2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

GestRight: Understanding the Feasibility of Gesture-driven Tele-Operation in Human-Robot Teams

Kevin Rippy[1], Aryya Gangopadhyay[2], and Kasthuri Jayarajah[3,4]

***Abstract*— In this paper, we propose *GestRight*, a real-time system for gesture-based tele-operation of a mobile robot. For field use (e.g., smart factory settings, search and rescue missions, etc.), relying on tablet-based controls or joysticks are limiting which has led to the recent interest in hands-free operation of these assistive robots. In this work, we design three gesture-based schemes, namely, *fist*, *touch*, and *wheel*, represent three levels of precision–intuitiveness tradeoffs for low-level navigational control of mobile robots. *GestRight* includes a head-mounted device that captures hand joint data for accurate gesture recognition which is then translated to motion commands at an edge server. Through a user study involving seventeen participants, we present quantitative insights in comparison to traditional modes of control. Specifically, we evaluate *GestRight* in terms of the ease of navigational control, task time, and amount of errors/corrective actions required, run extensive statistical analyses, and provide a series of design recommendations for gesture-driven teleoperation systems. Our results show that gesture based schemes perform as well as traditional modes of control in contrast to participants' self-reports on how successful they felt in controlling the robots.**

I. Introduction

Human-robot collaboration is an active area of research with a particular emphasis on supporting partnerships between humans and robots to safely and efficiently work together and complete shared missions. While there has been significant progress in designing completely autonomous robots that can handle complex tasks on their own, (a) their inability to generalize to entirely unseen environments, (b) the need for large amounts of training data for specific environments, and (c) situations that still require human decision (e.g., search and rescue missions), or where human presence is restricted (e.g., during circumstances such as the pandemic), favor solutions such as tele-operation.

While the most common form of teleoperation is through vendor-provided controllers such as tablets (e.g., see Fig. Im) and joysticks, recent work has explored the efficacy of design choices for visual interfaces (e.g., control over windows/sizing [1], video-centric vs. map-centric perspectives [2], constrained positioning/point-and-click [3], etc.), and more importantly, the use of more *natural*, *nonverbal* ways of communicating such as through gaze [4] and gestures [2], [5]–[7]. While the majority of works on gestures investigate the technical feasibility of recognizing gestures, in this work, we focus on the complementary goal of understanding factors that influence user experience. In particular, we seek to understand the effectiveness of using gestures for low level navigational commands in terms of both objective (e.g., task completion times, number of corrective actions required, etc.) and subjective (e.g., memorability, ease of use and user perceived cognitive load and success) measures.

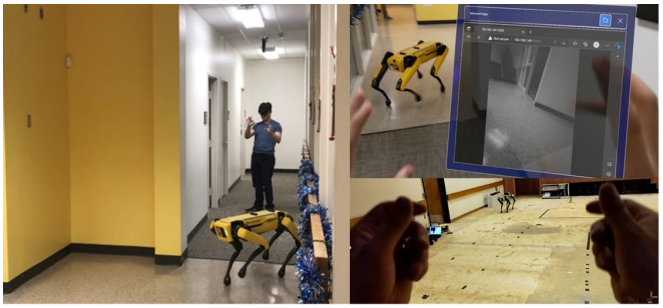


Fig. 1: *GestRight* system in operation.

Design. In this work, we present the details of our *GestRight* system. *GestRight* leverages the availability of fine-grained hand joint tracking available through multiple time-of-flight sensors on AR/VR headsets for recognizing gestures for low-level navigation of mobile robots. As an added advantage, the AR/VR display also doubles up to optionally show real-time visual feeds from the robot's point-of-view to support remote operation. Figure 1 illustrates the human operator wearing the Microsoft Hololens 2 headset *walking* the Boston Dynamics Spot robot (*Left*). The *Top-Right* view shows the point-of-view (POV) of the wearer optionally overlaid with the POV image stream from the Spot's front cameras. The *Bottom-Right* view shows the POV of the wearer during our user study performing a *steering wheel*-like action. The current *GestRight* system supports up to six basic commands, move forward, backwards, strafe left/right, and turn left/right, through three different gesture schemes. The **Fist-**based, **Touch-**based, and **Wheel-**based schemes offer varying **precision–memorability properties** [8] (see Table II). For instance, the *Wheel*-based operation resembles the familiar steering wheel of modern cars which is *highly* memorable (as reflected in survey responses from our participants). On the other hand, the *Touch*-control requires users to remember specific fingers to touch against the thumb for specific manoeuvres which users find harder to learn.

Key Hypotheses. Through an Institutional Review Board-approved user study with seventeen participants, we investi-

[1]K. Rippy is a Research Assistant with the Center for Real-time Distributed Sensing and Autonomy at the University of Maryland, Baltimore County, and this work was done while he was an undergraduate student at the university. krippy1@umbc.edu
[2]A. Gangopadhyay is a faculty member with the University of Maryland, Baltimore County, US. gangopad@umbc.edu
[3]K. Jayarajah is a faculty member with the New Jersey Institute of Technology, US. kj373@njit.edu
[4]Corresponding author

gate the feasibility of gesture-driven tele-operation for low-level navigational control of the legged platform. Specifically, we seek to find evidence to support/reject the following *Key Hypotheses:* (a) **H1:** among the various gesture-based control schemes with differing precision–memorability properties, users favour increased precision as it enables more deliberate control, (b) **H2:** gesture-based control schemes offer navigational control on par with traditional interfaces such as tablets, and (c) **H3:** users' prior familiarity with the platforms used (e.g., AR interface, mobile robot, or navigation control through driving) influences users' performance on the task, perceived cognitive load, and perceived success in controlling robots with gestures. To this end, we share quantitative insights through extensive statistical analyses.

