applied sciences

MDPI

Article

Interface Transparency Issues in Teleoperation

Luis Almeida [1,2,*], Paulo Menezes [2] and Jorge Dias [2,3]

1 Ci2—Smart Cities Research Center, Polytechnic Institute of Tomar, 2300-313 Tomar, Portugal
2 University of Coimbra, Institute of Systems and Robotics, 3030-290 Coimbra, Portugal; PauloMenezes@isr.uc.pt (P.M.); jorge@deec.uc.pt (J.D.)
3 Khalifa University Center for Autonomous Robotic Systems (KUCARS), Khalifa University of Science and Technology (KU), Abu Dhabi 127788, UAE
* Correspondence: laa@ipt.pt

Received: 29 July 2020; Accepted: 3 September 2020; Published: 8 September 2020

Abstract: Transferring skills and expertise to remote places, without being present, is a new challenge for our digitally interconnected society. People can experience and perform actions in distant places through a robotic agent wearing immersive interfaces to feel physically there. However, technological contingencies can affect human perception, compromising skill-based performances. Considering the results from studies on human factors, a set of recommendations for the construction of immersive teleoperation systems is provided, followed by an example of the evaluation methodology. We developed a testbed to study perceptual issues that affect task performance while users manipulated the environment either through traditional or immersive interfaces. The analysis of its effect on perception, navigation, and manipulation relies on performances measures and subjective answers. The goal is to mitigate the effect of factors such as system latency, field of view, frame of reference, or frame rate to achieve the sense of telepresence. By decoupling the flows of an immersive teleoperation system, we aim to understand how vision and interaction fidelity affects spatial cognition. Results show that misalignments between the frame of reference for vision and motor-action or the use of tools affecting the sense of body position or movement have a higher effect on mental workload and spatial cognition.

Keywords: human-centered computing; cognitive human-robot interaction; HCI design and evaluation methods; telerobotics; teleoperation; telepresence; immersion

1. Introduction

A telepresence robot presents a solution for doctors and health care workers to consult, handle, or monitor people in remote places or in contaminated areas, avoiding self-exposure. For instance, in the present coronavirus pandemic (COVID-19) or in recent epidemics like Ebola virus disease (EVD), a physician could teleoperate a robot and, through it, look around, move, or communicate safely in a contained environment. The robot's capabilities of perception, manipulation, and mobility allow performing some disaster-response tasks [1–3]. Telerobotics is already present in areas like surgery (e.g., the Da Vinci Robot) [4], remote inspection, space exploration [5,6], underwater maintenance, nuclear disposal, hazardous environment cleaning, and search and rescue. However, most of these robotic interventions in critical tasks still rely on the human's control capabilities. Teleoperated robots quite often include semi-autonomous functionalities to assist operators. To this end, cognitive human-robot interaction architectures are being used to minimize the control workload, improve the task performance, and increase safety [7,8]. Thus, the design of such cognitive robotic systems can integrate the knowledge of human perceptual factors to predict the intended actions and needs of operators.

Human's actions depend on the perception of the environment, while the decisions rely on correct the recognition and interpretation of sensory stimuli [9]. For this, several sensory modalities contribute information, providing cues for the perception of space and motion. These perceptual cues are continuously acquired and matched with our mental models. While some cues are consistently matched, others do not fit. Consequently, our brain tries to solve these conflicts to avoid compromising consequent actions, e.g., we expect that a known object seems bigger when it is near us and smaller if it is further away. We also do not expect to "see" a radio's loudspeaker somewhere and "listen" to the respective sound coming from a different direction. Another example of unsolved conflict can occur while on a train looking for a parallel train, where we cannot distinguish which train is moving, whether it it is ours or the other one. Surprisingly, we deal well with other conflicting situations: when we are combing our hair in front of a mirror, the hand that really "touches" us is not the one that we are "seeing": it is the reflection of the hand; this results from a learning process because children do not know how to comb in front of a mirror. In short, real-world representations are built on sensory information and cognition to understand them, using thoughts and experience (bottom-up and top-down processes complement each other).

Currently, a person can act in a remote environment through a teleoperation system; however, the use of any type of mediation requires training. This interaction with the system can be simple, if the control interface remains intuitive and natural. Ideally, if the operator feels as though he/she is in a remote location, he/she will act as naturally as though he/she were physically there. The illusion of telepresence can be induced through a proper action-perception loop supported by technology [10]. A person experiencing sensory stimuli similar to those in the remote environment and acting in line with them can actually sense "being there" [11]. Such a feeling depends on the capability of the system to accurately display remote environment properties and provide enough information about the remote agent and the responsiveness of the system to motor inputs, i.e., immerse the person in media [12]. Sheridan [13] suggested a distinction between (virtual) presence and (experiential) telepresence: presence refers to the experience of being present in a virtual world and telepresence to the sense of being in a mediated remote real environment. The term telepresence was initially introduced by Minsky in the teleoperation context. It refers to the phenomenon that a human operator develops a sense of being physically present at a remote location through interaction with the system's human interface [14]. Telerobotics appears as a subclass of teleoperation where the human operator supervises and/or controls a remote semi-automatic robotic system. The teleoperation of robots involves two major activities: remote perception and remote manipulation. In [15], to improve teleoperation, we proposed the use of immersive technologies to allow operators to perceive the remote environment as though being there, where the control of the robot was as simple as controlling their own bodies. Thus, operators could sense the robot's body as their own (embodiment [16]), simplifying the navigation control. The use of head-mounted displays (HMDs) to naturally control the point of view of the remote robot's camera and display control instruments was further explored in [17] to allow operators to feel that they were controlling the robot from inside. The sense of presence was improved using a "virtual cockpit" while the operators' faults and mental workload were minimized. Nevertheless, we realized that factors were compromising the sense of telepresence and degrading task performance in these works and other author's related works [18]. It is not simple to dissociate the flows that exist in a teleoperation system (visual feedback, haptic, control, etc.) because impairing one of them would make the robot uncontrollable. Therefore, the present research aims to be a contribution to human perceptual factors with an impact on the teleoperator's behavior. Some factors degrade human performances such as low video stream bandwidth, frame rates, time lags, frame of reference, lack of proprioception, two-dimensional views, attention switches, or motion effects [8,16,19]. Some related works [20] analyzed the global effect of these factors on physical workload and on task performance without discriminating their weights in the system. Other authors focused on the effect of a specific factor such as the field of view [21] or focused on virtual environments [22]. The present

study aims to decouple the flows of an immersive teleoperation system to understand how the vision and interaction fidelity levels affect spatial cognition.

The immersive experience can be enhanced through the replication of real visual feedback, thus supporting the operator's traditional hand-eye coordination [23,24]. Based on Slater's findings [16], we also aim at consistency between actions (movements) and the multimodal feedback (e.g., visual and haptic). Pursuing this goal, we expect to provide enough spatial and motion cues of the remote environment to allow operators to perceive and behave naturally. Therefore, we contribute to the minimization of cognitive workload and performance improvement.

We propose simplifying the operation control by exploring and combining telerobotics with telepresence. We argue that if with the use of media devices, the operators experiment with the sensation of being in the remote environment, then the task performance becomes as natural as being there. This research proposes an immersive interface approach for teleoperation and evaluates it against traditional remote control systems. The introduction of telepresence systems influences and enlarges the range of applications that can benefit from its use. Nevertheless, given some current technological limitations and taking into account the results from studies on human factors, a set of recommendations for the design of these systems is presented. The evaluation of any interactive system is a crucial step in the development process. This evaluation can be divided into two parts: user task performance and usability and user acceptance. Although the ultimate goal is to maximize the performance metrics, given the very central role of the user, such performance is very dependent on not only how the user is able to execute the task, but how the system influences the perception of that task and if that may influence its execution positively. Simplification, effort reduction, and intuitiveness are some keywords that positively influence the user and therefore his/her performance. To this end, some guidelines on which aspects are important to evaluate in a teleoperation system are also presented.

Structure of the article: Section 2 provides insights about skills and performance in human activity. Section 3 describes the human role in teleoperation, namely the control and feedback flows with the human in the loop operating a remote device. Section 4 analyzes the technological constraints and human perceptual factors with relevance for immersive teleoperation design, providing a set of recommendations for the construction of these systems. Section 5 presents an evaluation methodology for immersive interfaces in teleoperation. As an example, an immersive interface is proposed, and the influence of perception on task execution is analyzed. Section 6 presents the quantitative and qualitative results and a discussion of the proposed immersive interface. Section 7 presents the findings and conclusions.

2. Skills and Performance

Human activities have been distinguished by Rasmussen into three types: knowledge-, rule-, and skill-based [25]. In fact, complex activities may involve knowledge, rules, and skills with some level of alternation or simultaneous execution among them. Most of the skills, in particular those that involve motor coordination, are acquired via training, and during this process, the human brain establishes the relationships between low level sensory signals and muscle actions, and higher level goals, rules, and knowledge. The skill acquisition may be as difficult as it is to distinguish the relevant sensory information from signal noise and how complex the relationship between that information and the proper actions to be executed is [9,26].

Picking the example of writing some text, if a person perfectly masters the typewriting technique, the focus is on the search for which words to express the idea. However, on the other side, for someone who is not used to keyboards, the focus will be on searching for the keys to compose the words. In the latter case, as the foreground activity is the search for the keys, there will be an increased difficulty in producing the text directly from ideas. In this case, probably, it would be better to hand write it first to avoid distraction and the consequent ruining of the establishment of the intended line of thought. Either way, the performance will be lower, either in terms of ideas about the text or the time to produce it.

Skills are related to actions that we may delegate to more reactive levels of our brain, where actions are produced as answers to sensory signals without any conscious intervention [25]. There may exist one intention that establishes some behavior of reference that serves as a guide to the production of actions in response to sensory signals. Therefore, it is common to propose closed-loop-based models to study and explain these automatic behaviors [27]. There are however alarm mechanisms that are activated when excessive errors or mismatches are detected in these signals. When these happen, higher complexity layers are called to intervene to select a change in behavior, induce a correction, or a simple parameter adjustment that will bring the actuation back to "normality". In the typewriting example, this may happen when a finger strikes two keys at once and the touch sense reports that to the brain. This normally triggers a reaction of confirming the typed characters or words.

We may say that during skill acquisition, we start by the understanding of how to act on the controls or objects. In the second phase, the focus is on what we intend the object or system to do under our action while keeping some level of surveillance on the grasping of the object or command that enables us to detect slippage, abnormal resistance, or other aspects that may indicate the existence of some failure in the action or in the means of interaction. As the task execution becomes automatic and does not require any particular attention to it, the cognitive level is free to focus on more abstract levels [25], e.g., the text to be written. As another example, we may consider a taxi driver that is concentrated on choosing the best way to reach a given destination and apparently not paying any special attention to the other cars or the road limits. Nevertheless, as soon as the front car stop lights are lit, an obstacle appears on the road or the motor does not respond as usual to the gas pedal actuation, and that detail is brought immediately to the primal attention levels.

As skills are acquired for particular ways of performing activities, introducing changes in the way the activity is performed or sensed may result in performance reduction, at least until a re-adaptation is possible and completed. Again with the keyboard example, if it becomes less precise and typed keys produce zero or multiple characters, this requires the user to confirm the output constantly and check if the text is correct as expected. Similarly, a faulty screen where some parts of it show black bars instead of the content forces the user to make sure that all pressed keys are done in the correct order as some of the outputs cannot be checked visually a posteriori. Similar effects may be introduced by automated mechanisms that are intended to help the user, e.g., snap to grid in drawing programs or spelling correctors in text editors, which change fine drawing movements into unintended locations or foreign words into inappropriate and out of context ones.

As demonstrated by the above, the existence of imprecise or disturbing factors may reduce the performance of any task execution in particular, but is not limited to those that require motor-related skills.

3. The Human Role in Teleoperation

A teleoperated system can be seen as a control loop, as in Figure 1, where the operator plays the role of selecting the appropriate signals and acts on the controlled element to execute the desired task.

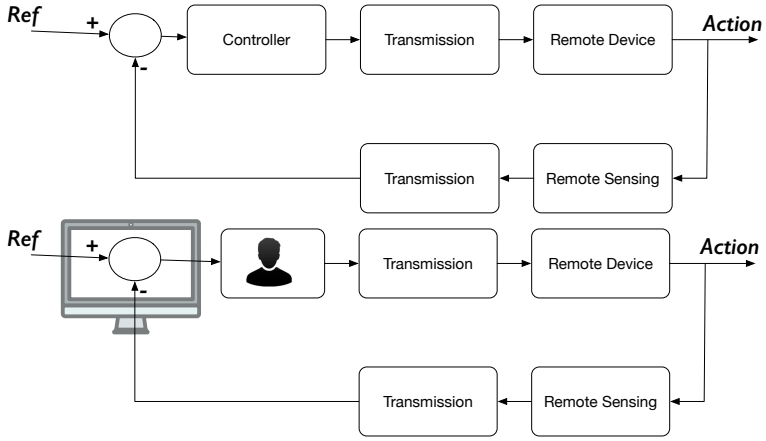


Figure 1. A remote control loop schema (top). A modified control loop where an operator visualizes the intended reference and remote signals to generate the control signals to operate the remote device (bottom).

