Int. J. Human–Computer Studies 177 (2023) 103079

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Enhanced Interactivity in VR-based Telerobotics: An Eye-tracking Investigation of Human Performance and Workload

Federica Nenna [a,b,*], Davide Zanardi [a], Luciano Gamberini [a,b]

[a] Dipartimento di Psicologia Generale, Università degli studi di Padova, Padova, Italy Via Venezia 8, 35131, Padova, Italy
[b] HIT | Human Inspired Technology Centre, University of Padova, Padova, Italy Via Luzzatti 4, 35131, Padova, Italy

ARTICLE INFO

Keywords:
Virtual Reality
Human-Robot Interaction
Human performance
Mental Workload
Eye-tracking
Telerobotics

ABSTRACT

Virtual Reality (VR) is gaining ground in the robotics and teleoperation industry, opening new prospects as a novel computerized methodology to make humans interact with robots. In contrast with more conventional button-based teleoperations, VR allows users to use their physical movements to drive robotic systems in the virtual environment. The latest VR devices are also equipped with integrated eye-tracking, which constitutes an exceptional opportunity for monitoring users' workload online. However, such devices are fairly recent, and human factors have been consistently marginalized so far in telerobotics research. We thus covered these aspects by analyzing extensive behavioral data during simulated guidance of an industrial robot in VR through a pick-and-place task. Users drove the robot via button-based and action-based controls and under low (single-task) and high (dual-task) mental demands. We collected self-reports, performance and eye-tracking data. Specifically, we asked i) how the interactive features of VR affect users' performance and workload, and additionally tested ii) the sensibility of diverse eye parameters in monitoring users' vigilance and workload throughout the task. Users performed faster and more accurately, while also showing a lower mental workload, when using an action-based VR control. Among the eye parameters, pupil size was the most resilient indicator of workload, as it was highly correlated with the self-reports and was not affected by the user's degree of physical motion in VR. Our results bring a fresh human-centric overview of human-robot interactions in VR, and systematically demonstrate the potential of VR devices for monitoring human factors in telerobotics contexts.

1. INTRODUCTION

Recently, the industrial sector has witnessed a massive shift of general interest from machines to humans, making the latter the core of the current industrial evolution. The manifest of Industry 5.0 is indeed "human-centric manufacturing" (Lu et al., 2022), which places the worker's well-being at the center of the production process. Such a framework was proposed in response to the challenges smart and intelligent manufacturing technologies have posed. Collaborative robots, or cobots, are one practical example of this technology (Faccio et al., 2022, Krüger et al., 2009, Wang et al., 2017), which removes boundaries and allows humans to interact with machines in a fluid-tight manner, even remotely. The close integration of humans, automated and intelligent robots, and digital platforms increased the skill demands for the operators, requiring them to cope with cognitive loads to work efficiently (Doolani et al., 2020). Therefore, the aims of Industry 5.0 are to focus more on the human workers, with their individual needs and capabilities. Nonetheless, practical human-centric applications are still limited, and researchers still tend to pay greater attention to the technical aspects of industrial systems rather than human and cognitive factors related to workers (Grandi et al., 2020; Smith and Sepasgozar, 2022).

Some researchers have proposed and tested various approaches for assessing human factors, such as user experience, stress, fatigue and mental workload, in interactions with robotic systems (e.g., Chien et al., 2018; Villani et al., 2020). Furthermore, recent methodological advances support multimodal assessments, combining various behavioral and physiological tools to understand humans and their workload in the field fully (Dehais et al., 2020; Matthews et al., 2015). However, it is important to acknowledge that measuring workload in highly interactive environments is not easy (Dehais et al., 2020). Workload is commonly quantified via self-reports (e.g., NASA-TLX questionnaire, Hart and Staveland, 1988), which show diverse disadvantages: their administration inevitably disrupts the task flow, and they are prone to

* Corresponding author: Federica Nenna, University of Padova, Department of General Psychology Via Venezia 8, 35131, Padova, Italy, Phone: +39 3402146430
E-mail address: federica.nenna@phd.unipd.it (F. Nenna).

https://doi.org/10.1016/j.ijhcs.2023.103079
Received 15 November 2022; Received in revised form 15 May 2023; Accepted 16 May 2023
Available online 23 May 2023

inter-subject variability and to the individual's ability to self-assess (Carswell et al., 2005). Furthermore, even when one uses multiple methods to evaluate a person's psychological state, such as self-reports and psychophysiological measurements, the tools used can be more or less intrusive for the task and environment. For example, measuring heart rate or electrodermal activity may require additional bands or electrodes, which can complicate the setup and cause data transmission or synchronization problems, even with fully wireless equipment. Therefore, although multimodal assessments can increase our understanding of human behavior, it is important to consider these limitations when implementing them in practical contexts.

Therefore, the question arises of whether and how workload can be assessed continuously, without task disruptions or physical obtrusions to the setup, and more objectively. Eye-tracking systems can help fill this gap because they are increasingly portable and affordable and can capture workload-related eye behavior in the field (Novak et al., 2015). Also, the most recent virtual reality (VR) headsets are equipped with an integrated eye tracker that allows for the measurement of eye-related physiological indexes in virtual environments without disrupting users' actions. This is an exceptional opportunity for continuous workload monitoring during VR robotic teleoperation.

Not without reason, the current study focuses on VR-based robotic teleoperations, which are gaining significant traction in industrial contexts (Franzluebbers and Johnsen, 2019; Martín-Barrio et al., 2020; Rosen et al., 2018). Compared to traditional teleoperation means, VR enables immersion in realistic environments and the use of teleoperation modalities that reproduce the manual features of cobots: those modalities include gestures or physical actions (e.g., Martin Barrio et al., 2019), going beyond the conventional keyboard, mouse, and joysticks (e.g., You and Hauser, 2012; Mavridis et al., 2015). The strength of action-based controls is that they leverage natural and embodied controls, allowing users to manipulate a replica of the robot without buttons, performing physical actions or gestures similar to those used when manipulating physical robots, which significantly streamlines robot programming. However, button-based controls are commercially widespread and therefore more familiar to most. Given the fair novelty of VR-based robotic teleoperation, how the VR system control affects work efficiency and human factors (i.e., performance, workload) remains unclear.

Besides a novel computerized methodology in industrial engineering and robotic teleoperation, VR is an exceptionally valuable tool for assessing human factors in simulated industry. It has been demonstrated that VR allows users to gain information on the qualitative experience while they interact with a product (e.g., Rebelo et al., 2012). Furthermore, extensive data can be gathered during VR experiences or operations, including time series data on the VR headset and controllers' position and rotation as well as timestamps of interactions between physical and virtual objects. This data can be used online or offline. Unity is a popular application for programming virtual environments and managing data streams and has also been used in simulated industrial robotics (Crespo et al., 2015; Naranjo et al., 2020; Nenna et al., 2022). As demonstrated in our previous study (Nenna et al., 2022), data generated from user interactions can provide information on the system and human states, such as the exact position and rotation of a robotic arm, the task efficiency, and the users' fatigue levels. Additionally, with the latest eye-tracking-equipped VR devices, it is even possible to monitor workers' eye parameters, such as pupil size and other eye parameters which are known to change with workload and would therefore be extremely useful in view of VR-based human-centric telerobotics (Novak et al., 2015; Nenna et al., 2022).

More specifically, in our previous study (Nenna et al., 2022), we demonstrated the utility of virtual interfaces for controlling industrial robots. By adopting a multi-method, user-centric approach, we designed a simple pick-and-place task that participants performed with both a physical industrial robot UR10e and its virtual counterpart, while varying the levels of task demands, including single and dual tasks. The findings underscored the significant potential of virtual simulations in enhancing users' mental well-being and industrial production, particularly for complex and demanding tasks. We also emphasized the importance of multidimensional assessments encompassing human performance, self-reports, and pupillometric measures in realistic work settings. Building upon these insights, our current research aims to delve deeper into the study of teleoperators' behavior within virtual environments, specifically investigating the impact, efficiency, and intuitiveness of the various interaction possibilities offered by VR in the area of telerobotics.

In this study, we thus seek to investigate 1) how the degree of a control system's interactivity affects users' performance and workload when they guide an industrial robotic arm in VR and 2) the sensitivity of various eye parameters collected via VR-integrated eye tracker in monitoring users' workload. For these purposes, we utilized the previously validated virtual environment that accurately reproduces the robot UR10e and the same pick-and-place task employed in our previous study (Nenna et al., 2022). As a novel addition, we implemented two distinct control systems, i.e., a button-based and an action-based one, and looked at how the level of interactivity of the VR interface affects users' performance and workload. Moreover, we extended our investigation to various eye parameters, such as PERCLOS, blink frequency, and duration, in addition to the previously employed pupillometry (Nenna et al., 2022). By doing so, we expanded the scope of workload assessment, providing valuable insights into users' vigilance too. All participants thus performed the pick-and-place task using both the button-based and action-based control systems, while experiencing various levels of mental demand, including single-task and dual-task conditions. We assessed participants' performance by measuring operation time and error rate during the task. Additionally, we evaluated their mental workload through traditional self-reports and by capturing eye-tracking indexes using the VR headset. Our research questions and hypotheses are outlined as follows:

RQ1. **task load**. As a methodological control, we expect the dual-task level to affect participants' performance (slower operation times and higher error rates) and workload (higher NASA-TLX score, higher pupil size variation, lower PERCLOS, and shorter and fewer blinks).

RQ2. **control system**. By leveraging the use of natural and embodied controls, we hypothesize that an action-based system potentially represents a more efficient and intuitive and less demanding solution for guiding robots than a button-based one in VR.

RQ3. **sensitivity of VR-embedded eye-tracker metrics to workload**. We here further investigate which eye parameter collected via VR headset is most sensitive to workload changes, especially considering the increased motion during action-based interactions. We particularly examine correlations between self-reported workload and several eye parameters, pupil size, PERCLOS, blink frequency, and duration, assuming higher correlations for those eye parameters that are more sensitive to workload.

2. STATE OF THE ART

2.1. Comparative literature on control systems for robot teleoperation

Robotic systems can be teleoperated via various control systems, allowing for a lower or higher degree of interactivity, which we here refer to, respectively, as low-interactivity control systems (LICS) and high-interactivity control systems (HICS). LICSs include keyboards, mice, and joysticks (You and Hauser, 2012; Lu and et al., 2008), which are the most common and therefore familiar to the most people. In contrast, all control systems allowing for physical and often direct interactions with a machine are here considered HICSs. For instance, by using motion capture technology, it is possible to leverage human

gestures as a control system to guide mobile robots: users could physically indicate with their hand the direction in which the robot will move (Cicirelli et al., 2015). Similarly, it is possible to manipulate robots in VR through physical and direct interactions (e.g., Martín-Barrio et al., 2020). A different case is a master-slave framework, which we will refer to with the alternative term "leader-follower". The leader-follower framework consists of controlling a teleoperated robot through the direct manual manipulation of a second robot (e.g., Vozar, 2013). Further examples of HICSs are kinesthetic or 3D haptic devices that provide the users tactile feedback (Martin and Hillier, 2009), which allows for the manipulation of 3D objects in virtual environments and is therefore particularly useful in teleoperation tasks (Berkley, 2003).

Among the studies to assess directly the effects of various degrees of teleoperation system interactivity on users' performance, many have demonstrated HICSs' advantages over LICSs. For example, Vozar (2013) compared leader-follower and joystick teleoperation for driving a customized robot and moving boxes in a confined space. Leader-follower control resulted in better performance, less time, and higher user satisfaction than joystick teleoperation. Gliesche et al. (2020) tested nurses' teleoperation performance using a haptic device or a keyboard and mouse to guide a 7-degree-of-freedom robot manipulator in a desktop pick-and-place task. Haptic device use resulted in faster task completion times than the mouse and keyboard.

The following studies also demonstrated how HICSs outperformed LICSs in VR. For example, Franzluebbers and Johnsen (2019) evaluated users' performance teleoperating a pair of 7-degree-of-freedom robotic arms in a pick-and-place task in VR using two control systems: a stationary 3D mouse and VR controllers that tracked participants' movements. The use of VR controllers resulted in faster task execution times than that of the 3D mouse. Similarly, Martin-Barrio et al. (2020) compared control systems for teleoperating the Kyma robot in VR, including controller, leader-follower, and physical gestures. Physical gestures were preferred over the other control systems and resulted in higher accuracy and faster operation times than the controller. However, participants reported a lower workload when using the controller and direct manipulation than when using the leader-follower modality.