To summarize, we make the following *Key Contributions*:

1) We present the details of three different gesture-control schemes with varying precision–memorability properties.
2) Conduct a user study with seventeen participants and statistically investigate objective and subjective measures.
3) We present a comprehensive discussion of gesture-based teleoperation.

II. Related Work

This section reviews and summarize related literature.

Gesture Recognition in Human-Robot Interaction: Non-verbal communication with robots, especially using gestures for more naturalistic ways for interaction, has received wide attention within the human-robot-interaction community [9]–[14]. Earlier works have focused on using camera-based technologies for recognizing gestures. For instance, Waldherr et al. [9] use the camera feed from the robot's interface to compare template-based and neural-network based methods to recognize gestures (e.g., arm motions) reliably. Lee et al. [10] and Yang et al. [12] study whole body gestures in conjunction with other technologies such as speech recognition and face recognition with explorations into the efficacy of using Hidden Markov Models (HMMs). Brethes et al. [11] further explore image segmentation-based approaches for face and hand tracking. More recent works have invested efforts in using newer technologies other than images/videos for gesture interpretation such as wearable [13] and depth [14] sensing. In our work, we focus our attention towards exploiting AR/VR headsets which capture fine-grained, joint-level 3D positions of the hand for recognizing gestures to control a distant robot. While we use template-based matching to capture low-level navigation commands, our research focus is on understanding the variabilities of different gesture schemes in terms of the usability of the system, as opposed to coming up with the most accurate recognition techniques.

Interfaces for Robotic Teleoperation: Situations where autonomous exploration has not been feasible yet, teleoperation by a manual operator is an alternative. Early works [1], [2], [15], [16] have explored the feasibility of a variety of interfaces. Several works have explored the impact of the display configurations (e.g., control over windows/sizing [1], video-centric vs. map-centric perspectives [2], constrained positioning/point-and-click [3], etc.) on user performance. Works such as [17]–[19] have investigated various aspects of teleoperation such as the impact of personality traits in a desktop-based teleoperation setting [17], mimickry-based manipulation and an (EMG)-based approach for operating robotic arms [18], [19]. Similar to our work, recent works have explored using gestures for teleoperation; Hu et al. [2] explore the feasibility of gesture control, however, focus more on the technical aspects of using image sequences for extracting gestures (e.g., segmentation, morphological filtering, etc.) and Ajili et al. [5] investigate procedures for whole-body gesture recognition in the presence of human movement. More recently, De et al. [6] control drone motion using gestures estimated from skeletal data. While these works focus on the accuracy of the gesture recognition pipelines, in our work, we investigate the orthogonal question of, given an accurate gesture recognition system, how do they perform in terms of effectiveness and usability in teleoperation scenarios, in comparison to traditional modalities such as a tablet-based visual interface. Most recently, Godoy et al. [19] investigated EMG-based gesture recognition for manipulation tasks such as grasping with electrodes attached to the user's hands; our work adds to this body of work by exploring the feasibility of gesture-based operations using 3D hand tracking data in the context of navigational control.

Closest to our work, Lee et al. [4] investigated the feasibility of using *gaze*, another mode of nonverbal communication, for teleoperation. Similar to our case, they designed a Hololens 2 based system that gauges where a user is looking at to control the motion of a Robotnik Summit-XL platform. Our experiments confirm their observations that novel, non-verbal modalities can in fact perform as comparable to traditional control mechanisms such as with tablets and joysticks.

III. Designing *GestRight*

In this section, we provide implementation details of our *GestRight* system. Figure 2 illustrates the overall architecture. Three-dimensional hand joint data captured through the Microsoft Hololens 2 headset [1] is offloaded to an edge server for real-time processing of gestures and translation to low-level motion commands for the Spot platform. We now describe the functions at each of the three devices.

A. Head-Mounted Display

The Hololens 2 device offers a complete suite of tracking technologies including head tracking, eye tracking, hand tracking, and point-of-view images and depth. For *GestRight*, we extend an existing publicly available client–server implementation[2]. The spatial information (head, eye and hand tracking) data is streamed at an effective rate of ≈ 59 Hz over a private WiFi network while the images and long

[1] https://www.microsoft.com/en-us/hololens/hardware
[2] https://github.com/jdibenes/hl2ss/tree/main

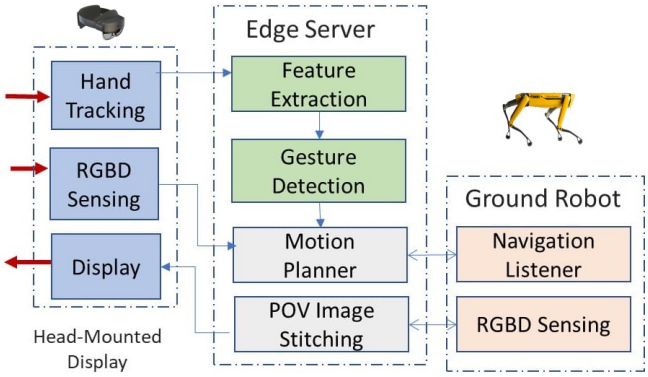


Fig. 2: Overview of the *GestRight* Framework

throw depth frames are transferred at $\approx$ 2.2 FPS and $\approx$ 0.5 FPS, respectively. The latter two data streams are captured for ground-truth, and are not required for the gesture-based operation. Optionally, we stream the real-time visual feed from the Spot platform to the wearer's display (details in Section III-C).