The task can be one of following a trajectory, a pick-and-place operation, or a much more complicated one. As in any control loop, the objective is to have the system follow some reference signal, and that is accomplished by comparing that input signal with some measurable observation of the system output. As the operator plays the role of the controller, he/she needs to receive the information about the system variables that he/she intends to control [28,29]. Frequently, the information received is very limited and corrupted by different types of noise. It is here that humans normally play still irreplaceable roles given their ability to make correct inferences from incomplete data. Humans can still extract the necessary information in the presence of the most variable conditions, where typically automated mechanisms can only be tuned to specific working conditions and fail if these conditions are not met. Humans have the ability to use temporal signals, to build mental image mosaics, and to use them to make and execute plans [9]. This comes with a price: it is very hard to follow a memorized plan and simultaneously keep track of multiple information sources. In fact, the required concentration level induces important fatigue, and as it rises, it reduces our attention and concentration capabilities. Humans have some difficulties in noticing changes in slowly varying stimuli even if their intensity is high [30]. Actually, the exposure to the same stimuli for some period of time induces a process of desensitization where a constant or repetitive stimuli loses importance and becomes progressively integrated as part of the "background". On the other side, we are particularly good at learning automated skills; thus, operations like driving a car become simple reactive tasks where the conscious level is mostly left for the high-level plan, whereas lane keeping, velocity control, distance to other vehicles, etc., are performed mostly at the subconscious (reactive) level. It is well known that the reactive execution is simpler and less "energy consuming" than tasks that require reasoning. Similarly, the reactions to well distinguishable visual, auditory, or tactile stimuli generate much less fatigue than being focused on the detection of minor changes of the same stimuli. We are equipped with a set of complementary sensing mechanisms that together enable us to pay attention to most of the important events in our lives with minimum effort. For example, although the human eyes only produce high resolution information at their fovea, to minimize the amount of normal processing workload, their peripheral vision can detect sudden changes and motion. These motion field-generated stimuli, as well as auditory stimuli are enough to direct the visual attention to the right spot in the surrounding space when needed. This means that although we frequently browse the surrounding areas around us, we rely more on particular event detection to attract and direct our attention, instead of using any type of constant and exhaustive scanning process. This has indeed two main advantages: it saves energy and allows performing focused tasks while still being able to detect nearby events, in particular those that may represent any type of danger and thus require immediate action.

Why are the above questions important when we are talking about teleoperation-based tasks? When we talk about executing tasks, we always define, implicitly or explicitly, success goals or success criteria. Once the goal is defined, the execution time, the amount of work done, or other task performance measures are used for evaluation and compare operator skills or the usability metrics of different systems. These analyses are very important as they are one of the few ways of evaluating interactive systems quantitatively and may complement other types of more subjective results obtained from user judgments or opinions. Nevertheless, these evaluations can only be done after the whole teleoperation system is built and set up. Furthermore, their analysis may reveal that the system is not adequate for the pre-established goals, but does not necessarily provide any guidance in discovering which are the elements that are responsible for the failure. From this, it seems clear that some set of recommendations and guiding rules can be very handy to anticipate the possible problem sources and avoid them during the early design phases. Some of these recommendations can be obtained from known studies from areas like psychophysics, neuroscience, physiology, or ergonomics and establish limits for various operation's parameters. Subsequently, an analysis of engineering models of the system to be designed may provide predictions for relevant parameters that may then be verified if they are inside the acceptable ranges defined by the former recommendations. Other recommendations can come out of heuristic experience and still provide valuable guidance to anticipate and avoid any possible problems or performance issues. The following section presents a set of recommendations for the design of teleoperation systems.

3.1. Human Factors, Tasks, and Telepresence

Human beings have a remarkable ability to adapt to unexpected constraints and still carry on with their actions [30]. Skilled people can perform tasks, including teleoperation, even when there are momentary failures in certain feedback channels, such as visual feedback. This means that the person has memorized all the operations and can predict the output of his/her actions under certain ranges of disturbances in the visual feedback. In the absence of the typical feedback modality, other senses are used to get the necessary feedback; memories are triggered, and adjustments can be made (e.g., haptic feedback, touch). Even in cases where the sequence of actions is to be memorized and the person is expected to be able to execute the sequence of actions without any kind of external stimuli, he or she uses various sensorial cues (e.g., proprioception and time notion) to adjust the performance and correct any motor action deviations. With teleoperation, a user aims to transfer his/her abilities to a remote agent to perform tasks like navigation, perception, and manipulation. If during this process, he/she may have the impression of being at the remote site, these actions will benefit from first person perception and cognitive mechanisms, and therefore improve the achieved performance. This sense of being at the remote site may be defined as telepresence [13]. Damasio [31] and Metzinger [32] mentioned that there is a close link between self-experience, selfhood, and the first-person perspective. Metzinger also referred to "The Consciously Experienced First-Person Perspective" to support more complex forms of phenomenal content, such a conscious representation of the *relation* between the person and varying objects in his/her environments.

Issues affecting telepresence: A user can perceive a remote environment and navigate or manipulate objects through a teleoperated robotic system; however, several issues can affect such tasks, degrading the sense of telepresence [8,19]:

Field of view (FoV): Observing the remote environment through a camera's video stream reduces the peripheral vision of the user, and it can negatively affect the spatial perception, compromising manipulation and navigation abilities.

Knowledge of the robot's orientation: Users need to know the position and orientation of the robotic agent in the remote place, as well as the robot's topology (e.g., arm position, body size, and pitch and roll angles).

Camera's view point and *frame of reference*: The placement of the cameras may affect visual perception by providing unnatural views to the user (i.e., compromising pose and position estimation).

The egocentric (first-person) perspective comes up as the view for someone present in a space, enhancing, therefore, the sense of telepresence. However, sometimes, an exocentric camera view (third person) may present advantages in the execution of a specific task.

Depth perception: Viewing the remote scenario through a monocular camera can limit the acquisition of significant depth information. The projection of 3D depth information onto a 2D display surface foreshortens or compresses depth cues (e.g., distance underestimation).

Video image quality: Factors like low image resolution, reduced frame rate, or reduced number of colors can make user's remote spatial awareness, target localization, and consequently, response time difficult.

Latency: The time lag verified between the operator's input control action and the observed system response determines his/her control behavior. The aim is a continuous and smooth control operation; however, when the latency increases, the operator adopts the "move and wait" control strategy.

Motion effect: Performing manual tasks on top of a moving platform can be quite challenging and ultimately can induce the operator's motion sickness. Vibrations and disturbances of the visual feedback can make the operator's input controls challenging.

4. Technological and Human Perceptual Factors: Effect on Skill-Based Behavior and Immersive Teleoperation Design

As mentioned in Section 2, the activities of human operators rely on three types of performance, namely skill-, rule-, and knowledge-based behavior. The higher the involvement of rule- and knowledge-based layers, the higher is the cognitive workload requested of those operators. Perceptual disturbances can lead to a higher intervention of more rational behaviors, contrasting the reactive mode of skill-based behavior [25]. Thus, it is important to keep activities at the skill-based behavior level where perceptual signs are essential to lower the workload. The present section discusses the technological and human perceptual factors with relevance for immersive teleoperation design to minimize workload.

Given that human vision provides over 70% of all the sensory information used in the interaction with the world, we start by presenting the ideal display specifications in Table 1, considering the human eye limits. The comparison of these references with those of current visual mediation systems allows the identification of critical factors for teleoperation.

Table 1. Ideal display specifications to match the human eye limits.

Display Propriety	Range Value	Ref.
Latency	<7–15 ms	[33,34]
Angular Pixel Density	>60–200 pixels/degree	[35,36]
Field of View	210° (H) × 135° (V)	[37]
Frame Rate	>1800 Hz	[38]
Color	10 Millions	[39]
Dynamic Range	$1{:}10^9$	[40]

Reliable perception is determinant for HRI task performance, the visual feedback being a major source of information. Actually, users perform better in simple or visually less complex environments (e.g., structured spaces, interactions with few objects and at similar depths, few concurrent tasks) [41,42]. Related works aiming to minimize workload through visual features' enhancement show positive results [43–45]. As the task becomes visually more demanding, the involvement of other sensory channels, such as tactile, haptic, and auditory, can contribute to maintaining performance [46]. However, these studies also demonstrate that technological mediation of the visual sensory channel can limit visual features and consequently degrade the user's performance and increase the user's workload. We present some findings concerning visual interface design that aim to mitigate the

mentioned problems. Table A1, at the end of this paper, presents a summary of the specs for mediation teleoperation technology based on human capability requirements and based on related works' specs findings for a given a task.

4.1. The Human Eye and the Real Scene Resolution

Human visual acuity, that is the power to resolve fine details, relies on optics and neural factors. A person with "normal" visual acuity, i.e., 6/6 vision in meters (or 20/20 in feet), can discriminate the letter E in the Snellen chart at a distance of six meters. Given that the size of these optotypes is subtended in a visual angle of five arc minutes and the eye can resolve the gap between the five horizontal lines of the E letter, i.e., 1/5 of the arc, the human visual acuity is one arc minute.

At a distance of six meters (or 20 feet), the eye can detect an interval lower than 1.75 mm between two contours. The discrimination of a line through optics suggests at least three photo-detectors (stimulated, not stimulated, and stimulated), so for each arc minute, there is a photo-detector. Thus, as one arc minute is equal to 1/60 degrees, there are at least 60 photo-receptors/degree in an angle subtended at the fovea [35] (an analogy with a camera could be 60 pixels/degree).

The retina of the eye is composed of *cone* cells that are sensitive to color and *rod* cells that are sensitive to low light (about seven million cones and 120 million rod cells). It includes the fovea, which is a small pit region of the retina (1.5 mm wide) with a high density of cone cells (exclusively) and where the visual acuity is higher. Fixating on an object implies the movement of the eyes that makes the image fall into the fovea. The result is the sharp central vision essential for human tasks like reading or driving.

Resolution acuity enables identifying two very close lines or contours, and there are people whose vision exceeds 6/6. Human grating resolution ranges from one to 0.3 arc minutes [36,47]. Related also, *detection acuity* refers to the ability to detect small elements in a well contrasted scenario, e.g., black points on a white background. Studies show that the human eye can detect a single thin dark line against a white background uniformly illuminated and subtended in just 0.5 arc seconds [48,49] or 0.0083 arc minutes. It is approximately 2% of the diameter of the fovea's cones, which is a mechanism of the visual cortex and not exclusively based on the structure of the retina. Detection acuity seems to result from a spatial temporal averaging process involving the retina's peripheral region (even where rods and cones present a decreasing density) [50]. Considering that the eye's fovea has at least 60 photo-receptors/degree, an ideal display would have, at least, a resolution of 60 pixels/degree to stimulate the cells of the fovea. Of course, in the fovea's periphery, there is a decrease of photo-receptors; however, no one knows at what region of the display the user will look. Therefore, hypothetically, the ideal display would provide a resolution of 12,600 $\times$ 8100 pixels to approximate human's vision (considering both the eye's FOV, 60 pixels/degree $\times$ (210° (H) $\times$ 135° (V)) [37]).

Assuming that an immersive remote visualization system works like a *video see-through HMD*, whose cameras follow the exact user's head orientation, it is possible to analyze such a system without image transmission issues. Let us consider a setup composed of an HMD attached to cameras that stream video from the real world in front. *Video see-through HMD* may acquire images of the real scene using miniature digital cameras and present them through the HMD's LCD or OLED displays. Then, the viewing optics of the HMD adapt these images to the human eye, using an eyepiece lens. Ideally, a *video see-through HMD* should present images with resolutions similar to the real ones, but in practice, that may not be the case. Therefore, the perceived resolution of the real scenario is limited by the resolution specifications of either of these three system components: the video cameras, the HMD display, or the HMD's viewing optics.

4.2. Resolution

Generally, higher display resolution leads to better teleoperation performances. Resolution impact is not as noticeable as other limiting factors like latency, but it is significant. Evaluating teleoperation driving tasks at different display resolutions (1600 $\times$ 1200, 800 $\times$ 600, and 320 $\times$ 200 pixels per screen),

Ross et al. [51] observed a 23% reduction in the rover's average speed and an increase of 69 times in the average time that the operator stopped the rover for planning or other reasons while comparing the highest and the lowest resolutions.

The path decision was negatively affected by low resolution conditions, as the teleoperator drove slower due to difficulties in distinguishing obstacles. A quality assessment suggested that high resolution improves operator confidence and contributes to the sense of realism and presence. An earlier study analyzed the impact of resolution in the periphery of an HMD in three conditions (64×48, 192×144, and 320×240 pixels) for a simple search and identification object task [52]. Watson et al. found that the lowest resolution was significantly worse than the two higher resolutions considering the criterion of accuracy and search time. Commodity game industry HMDs keep pushing resolution to higher levels of realism and fidelity (e.g., the HTC Vive display has 1080×1200 pixels per eye, while the new model, HTC Cosmos, presents a resolution of 1440×1700 pixels per eye). In [53], two HMDs with different display resolutions were evaluated for echography examinations in a pre-clinical and clinical study: the HM2-T2 (Sony Corp.) with a dual display with 1280×720 pixels per screen and the Wrap 1200 (Vuzix Corp.) with a resolution of 852×480 pixels for each eye. The study showed that the image quality and the diagnostic performance were significantly superior using the HMD with the highest resolution. Laparoscopic surgery using a high resolution HMD enabled superior image quality and faster task performances as opposed to a low resolution HMD model [54].