Moreover, some studies have not shown any advantages of LICSs over HICSs. Rouanet et al. (2009) had participants teleoperate a zoomorphic robot in a domestic environment to find an object using three control systems: touchscreen-based buttons, a virtual keyboard on a 2D screen, and arm movements tracked by a handheld controller. Although no performance differences were observed between the input modalities, participants preferred the touchscreen-based input modality over the other two. In the experiment by Grabowski et al. (2021), the participants controlled a mobile robot with two arms in a virtual environment. They either used VR controllers or physically walked to the target position to teleoperate the robot and moved their arms to control the robot's arms. The results showed that using VR controllers led to faster and more accurate task completion than active walking.

In summary, research shows that although the effects' magnitude depends on the task and technology used, using HICSs generally leads to better performance than using LICSs, with few exceptions (Rouanet et al., 2009; Grabowski et al., 2021). Nonetheless, limited research has been conducted on the interactivity levels' impact on cognitive workload. We aim to address this gap by examining the impact of control system interactivity levels on user performance and cognitive load in a simulated industrial environment (RQ2), which will help us understand discrepancies more thoroughly.

2.2. Eye-tracking evaluations in robotics and teleoperations

Eye-tracking metrics were often used to predict workload in human-machine interactions (e.g. during driving or in air traffic control operations, McIntire et al., 2014; Ahlstrom and Friedman-Berg, 2006) and in fewer cases also in the robotics domain (e.g., Novak et al., 2015; Nenna et al., 2022). Several studies have shown that pupil diameter increases with increasing cognitive workload (Kahneman, 1973; Marinescu et al., 2018; Pomplun and Sunkara, 2019). Pupil diameter variations are related to the activity of the parasympathetic nervous system, which is involved in regulating arousal (Köles, 2017). However, light significantly affects pupil size variations as well, making it crucial to maintain constant lighting in the setting, pre-process pupil data, and apply proper baseline correction to exclude pupil size variations that are possibly unrelated to the user's cognitive activity (Mathôt et al., 2018; Nenna et al., 2022). Evidence of the relation between pupil diameter and workload in robotics has been repeatedly found in the surgical domain (Wu et al., 2020; Zheng et al., 2015). In the industrial field, on the other hand, we previously demonstrated that pupil diameter varies with task load in participants physically driving a robot through a pick-and-place task (Nenna et al., 2022). This was true when they operated the physical robot (UR10e) and its digital counterpart in VR. Additionally, smaller pupil size variations were observed when participants interacted with the virtual robot, suggesting that it was preferable to the physical one because it allowed users to save mental resources.

Furthermore, PERCLOS is a robust measure of vigilance for humans interacting with machines, particularly in the automotive area (e.g., Du et al., 2022). PERCLOS can be defined as the percentage of time that the eyelids cover the eye area by more than 80% and can be determined from continuous data on eye openness. Literature on this metric showed that higher levels of fatigue and lower vigilance are associated with a higher PERCLOS (for a review, see Marquart et al., 2015). However, to the best of our knowledge, there is no study demonstrating PERCLOS's reliability as a measure of vigilance or fatigue in the robotic context. Indeed, Wu et al. (2020) found that only pupil diameter and gaze entropy differed between task difficulty levels that and PERCLOS did not show any significant variation across the conditions.

Blinks can also be informative of one's workload level (Fogarty and Stern, 1989; Marquart et al., 2015) and/or fatigue (Kim et al., 2022). For example, blink frequency has been shown to be inversely related to the level of mental load (Holland and Tarlow, 1972; Zheng et al., 2012; Borghini et al. 2014) and to performance in a static simulated air traffic vigilance task (McIntire et al., 2014). Similarly, during air traffic control operations, a decrease in blink duration was observed with increased visual workload (Ahlstrom and Friedman-Berg, 2006) whereas blink duration increased with performance deterioration in a vigilance task (McIntire et al., 2014). A possible explanation is that under high mental demand, users tend to inhibit blinks to reduce the risk of missing incoming information (Fogarty and Stern, 1989). Supporting evidence was found in a simulated laparoscopic task: fewer and shorter blinks occurred with higher workload, as self-reported in the NASA-TLX score (Zheng et al., 2012). More recently, Guo et al. (2021) evaluated the mental workload during a space robot teleoperation: participants controlled a robotic arm via desktop and joystick under varied latency and time pressure, which are known to affect workload. With increasing time pressure, blink frequency decreased and pupil size increased whereas the researchers observed no substantial differences across latency manipulations.

Some studies have also shown that eye-tracking measures can predict workload during human-robot interactions. For instance, Novak et al. (2015) used machine learning to test a continuous workload inference via various eye-tracking indexes rather than a discrete classification. Participants used the ARMin robot only to hit targets on a screen containing correct equations. The researchers found that eye-tracker metrics could reveal increases in workload during a task using the ARMin robot, with pupil dilation being the most sensitive indicator of workload and blink and fixation frequencies being the most sensitive to users' effort. Gao et al. (2013) found similar results when comparing the predictive ability of various eye measurements (including blink parameters and pupil dilatation) in assessing self-reported workload during digital nuclear power plant operations. Single measures were not reliable in assessing overall mental workload, but integrating all measures within a predictive model allowed for accurate assessment. Additionally, blink

rate was more sensitive to workload whereas pupil size was more sensitive to error-related attention and arousal.

Overall, existing research suggests that eye tracking is a useful tool for measuring mental load. Drawing on this evidence, we aim to expand the application of this methodology to the field of VR telerobotics (RQ3) to study and compare various interactive systems. Additionally, although contemporary researchers often use fragmented metrics, we plan to combine various eye-tracking measures to enhance our findings.

3. METHODS

3.1. Sample

24 participants, 11 women and 13 men ($M_{age} = 26.16$; $SD_{age} = 1.85$), voluntarily participated in the experiment after providing informed consent. Inclusion criteria were having normal or corrected-to-normal visual acuity (via contact lenses), normal color vision, no current or past neurological or psychiatric problems, and being right-handed. Due to our university's internal regulations in response to the pandemic, access to laboratories and university facilities was limited to students and researchers; therefore, our sample mainly consists of individuals from an academic population. The local ethics committee approved the experimental protocol, and we conducted the study following the principles of the Declaration of Helsinki. We excluded two participants for technical issues with the eye tracker, and another participant withdrew from the experiment because they reported visual difficulties in VR. Finally, we excluded three participants from the analysis for having an error rate greater than 50% in the arithmetic task. The final sample comprised 18 participants, 9 women and 9 men (Mage = 26.33; SDage = 2.02). Achieving gender balance was particularly important because gender-related disparities in task performance had been shown in the same VR scenario (Nenna and Gamberini, 2022).

3.2. Technical setup

An HTC Vive Pro Eye headset (resolution: 1440 × 1600 pixels per eye, refresh rate: 90 Hz, field of view: 110°) was connected to an MSI laptop (model GT63 Titan 8RF, processor Intel Core i7-6700HQ, RAM 16Gb). This head-mounted display has an embedded eye-tracking system (sampling frequency: 120 Hz, calibration: 5 points), which allows for continuous recording of eye parameters. Based on the present research questions, the virtual environment (Fig. 1) was rearranged from the one Nenna et al. (2022) developed, which was programmed in Unity (version 2019.4.18f1). Specifically, our virtual robot closely reproduces the UR10e cobot: it precisely imitates its physical attributes (e.g., it has multiple joints, an effector, and a workstation), it allows users to define a precise work area beyond which it 'will not extend its movements, and it has interactive capabilities that enable users to interact with it through direct contact and button-based remote control. After each experimental session, a large data log was automatically saved on the MSI laptop, including time series data of position and rotation of the VR headset and controllers, all interactions between the user and the robot, and the pick-and-place task accuracy.

3.3. Task and procedure

All participants provided informed consent before starting the experiment. Thereafter, they filled out questionnaires about their demographics (i.e., age, gender) and VR expertise, which allowed us to describe our sample more accurately; then, they expressed general preferences for virtual robot control systems. Specifically, we asked, "If you had to guide a robotic arm in VR, which of the following control modalities would you prefer?". The possible answers were "controller buttons" and "physical actions". Afterward, all participants underwent a training session to familiarize themselves with the tasks, in which they performed a few trials of each task. All instructions were presented in text format in the virtual environment. Once the participant reported having understood all the tasks, a 5-point calibration of the eye-tracking system was conducted and the experiment started.

During the experiment, participants performed 5 tasks, each of which comprised 40 trials that we re-adapted from previous research (Nenna et al., 2022): (1) an arithmetic task, (2) a pick-and-place task executed via controller buttons (button-based control systems) (3) and physical actions (action-based control system), (4) a dual-task performed via controller buttons (button-based control systems), (5) and physical actions (action-based control system). Unlike in our previous study (Nenna et al., 2022), two diversified interaction modalities were developed and tested (i.e., the button-based and the action-based one) to investigate specifically the effects of the enhanced interactivity provided in VR interfaces for telerobotics. These tasks were presented in a random fashion, and a NASA-TLX questionnaire was administered at the end of each task. Participants could also take a break after each NASA-TLX questionnaire; in that case, the eye-tracking system was recalibrated before participants started the next task. After completing tasks, participants were asked again about their preferred virtual robot control system and the experiment ended.

In the arithmetic task (1), participants mentally summed a series of four numbers presented in text format in the virtual environment. The numbers appeared on a virtual panel that always followed the participant's head movements and was placed in the upper part of his/her view in a way that it would not cover the worktable or the robot's effector and was always inside the participant's functional field of view. Thereafter, they reported the result of the arithmetic operation on a virtual keyboard by using the controller buttons. They were randomly presented numbers between 1 and 10, with a time interval of 2.5 sec ± 0.3 sec between them. For the pick-and-place task, we used the same paradigm employed in our previous work (Nenna et al., 2022), in which they asked participants to guide the robotic arm to pick a bolt from the

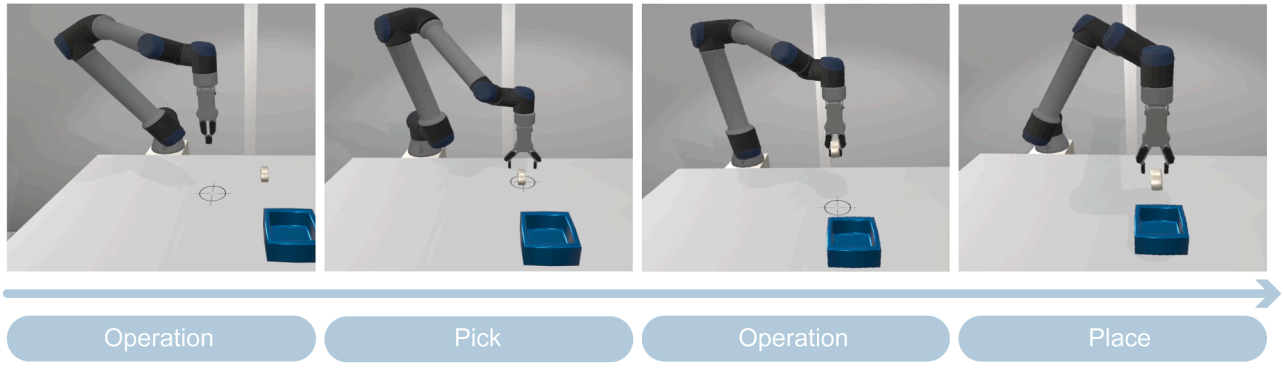

Fig. 1. Overview of the pick-and-place task. In the Operation phases, the participant is driving the robot toward the bolt or the box. In the Pick phase, the participant executed the command for picking the bolt, and in the Place phase, for placing the bolt inside the box.

workstation and place it into a box (Fig. 1). Although the task design remained unchanged, we here introduced two control systems: in the button-based pick-and-place task (2), participants used the pad button on the right controller to move the robot left, right, forward, and backward on the work table; in the action-based pick-and-place task (3), they were allowed to reach for the virtual robot with their right hand, grasp it by pressing the grip button on the right controller, and then move it to the desired position by simply moving their arm. The latter condition reproduced the direct manipulation feature of cobots. In both conditions, after placing the robot in the right location, participants pressed the pad button on the left controller to pick or place the bolt. Therefore, participants used the VR controllers with both control systems, but only in the action-based condition were they allowed to interact with the virtual robot physically. Finally, in the dual task, the pick-and-place task and the arithmetic task were concurrently performed, once using the button-based (4) and once using the action-based (5) control system. The series of numbers presented for the arithmetic task covered the whole pick-and-place task duration, and the result was reported only after the bolt was placed into the box. Fig. 2 depicts all task conditions.