B. Edge Server

We implemented the edge server on a laptop with an Intel Core i5-12450H processor, NVIDIA GeForce GTX 1650 GPU, and 8 GB RAM. It acts as a master node running on Robotic Operating System (ROS) Noetic. The edge server performs two key functions: (a) extract features from the 3D hand joint data, implements three different gesture schemes, and translates them to ROS-based motion commands for the Spot platform, and (b) receives streams of real-time visual feeds from the four cameras aboard Spot, stitches them, and exposes it through a webserver. The visuals can then be accessed through a web view on the Hololens 2 device (or any other device that supports web browsing).

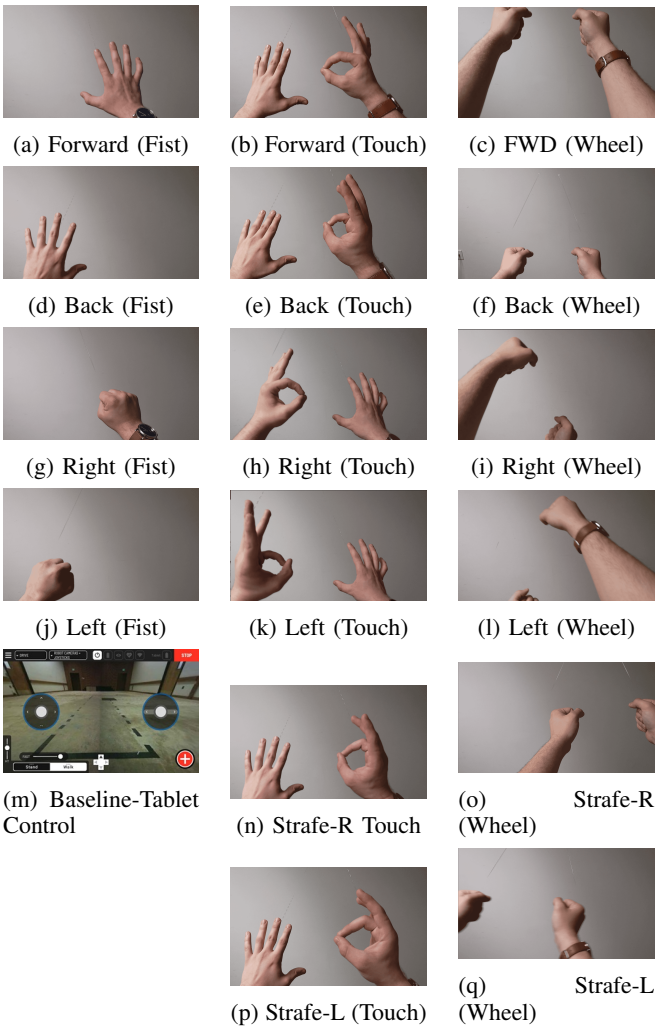
Hand Feature Extraction and Gesture Recognition: *GestRight* implements three different gesture schemes with different precision–memorability properties, building on the definitions of Nacenta et al. [8]. The schemes are defined as: (a) *Fist*-based - a simple control using opening and closing of palms, (b) *Touch*-based - a scheme that allows fine-grained control of the robot using a combination of fingertip touches, and (c) *Wheel*-based - a scheme that resembles the everyday task of driving a car. In Table I, we tabulate the dictionary of controls and their corresponding poses for each of the schemes.

1) Fist*-based control (Figs. Ia, Id, Ig, Ij):* This scheme supports 4 controls, forward/backward/turn right/turn left. The presence of the right hand controls the forward (open fist) and turning right (closed fist) motions, and the presence of the left hand controls the backward (open fist) and turn left (closed fist) motions. When both hands are present within the AR/VR headset's viewing angle, the right hand pose is prioritized.

Criteria for open/close: A hand is declared as "open" if the summation of the euclidean distance of (1) the index finger's tip to the distal joint, (2) the distal joint to the proximal joint, and (3) the proximal joint to the metacarpal joint is within a threshold d_{open} of the distance from the tip to the metacarpal joint. A higher d_{open} leads to a looser definition of open, and hence less precise control of the scheme. We set $d_{open} = 10\%$ in our experiments as a good trade-off between precision and usability.

2) Touch*-based control (Figs. Ib, Ie, Ih, Ik, In, Ip):* This scheme differentiates between fingertip "touches", specifically, the touching of the thumb against different fingers. They are defined as: **Move forward:** right index finger, **Move backward:** right pinky finger, **Strafe right:** right middle finger, **Strafe left:** right ring finger, **turn right:** left index finger, and **turn left:** left pinky finger.