4.3. Frame Rate

The frame rate (FR) is the number of images displayed per time unit, indicating the image refresh rate of the system (frames per second (fps) or Hz). Studies reveal that generally, a low frame rate degrades operators' performance. Massimo and Sheridan [55] analyzed the efficiency of moving a robot arm to a target via a camera view, on a placement style task (accuracy and speed peg in hole task), for three frame rate levels (3 fps, 5 fps, and 30 fps). They found that increases in the frame rate improved the teleoperation efficiency significantly, and for levels below 5 Hz, there was a significant performance deterioration. In Chen [56], the participant experienced a significant performance degradation in a simulated target acquisition task with a frame rate of 5 Hz. Ware and Balakrishnan [57] assessed the impact of different frame rates (60 Hz, 15 Hz, 10 Hz, 5 Hz, 3 Hz, 2 Hz, and 1 Hz) on a 3D target acquisition task (Fitts' law style task) and also concluded that lower frame rates, 5 Hz or below, decreased users' performance; however, they suggested that 10 Hz was enough for the tested task. They also noticed that the effect of frame rate and the lag were closely related. The impact of frame rate in the sense of presence was analyzed in Meehan et al. [58] for a virtual environment (VE) placement task at 10 Hz, 15 Hz, 20 Hz, and 30 Hz. They found that presence increased significantly from 15 Hz to 20 Hz and kept growing from 20 Hz to 30 Hz. They also reported that lower frame rates, besides impairing the sense of presence can also cause balance loss and a heart rate increase. Claypool [59] found that first-person-shooter video game users significantly improved their target acquisition task performances while the frame rate varied from 3 Hz to 60 Hz. In Ross's rover teleoperation driving task [51], several display rates were evaluated at 10 Hz, 15 Hz, 20 Hz, 25 Hz, and 30 Hz. Negative impacts were more noticeable at lower frame rates, namely on speed and motion perception. There was a drop of 37% in the average rover's speed when the FR condition changed from 20 Hz to 10 Hz. Participants felt comfortable teleoperating at 25 Hz and started experiencing an initial discomfort at 20 Hz due to flickering. They found it difficult to perceive the rover's speed at frame rates under 15 Hz. Chen and Thropp [60] conducted a survey involving 50 studies, and they found that generally higher FR and a small standard deviation of the frame rate benefit users' psychomotor performances ($\geq$17.5 Hz for placement tasks, $\geq$10 Hz for tracing tasks, $\geq$16 Hz for navigation and tracking targets). In summary, they concluded that 15 Hz is a reasonable threshold for most of the tasks, including perceptual and psychomotor. The sense of presence and that of immersion benefit from higher values of FR (60 Hz to 90 Hz), which enable improved realism [61,62].

4.4. System Latency

Latency in teleoperation refers to the elapsed time between a user's action and the consequent system's observed response. Several components of the system can contribute to this overall time lag: user-input device lag, motor actuator lag, video acquisition and display lag, synchronization lag, communication time delay, etc. This occurs in the remote agent control flow and the visual and/or haptic feedback flow.

The operator's control actions, such as head and hand motions, are mapped into the robot's commands using pose-tracking devices and hand controllers. These commands are then transmitted through communication links and executed remotely by the robot's actuators. Considerable delays can occur from the command generation to its execution, due to motor actuator lag and or the latency of the network. The execution of the commands is viewed through images of the remote scene acquired by cameras and delivered to the operator using the network link. This transmission adds more time delays because of the network latency and the time required to transfer the data images. The time to send the command itself is influenced by bandwidth limitation, and although it is comprised of a few small data packets, it can take a long time when the emitter is far from the receptor (e.g., 2.56 s in the two way light time latency between the Earth and Moon [63]). A head-mounted display (HMD) can also introduce some delays. Immersive technologies related to VR refer to motion-to-photon latency ($l_{m2photon}$) as the time delay between the movement of the operator's head (t_{mov}) and the change of the display screen reflecting the operator's movement (t_{disp_mov}) [34], $l_{m2photon} = t_{disp\_mov} - t_{mov}$.

Lag is a crucial element to allow the operator's brain to think that he/she is physically interacting in the remote place, experiencing the sense of telepresence. For that, immersive interfaces should provide stimuli that trick the sensory system, ensuring consistency between vision, internal sensory information (e.g., vestibular, proprioceptive), and cognitive models. The non-compliance of expectations can break the experience of presence, negatively affecting the task performance and causing motion sickness and nausea [64,65]. For example, when a user stops rotating his/her head, he/she expects not to see moving images on the display, or when he/she starts head motion, he/she is supposed to see images moving immediately and not after a delay. Research [34] has found that the *just noticeable difference* (JND) for latency discrimination should remain below 15 ms. The response of the system to head motions should be as fast as the human vestibulo-ocular reflex, the response of which ranges from 7 ms to 15 ms [33]. During head motion, this reflex stabilizes the retinal image at a chosen fixated point by rotating the eyes based on vestibular and proprioceptive information compensating the motion. In fast movements of the eye, such as *saccade movements* (up to 900°/s), the retinal image may become blurred. To avoid processing unclear information, the brain has a mechanism, called a *saccadic mask* [66], that temporally suspends the visual processing so that the motion of the eye or the gap in visual perception is unnoticed. *Saccade movements* enable fovea rendering approaches, optimizing display resources, and additionally, the *saccadic mask* suggests some tolerance in current VR eye-tracking-to-photon latency [67,68].

NASA researchers [69] studying HMDs to support synthetic enhanced vision systems in flight decks found that for extreme head motions, such as higher than 100 deg/s, the system latency requirement must be below 2.5 ms. Anyway, for demanding tasks requiring headset displays with a large field of view and high resolutions, the recommended system latency is less than 20 ms. For example, Oculus RIFT developers recommend not exceeding 20 ms for motion-to-photon latency in a VR application [70]. In recent developments, they enabled reducing its tracking subsystem latency to 2 ms, allowing more time for the overall system lag. The Oculus Rift CV1 [71] and HTC Vive [72] headsets have a refresh rate of 90 Hz, meaning that they can update their displays every 11 ms.

Human performance studies show that a person can detect latency from 10–20 ms [65]. Lags occur in teleoperation of mobile robots or robotic arms when information is transmitted across a communication network (i.e., end-to-end latency). When the latency is higher than 1 s, the operator tends to change his/her control approach to "move and wait", rather than continuously commanding, predicting, and trying to compensate the delay [73].

The time delay factor may differently affect the performance of a specific task. The negative effect extends to over-actuation for the variable delay, affecting robot-to-operator direction information flow more than the other way. In a telemanipulation task related to laparoscopy surgery, negative effects were observed in the system usability and the performance of experienced surgeons for a delay $\geq$105 ms [74]; in another task related to telesurgery (precision, cutting, stitching, knotting), a degradation in accuracy, precision, and performance was reported for a delay $\geq$300 ms [75]; in a driving simulation task, a performance degradation was observed for a delay $\geq$170 ms [76]; in social interaction with mobile telepresence robots, the recommendation for teleoperation commands' latency is under 125 ms [7]; in a real teledriving task (six wheel all-terrain rover), evaluating the average speed and the average time stopped for path decision, the negative effect appeared for latency $\geq$480 ms [51]; in a car teledriving task on city roads at 30 Km, while analyzing the tracking line, obstacle detection, and performance, problems for delays $\geq$550–600 ms were reported [77,78].

The mitigation of the effects of latencies related to visual feedback in teleoperation systems may involve a predictive display. The goal is to provide immediate visual feedback to the operator, displaying a scene model animated by the commands. Meanwhile, and in parallel, this scene model is updated with measures acquired by the remote teleoperator sensors [79,80].

4.5. Field of View

The field of view (FOV) refers to the size of the visual field observed. Manipulation studies typically compare narrow viewpoints with wide panoramic views, revealing that narrow FOVs result in difficult navigation [81]. For example, they can limit the acquisition of contextual information about the space around a rover, compromising the perception of distance, size, and direction, having difficult cognitive map formation, and being more demanding of operator memory and attention.

In [51], the FOV had a significant impact on rover teleoperation tasks. Different horizontal FOVs were tested, 40°, 60°, 120°, and 200°, and it was found that an horizontal field of view (HFOV) of 200° allowed average speeds 40% higher than with an HFOV of 40°. The average stopped time with an HFOV of 40° was two times greater than with an HFOV of 120°. Path decisions were compromised at lower FOVs, and operators reported that a wide field of view benefited the situational awareness while enabling higher speeds. In [82], the field of view was manipulated at two levels, 30° (narrow FOV) and 60° (wide FOV), while navigating a virtual UGV. Widening the FOV resulted in a superior performance benefit, leading to lower times to complete obstacle navigation tasks, decreasing the number of collisions and the number of turnarounds, and having higher piloting comfort. In [83], three different levels of FOV (48°, 112°, and 176°) wearing an HMD while walking, avoiding obstacles, estimating distances, and recalling spatial characteristics were analyzed. He reported that users' walking was more efficient with a wider FOV, but did not find significant effects on distance estimation (spatial understanding), on user's balance, nor on recalling the characteristics of the environment (i.e., memory recall).

Wider to moderate fields of view tend to contribute positively to performance. Subjectively, FOV significantly improves teleoperators' situational awareness and perception of robot position and motion. However, moderation is advised as a wider FOV can increase motion sickness [81,84,85]. The causes of motion sickness can be related to the fact that a wider FOV increases ocular stimulation and motion in operators' peripheral vision.

Video cameras have been used in robot navigation to perceive the environment, although this process can suffer from the "keyhole" effect [8,86]. This means that just part of the environment can be acquired and presented to the human operator. Usually, the operator overcomes this situation with extra effort, manipulating the cameras to survey the environment and gaining similar scene awareness to direct viewing.

Human eyes provide a horizontal field of view of 210° by 135° in the vertical [37]; however, stereo central vision relies on the overlap of the FOV of both eyes, and it is around 114°. Thus, as a result of $210°/2 + 114°/2$, an ideal HMD should provide a view of 162° horizontally for each eye [38].

Robotic remote operations include remote spatial and motion perception, navigation, and remote manipulation. A limited field of view (FOV) degrades the remote perception in several ways; however, it affects tasks differently. Tasks like navigation, self-location identification in space, and target detection are negatively affected due to the video feedback constraints [87]. Manipulation tasks involving action in a limited space can overcome the FOV limitation by gathering new points of view of the scene with more or less extra effort.

Operators that perform navigation tasks based on a fixed camera mounted on a mobile robot tend to perceive the environment through a stack of images for which the points of view correspond uniquely to the robot's path. This "cognitive tunneling" effect present in egocentric visualization approaches contrasts with exocentric systems in which the frame of reference (FOR) remains unchangeable, requiring less mental transformation. In navigation tasks, operators usually focus their attention on a destination point without worrying much about the surrounding environment. Either in navigation or manipulation tasks, a limited FOV can cut crucial distance cues and limit user depth perception [88]. However, Knapp [21] proved that an HMD's limited field of view has no effect on perceived distance, reported in virtual environments.

Navigation performance research has identified several problems related to restricted FOV: drivers with difficulties in judging vehicle speed, object perception, time to collision, obstacle location, and start procedures to curve. The peripheral vision contributes to speed perception, lane following, and lateral control avoidance. Wider FOVs are common solutions in teleoperation indirect navigation. This broadens the view of the scene either with wide angle cameras or using extra cameras on-board to cover the surrounding space. Nevertheless, such FOV increments, particularly when based on wide angle cameras, might induce a faster perception of speed. This effect is related to scene compression and quite often leads the operators reducing their speed. Researchers have pointed out that the scene distortion and resolution decrements, related to scene compression, may increase operator cognitive workload, degrade object localization tasks, and cause motion sickness symptoms [85]. These authors conducted a direct viewing driving and an indirect video driving of a military car. Three internal tiled LCDs provided a 100° panoramic view returned by a camera array mounted in the front roof of the car. They tested three sets of camera lenses, providing 150° (near unity), 205° (wide), and 257° (extended) camera FOVs. They found that map planning performance and spatial rotation improve with a wider FOV. Wider FOVs had an effect on spatial cognitive functions similar to peripheral cues for direct viewing. The best performances were achieved using vision displays with the FOV close to direct vision, using systems that enabled electronic adjustment of the FOV.

4.6. Depth Perception

The majority of human interactions depend on space and motion perceptions. This ability enables navigation in an environment and handling objects. *Depth perception* is a filtering process that enables us to perceive the world's tridimensionality. Humans can identify information in images and correlate it with depth in the scene, using both psychological and physiological cues [89]. For example, if an object X partially covers another one Y, then object X is inferred as the nearest to us (i.e., occlusion). These different depth cues can be classified into three groups according to the sensory information: *oculomotor cues*, *monocular cues*, and *binocular cues*.

Oculomotor cues for depth/distance are based on:

- *vergence*, which results from the sense of inward movement that occurs when the two eyeballs rotate to fixate near or distant objects, and
- *accommodation*, the change in the eye lens's shape required to focus on the object at different distances. In fact, accommodation is a monocular cue, based on the kinesthetic sensation of stretching the eye's lens and sensing the tension of the eye's muscles [30].

Monocular cues provide distance/depth information by viewing the scene with one eye only. Besides accommodation, they include:

- *pictorial cues*, which identify depth information in a single two-dimensional image, and
- *movement-based cues*, which extract depth from the perceived movement.

