) compared to the front row panels (C7, C8, and C9). When in the case of misinterpretation of the middle row panels, human observers tended to predict the target panel

as one of the front row panels, other than the third row panels (C1, C2, and C3). This may be explained by the imaginary "attentional cone" suggested by Williams et al. (2019) [16]. Because of the conic nature of this region, when human observers had to choose between panel candidates, they were biased towards panels which are located closer to the pointing subjects, rather than distant panels.

Meanwhile, the highest recognition accuracy of the robot reached up to 0.725 of F1 score in predicting panel C7, which was 9.3% higher than the highest F1 score (0.657) of the human in predicting panel C8. Nevertheless, considering the lowest recognition accuracy of the panel C4 which drops to an F1 score of 0.416, the results indicate that the high-level pointing tasks are still challenging for robots.

Deviation from target showed more detailed results. The mean value of nine ceiling panel pointing tasks was 0.449m, which was 60.01% lower than the mean deviation of the ceiling panel pointing tasks in the first experiment (**E1**). This performance enhancement is mainly due to the pointing calibration. Without the pointing calibration, the mean deviation of the second experiment increases up to 1.694m, which is 33.59% higher than the mean deviation of the first experiment (**E1**). The pointing calibration showed considerable correction in the vertical component, as in Table 5-5, where the deviation decreased by 75.73% compared to the raw data, while the overall deviation decreased by 73.49%.

Table 6-1. Comparison of the deviation from target (Mean ± SD) with and without pointing calibration.

	No pointing calibration	**With pointing calibration**
Deviation from target [m]	1.694 ± 0.476	0.449 ± 0.505
Horizontal component [m]	0.396 ± 0.274	0.151 ± 0.156
Vertical component [m]	1.619 ± 0.496	0.393 ± 0.504

Finally, the results showed a more consistent and higher spatial referring ability at the pointer-robot angle of 90° compared to the other angles, where the ability drops when the robot moves away from the pointing subject. This is because this work performed deictic gesture recognition based on pose estimation. For the pose estimation-based deictic gesture recognition, three right arm joints are mainly detected and utilized: shoulder, elbow, and the wrist joint. Therefore, the more these joints come into the camera view without depth occlusion, the more likely the system will detect the exact position of the joints. This led to enhanced deviation at the angle of 90°, which has a comparatively low possibility of depth occlusion while stretching the right arm for pointing.

Considering the results mentioned above, we see three ways to improve the current spatial referencing method for application in collaborative construction robots. First, we could give the robot dimensional and scaled spatial information with interaction modalities (i.e., speech). This allows the robot to reason about the region of interest with additional criteria, thus enhancing perception accuracy. Presenting the spatial information with a form of region could be considered as well. Deictics

are often thought of as referring to an object but can also be used to refer to a region of space [27]. This method provides interpretability and predictability to the user intent and has a collateral benefit of correction. Lastly, as suggested by Medeiros et al. [36], visual feedback makes a difference in the accuracy of the pointing task. In particular, we could enhance human's ability to indicate the target with a smaller visible area by receiving visual feedback from robots.

Chapter 7. Conclusion

This work explored the challenges of the latest deictic gesture-based spatial referencing method to evaluate the potential of the deictic gestures in panel installation tasks in construction. Despite the intuitiveness and naturalness of the deictic gestures, they are known to be inherently imprecise and therefore challenging for both human and robot to interpret the exact region of interest, especially in a large-scale 3D construction environment, cluttered in many situations. In this context, we selectively overviewed the spatial referring abilities of both human and robot through two experiments of panel pointing with different levels of precision. In the first experiment, the results presented a significant performance drop in the ceiling and angled targets, while the overall recognition accuracy was acceptable. These tendencies imply that pointing the distanced targets in a large-scale environment is challenging for a human pointer, especially for targets located on ceiling surfaces. This work concluded that these tendencies result from the variations on visible area by the target locations. This work tried to further the knowledge on spatial referencing through the second experiment, which involved evaluation of both human and robot observers with various positional relationships with the pointing subject. Through the second experiment, this work showed that it may be difficult to interpret and predict targets solely on the deictic gestures in this level of precision not only for robots, but also for humans. For these ceiling panel point-

ing tasks with high-level of precision, the recognition accuracy of the robot dropped by 30.18% compared to the low-level tasks. The results of the second experiment presented that humans tend to have consistency in the spatial referring abilities regardless of the observation positions, while the robots showed statistically significant differences pertaining to the pointer-robot distances. The results also showed the performance enhancement of the target deviation by 73.49% through pointing calibration.