Criteria for touch: Two fingers are declared as "touching"



(a) Forward (Fist) (b) Forward (Touch) (c) FWD (Wheel)
(d) Back (Fist) (e) Back (Touch) (f) Back (Wheel)
(g) Right (Fist) (h) Right (Touch) (i) Right (Wheel)
(j) Left (Fist) (k) Left (Touch) (l) Left (Wheel)
(m) Baseline-Tablet Control
(n) Strafe-R Touch
(o) Strafe-R (Wheel)
(p) Strafe-L (Touch)
(q) Strafe-L (Wheel)

TABLE I: Three different schemes for navigation control of the Boston Dynamic Spot platform used in the study

Scheme	Precision	Memorability	Params.
Fist	Medium	Medium	$d_{open} = 10\%$
Touch	High	Low	$d_{touch} \leq 0$
Wheel	Medium	High	$d_{vertical} \geq 0$, $a = \pm 20°$, $d_{turn} = \pm 20cm$

TABLE II: Precision–Memorability properties of the three gesture schemes.

if the euclidean distance, $d_{touch} = d2 - d1$ is less than zero where $d1$ is the first finger tip and the thumb tip and $d2$ is the distance between the first finger tip and its corresponding distal joint. Comparing the distance to the length between the tip and the distal joint generalizes the scheme to work across people with different hand sizes (as is evident from our user study described later in Section IV.

3) Wheel-*based control (Figs. Ic, If, Ii, Il, Io, Iq):* This scheme is designed to mimick the controls of the steering wheel - a familiar, everyday control scheme. The user puts out both their hands in front of them as if they are holding a hypothetical wheel, and rotate the wheel for turning right/left. Unlike the real wheel, here, the forward, backward, strafe right and strafe left motions are achieved by moving the hypothetical wheel up, down, to the right and to the left, from "center" of the wheel, or the neutral position.

Calibrating the neutral position: The user initially holds out their hand at a relaxed position, and touch the right index to the right thumb. This starting point is saved and subtracted by the head position at that time, to give a point relative to the head position. The coordinate at the center of the two palms is calculated as the *center of the wheel*. This calculation is also generalizes to support users of varying heights.

Criteria for Forward/Backward: If the vertical position (captured by the $Y-$axis of the joint data) of the wheel position is higher than a threshold $d_{vertical}$ is higher (or lower) than the neutral position (normalized zero position), then it is declared as Forward (or Backward).

Criteria for Strafe Right/Left: We compute the forward vector of the user (or the MS Hololens 2 device, more precisely, $V_{forward}$), and the head pose vector subtracted from the center-point of the user's hands (V_{head}). We project the two vectors to the $X-Z$, and calculate the angle between them. If the angle is greater (or less) than a threshold, a, we declare a "Strafe Right" (or "Strafe left) action. In our studies we set $a = 20$ and $a = -20$ for both cases, respectively.

Criteria for Turn Right/Left: Depending on whether the user's right or left hand is at a higher position vertically than the other, a Turn Right or Left action is declared. Again, a distance threshold d_{turn} allows for tuning, and experimentally, we find $d_{turn} = 5cm$ to be a good trade-off between precision and usability.

Image Stitching: We implemented a Flask-based web application [3] building on the Spot platform's image service[4]. The Flask app then enables any other device with web browsing capabilities to be able to view the stitched frontal camera views over the Transmission Control Protocol (TCP).

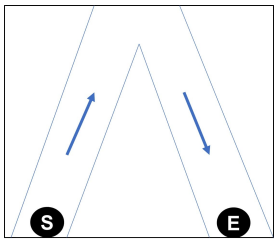
C. Mobile Robots

We consider the Boston Dynamics Spot platform as the mobile robot that is tele-operated. The Spot platform runs as a ROS slave whose navigation listener listens to the $/cmd\_vel$ topic for linear and angular changes to velocity as *Twist* objects. Throughout our studies, we set linear motions at 2 m/s and angular motions at 50 deg/s. We point to the astute reader that the gestures are processed at $\approx$ 59 Hz meaning that at each command, the robot moves by a maximum of $\approx$ 3.3 cm (linear) or $\leq$ 1 degree.

Additionally, we also subscribe to the image streaming topics for stitching the Spot's Point-of-View to be relayed back to the human controller.

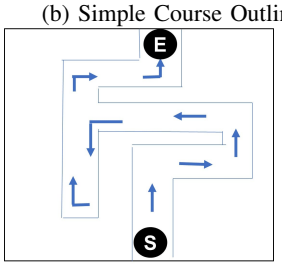
Code and Data Sharing: We make our codebase for gesture recognition and navigation control (implemented in Python) for the *GestRight* system available publicly[5]. The data collected through our user study (described next) will also be made available to the wider research community on request.

(a) Simple Course



(b) Simple Course Outline

(c) Complex Course



(d) Complex Course Outline

Fig. 3: Courses used in the main study where participants were asked to tele-operate the Spot platform. The **Simple** course used for practice has a single turn while the **Complex** course has seven tight turns required by the participants. **S** and **E** refer to the Start and End points of the two courses.

IV. User Study

In this section, we present the details of the user studies used in evaluating the effectiveness of the gesture-based control schemes *GestRight* supports.

A total of **17 participants** (11 males, 6 females) with age ranging from 18–43 participated in this IRB-approved study. The participants were recruited widely through solicitations via institutional email and no monetary compensation was provided. Participants wearing any form of medical implants such as pacemakers were not considered in the study as a safety precaution. Participants gave written informed consent following a briefing of the study. During the registration process, the participants also provided information on their prior experience with using AR/VR platforms, driving, and working with mobile robots. In all our studies, the participants wore the Microsoft HoloLens 2 AR/VR headset

[3] https://flask.palletsprojects.com/en/3.0.x/

[4] https://dev.bostondynamics.com/python/examples/get_image/readme

[5] https://github.com/Connected-and-Autonomous-Systems-Lab/GestRight

which captures fine-grained, joint-level data streams of the wearer's hands in addition to head motion, eye gaze vector as well as first person imagery and depth data. Additionally, we measured the Galvanic Skin Response and heart rate of participants using the Shimmer[6] sensor. For best contact and reduced motion artifacts during the studies, we placed the electrodes over the participants' shoulder, on the back. However, we point to the reader that we don't present analyses from the objective physiological measures from this sensor in this paper.