Pictorial cues include (a) *occlusion*, (b) *relative height*, (c) *relative size*, (d) *perspective convergence*, (e) *familiar size*, (f) *atmospheric perspective*, (h) *texture gradient*, and (i) *shadows*. The mentioned cues consider a stationary observer; however, new depth cues arise when a person rotates his/her head or starts walking. *Motion-produced cues* include (1) *motion parallax* and (2) *accretion and deletion*. *Motion parallax* results from the fact that when we move, nearby objects seem to move faster than the more distant ones. As we move and due to perspective projection, the displacement of the projection of the objects in the retinal image depends on the distance. *Accretion* and *deletion* result from sideways movements of the observer, leading some object parts to become hidden, and others visible. These cues, based on both occlusion and motion parallax, arise when superimposing surfaces seem to move in relation to one another. Two surfaces, at different depths, can be detected using these cues.

Binocular cues use information from both eyes to provide important distance/depth information. *Binocular disparity* (*binocular parallax*), i.e., the difference of corresponding image points in both eyes, combined with the geometry of the eyes (convergence) enable localizing a tridimensional point (triangulation). *Stereopsis* is the feeling of depth that results from information provided by binocular disparity.

In teleoperation, by changing the point of view of remote monocular or supporting stereo vision allows operators to get important cues for depth perception, crucial for navigation and telemanipulation. In telesurgery or laparoscopic tasks, the observation through an indirect method affects the surgeon's depth perception and diminishes his/her eye-hand coordination ability. Depth perception problems result from accommodation and convergence inconsistencies, the lack of shadows in endoscopic video images, and the lack of movement parallax and stereovision. Eye-hand coordination problems are due to the distance location of the monitor and because the video images of the doctor's hand movements appear mirrored, rotated, and amplified [90]. To overcome these contingencies, surgeons must have intensive, long, and specialized training. Technological solutions [91,92] include flexible and motorized endoscopes that compensate misorientation and provide movement parallax through probe positioning controlled by the surgeon's head movements.

4.7. Frame of Reference

Many teleoperation visuomotor tasks such as telemanipulation or telesurgery can be highly demanding for the operator because the frames of reference for vision and action are misaligned [93,94]. The operator's control actions are harder when there is no alignment between his/her point of view, the input device, and the coordinate frame of the robot. Any mental transformation, such as rotation or translation, imposed by the input/output interface to control the robotic system can increase mental workload and consequently decrease task performance. Thus, the interface should be intuitive enough, so when the operator acts on the input device, the robotic agent should move in the expected direction coherently with the input gesture. Moreover, the operator should observe or sense the robotic agent moving in the expected direction. Researchers have shown that the angular misalignment between the visualized axis of rotation and the controller's hand axis rotation can cause control response delays and decrease accuracy in telemanipulation tasks [95,96]. Human path following performances degrade for non-orthogonal angles relating control and visual frames [97]. The eye-hand coordination tasks should be as natural as if the operator were physically in the remote place. In teleoperation systems, a one-to-one correspondence between control and display devices is desirable, so for a given control movement, the consequent change in the image display should appear to be in the same relative place [10]. Humans can easily adapt to a lateral displacement between the expected point of view of a movement and the observed movement through a 3D perspective display, if less than 15° [44]. For higher angles >15°, the adaption decreases.

One way to achieve the alignment between the operator's point of view, the input/output device coordinates, and the robot's coordinate frame is to design the teleoperator agent anthropomorphically,

or ensure that the extent of the control action is easily perceived and mapped through the feedback of the agent (e.g., ensure that a steering wheel rotation is proportional to the robot's linear displacement in the diagonal). The egocentric visual feedback suggested in the present work aims to fulfill one of these natural and one-to-one correspondences (visual, haptic, proprioceptive).

4.7.1. Egocentric or First-Person View

Why provide different views of the remote site? Depth cues provided by human's binocular system are important for the perception of the environment. In the absence of pure stereo vision, different points of view can contribute to the visual perception. Additionally, mental models are matched against reality through egocentric information arriving from the different sensory modalities [32]. For example, an observer expects to see the scene's elements changing when he or she moves. He/she expects that parts of the scene initially occluded become visible and others become occluded. Moreover, it is expectable that the views of a nearby object or part of it present higher changes in images than those more distant. Thus, all these reality cues need to be replicated by the egocentric visualization system because the human brain uses them to estimate distances.

In traditional teleoperation systems, the operator controls the remote camera's orientation manually. He/she has to control two or three degrees of freedom of the camera and, additionally, several degrees of freedom of the robot, which increases the operator's efforts. To address these problems, we suggest a virtual cockpit for the operator where he/she experiments with the sensation of controlling the robot inside of it, i.e., telepresence [17,98]. Operators can browse the vehicle's surroundings as if they were aboard it and simultaneously realize that the controls and instrumentation panel at his/her disposal present coherent feedback. This is implemented using a head-mounted display (HMD) whose orientation is used to control a pan and tilt unit (PTU) that supports the remote camera. With this solution, the user's head movements are implicitly transposed into camera movements. The approach enables the operator to look around, just by naturally rotating his/her head, viewing what was supposed to be seen in the remote place with that movement. By superimposing real video feeds with virtual elements, synchronous with the operator's point of view, we contribute to the visual perception of being immersed in the remote place.

4.8. Working Memory

Spatial memory is part of our cognitive system. Those tasks, involving higher demands from human short- and long-term memory, typically result in higher mental workloads. *Working memory* refers to the mind's capability to maintain and manipulate important information to perform complex tasks such as reasoning, comprehension, and learning. Badlley's model [99] evolution proposes a four component model: a central executive system along with three short-term storage subsystems, the visuo-spatial sketchpad (processes visual-spatial information), the episodic buffer (with limited capacity), and the phonological loop (processes verbal-acoustic information). For example, route planning or thinking about time involve the visuo-spatial sketchpad subsystem and the central executive system. Maintaining information in the working memory involves a rehearsal process [100,101]; however, working memory overload or significant distractions can interrupt this process. For example, operators that rely on working memory to label objects or to recall the spatial positions and orientation of the objects tend to experience higher cognitive workloads. Moreover, working memory data are necessary for mental transformations involving visual and proprioceptive frames of reference (rotations and translations). The human cognitive system uses the *working memory* to retain new information temporally for processing or to store in long-term memory. It is shown that working memory has a limited capacity and holds the information for a short time (a few seconds). The manipulation of this information enables reasoning and decisions and shapes event behavior. Miller [102] introduced the "chunks" concept, referring to any set of information to be associated with long-term memory. Moreover, he defended that human working memory can hold just 7 ± 2 "chunks" or "items" of information. More recent studies point to a number of "chunks" dependent on the type of information.

Managing multiple and separate frames of reference, controlling multiple actuators, distributing attention over multiple instruments, memorizing the spatial position of multiple targets due the limitation of the field of view, register numbers with several digits, or changing sequential procedures frequently due to external unpredictable events are some examples that can compromise tasks. As a conclusion, teleoperation systems should avoid making a person recall more than 7 ± 2 "chunks" to evolve through a task.

5. Immersive Interface Testbed: An Evaluation Example

Traditional interfaces for teleoperated robots integrate multiple displays and controls, requiring specialized training by the operators. Such additional effort is not easy for a non-specialized operator, making them skeptical about technological advances. As noted in the previous sections, common commands, or perceptual actions, not properly executed or reflected by the system, can become distracting factors. Additionally, letting operators experience the sensation of being in the remote environment (telepresence), with minimum disturbances of the mediation technology, makes the task performance more natural. Simpler operation control is achieved through the combination of telerobotics and telepresence. We propose an immersive interface testbed that enables the manipulation of perceptual factors and an analysis of their impact on immersion, presence, task performance, and workload. Ensuring perceptual factors that keep the operator's activities at the skill-based behavior level, with less intervention from higher decision levels, saves energy and enables focused tasks [25]. The approach suggests a familiar and traditional workspace where the operator can perform manipulations while disposing of a wide egocentric field of view and precise control of the actuator. The solution includes: providing a *first-person view* (*FPV*) of the workspace scene; enriching the operator's spatial perception, by virtually disposing of the mediated sources of information in conforming with natural layouts; providing consistency between action-movement and the human's sensory system (visual, haptic); transforming some explicit commands into implicit ones; providing a natural point of view of the remote site. These approaches contribute to reducing physical and cognitive workload while improving task performance. The following text describes the evaluation methodology for the proposed teleoperation immersive interface compared with the traditional approaches.

5.1. Participants

The experiments were performed at the Polytechnic Institute of Tomar and the Institute of Systems and Robotics, Coimbra, Portugal, with 25 participants, two females and 23 males. The participant group included students and researchers in fields such as engineering and computer science, with an overall average age of 30.3 years and a standard deviation of 8.65 years. All participants reported normal or corrected to normal vision. None of them had prior knowledge of the experience or technologies involved. Participation was voluntary, and ethical research principles were observed.

5.2. Experiment Design

To evaluate the effect of immersive technologies in teleoperation spatial perception, we designed several evaluation procedures where participants were invited to accomplish several hand-eye coordination tasks while their performances were analyzed. We compared immersive visual feedback against traditional visual feedback based on a traditional monitor and a remote fixed camera (Figure 2 exemplifies the setups used for the pick-and-place task). The proposed immerse visual feedback uses a head-mounted display (HMD) for visualization purposes while, the pose of the user's head is used to control the remote camera orientation, gathering different points of view.

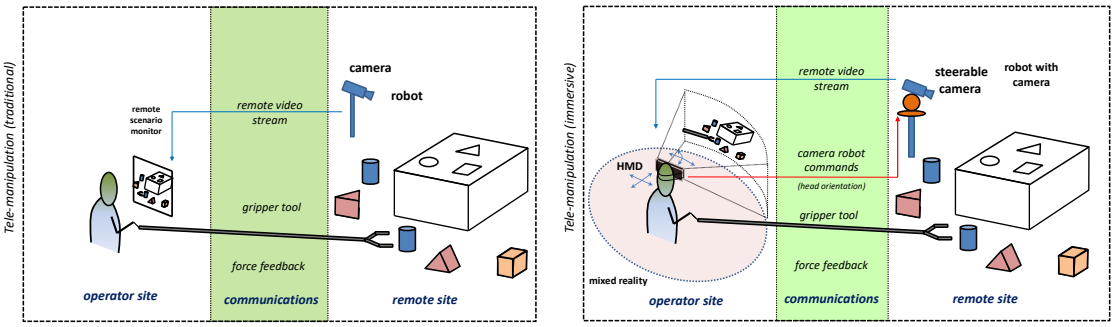


Figure 2. Traditional teleoperation setup (remote visualization through a fixed monitor and camera) vs. immersive interface (point of view transfer; the head-mounted display (HMD) controls the remote camera).

The developed setups manipulate the type of visualization of the remote environment and the control of the remote camera orientation for several tasks (see Table 2).

Table 2. The two different test setup combinations for semi-teleoperation tasks. Fixed view, first-person view (FPV).

Test	Display	Camera Orientation	Stick/Arm Control
1	Traditional 1 Monitor	Fixed	stick/gripper haptic feedback
2	Immersive via HMD (FPV)	head orientation from HMD IMU	stick/gripper haptic feedback

Each participant performed a given task using both setups. The evaluation consisted of analyzing a set of related performances (quantitative measures), recorded during the experiment, and the answers to a short questionnaire (subjective measures) given after each trial. Statistical significance was assessed using repeated measures (within subjects) ANOVA analysis.

However, given that the immersive visual interface and the traditional visual feedback interface differ in terms of setup components, a control experiment was conducted to assess the disturbance factor introduced by each component of the interface.

For a typical pick-and-place task, users should grab five blocks of different shapes and insert them into a box through the respective shape hole. This task was accomplished standing in front of the workspace, using the right hand to manipulate the blocks and using different media interfaces that introduced displacements between visual and haptic frames.

5.3. Apparatus

The study was conducted using two types of visualization interfaces to perceive the workspace: a *traditional interface* and a *immersive interface*.

The *traditional interface* consisted of one 22″ LCD monitor, the Samsung SyncMaster 2233RZ with a native resolution of 1680 × 1050 and vertical refresh rate of 120 Hz. This monitor displays images acquired through a webcam with a wide field of view, in the remote workspace from a single fixed point of view. The webcam used is the HD Logitech C270 with an optical resolution of 1280 × 960 pixels (1.2 MP), configured to capture video (4:3 SD) with 800 × 600 pixels, a focal length of 4.0 mm, and a field of view (FOV) of 60°.

The *immersive interface* consisted of one HMD, the Oculus Rift DK2, with a display resolution of 960 × 1080 per eye, an OLED screen of 5.7″, a refresh rate of 75 Hz, a persistence of 2 ms, 3 ms, full and viewing optics with a field of view of 100°. Internal tracking includes gyroscope, accelerometer,

and magnetometer sensors with an update rate of 60 Hz. It refines HMD's pose with a positional tracking sensor based on a near-infrared CMOS sensor. The HMD controls the pan and tilt unit coupled with a webcam, the full HD Logitech C920, with an optical resolution of 1920 × 1080 pixels, configured to capture video with 800 × 600 pixels and an FOV of 78°. The pan and tilt unit supports this camera and orients it according to the HMD's pose orientation, that is conforming to the user's head orientation. The Computer-Controlled Pan-Tilt Unit (PTU 46-17.5) from Directed Perception has two-step motors for the pan and tilt movements: load capacity over four lbs, speeds over 300°/s, resolution of 3.086 arc minutes (0.0514°).