3.4. Measurements

3.4.1. Pick-and-place performance

We measured the operation times as the time elapsed during the robot's movements (start: first movement of the robot, end: last movement before the pick/place action). Considering that the pick action required greater precision to align the robot effector with the bolt to pick compared to the place action, we analyzed independently users' performance in the pick and place phases. We removed trials whose duration exceeded 4 SD from the average duration because they represented very unrealistic operation times (pick phase: 1.43% removed, range 13.45 sec-46.99 sec; place phase: 0.08% removed over 2.827, range 14.91 sec-42.94 sec). We did not consider the same trials to analyze the other independent measures, either. Additionally, we measured the error rate in the pick-and-task place independently for the pick and place phases. Particularly, the pick and place automations were executed only if the left pad button was pressed while the robot was perfectly positioned above the bolt in the pick phase and above the box in the place phase ("correct" event). If at the first attempt of button pressing the robot was not in line with the bolt/box, the event was registered as "incorrect". The participant then had to relocate the robotic arm in the right position to initiate the automation. The percentage of "incorrect" events registered for each action informed the error rate for the pick and the place actions.

3.4.2. Arithmetic performance

In the arithmetic task, we measured the arithmetic input time as the time elapsed from the end of the last number presentation to the moment the participant sent the result of his/her arithmetic calculation through the controller. This measure can be informative of the mental effort deployed for finalizing the mental calculations in each condition. When computing it, we only considered trials in which the correct sum was sent because when participants lost track of the sum during the arithmetic task, they often quickly inserted a random number, resulting in casual input time. In contrast, trials in which participants sent the correct result are more likely to be the product of meaningful cognitive processes. Finally, we computed the error rate in the arithmetic task as the percentage of wrong sums reported, providing an understanding of the dual-task-induced interference with the main task (pick-and-place task).

3.4.3. Eye-tracking measures

Based on previous studies (Guo et al., 2021; Nenna et al., 2022; Novak et al., 2015; Wu et al., 2020), we computed pupil size variation, PERCLOS, blink frequency, and duration as workload-related

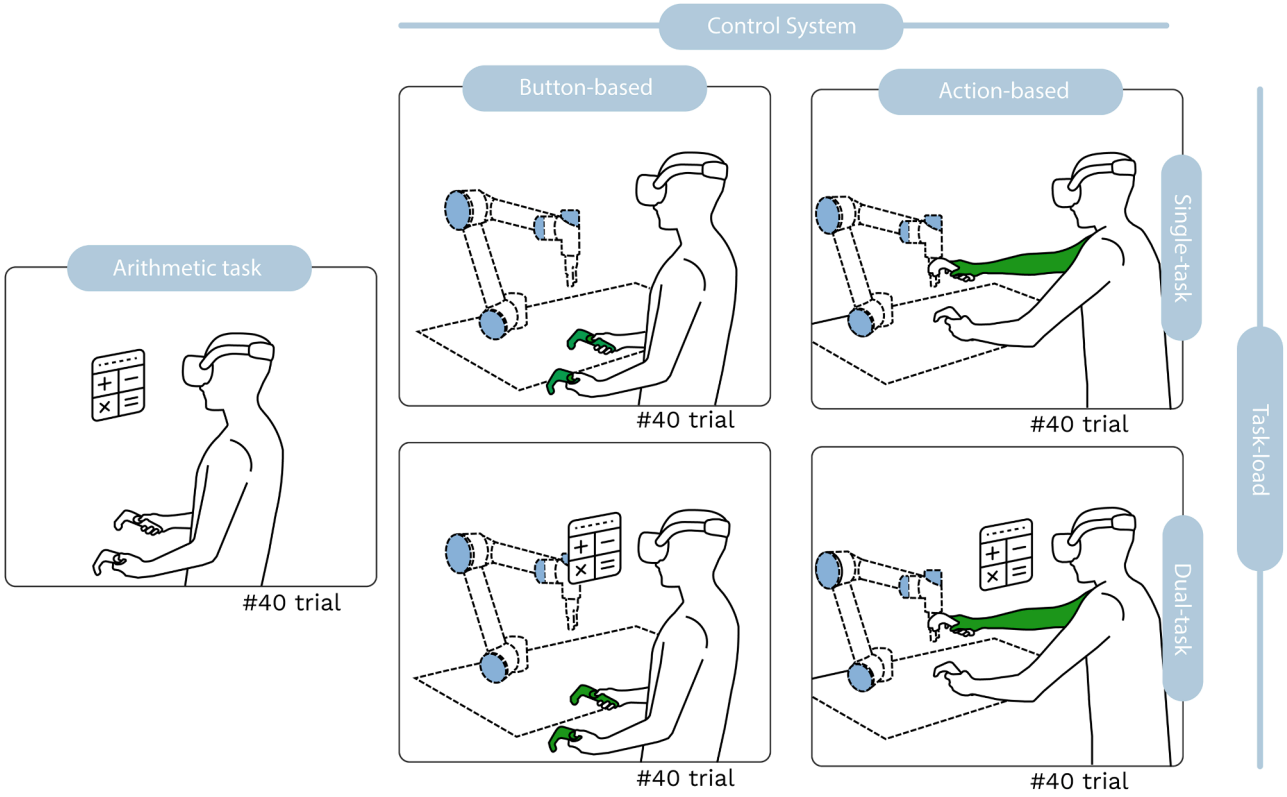

Fig. 2. Experimental design. All participants executed an arithmetic task as a baseline, and a pick-and-place task under low (single-task) and high (dual-task) mental load. The pick-and-place was additionally executed under two control system conditions: low-interactivity and high-interactivity. All participants executed all five tasks.

eye-tracking indexes. Specifically, pupil data preprocessing followed the same procedure Nenna et al. (2022) used. We averaged the pupil size values across the left and right eye and applied a median filter. We ensured that none of the trials or the participants had more than 35% missing data within the analyzed time windows and then applied a subtractive baseline correction on the first 4 data points at the trial level (corresponding to about 200 ms on average). Therefore, we only assessed pupil size variations compared to the baseline period (Mathôt et al., 2018). For the pick-and-place task, we analyzed variations in pupil size within the pick and the place actions independently. To account for varying lengths of the time series resulting from different operation times, we used the dynamic time warping technique (Berndt and Clifford 1994; Keogh and Pazzani 2001; Nenna et al., 2022). Specifically, the length of each pick and place operation was standardized to fit 30 data points. Similarly, for the arithmetic task, we selected four time windows, one for each number presented, and applied dynamic time warping to standardize their length (which ranged between 2.3 and 2.7 sec). Besides pupil size variations, which researchers have also analyzed in previous investigations (Nenna et al., 2022), we also used the eye openness data stream outputted from the HTC Vive headset to compute PERCLOS and blinks. We calculated PERCLOS as the percentage of time during which the eyelids covered more than 80% of the pupil (Wu et al., 2020) in four time windows, each including 10 trials. Blinks were detected as eye closures lasting a minimum of 70 ms and a maximum of 500 ms (Benedetto et al., 2011). If the eyes were closed for less than 70 ms, it was considered a technical issue of the eye tracker likely losing pupil tracking for some frames (Faure et al., 2016). Blink frequency was operationally defined as the blink rate per minute.

3.4.4. Self-reports

We administered the NASA-TLX questionnaire (Hart and Steveland, 1988) after each task directly in VR to measure self-reported workload. Once before starting the experiment and once in the end, we additionally administered a question asking the participants to express their individual preference for guiding a robot in VR either via controller buttons or physical actions. With these questions, we intended to determine whether the individual preferences for one or the other control system would change after participants tested the button-based and the action-based control systems. Furthermore, immediately after the last task, we measured the levels of cybersickness using the Simulation Sickness Questionnaire (SSQ, Kennedy et al., 1993) and the levels of presence via the MEC-Spatial Presence Questionnaire (MEC-SPQ, Vorderer et al., 2004).

3.5. Statistical Analysis

Studies have shown gender-based differences in VR-based telerobotics task performance (Nenna and Gamberini, 2022), so we ran a pre-analysis to control for gender effects in all measurements. Only the operation times in the pick-and-place task were affected by age, aligning with previous research (Nenna and Gamberini, 2022). However, because this goes beyond the primary aims of the present work, we will not delve into the discussion of gender results; pre-analysis results have been deposited in a public data sharing platform (Nenna, 2023).

We analyzed of all measurements using generalized linear models (GLMs from *lme4* package, Bates et al., 2014) in RStudio (Team, 2022). For all of them, were first fitted the data using the function *descdist()* of the package *fitdistrplus* (Delignette-Muller and Dutang, 2015). Then we chose the appropriate models according to the data distribution. The participant was always set as a random effect, and we applied the Bonferroni correction when interpreting the post hoc contrasts within the significant interactions. With our sample of 18 participants and for a medium effect size (d=0.5), the power of the eye tracking and NASA-TLX questionnaire's analysis was greater than 0.95 and the power of all performance measures' analyses ranged from 0.59 to 0.67.

3.5.1. Performance measures

For each performance measure in the pick-and-place task, we computed a GLM including the factors task load (single-task, dual-task) and control system (button-based, action-based). To analyze performance measures in the arithmetic task, we instead ran a GLM over the factor task (single task, button-based dual task, action-based dual task).

3.5.2. Eye-tracking measures

Models analyzing pupil size variation during the pick-and-place task included the factors task load (single-task, dual-task), control system (button-based, action-based), and window (1, 2, 3, 4, 5, 6). The factor window allowed us to consider pupil size changes in the time course at the trial level. When analyzing the pupil size variation throughout the arithmetic task, instead, we ran a model including the factors task (arithmetic task, button-based dual task, action-based dual task) and arithmetic operation (start, 1st sum, 2nd sum, 3rd sum). For this analysis, we only considered the first 3 arithmetic operations to compare the single task with the dual tasks. In contrast, the statistical models analyzing PERCLOS, blink frequency, and duration included the factors task load (single-task, dual-task), control system (button-based, action-based), and window (1, 2, 3, 4). Each window included 10 trials (window 1: trials 1-10; window 2: trials 11-20, etc.) and allowed us to examine changes in eye parameters in the time course at the task level.

3.5.3. Self-reports

We analyzed the NASA-TLX questionnaire regarding the following factors: task load (single task, dual task), control system (button based, action based) and items (mental demand, physical demand, temporal demand, performance, effort, frustration). We performed post hoc contrasts specifically between the levels of task load and control system in each of the questionnaire's items. We also assessed relations between the NASA-TLX score (the overall score and the score on the individual NASA-TLX items) and each eye-tracking measure via Pearson's linear correlation tests. Furthermore, we reported the response rate regarding the individual preference for action- vs. button-based control systems expressed before and after the experiment as well as the descriptive statistics of the scores reported on the MEC-SPQ. We analyzed the responses on the SSQ as Kennedy et al. (1993) and Bimberg et al. (2020) proposed.