In particular, we seek to find evidence to support/reject the following *Key Hypotheses:*

1) H1: Among multiple possible gesture-based control schemes with differing precision and intuitiveness properties, users favour increased precision.
2) H2: Gesture-based control schemes offer navigational control on par with traditional interfaces such as tablets.
3) H3: Users' prior familiarity with the platforms used (e.g., AR interface, mobile robot, or navigation control through driving) can impact user's perceived success in controlling robots with gestures.

The study was split into three phases, and conducted across three locations: lab setting, a Simple Course and a Complex Course (see Figure 3).

Phase 1: Learning- During this phase, we introduce the three different control schemes that *GestRight* supports in a lab setting. The experimenter first demonstrated the forward, backward, left, right and strafing manoeuvres using their hands/fingers, asks the participants to repeat it, and clarifies as many times as each participant requested. We also asked an open-ended question where the participants are asked to perform the above set of commands in their most natural way. At the end of this phase, the participants were provided a questionnaire that asked them to rate the ease of use of each type of control scheme on a Likert scale (1-hard, 7-easy), and compare the ease of learning (for the three schemes *GestRight* supports) or authoring (for the user-defined scheme) of the different schemes.

Phase 2: Practice- In this phase, we introduced the Boston Dynamics Spot platform to the participants. They were asked to use the *GestRight* gesture control schemes to control and move the Spot robot along a **Simple Course** (see Figures 3). The simple course mostly consisted of forward moves with only involving a single turn. To allow for more open ended practice, we also marked a few obstacles (i.e., a stool and desk) towards the end of the course for participants to navigate the robot around them.

Phase 3: Tele-operation- In the final phase, we asked the participants to move the Spot robot along a **Complex Course** consisting of seven turns that required tighter, more disciplined, use of the gestures. The default control of the platform via the Tablet provided by the vendor was also used by the participants as the baseline. In both the second and final phases, we randomized the order of the control scheme provided in order to eliminate any ordering effects. The participants responded to the NASA TLX questionnaire [20] at the end of each control scheme, and completed a post-study questionnaire after all the trials were completed. In the post-study questionnaire, the participants were asked to rate each scheme in terms of (a) how easily they could recall the motion commands, (b) how easily they were able to control the robot's movements, and (c) how *fun* each scheme was to work with on a four-point scale (1-hardest, 4-easiest, 1-least fun, 4-most fun). Finally, they were also open-ended questions on which schemes they thought led to more errors and any feedback for improving the system usability.

[6]https://shimmersensing.com/

Question	Fist	Touch	Wheel	Baseline
Learning				User Defined gestures
Ease of use	5.73 (1.22)	4.93 (1.49)	5.47 (1.92)	5.8 (1.01)
Ease of learning/creating [R]	2.73 (0.79)	2.6 (1.12)	2.93 (1.03)	3.07 (0.59)
Tele-operation				Tablet control
Ease of recall [R]	2.67 (0.98)	3.22 (0.89)	2.6 (1.18)	3 (0.85)
Ease of use/navigation [R]	2.64 (0.84)	3.13 (0.83)	2.2 (1.21)	2.93 (0.80)
Fun [R]	2.6 (1.12)	3.6 (0.63)	2.93 (1.16)	2.93 (0.88)
Perceived Cognitive Load				
Mental demand	3.56 (1.72)	3.17 (1.47)	4.82 (1.47)	2.35 (1.66)
Physical demand	3.28 (1.71)	2.56 (1.54)	4.47 (1.87)	1.65 (1.37)
Pace	3.83 (1.69)	3.59 (1.66)	3.88 (1.65)	3.18 (1.88)
Perceived success in accomplishing the goal	4.05 (1.7)	4.82 (1.81)	3.52 (1.62)	5.47 (1.77)
Perceived effort required to accomplish the goal	4.16 (1.76)	3.64 (1.53)	5 (1.5)	2.47 (1.41)
Insecurity, discouragement, annoyed	3.38 (1.75)	2.35 (1.32)	3.41 (1.58)	1.94 (1.43)

TABLE III: Self-reported average ratings from the questionnaires. Standard deviations within brackets. Items marked as [R] are questions where participants were asked to *rank* the four different control schemes.