5.4. Evaluation Procedure

To analyze the effect of immersive and egocentric visual feedback, participants were asked to perform three hand-eye coordination tasks using two visual interfaces: traditional interface: fixed monitor, a single point of view with a wide view scene ($Fix_{Display} + Fix_{Cam}$) vs. the immersive interface: HMD and controllable point of view ($Rift_{Display} + Mov_{Cam}$).

Task 1, touch: Users should press on several key blocks using a 1 m stick according to a random sequence defined by the computer (the time to accomplish the task was measured and block touch mistakes) (see Figure 3).

Task 2, pick-and-place: Users should grab five blocks of different shapes using a manual gripper (80 cm long) and insert them into the respective hole of a wood box (the time to accomplish the task was measured; insertion difficulties due to spatial perception errors or handling were interpreted as delays) (see Figure 4 where blocks with different shapes like cylinders, cubes, and triangular prisms only enter in the right hole shape; time recording starts with the first block grab and ends with the last insertion).

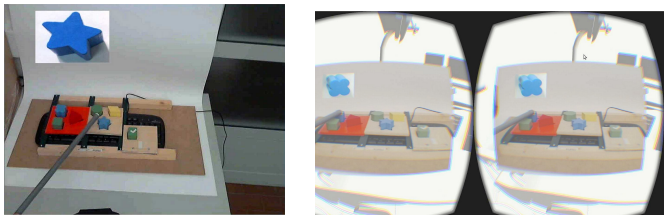
Task 3, path following: Users should follow a predefined 3D path with a metal loop at the tip of a stick (the time to accomplish the task was measured and the number of hits loop/wire recorded). This setup consists of one metallic pipe with curved and straight paths, where the user should move the metal loop along the pipe, avoiding electric contact between both (see Figure 5).

Participants started randomly, either with Interface 1 or 2, to mitigate the effect of the learning factor. Task 2 and Task 3 were performed in a random order for the same reason.

The evaluation consisted of analyzing a set of performance-related parameters, which were collected during the experiments, and the answers given to a short questionnaire after each trial. The collected parameters, the questionnaire, and their analysis are presented in the remainder of this section.

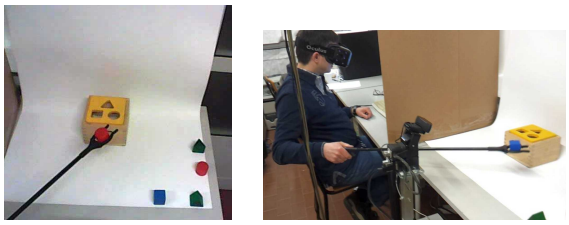
The procedure can be summarized as:

1. The participant is instructed about the task objectives and procedures.
2. Execute the trial (Tasks 1, 2, 3) with one of the two setups randomly selected.
3. Fill in the questionnaire about the user experience.
4. Repeat until four trials are complete.



(**a**) (**b**)

Figure 3. Task 1 setup used to compare performances. The user using a hand stick presses a sequence of key blocks defined by the computer at random. Interface view shows the next block to touch at the upper left corner of the display, after each correct touch. (**a**) Traditional interface vs. (**b**) immersive interface.



(**a**) (**b**)

Figure 4. Task 2 setup used to compare performances: The user grabs blocks with a hand gripper and places them into a wood box through the corresponding hole, using (**a**) the traditional interface ($Fix_{Display} + Fix_{Cam}$) vs. (**b**) the immersive interface ($Rift_{Display} + Mov_{Cam}$).

(**a**) (**b**)

Figure 5. Task 3 setup: The user follows a predefined 3D path holding a stick with a metal loop through a thick metallic pipe avoiding contact. An LED light signals the electric contact. (**a**) Traditional interface vs. (**b**) immersive interface.

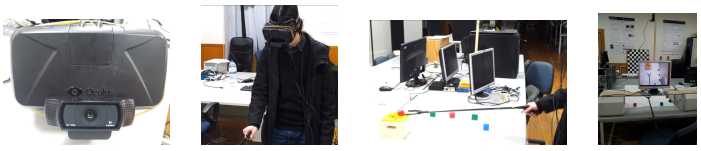
The procedures of the control experiment consisted of a typical pick-and-place task using different media interfaces randomly following the repeated measures approach (within subjects); see Figure 6. Users were asked to perform the task using:

- Direct vision (stereo binocular) and his/her bare hand (the baseline);
- Direct vision (stereo binocular) and a manual gripper;
- HMD indirect vision (mono biocular) and his/her bare hand;
- HMD indirect vision (mono biocular) and a manual gripper;
- Monitor indirect vision (mono biocular) and a manual gripper;

In the *"direct vision (stereo binocular)"* setup, users used direct visualization to grab the blocks and insert them into the respective holes of a wood box, either using their hand or a manual gripper (80 cm long).

In the *"HMD indirect vision (mono biocular)"* setup, users used an HMD with a monocular video camera fixed in front of it, creating a video see-through HMD. The users used their hand or a manual gripper to manipulate the blocks and could move their head freely (position and orientation). Because the camera and HMD display have different resolutions and FoVs, the 1:1 magnification was not suitable. To reflect the true size of the objects, the user adjusted the viewing image to half of the original size (looking either directly at the scene or through the "video see-through HMD").

In the *"monitor indirect vision (mono biocular)"* setup, the user performed the pick-and-place task standing in front of the workspace using a manual gripper to manipulate the blocks and while looking through a fixed LCD monitor that was displaying a video streaming of the scene using a monocular camera. The users were seated with a LCD monitor at their eyes' height (monitor hanged) so they could manipulate under the monitor with the manual gripper.



(**a**) (**b**) (**c**) (**d**)

Figure 6. (**a**) "HMD indirect vision (mono biocular)": a video see-through HMD; (**b**,**c**) user performing a pick and place task using the setup "HMD indirect vision (mono biocular) + gripper"; (**d**) user performing a pick and place task using the setup "fixed monitor vision (mono biocular) + gripper".

5.5. Measurements and Questionnaires

The usability evaluation was performed in two parts: performance-related measures and user subjective evaluation using a questionnaire.

Regarding the analysis of the performance, we measured the following variables directly from the instrumented object and/or using a third observer to keep records.

Time: task completion time for the procedure. This integrates delays due to depth perception errors.

Hits: collisions with objects due to erroneous spatial perception.

Path following precision: 3D path following precision with the tip of the stick (electric contact and time to accomplish)

The goal is to identify task performance manipulation errors due to impaired visualization, shaking, etc.

For the subjective evaluation, a questionnaire was created inspired by the IBM Computer Usability Satisfaction Questionnaire [103] and based on the presence questions of Slater, Usoh, and Steed [104,105]. The participant feedback classified, on a seven point Likert scale, factors like usability, easiness, control precision, fatigue, realness, telepresence, and embodiment feeling. Questions were translated into the Portuguese language to simplify their understanding. The eight questions to answer were divided into two groups as follows:

Usability and task load questions:

Q1: I visualized the workspace ... (1 = without any difficulties, 7 = with difficulties)
Q2: Was the task tiring? (1 = Not tiring, 7 = Very tiring)
Q3: I managed to manipulate objects quite accurately (1 = Not at all, 7 = Very much)
Q4: The workspace visualization did not difficult object manipulation (1 = Disagree, 7 = Agree)

Immersion presence questions:

Q5: I forgot that I used an indirect technological visualization device (1 = Disagree, 7 = Agree)
Q6: I had a clear perception and total control of stick's movements? (1 = Not at all, 7 = Yes totally)
Q7: I perform better when: (1 = I move my head, 7 = I do not move my head)
Q8: I know where the objects are because I can touch them. (1 = Disagree, 7 = Agree)

6. Results and Discussion

6.1. Results

The performance of the user is measured in terms of time spent and mistakes occurring while executing the task, for each of the proposed interfaces. Additionally, the questionnaire score results enabled a qualitative evaluation.

Task performance related measures:

Figure 7 depicts the mean task-time performance and standard deviation for the key block sequence touches while using the traditional interface and the immersive interface, Task 1 setup. Keyb_Seq. 1 corresponds to touching the first four blocks, Keyb_Seq. 2 to touching the second four

blocks, and Keyb_Seq. 1+2 the time spent to correctly touch the eight blocks. The second trial Keyb_Seq. 2 is faster due to the learning process.

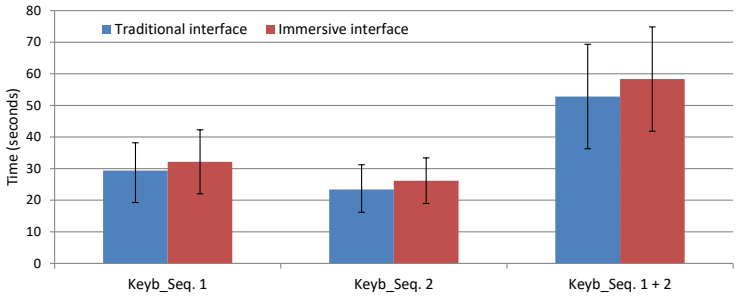


Figure 7. Task 1: mean task-time performance of participants while pressing on a sequence of key blocks determined by the computer: Keyb_Seq. 1 (first round), Keyb_Seq. 2 (second round), and Keyb_Seq. 1+2 (sum of both sequences).

These task performance measures are also available in Table 3, where for Task 1, μ stands for mean task-time performance in seconds and σ is the standard deviation, and the respective subscript t stands for traditional and i for immersive. According to the one-way repeated measures ANOVA (analysis of variance) tests, the round comparisons of Task 1 are not statistically significant:

$$\begin{array}{ll} \text{Keyb\_Seq. 1} & F_{1,8} = 0.212,\ p = 0.657; \\ \text{Keyb\_Seq. 2} & F_{1,8} = 0.343,\ p = 0.574; \\ \text{KeybSeq. 1+2} & F_{1,8} = 0.281,\ p = 0.609; \end{array}$$

where F stands for the F-statistic, p the p-value and should be <0.05 for significant comparisons. Basically, in Task 1, the immersive interface has a poor performance when compared with the fixed camera setup, although it is not a significant difference. Even in trials where participants had some manipulation training first with the $Fix_{Display} + Fix_{Cam}$ setup, there were no changes. An explanation of this fact is that the tested workspace is small, fits all in the field of view of the user, and he/she does not feel the need to move his/her head to get better views.

Figure 8 depicts the mean performance times (seconds) and standard deviation for the Task 2 setup. It shows that users perform the pick-and-place task faster while using the immersive interface than when using the traditional interface. The values are also presented in Table 3, and the comparison involving 20 participants is statistically significant. It was assessed using the one-way repeated measures (within subjects) ANOVA analysis: $F_{1,19} = 7.95$, $p = 0.0109^*$ (the asterisk indicates a significant comparison, $p < 0.05$).

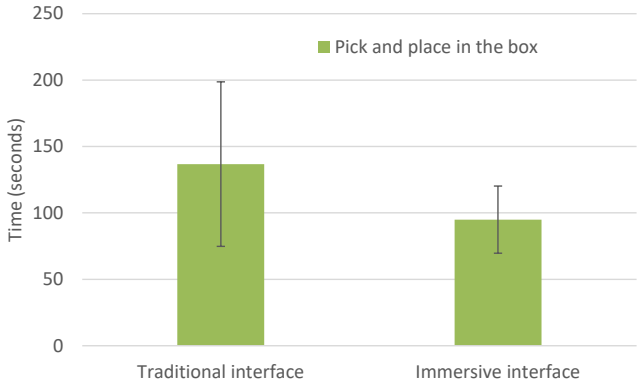
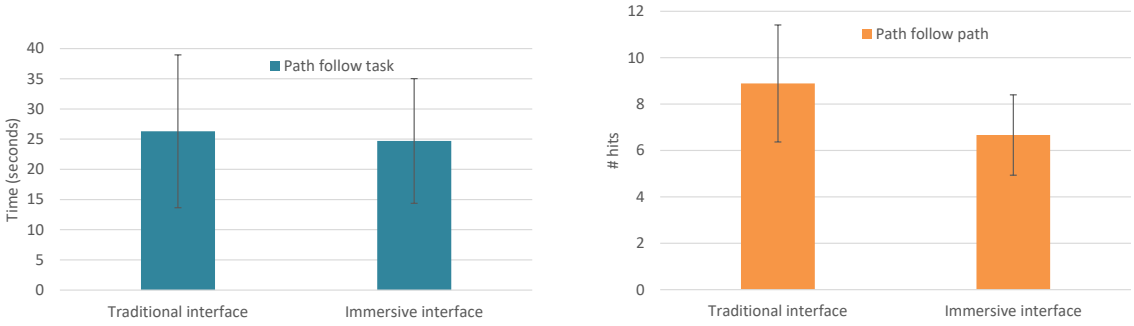


Figure 8. Task 2: mean task-time performance for picking and placing blocks in the box's holes.

Task 3's mean time performance measurements are presented in Figure 9a and in Table 3. The mean time performance using the traditional interface is similar to the immersive interface,

with $F_{1,14} = 0.037$ and $p = 0.85$. Nevertheless, for this task, a lower number of hits indicates better performances, and it occurs while using the immersive interface in opposition to the traditional interface; see Figure 9b. It has a statistically significant difference: $F_{1,16} = 4.747$, $p = 0.044^*$.