The primary contribution of this paper is in the evaluation of the deictic gesture-based spatial referring ability of the robot in a large-scale human-robot collaborative environment, furthering the application of the deictic gestures and robotics to panel installation work. However, this work has limitations, since this work assumes that the robot has a full understanding of its surrounding environment and workspace geometry, which is completely accurate with no measurement errors. Future works are needed to investigate the spatial referencing system's behavior when the robot has to learn its surrounding environment (i.e. with simultaneous localization and mapping).

Bibliography

[1] M. Gharbia, A. Chang-Richards, Y. Lu, R.Y. Zhong, H. Li, Robotic technologies for on-site building construction: A systematic review, J. Build. Eng. 32 (2020) 101584. https://doi.org/10.1016/j.jobe.2020.101584.

[2] F. Barbosa, J. Woetzel, J. Mischke, Reinventing Construction: A Route of Higher Productivity, 2017.

[3] M. Schwartz, Solving The Construction Labor Shortage Through Ingenuity, (2012). https://www.forbes.com/sites/forbestechcouncil/2021/10/15/solving-the-construction-labor-shortage-through-ingenuity/?sh=97d97233b7b3

[4] C.-J. Liang, X. Wang, V.R. Kamat, C.C. Menassa, Human–Robot Collaboration in Construction: Classification and Research Trends, J. Constr. Eng. Manag. 147 (2021) 03121006. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002154.

[5] C. Brosque, E. Galbally, O. Khatib, M. Fischer, Human-Robot Collaboration in Construction: Opportunities and Challenges, in: HORA 2020 - 2nd Int. Congr. Human-Computer Interact. Optim. Robot. Appl. Proc., Institute of Electrical and Electronics Engineers Inc., 2020. https://doi.org/10.1109/HORA49412.2020.9152888.

[6] Construction Robotics, (2021). https://www.construction-robotics.com/

[7] Construction Robots, (2021). https://www.constructionrobots.com

[8] Civ Robotics, (2021). https://www.civrobotics.com/

[9] Dusty Robotics, (2021). https://www.dustyrobotics.com/

[10] MX3D, (2021). https://mx3d.com/

[11] COBOD, (2021). https://cobod.com/

[12] X. Wang, C.-J. Liang, C.C. Menassa, V.R. Kamat, Interactive and Immersive Process-Level Digital Twin for Collaborative Human–Robot Construction Work, J. Comput. Civ. Eng. 35 (2021) 04021023. https://doi.org/10.1061/(asce)cp.1943-5487.0000988.

[13] K.M. Lundeen, V.R. Kamat, C.C. Menassa, W. McGee, Autonomous motion planning and task execution in geometrically adaptive robotized construction work, Autom. Constr. 100 (2019) 24–45. https://doi.org/10.1016/J.AUTCON.2018.12.020.

[14] B. Gromov, L. Gambardella, A. Giusti, Guiding Quadrotor Landing with Pointing Gestures, in: Springer Proc. Adv. Robot., Springer, Cham, 2020: pp. 1–14. https://doi.org/10.1007/978-3-030-42026-0_1.

[15] T. Obo, R. Kawabata, N. Kubota, Cooperative Human-Robot Interaction Based on Pointing Gesture in Informationally Structured Space, in: World Autom. Congr. Proc., Elsevier, 2018: pp. 103–108. https://doi.org/10.23919/WAC.2018.8430388.

[16] T. Williams, M. Bussing, S. Cabrol, E. Boyle, N. Tran, Mixed Reality Deictic Gesture for Multi-Modal Robot Communication, ACM/IEEE Int. Conf. Human-Robot Interact. 2019-March (2019) 191–201. https://doi.org/10.1109/HRI.2019.8673275.

[17] A. Jevtić, A.F. Valle, G. Alenyà, G. Chance, P. Caleb-Solly, S. Dogramadzi, C. Torras, Personalized Robot Assistant for Support in Dressing, IEEE Trans. Cogn. Dev. Syst. 11 (2019) 363–374. https://doi.org/10.1109/TCDS.2018.2817283.

[18] M.A. Zamani, H. Beik-Mohammadi, M. Kerzel, S. Magg, S. Wermter, Learning Spatial Representation for Safe Human-Robot Collaboration in Joint Manual Tasks, in: ICRA Work. Work. Is Better with Intelligent, Collab. Robot MATEs, 2018. https://www.researchgate.net/publication/338068215_Learning_Spatial_Representation_for_Safe_Human-Robot_Collaboration_in_Joint_Manual_Tasks

[19] S. Mayer, V. Schwind, R. Schweigert, N. Henze, The effect of offset correction and cursor on mid-air Pointing in real and virtual environments, in: Conf. Hum. Factors Comput. Syst. -

Proc., ACM, New York, NY, USA, 2018. https://doi.org/10.1145/3173574.3174227.