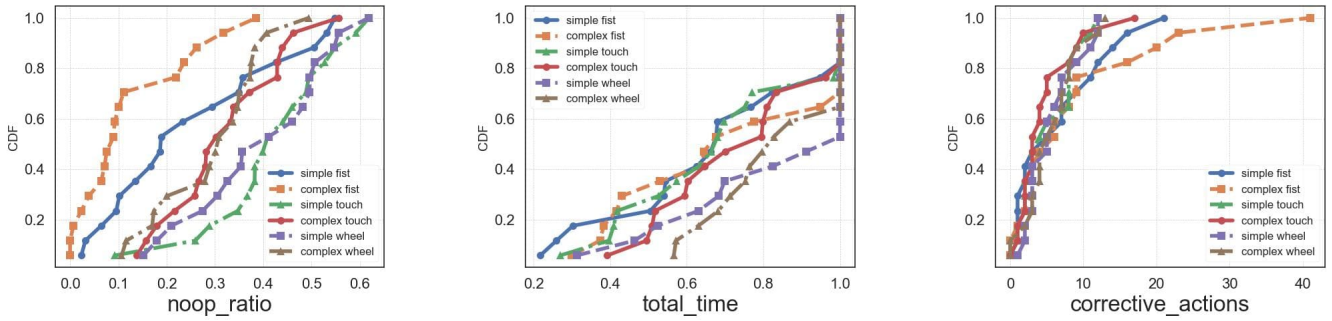
V. Quantitative Insights

In this section, we report on experimental results from the user study. In Section V-A, we first summarize the self-reported ratings from participants, and then in Section V-B, we quantify the performance of the trial runs. In Section V-C, we further investigate how past experiences with related technologies impact the perceived load and success in using gesture-based tele-operation systems.

A. Self-Reported Responses

Based on the self-reports collected during Phases 1 and 3 of the user study, we summarize participant ratings in Table III in terms of (a) ease of learning, (b) efficiency of use, and (c) perceived cognitive load and success, and make the following observations:

1) **Ease of Learning:** After the first phase (i.e., Learning), participants rated the control scheme that they suggested to be the easiest to learn or create (average rating of 3.07), and the *Touch*-based control to be the hardest to learn (2.6 on average). The *Wheel-based* control mimicks an everyday control that many participants were familiar with, and the *Fist*-control has a shorter dictionary than the other two schemes both of which were rated better than *Touch*.
2) **Effectiveness of Schemes:** Responses from immediately after completing all the trials on the *Complex Course* revealed surprising insights. Although the participants rated the *Touch*-scheme to be the hardest to



(a) Empirical CDF of the ratio of *null* actions

(b) Empirical CDF of the (normalized) total time taken to complete the courses

(c) Empirical CDF of the number of corrective actions taken by the participants.

Fig. 4: Empirical distributions of the $noop-ratio$, *total time taken for course completion and number of corrective actions taken.*

learn, after using the gestures for navigating the robot, they found it be the most effective. In fact, the *Touch*-control outperforms the traditional tablet-based control in terms of all three aspects: ease of recall (3.22 vs. 3.03), ease of navigation (3.13 vs 2.93), and most fun to use (3.6 vs. 2.93). We argue that the *Touch-scheme* provides two advantages over other schemes: (a) control is discrete - i.e., there are natural pauses between the different gesture poses which allows for more precise control (whereas *Wheel*- and *Tablet* are continuous with most poses being "actions"), and (b) *Touch*-control offers a larger dictionary of poses compared to *Fist*-control which is the only other discrete scheme.

3) **Perceived Cognitive Load and Success:** Reaffirming the previous result, the *Touch*-control is rated on par with the *Tablet*-control in terms of the lower mental and physical effort required. Participants also felt that they were most *successful* with these two schemes. On the contrary, although participants rated the *Wheel*-control to be the easiest to learn, they also found it to be the least preferred for use with the actual Spot platform. We ran a series of paired $t-$tests which confirms that these disparities are statistically significant. We find that the distributions of ratings for *Touch* and *Tablet* are not statistically significantly different (with p-values ranging from 0.07 to 0.50) for the NASA TLX items: mental demand, physical demand, pace, success and insecurity. At the same time, we find that the distributions of ratings for *Wheel* and *Tablet* are statistically significantly different (with p-values ranging from 6.450e(-5) to 0.002) on the same items.

Hence, we **do not reject the null hypothesis H1**.

B. Objective Measurements

While the previous section analyses participant-reported ratings, in this section we investigate the performance of the participants in terms of navigating the Spot robot through the two different courses, objectively. To this end, we measure the time they took to complete the courses, their efficiency in using the different control schemes, and the instances they needed to take corrective actions.

Completion Time: In Figure 4b, we plot the empirical cumulative distribution function (ECDF, on the $y-$axis) of the total time (on the $x-$ acis) participants took to complete the two courses, **Simple** and **Complex**, across the three different schemes *GestRight* offers. While the self-reports exemplified a strong preference of participants for the *Touch*-control as opposed to the *Wheel*-control, the data shows that there were no statistically significant differences. For instance, pairwise Kolmogrov-Smirnov (KS) tests on the distributions of total time taken of the three *GestRight* schemes, across both courses, return $p-$values ranging from 0.112 to 0.751 rejecting the null hypothesis that the participants' performance across the different gesture schemes are statistically significantly different. More interestingly, we also find that the KS-tests further show that the time taken for the gesture schemes are not different from the tablet-based control times - with $p-$ values ranging from 0.245 to 0.963. Hence, we **do not reject the null hypothesis H2**.