(**a**) (**b**)

Figure 9. Task 3: (**a**) mean task-time performance to follow a 3D path with a metallic loop avoiding contact with the guiding pipe; (**b**) mean hits.

Table 3. Mean performance measures summary of Tasks 1, 2, and 3.

			Traditional Interface	**Immersive Interface**
Task 1	mean time performances	Keyb_Seq. 1	$\mu_t = 29.39, \sigma_t = 8.78$	$\mu_i = 32.15, \sigma_i = 10.13$
		Keyb_Seq. 2	$\mu_t = 23.39, \sigma_t = 7.82$	$\mu_i = 26.18, \sigma_i = 7.22$
		KeybSeq. 1+2	$\mu_t = 52.78, \sigma_t = 16.55$	$\mu_i = 58.33, \sigma_i = 16.51$
Task 2	mean time performances	*	$\mu_t = 136.75, \sigma_t = 61.93$	$\mu_i = 94.95, \sigma_i = 25.28$
Task 3	mean time performances		$\mu_t = 26.75, \sigma_t = 14.31$	$\mu_i = 25.50, \sigma_t = 11.52$
	mean hits	*	$\mu_t = 8.8, \sigma_t = 2.52$	$\mu_i = 6.6, \sigma_i = 1.73$

Control task:

Comparisons between individual disturbance factors introduced by each mediation technology (e.g., visual, haptic) while in a pick-and-place box task (Figure 6) are presented in Figure 10 regarding the mean task-time performance.

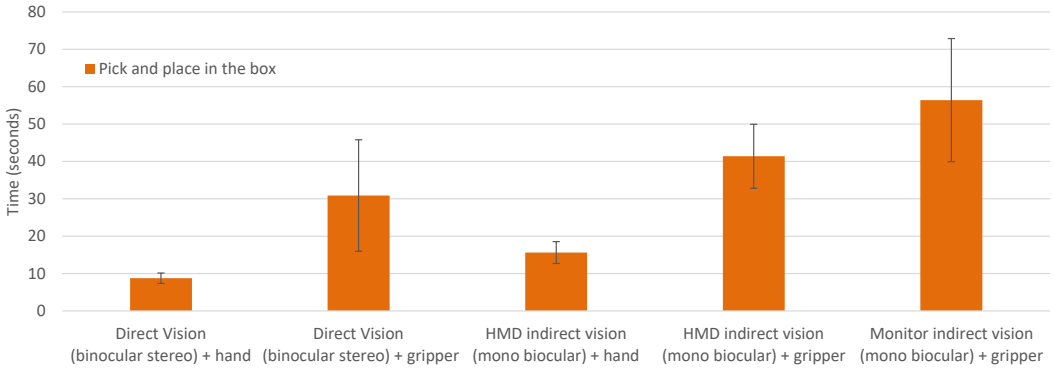


Figure 10. Mean task-time performance: pick and place in the box task. Comparison of individual disturbance factors introduced by each mediation technology (visual, haptic, shift between kinesthetic and visual feedback).

The one-way ANOVA test for the five factors shows a statistically significant comparison: $F_{4,32} = 24.05$, $p < 0.001^*$. The effect of the four disturbance factors on the performance time of eight participants was examined, regarding setup conditions "direct vision (binocular stereo) + hand" (F1), "direct vision (binocular stereo) + gripper" (F2), "HMD indirect vision (mono biocular) + hand"

(F3) and "HMD indirect vision (mono biocular) + gripper" (F4). Its one-way repeated measures ANOVA analysis reveals that the comparison is statistically significant: $F_{3,21} = 23.62$, $p < 0.001^*$. Additionally, post-hoc tests and the pairwise multiple comparisons show which factors' means are significantly different and are summarized in Table 4.

Table 4. Characterization of individual disturbance factors introduced by each mediation technology.

F1	F2	F3	F4	F5
Direct vision (binocular stereo) + Hand	**Direct vision (binocular stereo) + Gripper**	**HMD indirect vision (mono biocular) + Hand**	**HMD indirect vision (mono biocular) + Gripper**	**Monitor indirect vision (mono biocular) + Gripper**
$\mu = 8.75$	$\mu = 30.875$	$\mu = 15.625$	$\mu = 41.375$	$\mu = 56.4$
$\sigma = 1.39$	$\sigma = 14.91$	$\sigma = 2.92$	$\sigma = 8.56$	$\sigma = 16.47$
$F1F2F3F4 : F_{3,21} = 23.62$, $p < 0.001^*$				
	$F1F2 : F_{1,14} = 17.45$, $p < 0.001^*$	$F1F3 : F_{1,14} = 113.023$, $p < 0.001^*$	$F1F4 : F_{1,14} = 36.07$, $p < 0.001^*$	
		$F2F3 : F_{1,14} = 8.05$, $p = 0.013*$	$F2F4 : F_{1,14} = 2.98$, $p < 0.106$	
			$F3F4 : F_{1,14} = 64.71$, $p < 0.001^*$	

Qualitative evaluation based on the user questionnaire:

Figure 11 presents a comparison of the scores for each question while performing Task 2 and Task 3 ("pick and place in the box task" and "path following task") using either the traditional interface ($Fix_{Display} + Fix_{Cam}$) or the immersive interface ($Rift_{Display} + Mov_{Cam}$).

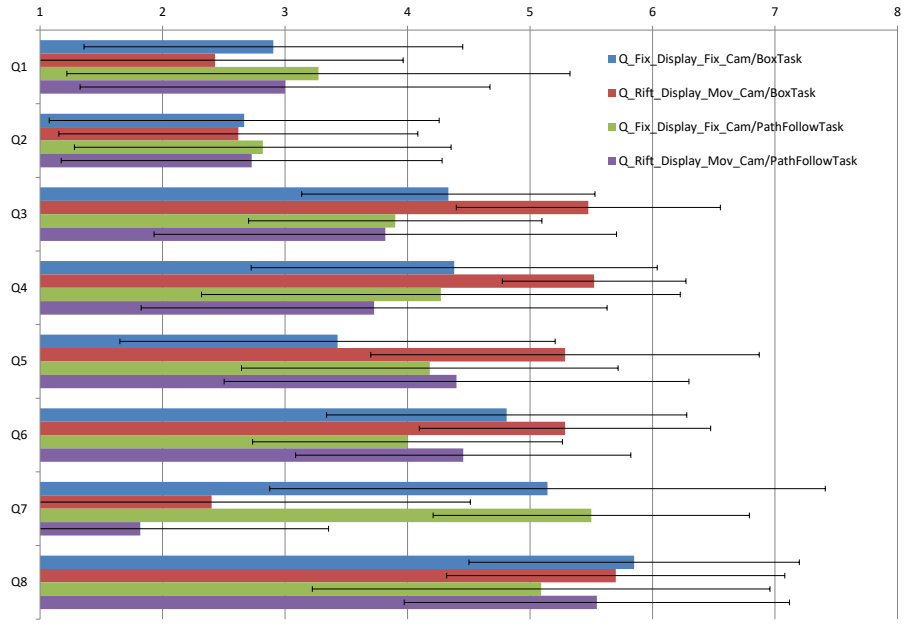


Figure 11. Comparison of mean scores from the user questionnaire feedback for the pick and place in the box task and 3D path following task, Likert scale: one to seven. Q1: I visualized the workspace (without any difficulties/with difficulties); Q2: Was the task tiring? (Not tiring/Very tiring); Q3: I managed to manipulate objects quite accurately (Not at all/Very much); Q4: The workspace visualization did not difficult object manipulation (Disagree/Agree); Q5: I forgot that I used an indirect technological visualization device (Disagree/Agree); Q6: I had a clear perception and total control of stick's movements? (Not at all/Yes totally); Q7: I perform better when: (I move my head/I do not move my head); Q8: I know where the objects are because I can touch them. (Disagree/Agree).

One-way ANOVA tests, without repeated measures, reveal the following statistically significant scores marked with an *:

Q1: $F_{3,60} = 0.7$, $p = 0.541$; Q2: $F_{3,60} = 0.044$, $p = 0.987$;
Q3: $F_{3,59} = 5.7$, $p = 0.0016^*$; Q4: $F_{3,60} = 4.04$, $p = 0.010^*$;
Q5: $F_{3,59} = 4.23$, $p = 0.0088^*$; Q6: $F_{3,60} = 2.49$, $p = 0.068$;
Q7: $F_{3,38} = 6.99$, $p = 0.0007^*$; Q8: $F_{3,58} = 0.64$, $p = 0.59$;

Given that all of the 21 participants performing Task 2 filled in the questionnaires and just a part of them performed Task 3, we opted to detail Task 2. Figure 11 (red and blue bars) summarizes the obtained scores for each question while performing Task 2 with both interfaces. These results were validated through one-way repeated measures ANOVA analysis, and the statistically significant scores are marked with an * :

Q1: $F_{1,20} = 2.016$, $p = 0.171$; Q2: $F_{1,20} = 0.022$, $p = 0.883$;
Q3: $F_{1,20} = 11.294$, $p = 0.003^*$; Q4: $F_{1,20} = 6.981$, $p = 0.015^*$;
Q5: $F_{1,20} = 26.544$, $p < 0.001^*$; Q6: $F_{1,20} = 2.016$, $p = 0.171$;
Q7: $F_{1,25} = 8.432$, $p = 0.007^*$; Q8: $F_{1,19} = 0.516$, $p = 0.481$;

6.2. Discussion

The experiments aim to understand the influence of an immersive visualization interface in relation to spatial and movement perception. To study these factors, we designed several tasks where users had to indicate a 3D position in space based on visual feedback mediated by technological means: a single monitor with a wide single view of the remote scene (traditional interface) versus multiple partial views of the scene naturally acquired while moving their head.

The experiment described in Task 1 demonstrates that, if the workspace of interest is all within the field of view of the HMD, there is no benefit using an immersive interface. The participants did not feel the need to move their head to accomplish the task of pressing key blocks, not gaining depth perception. The time performance was similar for both visualization interfaces.

Task 2 experiments show that an immersive interface outperforms the traditional interface when the workspace of interest is larger than the HMD field of view. Comparing a fixed wide view of the scene with a set of partial views of the scene provided by the head user movement, the last approach improves depth perception due to motion parallax. There is an enhancement of spatial and movement perception demonstrated by the accuracy and speed to accomplish the pick-and-place task. A consequence of the limited FOV of the remote camera (immersive interface condition) was that users moved their head more frequently. Nevertheless, this active spatial perception allowed users to focus their attention on one region of interest, minimizing the workload. All participants took advantage of the touch sense to localize objects and to perceive the height of the gripper tip. Users' common procedures consisted of moving the gripper in contact with the table until reaching the object. This way, the user compensates the lack of height perception through vision while moving the gripper or the tip of the stick.

In Task 3, we tried to evaluate the accuracy of the movement without the help of haptic feedback. Here, the time performances to follow a 3D path did not present a significant difference while comparing both interfaces; however, there was a significant improvement concerning the accuracy of the 3D movement. This is shown by the lower number of hits occurring while using an immersive interface.

Concerning the latency, there was no communication delay due to the proximity of the devices. However, special care was dedicated to tuning the two-step motor lag responsible for the pan and tilt movements of the camera, because faster movements of the head were tracked by the HMD, but were not properly executed by the motors (faster DC motors are advisable). The cameras of the system enabled the video frame rate (>=30 fps), thus fulfilling the requirements for handling the tasks.

The experiment referred to as the "control task" was initially executed to evaluate the weight of each mediation component in the immersive system. The no mediation task "direct vision (binocular stereo) + hand" became our ground truth. Users handled the block with their hands, looking directly with their eyes, and took $\mu_t = 8.75$ s ($\sigma_t = 1.39$) to accomplish the task. With the introduction of a tool ("direct vision (binocular stereo) + gripper"), users took $\mu_t = 30.87$ s ($\sigma_t = 14.9$). The gripper

disturbance caused an increase of 252% in time performance. Movement perception of the gripper in relation to the objects became an issue. Using see-through HMD only ("HMD indirect vision (mono biocular) + hand"), users took $\mu_i = 15.62$ s ($\sigma_i = 2.92$). The see-through HMD disturbance increased 78% in time performance. The limited FoV contributed to such disturbance, and participants quite frequently used their other hand to self-position their body in relation to the workspace. With the introduction of the gripper and see-through HMD ("HMD indirect vision (mono biocular) + gripper"), users took $\mu_i = 41.3$ s ($\sigma_i = 8.5$). The combined disturbance increased 372% in time performance.

Questionnaires showed that:

(Q1) users found it quite "easy to use both visualization interfaces", traditional and immersive ($\mu_t = 2.9$, $\sigma_t = 1.54$ vs. $\mu_i = 2.42$, $\sigma_i = 1.53$) (the subscript t stands for traditional and i for immersive).

(Q2) users did not consider the pick, insert, and place task (Task 2) as a "tiring" one using either interface ($\mu_t = 2.66$, $\sigma_t = 1.59$ vs. $\mu_i = 2.61$, $\sigma_i = 1.46$). Q2 did not present significant differences because the question addressed a common task.

(Q3) users felt that their "movement action" was more precise using the immersive interface (HMD providing an active point of view) ($\mu_t = 4.33$, $\sigma_t = 1.19$ vs. $\mu_i = 5.47$, $\sigma_i = 1.07$). This significant difference results from the gain in depth perception, a consequence of motion parallax. Better visual depth feedback helps to calibrate the arm proprioception system, enabling finer movements of the tool. By moving and seeing our arm, or seeing the tool as an extension of our limb in an unknown 3D space, it helps us to perceive the spatial dimension and localize objects.