[20] S. Mayer, J. Reinhardt, R. Schweigert, B. Jelke, V. Schwind, K. Wolf, N. Henze, Improving Humans' Ability to Interpret Deictic Gestures in Virtual Reality, in: Conf. Hum. Factors Comput. Syst. - Proc., Association for Computing Machinery, 2020. https://doi.org/10.1145/3313831.3376340.

[21] D. Yokoyama, N. Hama, N. Nakamichi, K. Sugihara, K. Watanabe, T. Yamada, Pointing-gestures' Angle Differences between a Standing posture and a Sitting posture, 2018 IEEE 7th Glob. Conf. Consum. Electron. GCCE 2018. (2018) 64–65. https://doi.org/10.1109/GCCE.2018.8574873.

[22] N. Dhingra, E. Valli, A. Kunz, Recognition and Localisation of Pointing Gestures Using a RGB-D Camera, in: Commun. Comput. Inf. Sci., 2020: pp. 205–212. https://doi.org/10.1007/978-3-030-50726-8_27.

[23] D. Yongda, L. Fang, X. Huang, Research on multimodal human-robot interaction based on speech and gesture, Comput. Electr. Eng. 72 (2018) 443–454. https://doi.org/10.1016/j.compeleceng.2018.09.014.

[24] D. Jirak, D. Biertimpel, M. Kerzel, S. Wermter, Solving visual object ambiguities when pointing: an unsupervised learning ap-

proach, Neural Comput. Appl. 33 (2021) 2297–2319. https://doi.org/10.1007/S00521-020-05109-W.

[25] D. Brand, A. Meschtscherjakov, K. Büchele, Pointing at the HUD: Gesture interaction using a leap motion, in: AutomotiveUI 2016 - 8th Int. Conf. Automot. User Interfaces Interact. Veh. Appl. Adjun. Proc., Association for Computing Machinery, Inc, 2016: pp. 167–172. https://doi.org/10.1145/3004323.3004343.

[26] M. Sousa, R.K. Dos Anjos, D. Mendes, M. Billinghurst, J. Jorge, Warping deixis: Distorting gestures to enhance collaboration, in: Conf. Hum. Factors Comput. Syst. - Proc., ACM, New York, NY, USA, 2019. https://doi.org/10.1145/3290605.3300838.

[27] A. Sauppé, B. Mutlu, Robot deictics: How gesture and context shape referential communication, in: ACM/IEEE Int. Conf. Human-Robot Interact., 2014: pp. 342–349. https://doi.org/10.1145/2559636.2559657.

[28] M.W. Alibali, Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information, Spat. Cogn. Comput. 5 (2005) 307–331. https://doi.org/10.1207/s15427633scc0504_2.

[29] B. Gromov, G. Abbate, L.M. Gambardella, A. Giusti, Proximity

human-robot interaction using pointing gestures and a wrist-mounted IMU, in: Proc. - IEEE Int. Conf. Robot. Autom., Institute of Electrical and Electronics Engineers Inc., 2019: pp. 8084–8091. https://doi.org/10.1109/ICRA.2019.8794399.

[30] M. Tölgyessy, M. Dekan, F. Duchoň, J. Rodina, P. Hubinský, L. Chovanec, Foundations of Visual Linear Human–Robot Interaction via Pointing Gesture Navigation, Int. J. Soc. Robot. 2017 94. 9 (2017) 509–523. https://doi.org/10.1007/S12369-017-0408-9.

[31] S. Navas Medrano, M. Pfeiffer, C. Kray, Remote Deictic Communication: Simulating Deictic Pointing Gestures across Distances Using Electro Muscle Stimulation, Int. J. Hum. Comput. Interact. 36 (2020) 1867–1882. https://doi.org/10.1080/10447318.2020.1801171.

[32] J. Delpreto, D. Rus, Plug-and-Play Gesture Control Using Muscle and Motion Sensors, Proc. 2020 ACM/IEEE Int. Conf. Human-Robot Interact. (2020). https://doi.org/10.1145/3319502.

[33] S. Walkowski, R. Dörner, M. Lievonen, D. Rosenberg, Using a game controller for relaying deictic gestures in computer-mediated communication, Int. J. Hum. Comput. Stud. 69 (2011) 362–374. https://doi.org/10.1016/J.IJHCS.2011.01.002.