4. RESULTS

4.1. Performance measures

4.1.1. Pick-and-place performance

Descriptive statistics are shown in Table 1, and results of the GLM in Table 2. Post-hoc contrasts on the operation times revealed significant differences between the button- and action-based control systems both

Table 1
Descriptive statistics of the pick-and-place performance

		Operation time (sec)		Error rate (%)	
		Pick Mean (SD)	*Place Mean (SD)*	*Pick Mean (SD)*	*Place Mean (SD)*
Task load	*Single task*	2.51 (1.88)	2.04 (1.18)	9.93 (9.03)	0.97 (1.50)
	Dual task	2.73 (2.00)	2.11 (1.28)	14.2 (9.91)	1.32 (2.57)
Control system	*Button-based*	3.67 (2.07)	2.83 (1.07)	16.8 (10.4)	1.11 (2.35)
	Action-based	1.57 (1.57)	1.31 (0.85)	7.36 (5.91)	1.18 (1.84)

Table 2
Results of pick-and-place performance measures

Measurement		Factor		
		Task load	*Control system*	*Task load Control system*
Operation times	*Pick phase*	$X^2 = 20.02$	$X^2 = 1462$	$X^2 = 0.56$ ns
	Place phase	$X^2 = 3.73$ ns	$X^2 = 1976.7$	$X^2 = 10.64$
Error rate	*Pick phase*	$X^2 = 5.91$	$X^2 = 22.27$	$X^2 = 1.99$ ns
	Place phase	$X^2 = 0.43$ ns	$X^2 = 0.01$ ns	$X^2 = 0.01$ ns

Stars indicate the significance level of the post hoc tests: * $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$

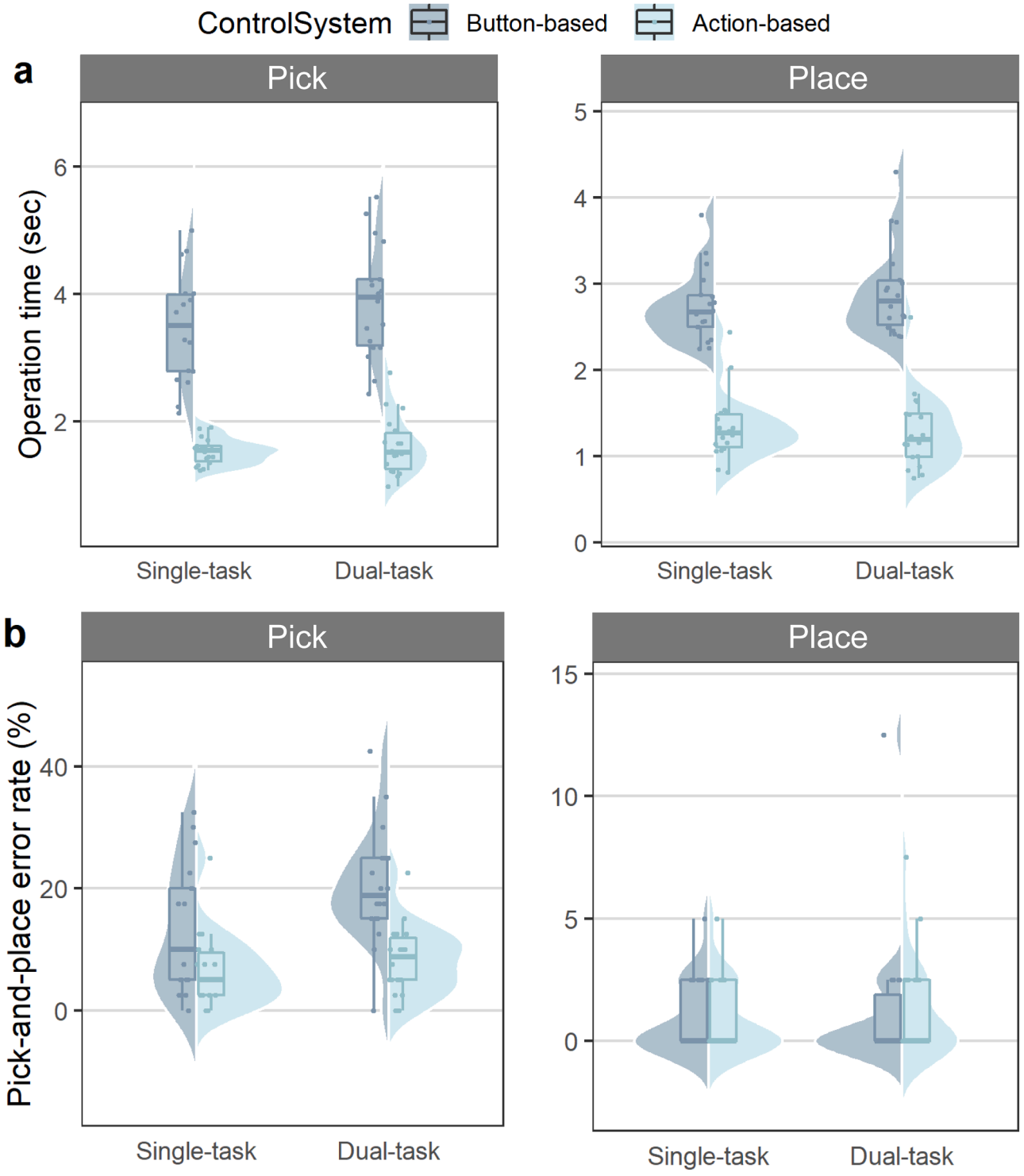

Fig. 3. Averaged operation time (a) and error rate (b) at the pick-and-place task. In each plot and condition, a boxplot and a half violin depict the data distribution. Each point corresponds to the averaged data of one participant.

Table 3
Descriptive statistics of the arithmetic performance

		Error rate (%)	Input time (sec)
Task	*Arithmetic single task*	6.55 (4.10)	3.37 (1.49)
	Action-based dual task	12.1 (7.52)	3.95 (2.16)
	Button-based dual task	20.4 (12.6)	4.38 (2.28)

under single ($p<.0001$) and dual task ($p<.0001$). After applying the Bonferroni correction, differences between the single and dual task were not significant in any of the control system modalities. Fig. 3 depicts the main results on the pick-and-place task performance.

4.1.2. Arithmetic performance

Descriptive statistics are shown in Table 3. The GLM on the arithmetic error rate indicated a significant main effect of the factor Task ($X^2 = 14.58$, $p < .001$). Compared to the single arithmetic task, the error rate was significantly higher only while executing the pick-and-place task via button-based ($p<.01$) but not via action-based controls ($p=.16$). Moreover, the error rate at the arithmetic task did not differ significantly between the two dual-tasks ($p=.39$). For the analysis of the arithmetic input time, instead, the main factor Task resulted in being statistically significant ($X^2 = 146.04$, $p<.0001$). Post hoc contrasts revealed that the arithmetic input time was significantly lower in the Single-task compared to both Dual-task conditions ($p_s<.0001$). Moreover, a significantly higher arithmetic input time was observed in the button-based dual-task condition compared to the action-based dual-task condition ($p<.0001$). Performance results for the arithmetic task are depicted in Fig. 4.

4.2. Eye tracking measures

4.2.1. Pupil size variation - Arithmetic task

As depicted in Fig. 5, the analysis of pupil size at the arithmetic task yielded significant main effects of both task ($X^2 = 1553.6$, $p <.0001$) and arithmetic operation ($X^2 = 1017.3$, $p <.0001$). Furthermore, two factors interacted significantly ($X^2 = 1346.3$, $p <.0001$). Post-hoc contrasts showed significant increases in pupil size when moving from start to 2nd sum only for the button-based dual task ($p < .001$) and from start to 3rd sum for all task conditions (all $p_s <.001$). Similarly, significant pupil size increases were observed when moving from the 1st sum to 2nd for the arithmetic task ($p<.01$) and the button-based dual task ($p<0001$), and from 1st sum to 3rd sum for all tasks (all $p_s <.0001$). When moving from 2nd to 3rd sum, only the dual task conditions yielded significant contrasts (all $p_s <.0001$).

4.2.2. Pupil size variation - pick-and-place task

Table 4 resumes the results of the statistical models. Post hoc of interest included the comparison between single and dual task in each control system condition and within each window. Specifically, pupil size variation was significantly higher in the dual task compared to the single task in both control system conditions and from window 2 to 6 specifically in the pick phase (all $p_s < .0001$) and in the button-based condition of the place phase (all $p_s < .0001$). Differently, in the action-based condition of the place phase, pupil size variation was higher in the dual task compared to the single-task in windows 4 ($p<.01$), 5 and 6 ($p_s < .0001$). Results of pupil size variation are depicted in Fig. 6.

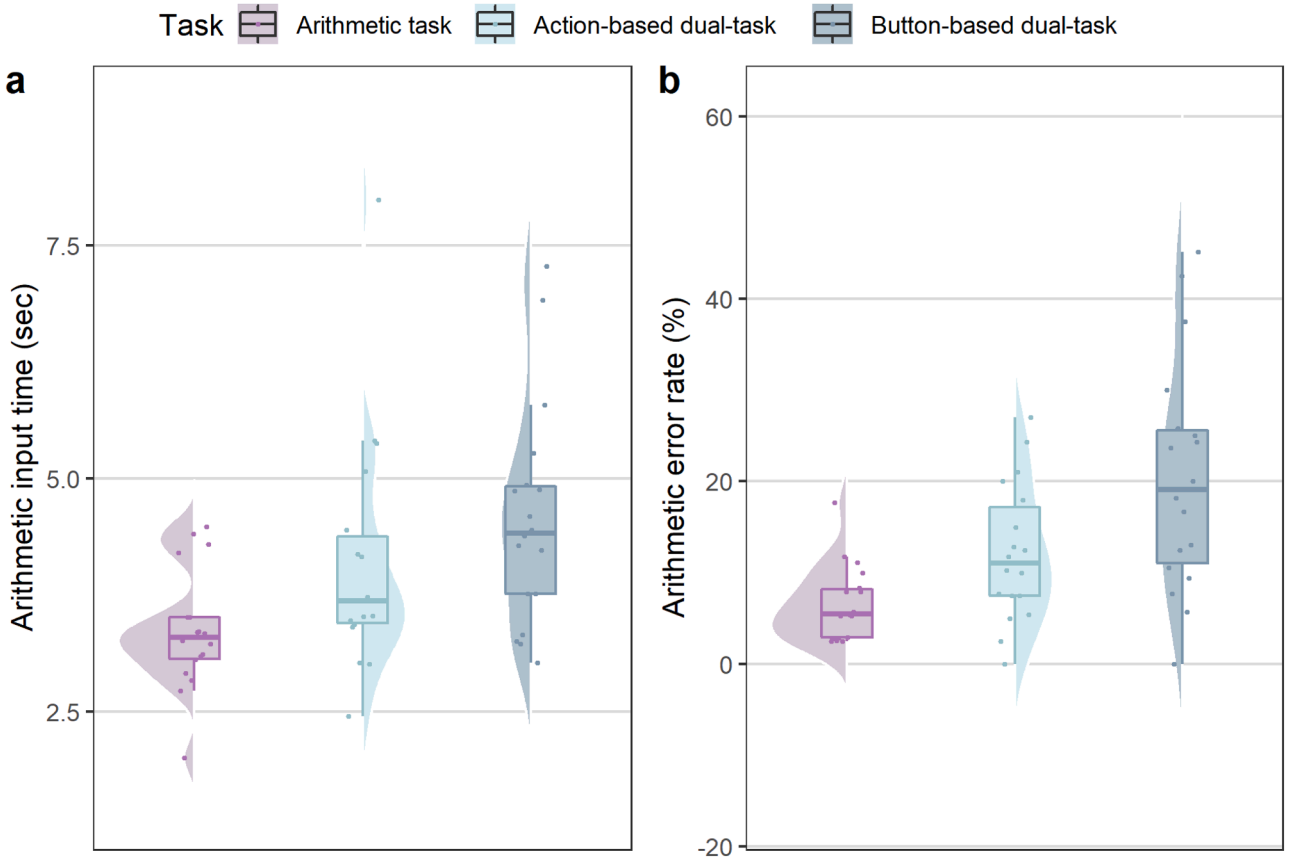

Fig. 4. Averaged input time (a) and error rate (b) at the arithmetic task. In each plot and for each task condition, a boxplot and a half violin depict the data distribution. Each dot corresponds to the averaged data of one participant. Abbreviation: DT = dual task.