(Q4) Inquiring about which interface enabled a better visualization to manipulate objects, it was clearly stated that it was the immersive interface ($\mu_t = 4.38$, $\sigma_t = 1.65$ vs. $\mu_i = 5.52$, $\sigma_i = 0.74$). Subjective scores were statistically significant, and quantitative time mean performances measurements confirmed these results. Task 2 itself is an easy one, and in terms of usability, the preference was for the immersive interface.

(Q5) Inquiring about if users were aware that "visualization" was being supported through an "indirect technological device", they answered that they forgot more easily when they were using the immersive interface ($\mu_t = 3.42$, $\sigma_t = 1.77$ vs. $\mu_i = 5.28$, $\sigma_i = 1.58$). It is understandable that users answered this way, as the immersive interface provides a more natural point of view. People tend to forget that their own eyes are not really in the remote space. This is a confirmation of the importance of the view point transfer proposal as a key element for achieving telepresence.

(Q6) Users also thought that with the immersive interface, they had "a clear perception and total control of stick/tool movements" ($\mu_t = 4.80$, $\sigma_t = 1.47$ vs. $\mu_i = 5.28$, $\sigma_i = 1.18$). This result might require more samples to become statistically validated; notice, however, that some of spatial perception arises from monocular cues already available in the traditional interface ($Fix_{Display} + Fix_{Cam}$). Pictorial cues in a single image and movements in the scene can provide depth information. This type of cue provides information about occlusion, relative height, relative size, perspective convergence, texture gradient, and shadows, enough to perceive the space.

(Q7) The vast majority of users were unanimous in stating that they "they perform better when they move their" heads ($\mu_t = 5.14$, $\sigma_t = 2.26$ vs. $\mu_i = 2.40$, $\sigma_i = 2.11$). Although this question does not make sense with relation to the traditional interface, because a fixed monitor ($Fix_{Display} + Fix_{Cam}$) does not provide a dynamic point of view, it still provides monocular cues related to depth. On the other end, by moving the user's head, the immersive interface adds to the mentioned monocular cues and the motion-produced cues, like motion parallax, accretion, and deletion. Besides the benefit of depth information from motion parallax, users can play with occlusion to match their mental models.

(Q8) Inquiring about the importance of the sense of touch during an interaction, users confirmed the importance of haptic feedback. The requirement of this type of feedback was evident for both interfaces, especially when depth information was unavailable (similar scores, $\mu_t = 5.85$, $\sigma_t = 1.34$ vs. $\mu_i = 5.7$, $\sigma_i = 1.38$). Visual localization of the objects was frequently confirmed through the sense of touch while reaching it. Visual images of the tool in movement enable knowing its position; however, the knowledge of its height can be difficult. To overcome this limitation, some users moved the tip of the tool in contact with the table, refining in this way the reference plane with haptic information provided by arm proprioceptive sensors.

Presence Question Q5-Q8 show that the immersive interface easily becomes transparent for the user, letting him/her feel that he/she is naturally perceiving the remote environment. Visual feedback is transparently mediated and, combined with motor-actions, contributes to the sense of telepresence.

The findings also show that the performances are better when the task workspace is in front of the user, in opposition to setting it on the right side. A visuomotor task where the natural frame of reference of vision is shifted with relation to the common frame of reference for hand/arm movement increases mental workload and consequently decreases task performance. Operators are required to do mental transformation to compensate for their manipulation inputs with relation to the corresponding tool action observed in the displays. The gesture coordination ends up relying more on visual feedback. For example, in our experiments, the pick-and-place task using a hand gripper (Task 2) with the traditional interface had the workspace on the right side; whereas, in the control experiment, the same task was performed using the traditional setup "monitor indirect vision (mono biocular) + gripper" (F5) with the workspace in front of the user. Comparing the mean task-time performance for the traditional interface in both experiments, we found that participants performed faster with the workspace in front than with it on their right side ($\mu = 56.4$, $\sigma = 16.47$ vs. $\mu_t = 136.75$, $\sigma_t = 61.93$). It is a significant difference with one-way ANOVA $F_{1,23} = 8.03$ and $p = 0.009^*$. Thus, users are used to the consistency between visual, proprioception (sense of body position), kinesthesia (sense of body movement), and vestibular feedbacks, and any inconsistency implies new skills.

The research on perceptual factors affecting user's control behavior is useful to improve direct control teleoperation interfaces, minimize control workload, improve task performance, and maximize safety. Additionally, it can contribute to designing semi-autonomous functionalities based on cognitive human-robot interaction architectures to assist operators. For example, during direct interaction such as telesurgery, semi-autonomous systems can prevent dangerous movements of the robotic arm, adapt the responsiveness of the system to the variability of perceptual factors, or adapt the interface to different users. Moreover, the interfaces or robot's behaviors can be adapted to the context.

7. Conclusions

This research shows that people can experience and perform actions in remote places, through a robotic agent having the illusion of being physically there. The sensation can be compelled through immersive interfaces; however, technological contingencies can affect human perception. Considering the results from studies on human factors, we provide a set of recommendations for the design of immersive teleoperation systems aiming to improve the sense of telepresence for typical tasks (ex. Table A1). The mitigation of issues like system latency, field of view, frame of reference, or frame rate contribute to enhancing the sense of telepresence. The presented example of the evaluation methodology allows analyzing how perceptual issues affect task performance. By decoupling the flows of an immersive teleoperation system, we start to understood how vision and interaction fidelity affects spatial cognition.

Task experiments with participants using traditional vs. immersive interfaces allowed quantifying the disturbance introduced by each component on the system. For example, taking as a reference a simple manual pick-and-place task, the introduction of a visual see-through HMD increased the time to perform it by 78%; the introduction of a manual gripper tool increased that time by 252%; and the combination of visual and tool mediation increased the overall time by 372%. Decoupling the flows

of an immersive teleoperation system allowed a separate analysis of visual feedback disturbances (e.g., limited FOV) without the influence of other factors that affect the frame of reference for motor-action. Our findings show that misalignment between the frame of reference for vision and motor-action or the use of tools that affect the sense of body position or sense of body movement leads to higher mental workload and has a higher effect on spatial cognition. Misalignment between kinesthetic and visual feedback increases the mental workload and compromises the sense of telepresence and the embodiment feeling. The mental workload to control the suggested video feedback component is considerably lower (in the immersive interface); however, the combination of both requires a higher effort (i.e., active visual mediation plus tools). Thus, a recommendation is to keep activities at skill-based behavior levels, where familiar perceptual signals are essential to lower the cognitive effort. Future work includes the evaluation of the traditional interface setup, considering the control of the remote camera orientation with a joystick, and the evaluation of the proposed immersive interface to control a robotic arm with haptic feedback.

Author Contributions: All authors contributed equally to this work. All authors read and agreed to the published version of the manuscript.

Funding: This research has been partially supported by Fundação para a Ciência e a Tecnologia (FCT), project UIDB/00048/2020.

Acknowledgments: This publication acknowledges the support of the University of Coimbra, Institute of Systems and Robotics, Coimbra, Portugal and the Polytechnic Institute of Tomar, Portugal.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Human capabilities vs. human capabilities through mediated technologies.

	Task	Analyzed Criteria	Ref	Resolution	Frame Rate (FR)	Latency	Field of View (FoV)	Frame of Reference (Camera Perspective)	Depth Cue	Display Type	Results
Human capability	Multi-purpose	-	Table 1	>60–200 pixels/degree	>1800 Hz	<7–15 ms	210° (H) × 135° (V)	Egocentric	Pictorial, motion parallax, binocular cues	-	
Teleoperation	Placement	Accuracy, speed, and performance	[8,60]	-	>15Hz	-	-	-	-	-	-
	Placement and grasping	Accuracy	[106,107]	-	> 25 Hz	-	-	-	-	-	-
	Tracking	Accuracy, perceived control, and stability	[108]	-	> 12 Hz	-	-	-	-	-	-
	3D Tracking	Accuracy and speed	[109]	-	> 33 Hz	-	-	-	-	-	-
Telemanipulation	Telesurgery: cutting, stitching, knotting	Accuracy, precision, and performance	[75]			<300 ms	-	-	-	-	-
Telemanipulation	Laparoscopy surgery	Usability and performance of experienced surgeons	[74]			<105 ms	-	-	-	-	-
Telepresence	Telepresence robot	Performance, usability, workload	[7]			<125 ms	>170° (H), wide or with pan/tilt	Egocentric	Pictorial, monocular, parallax motion cues	1 × monitor or HMD	Navigation and social interaction
Driving	A 6 wheel all terrain rover of 6.800 kg	Avg. speed and avg. time stopped	[51]	40 pixels/degree, 5 × 1600 × 1200	>25 Hz	<480 ms	200° (H) × 30° (V)	Egocentric	Pictorial, monocular, and motion parallax cues	5 × high-res LCD monitor, side-by-side, true size	Operator's situational awareness and perception of the vehicle's position and motion
Driving	A car driving on city roads at 30 km	Tracking line, obstacle detection, performance	[77,78]	5 × 640 × 480	>25 Hz	<550–600 ms	240° (H)	Egocentric	Pictorial, monocular, and motion parallax cues	3 × high-res LCD monitor side-by-side, true size, and HMD	–

References

1. MTR Corporation, Hong Kong. MTR Deploys New "Vapourised Hydrogen Peroxide Robot" to Further Enhance Disinfection of Stations and Trains. Available online: https://www.mtr.com.hk/archive/corporate/en/press_release/PR-20-020-E.pdf (accessed on 28 August 2020).
2. DeDonato, M.; Dimitrov, V.; Du, R.; Giovacchini, R.; Knoedler, K.; Long, X.; Polido, F.; Gennert, M.A.; Padır, T.; Feng, S.; et al. Human-in-the-loop Control of a Humanoid Robot for Disaster Response: A Report from the DARPA Robotics Challenge Trials. *J. Field Robot.* **2015**, *32*, 275–292. [CrossRef]
3. Murphy, R. Real Robots to Help Fight Ebola (spectrum.ieee.org news). Available online: http://spectrum.ieee.org/automaton/robotics/medical-robots/real-robots-to-help-fight-ebola (accessed on 28 August 2020).
4. Abolmaesumi, P.; Fichtinger, G.; Peters, T.M.; Sakuma, I.; Yang, G.Z. Introduction to Special Section on Surgical Robotics. *IEEE Trans. Biomed. Eng.* **2013**, *60*, 887–891. [CrossRef] [PubMed]
5. Baker, W.; Kingston, Z.; Moll, M.; Badger, J.; Kavraki, L.E. Robonaut 2 and you: Specifying and executing complex operations. In Proceedings of the 2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO), Austin, TX, USA, 8–10 March 2017; pp. 1–8. [CrossRef]
6. Maimone, M.; Cheng, Y.; Matthies, L. Two years of Visual Odometry on the Mars Exploration Rovers. *J. Field Robot.* **2007**, *24*, 169–186. [CrossRef]
7. Tsui, K.; Yanco, H. Design Challenges and Guidelines for Social Interaction Using Mobile Telepresence Robots. *Rev. Hum. Factors Ergon.* **2013**, *9*, 227–301. [CrossRef]
8. Chen, J.; Haas, E.; Barnes, M. Human Performance Issues and User Interface Design for Teleoperated Robots. *IEEE Trans. Syst. Man Cybern. Part Appl. Rev.* **2007**, *37*, 1231–1245. [CrossRef]
9. Ernst, M.O.; Bülthoff, H.H. Merging the senses into a robust percept. *Trends Cogn. Sci.* **2004**, *8*, 162–169. [CrossRef]
10. Sheridan, T.B. *Telerobotics, Automation, and Human Supervisory Control*; MIT Press: Cambridge, MA, USA, 1992.
11. Lombard, M.; Biocca, F.; Freeman, J.; IJsselsteijn, W.; Schaevitz, R. *Immersed in Media: Telepresence Theory, Measurement & Technology*; Springer International Publishing: Cham, Switzerland, 2015.
12. Bohil, C.J.; Alicea, B.; Biocca, F.A. Virtual reality in neuroscience research and therapy. *Nat. Rev. Neurosci.* **2011**, *12*, 752–762. [CrossRef] [PubMed]
13. Sheridan, T.B. Musings on telepresence and virtual presence. *Presence* **1992**, *1*, 120–126. [CrossRef]
14. Minsky, M. Telepresence. *Omni* **1980**, *2*, 45–51.
15. Almeida, L.; Patrao, B.; Menezes, P.; Dias, J. Be the robot: Human embodiment in tele-operation driving tasks. In Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN, 2014), Edinburgh, UK, 25–29 August 2014; pp. 477–482.
16. Slater, M.; Lotto, B.; Arnold, M.M.; Sanchez-Vives, M.V. How we experience immersive virtual environments: the concept of presence and its measurement. *Anu. Psicol.* **2009**, *40*, 193–210.
17. Garcia, J.C.; Patrao, B.; Almeida, L.; Perez, J.; Menezes, P.; Dias, J.; Sanz, P.J. A Natural Interface for Remote Operation of Underwater Robots. *IEEE Comput. Graph. Appl.* **2017**, *37*, 34–43. [CrossRef] [PubMed]
18. Martins, H.; Ventura, R. Immersive 3-d teleoperation of a search and rescue robot using a head-mounted display. In Proceedings of the 2009 IEEE Conference on Emerging Technologies & Factory Automation, Palma de Mallorca, Spain, 22–25 September 2009; pp. 1–8.
19. Prewett, M.S.; Johnson, R.C.; Saboe, K.N.; Elliott, L.R.; Coovert, M.D. Managing workload in human–robot interaction: A review of empirical studies. *Comput. Hum. Behav.* **2010**, *26*, 840–856. [CrossRef]
20. Janvier, M.; Durand, L.; Cardinal, M.R.; Renaud, I.; Chayer, B.; Bigras, P.; de Guise, J.A.; Soulez, G.; Cloutier, G. Performance evaluation of a medical robotic 3D-ultrasound imaging system. *Med. Image Anal.* **2008**, *12*, 275–290. [CrossRef] [PubMed]
21. Knapp, J.M.; Loomis, J.M. Limited Field of View of Head-mounted Displays is Not the Cause of Distance Underestimation in Virtual Environments. *Presence Teleoper. Virtual Environ.* **2004**, *13*, 572–577. [CrossRef]
22. Mania, K.; Wooldridge, D.; Coxon, M.; Robinson, A. The effect of visual and interaction fidelity on spatial cognition in immersive virtual environments. *IEEE Trans. Vis. Comput. Graph.* **2006**, *12*, 396–404. [CrossRef]
23. Schloerb, D.W.; Sheridan, T.B. Experimental investigation of the relationship between subjective telepresence and performance in hand-eye tasks. In *Telemanipulator and Telepresence Technologies*; Das, H., Ed.; International Society for Optics and Photonics; SPIE: Bellingham, WA, USA, 1995; Volume 2351; pp. 62–73. [CrossRef]