No Operation Ratio ($noop-ratio$): To measure the efficiency of use of the gesture poses, we distinguished deliberate actions or operations and poses that did not correspond to any navigation related actions. We computed a $noop-ratio$ which is the fraction of poses with no actions over the total number of gesture poses the participants performed, for both courses. We plot the CDF of the $noop-ratio$ in Figure 4a. Pairwise KS-tests between *Touch*-control and *Wheel*-control, contrary to participants' perceived load, shows that the two distributions are not statistically significantly different ($p-values$ of 0.751 on both simple/complex course). Interestingly, we see that the *Fist*-control, however, had a significantly larger $no-op$ ratio than the other schemes ($p-$values ranging from 0.0002 to 0.11). Further explorations into the action sequence the participants took reveals that in the case of *Fist*, participants preferred to issue a series of "Right" commands to rotate in place in order to turn "Left". This increases the number of total actions taken by the participants disproportionately

Scheme	AR/VR	Driving	Mobile Robots
Fist	0.10	-0.09	-0.20
Touch	**-0.73**	**-0.46**	**-0.56**
Wheel	**0.52**	**0.27**	**0.59**

TABLE IV: Pearson's correlation values of task time on the Complex course and participants' prior experience with AR/VR-based gaming, Driving and Mobile Robots.

leading to lower than usual $noop - ratios$.

Corrective Actions: As a measure of accuracy of control, we counted the number of instances the participants had to "correct" their course in the event of overshooting. For instance, when a participant attempts to turn "Left" by issuing a series of commands, it is often followed by a shorter sequence of the opposing action - in this case, "Right" - as a measure of correcting for an overshoot. We look for such signatures in the action sequence, and we plot the CDF ($y-$axis) of the number of corrective actions taken ($x-$axis) within each trial. Confirming our previous results, the three gesture schemes do not show any statistically significant difference in terms of the overshoots or mistakes participants made ($p-$values ranging from 0.456 to 0.963 on pairwise KS-tests).

C. Impact of User Background on Task Performance

In this section, we further investigate the influence of a participants' background such as their prior familiarity with AR/VR platforms, driving experience (recall that our *Wheel*-control resembles this everyday task), and mobile robots. To this end, we ran a series of Pearson correlation tests to understand the co-varying trends of each of these independent variables and the objectively measured task completion times. We tabulate the correlation values and **bolden** those that had a $p-$value $\leq$0.05 in Table IV.

Furthermore, we ran multiple-ANOVA analyses with these independent variables and the task time as the dependent variable, for each scheme separately, for the Complex course case. In all cases, we find that the three variables are statistically significant with $p-$values ≤ 0.05.

Hence, we **do not reject the null hypothesis H3**.

VI. Discussion

A. Easy to Learn vs. Ease of Use

The observations in Section V-A partially confirm our initial remark that the more familiar schemes will be easier to remember as the participants' ratings show after the Learning phase. However, we observe strong negative relationships between what was considered to be easier to learn (more memorable) in the first phase, and what was considered to be easier to use (more precise) confirming our hypothesis that users would favor precision over memorability, especially with increased familiarity. Further to the average ratings presented in Table III, we further find that the Pearson's correlation coefficients for the ratings for Easy to Learn and Easy to Use for the three schemes are: -0.70, -0.81, and -0.89, with $p-$values ≤ 0.01.

B. Gesture vs. Default Control

The three objective measurements further show that the participants' performance using all three gesture-based schemes were in fact on par with the tablet-based visual interface demonstrating the effectiveness of this modality for teleoperation. Furthermore, the self-reports even show that the participants' found the *Touch*-control to be easier to use, more fun to use, and easier to navigate than the tablet control.

C. Relationship to Experience with Related Technologies

Quite surprisingly, we see strong opposing relationships for the *Touch*-based and *Wheel-based* schemes. The strong negative correlation between experience in terms of using AR/VR for video games, driving and mobile robots, and the task time suggests that users with more experience with these related technologies took lesser time, on average, and vice versa, to complete the tasks. In comparison, those with higher experience with these technologies seemingly ended up taking longer duration to complete the trials when using the *Wheel*-based scheme. We believe that a possible explanation to this is that driving on the road typically involves larger maneuvers (or, radius of gyration), on fewer occasions, which is quite different from the Complex Course that the user studies were conducted on. Although the design of the Wheel resembles the steering wheel, the differences in conditions (outdoors vs. confined indoors) could have led to the participants' lower preference and performance for this scheme. An interesting future direction is thus to understand preferences for such varied environments which we defer as future work.

D. Open Questions

Our user study reveals several useful insights. While this is a promising starting point, our work poses several interesting future directions.

Fluency Estimation and Online Adaptation: While our scope for the current work was to investigate the feasibility of using gestures for teleoperation, our findings suggest several intricate differences between the different gesture schemes. Even though both the Simple and Complex courses were performed on the same day, there are still significant differences in performance between these two trials. For instance, in Fig. 4a, we see that while all participants had a $noop_ratio$ of less than 0.38 in the second trial (Complex), the same was larger in the first trial (0.52, Simple) denoting a marked improvement from the first to the next trials. Similar trends can be observed across all three schemes, for all three measures (ratio, time taken and corrective actions). A key question then is to understand how users **fluency** with the system can then change their experience and performance of using the system. Furthermore, while we set the parameters (see Table II) to be fixed values in our current experiments, the online adaptation of these parameters with improving fluency of the system users would be an interesting future direction.

Objective Measurements of Cognitive Load: As we see in our user study, the perceived workload is at odds with the

objective measurements such as task completion time and corrective actions needed. The various physiological sensing data that we ascertained during these trials using the Shimmer device will be useful in understanding discrepancies between self-reports, actual performance, and physiological states.

Remote Execution: As we alluded to earlier, while our experiments were conducted in indoor environments with the Spot robot within the line-of-sight of the operator, another pertinent direction to investigate is whether these observations generalize to other settings; notably, outdoors and remote operation where the operator is only able to observe Spot's POV streams.