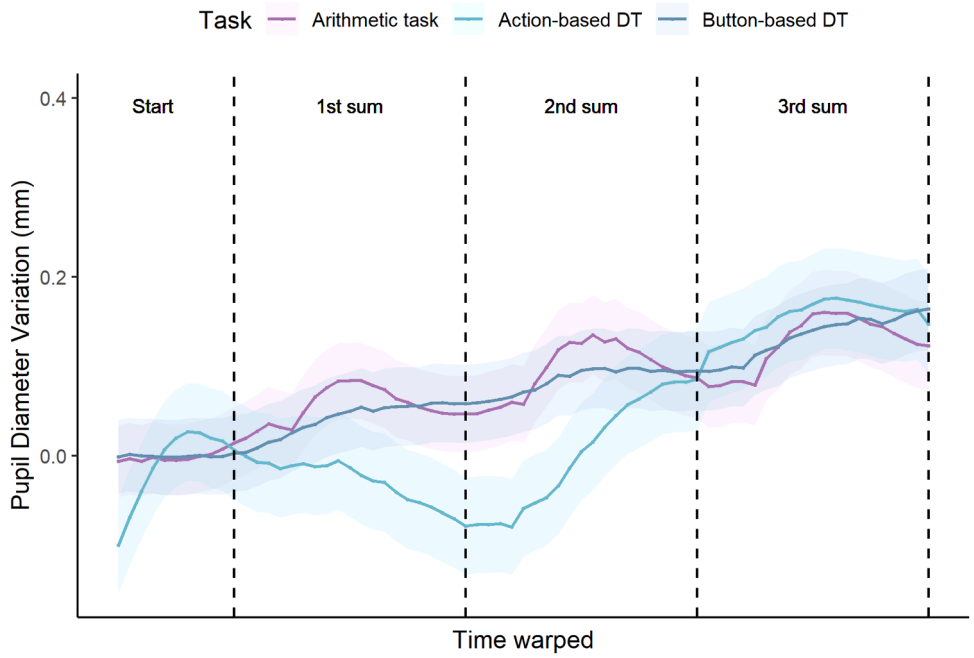

Fig. 5. pupil size variations throughout the arithmetic task, grouped by task (arithmetic task, action-based dual task, button-based dual task). Lines represent the averaged pupil size in each data point, and the shadows indicate the standard error. The x axis indicates the time standardized via dynamic time warping. Abbreviation: DT = dual task.

Table 4
Results of pick-and-place eye-tracking measures

Measurement	Factor											
	Task load		*Control system*		*Task load * Control system*		*Window * Task load*		*Window * Control system*		*Window * Task load * Control system*	
Pupil size variation - Pick phase	$X^2 = 2979.5$	***	$X^2 = 1137.7$	***	$X^2 = 20.37$	***	$X^2 = 981.76$	***	$X^2 = 1884.7$	***	$X^2 = 52.20$	***
Pupil size variation - Place phase	$X^2 = 903.0$	***	$X^2 = 2764.9$	***	$X^2 = 364.9$	***	$X^2 = 382.1$	***	$X^2 = 819.9$	***	$X^2 = 66.23$	***
PERCLOS	$X^2 = 16.55$	***	$X^2 = 0.03$	Ns	$X^2 = 4.71$	*	$X^2 = 0.27$	Ns	$X^2 = 2.06$	ns	$X^2 = 0.29$	ns
Blink frequency	$X^2 = 12.32$	***	$X^2 = 1.40$	ns	$X^2 = 10.52$	**	$X^2 = 0.26$	Ns	$X^2 = 1.37$	ns	$X^2 = 0.49$	ns
Blink duration	$X^2 = 1.29$	ns	$X^2 = 3.05$	ns	$X^2 = 6.69$	**	$X^2 = 2.38$	Ns	$X^2 = 2.03$	ns	$X^2 = 1.90$	ns

Stars indicate the significance level of the post hoc tests
* $p \leq .05$;
** $p \leq .01$;
*** $p \leq .001$)

4.2.3. PERCLOS

Results of the statistical tests are summarized in Table 4 and depicted in Fig. 7. Post hoc tests on the interaction between task load and control system revealed significant differences between single and dual task only in the action-based condition ($p<.0001$).

4.2.4. Blink parameters

Table 4 shows the results of the GLM on blink parameters, which are also depicted in Fig. 7. As regards the post hoc on the blink duration, after applying the Bonferroni correction, none of the contrasts reached the significance threshold. Differently, when analyzing blink frequency, higher blink frequency was observed in the single task ($M = 3.99$, $SD = 3.7$) than the dual task ($M = 2.41$, $SD = 2.43$). Furthermore, higher blink frequency was observed in the single- compared to the dual-task only in the action-based ($p<.001$) but not in the button-based condition ($p=0.54$).

4.3. Self-reports

4.3.1. NASA-TLX questionnaire

As depicted in Fig. 8, a significant main effect was observed both for task load ($X^2 = 212.99$, $p<.0001$) and control system ($X^2 = 12.79$, $p<.001$). Significant interaction effects were also observed between item and task load ($X^2 = 47.32$, $p<.0001$) and Item and control system ($X^2 = 11.10$, $p<.0001$). Post hoc contrasts revealed significant differences between single and dual task for the following items: mental demand ($p<.0001$), temporal demand ($p<.0001$), performance ($p<.05$), effort ($p<.0001$), and frustration ($p<.0001$). Differently, significant differences between the button-based and action-based conditions were only observed for the item frustration ($p<.01$).

4.3.2. Individual preferences for button- vs. action-based control systems

Before conducting the experiment, 73.68% of the participants indicated their preference for guiding the robot via the action-based control system, while 26.32% preferred the button-based one. Following the experiment, the proportion of users who favored the action-based control system rose to 89.47%, whereas the percentage of those who preferred the button-based system decreased to 10.53%.

4.3.3. Self-reported sense of presence and cybersickness

The responses at the MEC-SPQ (5-points Likert scale) indicated a quite high sense of presence ($M = 3.79$; $SD = 1.11$). Furthermore, the score computation of the SSQ revealed extremely low levels of cybersickness ($M = 12.72$; $SD = 15.52$); according to Kennedy et al. (1993)

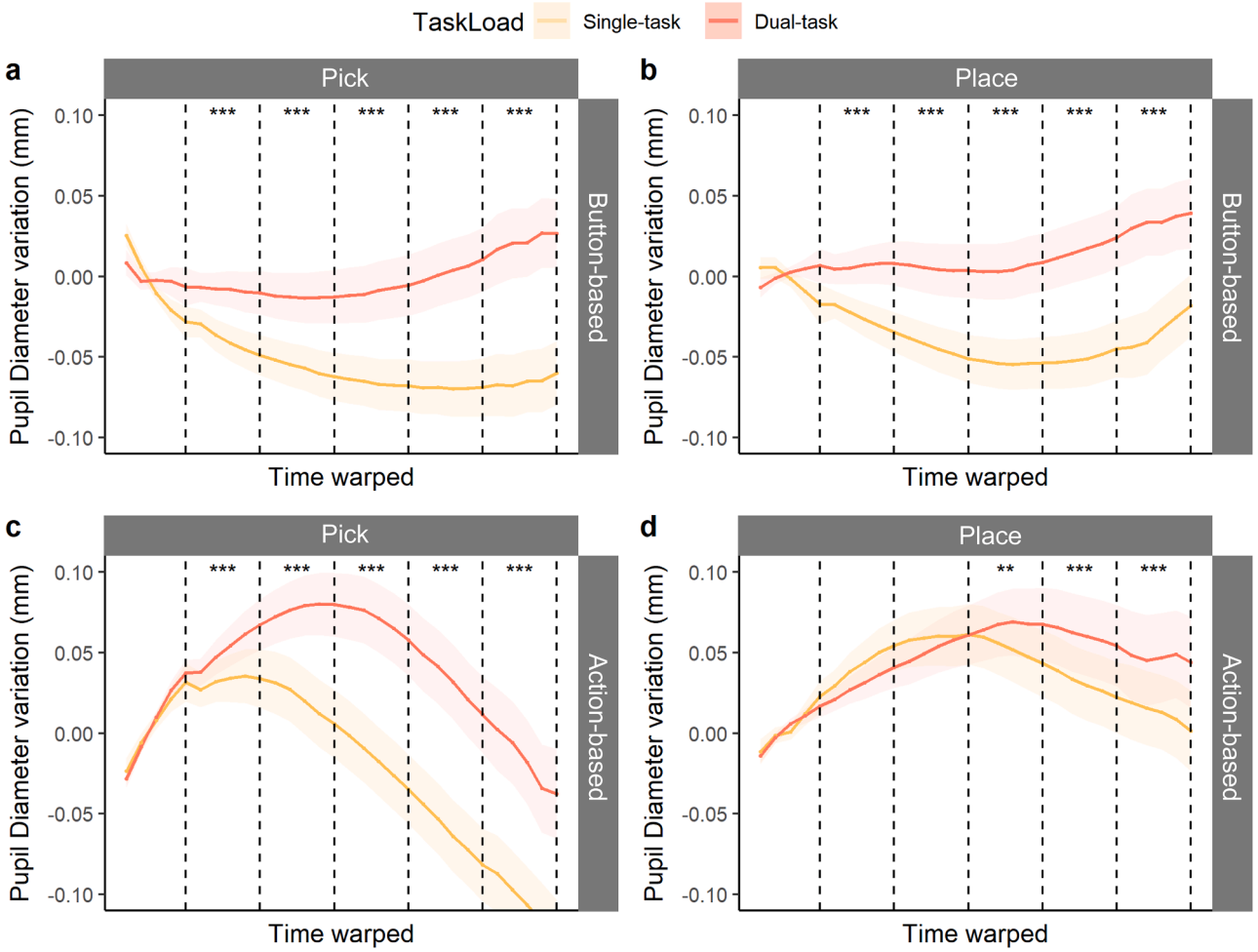

Fig. 6. Pupil size variations during the pick-and-place task relative to the task load conditions. The first row depicts the pick (a) and place (b) phases of the button-based condition. The second row depicts the pick (c) and place (d) phases of the action-based condition. The x axis indicates the time standardized via dynamic time warping. All the plots are complemented by stars indicating the significance level of the post hoc tests (*p ≤ .05; **p ≤ .01; ***p ≤ .0001).

and Bimberg et al.'s (2020) score computation, the maximum score at the SSQ exceeds the value of 230.

4.4. *Relations between self-reported workload and eye-tracking indexes*

Results of the correlation matrix are reported in Fig. 9.

5. DISCUSSION

In this section, we deeply discuss our main questions in the following subparagraphs, including the effectiveness of our task load manipulation in creating multiple levels of task demands (RQ1), the effects of the enhanced interactivity of VR (action-based controls) compared to button-based control systems on human performance and workload (RQ2), and each eye parameter's sensitivity to workload as recorded via the VR-embedded eye tracker throughout the tasks (RQ3).

5.1. *RQ1: Task load*

Our results support our hypotheses and the idea of cognitive interference between the arithmetic and pick-and-place tasks. Specifically, performance and self-reported workload demonstrated a higher demand for dual tasking whereas eye parameters only partially reflected task-load-related differences. As PERCLOS suggested, our participants further demonstrated higher levels of vigilance in the dual task than in the single task. Interestingly, a tendency of fatigue to increase throughout the task performance can also be inferred with PERCLOS and blink frequency variations. In the following subparagraphs (5.1.1, 5.1.2, 5.1.3), we unfold the effects of our task load manipulation on each of the investigated measures.

5.1.1. *Performance measures*

From a behavioral perspective, when performing the pick action concurrently with the arithmetic task, participants were slower and committed more errors than when they performed the pick action as a single task (Fig. 4). Similarly, in the dual-task condition, they were more error prone in the arithmetic task and it took longer for them to finalize the arithmetic sums compared to performing mental operations without additional tasks. Interestingly, the button-based condition revealed stronger dual-task behavioral effects, suggesting that it likely imposed a greater demand than the action-based condition.

5.1.2. *Self-reports*

The NASA-TLX results aligned with user performance (Fig. 8). The participants reported higher workload levels in the pick-and-place and arithmetic tasks together, indicating they perceived it as more demanding than the single task. They found dual tasking more mentally and temporally challenging, frustrating, and effortful than the single task while experiencing comparable physical demands. Additionally, participants rated their performance significantly better in the single task than in the dual task.