[34] P. Kumar, J. Verma, S. Prasad, Hand Data Glove: A Wearable

Real-Time Device for Human-Computer Interaction, Int. J. Adv. Sci. Technol. 43 (2012) 15–26. http://www.sersc.org/journals/IJAST/vol43/2.pdf

[35] Y. LI, J. HUANG, F. TIAN, H.A. WANG, G.Z. DAI, Gesture interaction in virtual reality, Virtual Real. Intell. Hardw. 1 (2019) 84–112. https://doi.org/10.3724/SP.J.2096-5796.2018.0006.

[36] A.C.S. Medeiros, P. Ratsamee, J. Orlosky, Y. Uranishi, M. Higashida, H. Takemura, 3D pointing gestures as target selection tools: guiding monocular UAVs during window selection in an outdoor environment, ROBOMECH J. 2021 81. 8 (2021) 1–19. https://doi.org/10.1186/S40648-021-00200-W.

[37] K. Nickel, R. Stiefelhagen, Pointing gesture recognition based on 3D-tracking of face, hands and head orientation, ICMI'03 Fifth Int. Conf. Multimodal Interfaces. (2003) 140–146. https://doi.org/10.1145/958432.958460.

[38] S. Abidi, M. Williams, B. Johnston, Human pointing as a robot directive, in: ACM/IEEE Int. Conf. Human-Robot Interact., 2013: pp. 67–68. https://doi.org/10.1109/HRI.2013.6483504.

[39] Z. Cao, G. Hidalgo, T. Simon, S.E. Wei, Y. Sheikh, OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2021) 172–186. https://doi.org/10.1109/TPAMI.2019.2929257.

[40] D. Sprute, R. Rasch, A. Portner, S. Battermann, M. Konig, Gesture-Based Object Localization for Robot Applications in Intelligent Environments, in: Proc. - 2018 Int. Conf. Intell. Environ. IE 2018, Institute of Electrical and Electronics Engineers Inc., 2018: pp. 48–55. https://doi.org/10.1109/IE.2018.00015.

[41] A. Sharma, R. Nett, J. Ventura, Unsupervised learning of depth and ego-motion from cylindrical panoramic video with applications for virtual reality, Int. J. Semant. Comput. 14 (2020) 333–356. https://doi.org/10.1142/S1793351X20400139.

국 문 초 록

최근 AI의 발전과 더불어 로봇의 지각 및 제어 능력의 비약적 발전에도 불구하고, 건설 로봇은 실제 현장에 배치되었을 때, 비구조화되고 비정형화된 작업환경으로 인해 빈번한 간섭을 받게 된다. 따라서 이러한 현장의 상황에 유연하게 대처하여 작업 순서 및 방식을 변형시키기 위해 인간의 현장 작업지시 (in-situ improvisations)가 필요하다. 선행연구에서는 직관적인 인간-로봇 상호작용을 위해 지시적 제스처를 통한 현장 작업지시의 기반이 되는 공간 참조 (spatial referencing)의 타당성을 검증하였다. 그러나 선행연구에서 진행한 성능 평가에는 두 가지 한계점이 존재한다: (1) 지시적 제스처를 통한 공간 참조는 좁은 공간 범위에 대해서만 개발되었으며, 건설현장과 같은 대규모 공간 및 비구조화된 공간에서는 제한적이다; (2) 인간과 로봇의 다양한 위치관계에 대한 고려 없이 성능 평가를 진행하였다. 이에 본 연구에서는 건설 작업 중 패널 설치 작업에서 지시적 제스처의 가능성을 검증하기 위해, 손동작을 통해 패널을 가리키는 과업에 대한 도전과제를 규명하는 것을 목표로 한다. 이를 위해 서로 다른 정밀도 범위를 요구하는 두 가지 실험을 통해 인간과 로봇의 공간 참조 능력을 평가하였다. 첫 번째 실험에서는, 전반적으로 높은 인식 정확도(0.864)를 보였으나, 천장 및 양 측면에 위치한 패널에서 성능이 감소하였다. 두 번째 실험에서는, 인식 정확도가 첫 번째 실험에 비해 30.18% 감소하였다. 두 번째 실험의 결과에서는 로봇이 인간에 비해 관측 위치에 대해 공간 참조 능력에 큰 편차를 보이는 것으로 나타났다. 또한, 지시 보정을 통해 목표

거리 오차가 73.49% 개선됨을 확인하였다. 본 연구는 건설현장과 같은 대규모 인간-로봇 협업 환경에서 로봇의 공간 참조 능력을 평가하였다는 점과 패널 설치 작업에서 지시적 제스처의 적용에 대한 가능성을 보여주었다는 점에서 그 의의가 있다.

주요어: 공간 참조; 인간-로봇 협업; 지시적 제스처
학 번: 2020-26076