VII. Conclusion

In this paper, we investigated the impact of precision/memorability properties of gesture-based teleoperation schemes on objective task performance (e.g., in terms of task time, proportion of "null" actions within individual trials, etc.) and user perceived workload and success. We ran extensive statistical analyses to test our hypotheses to explore the links between effectiveness of using gestures as compared to default modes of operation such as through visual interfaces. Based on our analysis, we concluded that even though gesture schemes that resemble everyday operations (such as driving) maybe easier to learn, users find schemes that offer more precise control of the robot to be more acceptable and easier to use. We learned that while users felt less successful in performing the tasks as compared to when using the tablet interface, data shows that the time they took to complete the task were comparable. We further concluded that the user's prior experience with driving and/or navigating mobile robots had a significant influence on whether they perceived their task to be successful. Overall, we conclude that gesture-based schemes remain attractive for teleoperation, and considerations for precision and memorability of the specific schemes play a critical role in user performance and success. Future research could investigate the long term effects of gesture-based teleoperation, specifically, towards exploring whether notions of precision/memorability can change as users become more familiar, or ***fluent***, with the system.

Acknowledgement

We acknowledge the support of the U.S. Army Grant No. W911NF21-20076. We thank Harsh Shroff for his help in collecting data and the reviewers for their suggestions.

References

[1] R. Olivares, C. Zhou, B. Bodenheimer, J. A. Adams, *et al.*, "Interface evaluation for mobile robot teleoperation," in *Proceedings of the ACM Southeast Conference (ACMSE03)*, vol. 112. Citeseer, 2003, p. 118.

[2] C. Hu, M. Q. Meng, P. X. Liu, and X. Wang, "Visual gesture recognition for human-machine interface of robot teleoperation," in *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, vol. 2. IEEE, 2003, pp. 1560–1565.

[3] D. Kent, C. Saldanha, and S. Chernova, "A comparison of remote robot teleoperation interfaces for general object manipulation," in *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, 2017, pp. 371–379.

[4] J. Lee, T. Lim, and W. Kim, "Investigating the usability of collaborative robot control through hands-free operation using eye gaze and augmented reality," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 4101–4106.

[5] I. Ajili, M. Mallem, and J.-Y. Didier, "Gesture recognition for humanoid robot teleoperation," in *2017 26Th IEEE international symposium on robot and human interactive communication (RO-MAN)*. IEEE, 2017, pp. 1115–1120.

[6] K. B. de Carvalho, D. K. D. Villa, M. Sarcinelli-Filho, and A. S. Brandao, "Gestures-teleoperation of a heterogeneous multi-robot system," *The International Journal of Advanced Manufacturing Technology*, vol. 118, no. 5, pp. 1999–2015, 2022.

[7] W. Zhang, H. Cheng, L. Zhao, L. Hao, M. Tao, and C. Xiang, "A gesture-based teleoperation system for compliant robot motion," *Applied Sciences*, vol. 9, no. 24, p. 5290, 2019.

[8] M. A. Nacenta, Y. Kamber, Y. Qiang, and P. O. Kristensson, "Memorability of pre-designed and user-defined gesture sets," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2013, pp. 1099–1108.

[9] S. Waldherr, R. Romero, and S. Thrun, "A gesture based interface for human-robot interaction," *Autonomous Robots*, vol. 9, pp. 151–173, 2000.

[10] S.-W. Lee, "Automatic gesture recognition for intelligent human-robot interaction," in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*. IEEE, 2006, pp. 645–650.

[11] L. Brethes, P. Menezes, F. Lerasle, and J. Hayet, "Face tracking and hand gesture recognition for human-robot interaction," in *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, vol. 2. IEEE, 2004, pp. 1901–1906.

[12] H.-D. Yang, A.-Y. Park, and S.-W. Lee, "Gesture spotting and recognition for human–robot interaction," *IEEE Transactions on robotics*, vol. 23, no. 2, pp. 256–270, 2007.

[13] P. Neto, M. Simão, N. Mendes, and M. Safeea, "Gesture-based human-robot interaction for human assistance in manufacturing," *The International Journal of Advanced Manufacturing Technology*, vol. 101, pp. 119–135, 2019.

[14] G. Canal, S. Escalera, and C. Angulo, "A real-time human-robot interaction system based on gestures for assistive scenarios," *Computer Vision and Image Understanding*, vol. 149, pp. 65–77, 2016.

[15] D. Labonte, P. Boissy, and F. Michaud, "Comparative analysis of 3-d robot teleoperation interfaces with novice users," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 40, no. 5, pp. 1331–1342, 2010.

[16] J. Kofman, X. Wu, T. J. Luu, and S. Verma, "Teleoperation of a robot manipulator using a vision-based human-robot interface," *IEEE transactions on industrial electronics*, vol. 52, no. 5, pp. 1206–1219, 2005.

[17] G.-E. Cha, W. Jo, and B.-C. Min, "Implications of personality on cognitive workload, affect, and task performance in remote robot control," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 4153–4160.

[18] Y. Wang, C. Sifferman, and M. Gleicher, "Exploiting task tolerances in mimicry-based telemanipulation," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 7012–7019.

[19] R. V. Godoy, B. Guan, A. Dwivedi, and M. Liarokapis, "An affordances and electromyography based telemanipulation framework for control of robotic arm-hand systems," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 6998–7004.

[20] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in psychology*. Elsevier, 1988, vol. 52, pp. 139–183.