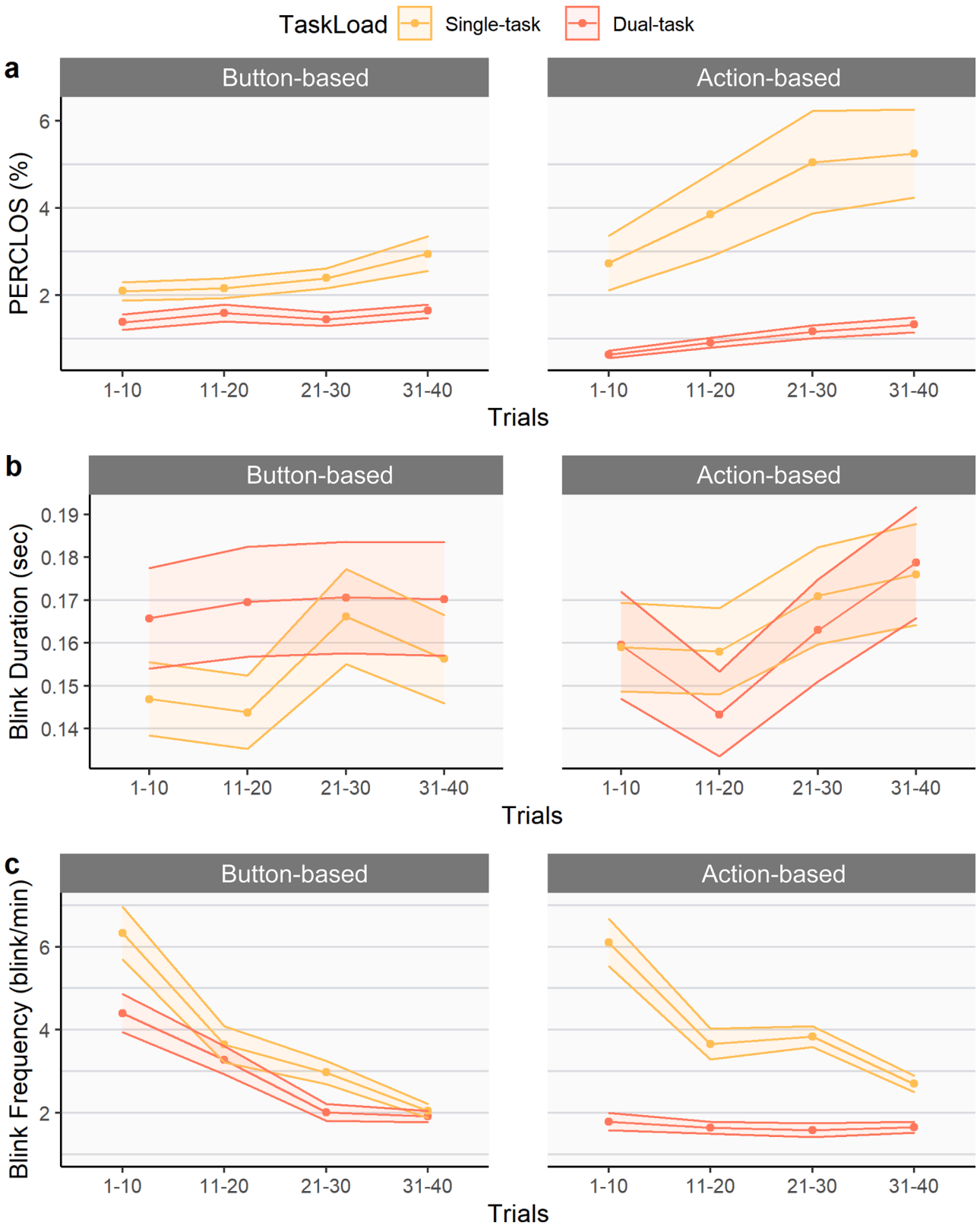

Fig. 7. Perclos, blink duration and frequency throughout the experimental sessions in each task load (single task, dual task) and control system condition (button-based, action-based).

5.1.3. Eye-tracking measures

Eye parameter analysis partially reflected self-reported workload. Pupil size (Fig. 5) increased throughout the arithmetic task from the start to the following arithmetic sums no matter the control system involved, indicating an increasing workload. Furthermore, pupil size variation during the pick-and-place task was significantly higher in the dual task than in the single task regardless of the control system deployed. This finding is fairly robust and aligns with the literature on

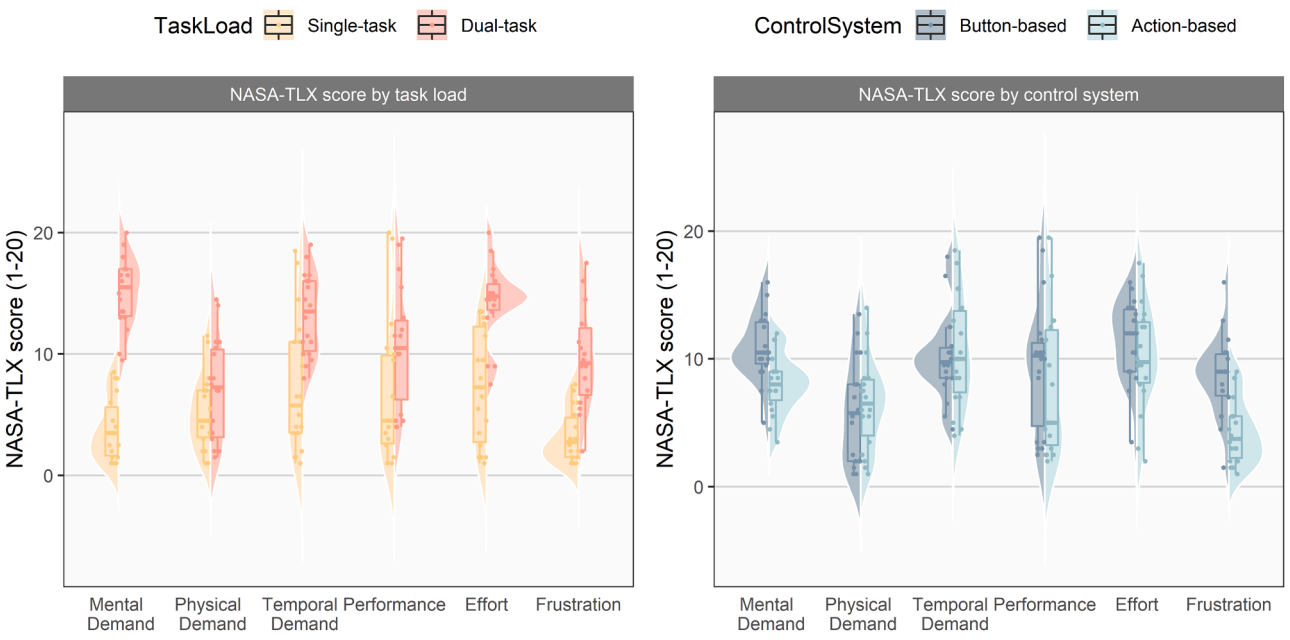

Fig. 8. Averaged NASA-TLX score in each item according to the task load and control system. In each plot and for each item of the questionnaire (x axis), a boxplot and a half violin depict the data distribution. Each dot corresponds to the averaged data of one participant.

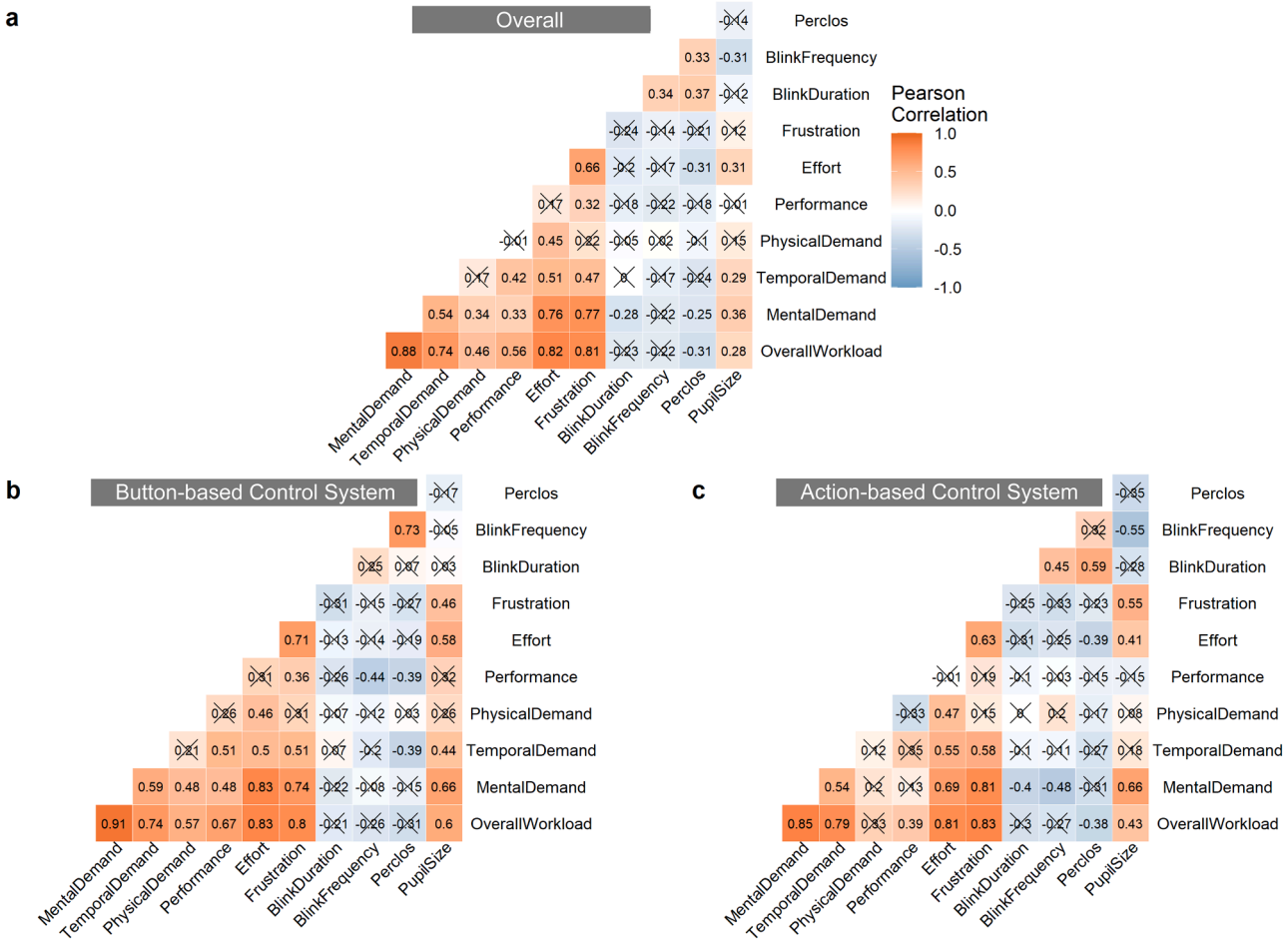

Fig. 9. Correlation matrixes between NASA-TLX scores (overall workload, mental demand, temporal demand, physical demand, performance, effort, frustration) and eye-tracking parameters (blink duration, blink frequency, perclos, pupil size). Fig. a) depicts correlations on the overall dataset, independently from the control system used to teleoperate the robot; figure b) shows correlations within the button-based condition; figure c) shows correlations for the action-based control system.

teleoperation and/or robotics reporting higher pupil size for higher task load (Nenna et al., 2022; Wu et al., 2020; Zheng et al., 2015). In contrast, PERCLOS and blink parameters did not capture task load differences as accurately as pupil size variations. Indeed, task load manipulation affected PERCLOS and blink frequency only in the action-based condition. One of the reasons for this result, as deeply discussed in paragraph 5.2, might be the various difficulties in use of the two control systems (i.e., button and action based). Specifically for the action-based condition, at the macro level, we observed a lower PERCLOS and blink frequency in the dual task, likely suggesting higher levels of vigilance than in the single task (Marquart et al., 2015). We also observed similar results on workload-related blink variations in previous teleoperation research (Zheng et al., 2012; Guo et al., 2021). A common assumption is that users are likely to inhibit eye closures to reduce the risk of missing salient information (Fogarty and Stern, 1989), and such interpretation seems to apply to the present task, too. At the micro level, PERCLOS values gradually increased throughout the single task from 2.74% on average in the first trials to 5.26% in the last trials, but they only increased from 0.63% to 1.32% in the dual task. Similarly, blink frequency decreased from 4.65 blink/min on average in the first trials to 2.11 blink/min in the last trials of the task. Even though it was not supported by a statistical significance, this trend might reflect changes in the level of fatigue (Marquart et al., 2015): as time passed, users got tired and their eye closures decreased. Another possible explanation is that performing the same monotonous task for some minutes can be tiring, thus affecting the level of vigilance in the task course (Körber et al., 2015).

5.2. RQ2: Control system

Our findings confirmed the hypothesis of better performance and lower workload with the use of the action-based compared to the button-based control system in VR. Participants also demonstrated higher levels of vigilance throughout the whole pick-and-place task executed via the button-based compared to the action-based control system. Performance, self-reports, and eye-tracking differences between the two conditions were prominent, and we thoroughly discuss them in the following subsections (5.2.1, 5.2.2, 5.2.3). Overall, this clear advantage of action-based controls might be related to embodied mechanisms involved in physical and direct operations, resulting in more intuitive control of virtual robot movements in 3D space. Hand-eye coordination is a primal embodied behavior that makes operations more affordable and natural. Guiding the robot via buttons, instead, requires transposing spatial intentions from a 3D view to 4 static directions over two axes (forward-backward, left-right), increasing the operation complexity.

5.2.1. Performance measures

Participants were significantly faster when executing the teleoperations via physical action in the pick and place phases: they saved about 2 sec on average in each pick action and 1.5 sec in each place action compared to when they used controller buttons. Furthermore, the error rate in the pick action decreased from about 17% to 7% when the participants switched from controller buttons to physical actions, but we did not observe this advantage in the place action. Again, this is possibly due to the ease of the place operation, in which the box in which the bolt was placed is larger than the bolt. Furthermore, looking at the arithmetic task performance, as compared to solely summing the presented numbers, the averaged error rate almost doubled when participants also performed the pick-and-place task via physical actions, and it even tripled when participants steered the robot via controller buttons. However, only the difference between error rates in the single arithmetic task and button-based dual task was statistically significant. It therefore seems that steering the robot via physical actions did not impose a degree of cognitive effort as high as steering the same robot via controller buttons. In general, there is strong evidence in favor of using action-based controls to guide the robotic arm. These findings are consistent with research that has shown better performance when the participants use HICSs during robotic teleoperations (Vozar, 2013; Franzluebbers and Johnsen, 2019; Gliesche et al., 2020; Martín-Barrio et al., 2020).

5.2.2. Self-reports

We observed a generally higher perceived workload and greater frustration in the button-based than in the action-based task. We also observed a tendency for higher mental demand in the button-based than in the action-based condition, which however did not reach the significance threshold. These results align with the expressed preference for action-based control systems. Even before we tested the teleoperation modalities, there was a clear tendency to prefer action-based control systems. This preference increased after the experiment: 89.47% of the tested sample reported preferring guiding the robot via physical actions, which was perceived as the less frustrating control system.

5.2.3. Eye-tracking measures

Notably, we did not intend to compare the control-system-related eye parameters directly because they might be strongly influenced by the various movement magnitudes involved in the action- and button-based conditions. This precaution was corroborated, for example, by the findings regarding pupil size variation: in the button-based condition, pupil size gradually increased from the start (window 1) to the end (window 6) of each action. In the action-based condition, we observed a quicker pupil size increase that reached its peak in windows 3 and 4 and then decreased (Fig. 6). This increase could occur either due to the larger physical motion involved in the action-based than in the button-based condition, which may have elicited higher arousal and activation, or to a constantly higher level of vigilance throughout the whole task session in the button-based condition, which may have flattened the pupil size variation. The latter interpretation seems further supported by the larger self-reported workload (Fig. 8) and by the constantly lower PERCLOS level observed in the button-based compared to the action-based condition (Fig. 7), which is known to be related to a higher level of vigilance (Marquart et al., 2015).

In the same line of interpretation, the PERCLOS difference between the single and dual tasks was more evident in the action-based than in the button-based condition, likely showing that executing the pick-and-place single task via physical actions was so easy that it required very low vigilance compared to executing the same task via controller buttons. Furthermore, we noticed that our task load manipulation affected PERCLOS and blink duration only in the action-based condition, not in the button-based condition. Furthermore, in the button-based control system, the task load manipulation similarly affected the pick and place actions whereas the action-based control system only affected the pick action. Taken together, it seems that no matter the task difficulty, participants always invested more mental resources when operating via controller buttons rather than physical actions. This might have prevented the emergence of different blink and PERCLOS trends in the single and dual tasks as well as different pupil size trends between the pick and the place actions, specifically in the button-based condition.

5.3. RQ3: Sensitivity of VR-embedded eye-tracker metrics to workload

As discussed in paragraph 5.1, our task load manipulation effectively produced two distinct workload levels. From this observation, we further assessed the sensitivity of each eye-tracking metric to workload variations by correlating the self-reported workload and each eye-tracking parameter. From a first glimpse of Fig. 9, regardless of the control system involved, pupil size and PERCLOS are particularly sensitive to changes in mental demand and effort whereas blink duration responds specifically to mental demand. However, the latter relation was not supported by the results shown in Fig. 7 because blink duration did not differ significantly between the single and dual tasks. In contrast, blink frequency did not show significant correlations with any workload dimensions on the overall task, suggesting that it might not be the best

indicator of workload in VR. Furthermore, no observed relations in the overall task exceeded R=0.36; they were therefore quite weak.

When we observed the two control systems independently, the positive but weak relations between workload and pupil size observed in the overall task became even stronger. Relations between pupil size and mental demand reached a correlation of R=0.66 in each control system condition, almost doubling the R coefficient observed in the overall task. A positive relation between frustration and pupil size, which was not observed in the overall task, additionally stood in both control system conditions. This could be explained by the strong positive relations between self-reported mental demand and frustration (R = 0.77): the higher the mental demand, the higher the frustration and the larger the pupil variation. Furthermore, it is worthwhile to comment briefly on Fig. 5, which shows trends in pupil size variations during the arithmetic task. Specifically, when participants performed the dual task via physical actions, there was greater variability in the pupil size trend than in the button-based and to arithmetic task conditions, in which the pupil increase was more linear throughout the task. Yet, we observed a pupil size increase across all conditions. Again, this result is indicative of the resilience of such a metric in measuring workload under higher (action-based) and lower (button-based) degrees of physical motion in VR. These findings regarding workload and pupil size align with robotics (Wu et al., 2020; Zheng et al., 2015) and virtual-robotics research (Nenna et al., 2022).

Whereas relations between pupil size and workload persist in the various control system conditions, PERCLOS and blink frequency better respond to workload fluctuations in the action-based than in the button-based control system. Although we observed only weak relations with temporal demand and performance considering button-based actions, PERCLOS showed stronger relations with overall workload and effort during action-based operations. Generally speaking, this result aligns with the literature (Marquart et al., 2015), demonstrating that PERCLOS is responsive to levels of vigilance and fatigue (which might be reflected in the effort dimension in the NASA-TLX), but it also contrasts with previous research in robotics that did not demonstrate significant relations between PERCLOS and workload (Wu et al., 2020). Furthermore, if blink frequency did not yield any significant relation with workload in the overall task, it showed a moderate relation with mental demand exclusively in the action-based condition. This finding aligns with literature showing inverted relations between blink frequency and mental demand (Zheng et al., 2012; Borghini et al., 2014). Overall, the sensitivity of pupil size to workload stands out compared to the other eye metrics, which is consistent with previous research (Novak et al., 2015).

6. LIMITATIONS

We recognize the following limitations. First, real-world teleoperation tasks require more complex and varied activities. In our study, the choice of a simple experimental task such as the pick and place was intentional to guarantee appropriate experimental control yet allow for a natural behavior with the least possible constraints. With the following studies, researchers might attempt to design more complex and diversified teleoperations.

Second, due to the health emergency spread during the data collection, we gathered our results from a sample of young users (mainly students) with no prior experience with robot teleoperations. Future studies should include participants who are more familiar with teleoperation tasks to determine whether our findings also apply to experts. Furthermore, our sample's average age was much lower than that of the European labor force, which was estimated to be 40 in 2019 (Statista, 2022). It would be interesting to see whether similar performance and workload trends apply to an older population, which is more representative of eventual final users. This would also help researchers assess how willing older users are to accept such technology in their work life, considering that they are typically unfamiliar with unconventional technologies, such as VR. Ultimately, evaluating our VR-based industrial scenario with experts in human factors and ergonomics would help validate its usability in working environments.

As a last point, the virtual robot used in the experimental study was not directly linked to the UR10e. This simulated environment allowed us to analyze human performance and workload during teleoperation activities successfully. Other studies have included digital twins instead, providing a more practical example of telerobotics although it lacks in-depth analysis of human workload mechanisms (e.g., Luo et al., 2021; Whitney et al., 2018; Lipton et al., 2017). In the future, we plan to link the VR scenario to the physical robot for a deeper analysis of real teleoperation activities. For example, if the VR scenario has real consequences for a physical robot, it may impact user performance, task strategies, and workload. Technical disruptions such as transmission latency and interruptions might also affect teleoperation performance and workload. Linking the VR scenario to the physical workstation will help analyze such instances.

7. CONCLUSIONS

At a glance, this study paves the way for new perspectives in the telerobotics sector, which sees eye-tracking-equipped VR as a valued resource in the ongoing 5.0 Industrial Revolution. Our most general intent was to bridge the lack of human factor-oriented research in the traditionally engineering-focused field of telerobotics. We therefore investigated human performance and workload with a mixed-approach methodology rather than evaluating the technical framework's feasibility. Furthermore, we generated our findings through a human-centered approach and in a virtual environment that promoted a high sense of presence and no cybersickness in the users. Our main results can be outlined as follows.

First, VR action-based control systems enable natural and embodied controls. They also capitalize on the innate human hand-eye coordination and therefore facilitate the benefits of collaborative robotics in virtual spaces. Ultimately, they resulted in more efficient and intuitive and less demanding solutions for robotic teleoperations, leading to improved performance and reduced cognitive load.

Second, the integration of eye tracking and VR can be highly advantageous for continuous workload monitoring without interrupting ongoing tasks. This finding aligns perfectly with the Industry 5.0 intent of digitalization that is centered on human needs. Of all the eye parameters investigated in this work, pupil size seems the most robust indicator of workload because the user's degree of physical motion does not affect it.

7.1. APPLICATIONS

Our findings have various potential applications. For example, the demonstrated advantage of action-based systems can inspire the design of future telerobotics platforms, which can benefit from rethinking their interaction and teleoperation modalities to exploit more human physical movements in VR.

Additionally, our findings on the effectiveness of eye-tracking metrics as a measure of workload in VR can be leveraged to develop biofeedback systems (e.g., Tan et al., 2014; Zargari et al., 2019). Information on users' performance and their oculometrics can be displayed via graphical representations in VR, supporting operators' tasks and better informing them of their workload and fatigue online. As a further step, it would be possible to develop VR platforms that automatically adjust their features and the required work pace based on the users' detected fatigue. In this way, the digital platforms would become totally human based and human tailored, allowing for customized work patterns in VR for each individual.

Finally, various eye parameters could be used as an interaction modality to guide or influence robotic systems in VR. For instance, research has shown that gaze can be used to manipulate virtual objects

(Monteiro et al., 2021; Yu et al., 2021). Therefore, it would be valuable to explore this feature's usefulness in our VR-based teleoperation framework and whether it may affect other workload or fatigue metrics, such as pupil size and PERCLOS.

Overall, we believe that such a research line, which complements telerobotics innovations with knowledge from human factors and cognitive ergonomics sectors, will streamline and improve interactions between humans and robots, thereby substantially contributing to industry and society.

CRediT authorship contribution statement

Federica Nenna: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. **Davide Zanardi:** Data curation, Formal analysis, Investigation, Methodology, Software, Writing – original draft. **Luciano Gamberini:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Ahlstrom, U., Friedman-Berg, F.J., 2006. Using eye movement activity as a correlate of cognitive workload. International journal of industrial ergonomics 36 (7), 623–636.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Benedetto, S., Pedrotti, M., Minin, L., Baccino, T., Re, A., Montanari, R., 2011. Driver workload and eye blink duration. Transportation research part F: traffic psychology and behaviour 14 (3), 199–208.

Berkley, J.J., 2003. Haptic devices. White Paper by Mimic Technologies Inc 1–4.

Berndt, D.J., Clifford, J., 1994. Using dynamic time warping to find patterns in time series. KDD workshop 10 (16), 359–370.

Bimberg, P., Weissker, T., Kulik, A., 2020. On the usage of the simulator sickness questionnaire for virtual reality research. In: 2020 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW). IEEE, pp. 464–467.

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., Babiloni, F., 2014. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. Neuroscience & Biobehavioral Reviews 44, 58–75.

Carswell, C.M., Clarke, D., Seales, W.B., 2005. Assessing mental workload during laparoscopic surgery. Surgical innovation 12 (1), 80–90.

Chien, S.Y., Lin, Y.L., Lee, P.J., Han, S., Lewis, M., Sycara, K., 2018. Attention allocation for human multi-robot control: Cognitive analysis based on behavior data and hidden states. International Journal of Human-Computer Studies 117, 30–44.

Cicirelli, G., Attolico, C., Guaragnella, C., D'Orazio, T, 2015. A kinect-based gesture recognition approach for a natural human robot interface. International Journal of Advanced Robotic Systems 12 (3), 22.

Crespo, R., García, R., Quiroz, S., 2015. Virtual reality application for simulation and off-line programming of the mitsubishi movemaster RV-M1 robot integrated with the oculus rift to improve students training. Procedia Computer Science 75, 107–112.

Dehais, F., Karwowski, W., Ayaz, H., 2020. Brain at work and in everyday life as the next frontier: grand field challenges for neuroergonomics. Frontiers in Neuroergonomics 1.

Delignette-Muller, M.L., Dutang, C., 2015. fitdistrplus: An R package for fitting distributions. Journal of statistical software 64, 1–34.

Doolani, S., Wessels, C., Kanal, V., Sevastopoulos, C., Jaiswal, A., Nambiappan, H., Makedon, F., 2020. A review of extended reality (xr) technologies for manufacturing training. Technologies 8 (4), 77.

Du, G., Zhang, L., Su, K., Wang, X., Teng, S., Liu, P.X., 2022. A Multimodal Fusion Fatigue Driving Detection Method Based on Heart Rate and PERCLOS. IEEE Transactions on Intelligent Transportation Systems.

Faccio, M., Granata, I., Menini, A., Milanese, M., Rossato, C., Bottin, M., Rosati, G., 2022. Human factors in cobot era: a review of modern production systems features. Journal of Intelligent Manufacturing 1–22.

Faure, V., Lobjois, R., Benguigui, N., 2016. The effects of driving environment complexity and dual tasking on drivers' mental workload and eye blink behavior. Transportation research part F: traffic psychology and behaviour 40, 78–90.

Fogarty, C., Stern, J.A., 1989. Eye movements and blinks: their relationship to higher cognitive processes. International journal of psychophysiology 8 (1), 35–42.

Franzluebbers, A., Johnson, K., 2019. Remote robotic arm teleoperation through virtual reality. In: Symposium on Spatial User Interaction, pp. 1–2.

Grandi, F., Zanni, L., Peruzzini, M., Pellicciari, M., Campanella, C.E., 2020. A Transdisciplinary digital approach for tractor's human-centred design. International Journal of Computer Integrated Manufacturing 33 (4), 377–397.

Gao, Q., Wang, Y., Song, F., Li, Z., Dong, X., 2013. Mental workload measurement for emergency operating procedures in digital nuclear power plants. Ergonomics 56 (7), 1070–1085.

Gliesche, P., Krick, T., Pfingsthorn, M., Drolshagen, S., Kowalski, C., Hein, A., 2020. Kinesthetic Device vs. Keyboard/Mouse: A Comparison in Home Care Telemanipulation. Frontiers in Robotics and AI 172.

Grabowski, A., Jankowski, J., Wodzyński, M., 2021. Teleoperated mobile robot with two arms: the influence of a human-machine interface, VR training and operator age. International Journal of Human-Computer Studies 156, 102707.

Guo, Y., Freer, D., Deligianni, F., Yang, G.Z., 2021. Eye-tracking for performance evaluation and workload estimation in space telerobotic training. IEEE Transactions on Human-Machine Systems 52 (1), 1–11.

Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Advances in psychology, 52, pp. 139–183. North-Holland.

Holland, M.K., Tarlow, G., 1972. Blinking and mental load. Psychological Reports 31 (1), 119–127.

Kahneman, D., 1973. Attention and effort, 1063. Prentice-Hall, Englewood Cliffs, NJ, pp. 218–226.

Keogh, EJ, Pazzani, MJ, 2001. Derivative dynamic time warping. In: Proceedings of the 2001 SIAM International Conference on Data Mining, pp. 1–11. Society for Industrial and Applied Mathematics.

Kennedy, R.S., Lane, N.E., Berbaum, K.S., Lilienthal, M.G., 1993. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. The international journal of aviation psychology 3 (3), 203–220.

Kim, S.Y., Park, H., Kim, H., Kim, J., Seo, K., 2022. Technostress causes cognitive overload in high-stress people: Eye tracking analysis in a virtual kiosk test. Information Processing & Management 59 (6), 103093.

Köles, M., 2017. A review of pupillometry for human-computer interaction studies. Periodica Polytechnica Electrical Engineering and Computer Science 61 (4), 320–326.

Körber, M., Cingel, A., Zimmermann, M., Bengler, K., 2015. Vigilance decrement and passive fatigue caused by monotony in automated driving. Procedia Manufacturing 3, 2403–2409.

Krüger, J., Lien, T.K., Verl, A., 2009. Cooperation of human and machines in assembly lines. CIRP annals 58 (2), 628–646.

Lipton, J.I., Fay, A.J., Rus, D., 2017. Baxter's homunculus: Virtual reality spaces for teleoperation in manufacturing. IEEE Robotics and Automation Letters 3 (1), 179–186.

Lu, Y., Zheng, H., Chand, S., Xia, W., Liu, Z., Xu, X., Bao, J., 2022. Outlook on human-centric manufacturing towards Industry 5.0. Journal of Manufacturing Systems 62, 612–627.

Luo, Y., Wang, J., Shi, R., Liang, H.N., Luo, S., 2021. In-device feedback in immersive head-mounted displays for distance perception during teleoperation of unmanned ground vehicles. IEEE Transactions on Haptics 15 (1), 79–84.

Marinescu, A.C., Sharples, S., Ritchie, A.C., Sanchez Lopez, T., McDowell, M., Morvan, H. P., 2018. Physiological parameter response to variation of mental workload. Human factors 60 (1), 31–56.

Martin, S., Hillier, N., 2009. Characterisation of the Novint Falcon haptic device for application as a robot manipulator. In: Australasian Conference on Robotics and Automation (ACRA). Citeseer, pp. 291–292.

Martín-Barrio, A., Roldán, J.J., Terrile, S., del Cerro, J., Barrientos, A., 2020. Application of immersive technologies and natural language to hyper-redundant robot teleoperation. Virtual Reality 24 (3), 541–555.

Marquart, G., Cabrall, C., de Winter, J., 2015. Review of eye-related measures of drivers' mental workload. Procedia Manufacturing 3, 2854–2861.

Mathôt, S., Fabius, J., Van Heusden, E., Van der Stigchel, S., 2018. Safe and sensible preprocessing and baseline correction of pupil-size data. Behavior research methods 50, 94–106.

Matthews, G., Reinerman-Jones, L.E., Barber, D.J., Abich IV, J., 2015. The psychometrics of mental workload: Multiple measures are sensitive but divergent. Human factors 57 (1), 125–143.

Mavridis, N., Pierris, G., Gallina, P., Moustakas, N., Astaras, A., 2015. Subjective difficulty and indicators of performance of joystick-based robot arm teleoperation with auditory feedback. In: 2015 International Conference on Advanced Robotics (ICAR). IEEE, pp. 91–98.

McIntire, L.K., McKinley, R.A., Goodyear, C., McIntire, J.P., 2014. Detection of vigilance performance using eye blinks. Applied ergonomics 45 (2), 354–362.

Monteiro, P., Goncalves, G., Coelho, H., Melo, M., Bessa, M., 2021. Hands-free interaction in immersive virtual reality: A systematic review. IEEE Transactions on Visualization and Computer Graphics 27 (5), 2702–2713.

Naranjo, J.E., Sanchez, D.G., Robalino-Lopez, A., Robalino-Lopez, P., Alarcon-Ortiz, A., Garcia, M.V., 2020. A scoping review on virtual reality-based industrial training. Applied Sciences 10 (22), 8224.

Nenna, F., Gamberini, L., 2022. The influence of gaming experience, gender and other individual factors on robot teleoperations in vr. In: 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, pp. 945–949.

Nenna, F., Orso, V., Zanardi, D., Gamberini, L., 2022. The virtualization of human–robot interactions: a user-centric workload assessment. *Virtual* Reality 1–19.

Nenna, F. (2023). [Enhanced Interactivity in VR-based Telerobotics: An Eye-tracking Investigation of Human Performance and Workload - Gender analysis] [Unpublished results]. https://osf.io/8dw29/?view_only=b89e16cf8ebd4f1b8e9edd429c8cc383.

Novak, D., Beyeler, B., Omlin, X., Riener, R., 2015. Workload estimation in physical human–robot interaction using physiological measurements. Interacting with computers 27 (6), 616–629.

Pomplun, M., Sunkara, S., 2019. Pupil dilation as an indicator of cognitive workload in human-computer interaction. Human-Centered Computing. CRC Press, pp. 542–546.

Rebelo, F., Noriega, P., Duarte, E., Soares, M., 2012. Using virtual reality to assess user experience. Human factors 54 (6), 964–982.

Rosen, E., Whitney, D., Phillips, E., Ullman, D., Tellex, S., 2018. Testing robot teleoperation using a virtual reality interface with ROS reality. In: Proceedings of the 1st International Workshop on Virtual, Augmented, and Mixed Reality for HRI (VAM-HRI), pp. 1–4.

Rouanet, P., Béchu, J., Oudeyer, P.Y., 2009. A comparison of three interfaces using handheld devices to intuitively drive and show objects to a social robot: the impact of underlying metaphors. In: RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication. IEEE, pp. 1066–1072.

Smith, K., Sepasgozar, S., 2022. Governance, Standards and Regulation: What Construction and Mining Need to Commit to Industry 4.0. Buildings 12 (7), 1064.

Statista. (2022, August 5). *Median age of the global labor force by region and gender 2019*. Retrieved September 27, 2022, from https://www.statista.com/statistics/996588/median-age-global-labor-force-region-gender/.

Tan, C.S.S., Schöning, J., Luyten, K., Coninx, K., 2014. Investigating the effects of using biofeedback as visual stress indicator during video-mediated collaboration. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 71–80.

Villani, V., Righi, M., Sabattini, L., Secchi, C., 2020. Wearable devices for the assessment of cognitive effort for human–robot interaction. IEEE Sensors Journal 20 (21), 13047–13056.

Vorderer, P., Wirth, W., Gouveia, F.R., Biocca, F., Saari, T., Jäncke, L., Jäncke, P., 2004. Mec spatial presence questionnaire 18 (2004), 2015. Retrieved Sept.

Vozar, S. E. (2013). *A Framework for Improving the Speed and Performance of Teleoperated Mobile Manipulators* (Doctoral dissertation).

Wang, X.V., Kemény, Z., Váncza, J., Wang, L., 2017. Human–robot collaborative assembly in cyber-physical production: Classification framework and implementation. CIRP annals 66 (1), 5–8.

Whitney, D., Rosen, E., Ullman, D., Phillips, E., Tellex, S., 2018. Ros reality: A virtual reality framework using consumer-grade hardware for ros-enabled robots. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 1–9.

Wu, C., Cha, J., Sulek, J., Zhou, T., Sundaram, C.P., Wachs, J., Yu, D., 2020. Eye-tracking metrics predict perceived workload in robotic surgical skills training. Human factors 62 (8), 1365–1386.

Yu, D., Lu, X., Shi, R., Liang, H.N., Dingler, T., Velloso, E., Goncalves, J., 2021. Gaze-supported 3d object manipulation in virtual reality. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–13.

You, E., Hauser, K., 2012. Assisted teleoperation strategies for aggressively controlling a robot arm with 2d input. In: Robotics: science and systems, 7. MIT Press, USA, p. 354.

Zargari Marandi, R., Madeleine, P., Omland, Ø., Vuillerme, N., Samani, A., 2019. An oculometrics-based biofeedback system to impede fatigue development during computer work: A proof-of-concept study. PLoS One 14 (5), e0213704.

Zheng, B., Jiang, X., Tien, G., Meneghetti, A., Panton, O.N.M., Atkins, M.S., 2012. Workload assessment of surgeons: correlation between NASA TLX and blinks. Surgical endoscopy 26 (10), 2746–2750.

Zheng, B., Jiang, X., Atkins, M.S., 2015. Detection of changes in surgical difficulty: evidence from pupil responses. Surgical innovation 22 (6), 629–635.