24. Tachi, S. Telexistence: Enabling Humans to Be Virtually Ubiquitous. *IEEE Comput. Graph. Appl.* **2016**, *36*, 8–14. [CrossRef]
25. Rasmussen, J. Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Trans. Syst. Man Cybern.* **1983**, *SMC-13*, 257–266. [CrossRef]
26. Swinnen, S.P.; Wenderoth, N. Two hands, one brain: cognitive neuroscience of bimanual skill. *TRends Cogn. Sci.* **2004**, *8*, 18–25. [CrossRef]
27. Blackler, A.; Popovic, V.; Mahar, D. Investigating users' intuitive interaction with complex artefacts. *Appl. Ergon.* **2010**, *41*, 72–92. [CrossRef]
28. Moore, J.W.; Fletcher, P.C. Sense of agency in health and disease: a review of cue integration approaches. *Conscious. Cogn.* **2012**, *21*, 59–68. [CrossRef]
29. Blanke, O.; Metzinger, T. Full-body illusions and minimal phenomenal selfhood. *Trends Cogn. Sci.* **2009**, *13*, 7–13. [CrossRef] [PubMed]
30. Goldstein, E.; Brockmole, J. *Sensation and Perception*; Cengage Learning: Belmont, CA, USA, 2016.
31. Damasio, A. *The Feeling of what Happens: Body and Emotion in the Making of Consciousness*; Harvest Book; Harcourt Brace: San Antonio, TX, USA, 1999.
32. Metzinger, T. *Being No One: The Self-Model Theory of Subjectivity*; MIT Press: Cambridge, MA, USA, 2003.
33. Amin, M.S. *Vestibuloocular Reflex Testing*; Medscape; 2016. Available online: https://emedicine.medscape.com/article/1836134-overview (accessed on 28 August 2020).
34. Mania, K.; Adelstein, B.D.; Ellis, S.R.; Hill, M.I. Perceptual Sensitivity to Head Tracking Latency in Virtual Environments with Varying Degrees of Scene Complexity. In Proceedings of the 1st Symposium on Applied Perception in Graphics and Visualization, Los Angeles, CA, USA, 7–8 August 2004; ACM: New York, NY, USA, 2004; pp. 39–47. [CrossRef]
35. Committee, V.F. Visual acuity measurement standard. *Ital. J. Ophthalmol.* **1988**, *II*, 1–15.
36. Carney, T.; Klein, S.A. Resolution Acuity is better than Vernier Acuity. *Vis. Res.* **1997**, *37*, 525–539. [CrossRef]
37. Andersen, S.R. The history of the Ophthalmological Society of Copenhagen 1900–50. *Acta Ophthalmol. Scand.* **2002**, *80*, 6–17. [CrossRef]
38. Cuervo, E.; Chintalapudi, K.; Kotaru, M. Creating the Perfect Illusion: What Will It Take to Create Life-Like Virtual Reality Headsets? In Proceedings of the 19th International Workshop on Mobile Computing Systems & Applications, Tempe, AZ, USA, 24 February 2018 ; ACM: New York, NY, USA, 2018; pp. 7–12. [CrossRef]
39. Judd, D.B.; Wyszecki, G. *Color in Business, Science, and Industry*, 3rd ed.; Wiley: New York, NY, USA, 1975; p. 553.
40. Boitard, R.; Mantiuk, R.K.; Pouli, T. Evaluation of color encodings for high dynamic range pixels. In *Human Vision and Electronic Imaging*; SPIE: Bellingham, WA, USA, 2015, Volume 9394, p. 93941K.
41. Chen, J.Y.C.; Joyner, C.T. Concurrent Performance of Gunner's and Robotics Operator's Tasks in a Multitasking Environment. *Mil. Psychol.* **2009**, *21*, 98–113. [CrossRef]
42. Darken, R.P.; Cevik, H. Map usage in virtual environments: orientation issues. In Proceedings of the IEEE Virtual Reality (Cat. No. 99CB36316), Houston, TX, USA, 13–17 March 1999; pp. 133–140. [CrossRef]
43. Ha Park, S.; Woldstad, J.C. Multiple Two-Dimensional Displays as an Alternative to Three-Dimensional Displays in Telerobotic Tasks. *Hum. Factors* **2000**, *42*, 592–603. [CrossRef]
44. Park, S.; Woldstad, J.C. Design of visual displays for teleoperation. In *International Encyclopedia of Ergonomics and Human Factors*, 2nd ed.; Karwowski, W., Ed.; CRC Press Inc.: Boca Raton, FL, USA, 2006.
45. Yeh, M.; Wickens, C. Display Signaling in Augmented Reality: Effects of Cue Reliability and Image Realism on Attention Allocation and Trust Calibration. *Hum. Factors* **2001**, *43*, 355–365. [CrossRef]
46. Folds, D.J.; Gerth, J.M. Auditory Monitoring of up to Eight Simultaneous Sources. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Nashville, TN, USA, 24–28 October 1994; Volume 38, pp. 505–509.
47. Thibos, L.N.; Cheney, F.E.; Walsh, D.J. Retinal limits to the detection and resolution of gratings. *J. Opt. Soc. Am. A* **1987**, *4*, 1524–1529. [CrossRef]
48. Klein, S.A.; Levi, D.M. Hyperacuity thresholds of 1 sec: theoretical predictions and empirical validation. *J. Opt. Soc. Am. A* **1985**, *2*, 1170–1190. doi:10.1364/JOSAA.2.001170. [CrossRef]
49. Strasburger, H.; Rentschler, I.; Jüttner, M. Peripheral vision and pattern recognition: A review. *J. Vis.* **2011**, *11*, 13. [CrossRef]
50. LaValle, S. *Virtual Reality*; Cambridge University Press: Cambridge, UK, 2019.

51. Ross, B.; Bares, J.; Stager, D.; Jackel, L.; Perschbacher, M. An Advanced Teleoperation Testbed. In *6th International Conference on Field and Service Robotics-FSR 2007*; Springer: Chamonix, France, 2007; Volume 42.
52. Watson, B.; Walker, N.; Hodges, L.F.; Reddy, M. An Evaluation of Level of Detail Degradation in Head-mounted Display Peripheries. *Presence Teleoper. Virtual Environ.* **1997**, *6*, 630–637. [CrossRef]
53. Takeshita, H.; Kihara, K.; Yoshida, S.; Higuchi, S.; ku Nagoya-shi Ito, M.M.; Nakanishi, Y.; Kijima, T.; Ishioka, J.; Matsuoka, Y.; Numao, N.; et al. Clinical application of a modern high-definition head-mounted display in sonography. *J. Ultrasound Med.* **2014**, *33*, 1499–504. [CrossRef]
54. Prendergast, C.J.; Ryder, B.; Abodeely, A.; Muratore, C.; Crawford, G.P.; Luks, F. Surgical Performance with Head-Mounted Displays in Laparoscopic Surgery. *J. Laparoendosc. Adv. Surg. Tech. Part A* **2008**, *19* (Suppl. 1), S237–S240. doi:10.1089/lap.2008.0142. [CrossRef]
55. Massimino, M.J.; Sheridan, T.B. Teleoperator performance with varying force and visual feedback. *Hum. Factors* **1994**, *36*, 145–157. [CrossRef] [PubMed]
56. Chen, D.J.Y.C.; Durlach, P.J.; Sloan, J.A.; Bowens, L.D. Human–Robot Interaction in the Context of Simulated Route Reconnaissance Missions. *Mil. Psychol.* **2008**, *20*, 135–149. [CrossRef]
57. Ware, C.; Balakrishnan, R. Reaching for objects in VR displays: lag and frame rate. *ACM Trans. Comput. Hum. Interact.* **1994**, *1*, 331–356. [CrossRef]
58. Meehan, M.; Insko, B.; Whitton, M.; Brooks, F., Jr. Physiological Measures of Presence in Stressful Virtual Environments. *ACM Trans. Graph.* **2002**, *21*. [CrossRef]
59. Claypool, K.T.; Claypool, M. The effects of resolution on users playing first person shooter games. In *Multimedia Computing and Networking*; SPIE: San Jose, CA, USA, 2007.
60. Chen, J.Y.C.; Thropp, J.E. Review of Low Frame Rate Effects on Human Performance. *IEEE Trans. Syst. Man Cybern.-Part A Syst. Hum.* **2007**, *37*, 1063–1076. [CrossRef]
61. Rift, O. PC SDK Developer Guide, Guidelines for VR Performance Optimization. Available online: https://developer.oculus.com/documentation/native/pc/dg-performance-guidelines/ (accessed on 28 August 2020).
62. Sanchez-Vives, M.; Slater, M. From presence to consciousness through virtual reality. *Nat. Rev. Neurosci.* **2005**, *6*, 332–339. [CrossRef]
63. Lester, D.; Thronson, H. Human space exploration and human spaceflight: Latency and the cognitive scale of the universe. *Space Policy* **2011**, *27*, 89–93. [CrossRef]
64. Meehan, M.; Razzaque, S.; Whitton, M.C.; Brooks, F.P. Effect of latency on presence in stressful virtual environments. In *IEEE Virtual Reality 2003, Proceedings*; IEEE: Los Angeles, CA, USA, 2003; pp. 141–148. [CrossRef]
65. Ellis, S.R.; Mania, K.; Adelstein, B.D.; Hill, M.I. Generalizeability of latency detection in a variety of virtual environments. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*; SAGE Publications Sage: Los Angeles, CA, USA, 2004; Volume 48, pp. 2632–2636.
66. Vallines, I.; Greenlee, M.W. Saccadic Suppression of Retinotopically Localized Blood Oxygen Level-Dependent Responses in Human Primary Visual Area V1. *J. Neurosci.* **2006**, *26*, 5965–5969. [CrossRef] [PubMed]
67. Albert, R.; Patney, A.; Luebke, D.; Kim, J. Latency Requirements for Foveated Rendering in Virtual Reality. *ACM Trans. Appl. Percept.* **2017**, *14*, 25:1–25:13. [CrossRef]
68. Sun, Q.; Patney, A.; Wei, L.Y.; Shapira, O.; Lu, J.; Asente, P.; Zhu, S.; Mcguire, M.; Luebke, D.; Kaufman, A. Towards Virtual Reality Infinite Walking: Dynamic Saccadic Redirection. *ACM Trans. Graph.* **2018**, *37*, 67:1–67:13. [CrossRef]
69. Bailey, R.E.; Arthur III, J.J.; Williams, S.P. Latency requirements for head-worn display S/EVS applications. In *Enhanced and Synthetic Vision 2004*; International Society for Optics and Photonics: Bellingham, WA, USA, 2004; Volume 5424, pp. 98–109.
70. Rift Oculus. PC SDK Developer Guide. Available online: https://developer.oculus.com/design/bp-rendering/ (accessed on 28 August 2020).
71. Rift Oculus. Oculus Rift CV1. Available online: https://www.oculus.com/rift (accessed on 28 August 2020).
72. HTC VIVE. Available online: https://www.vive.com/eu/ (accessed on 28 August 2020).

73. Lane, J.C.; Carignan, C.R.; Sullivan, B.R.; Akin, D.L.; Hunt, T.; Cohen, R. Effects of time delay on telerobotic control of neutral buoyancy vehicles. In Proceedings of the 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292), Washington, DC, USA, 11–15 May 2002; Volume 3, pp. 2874–2879. [CrossRef]
74. Kumcu, A.; Vermeulen, L.; Elprama, S.; Duysburgh, P.; Platisa, L.; Nieuwenhove, Y.; Van De Winkel, N.; Jacobs, A.; Looy, J.; Philips, W. Effect of video lag on laparoscopic surgery: correlation between performance and usability at low latencies. *Int. J. Med. Robot. Comput. Assist. Surg.* **2016**, *13*, e1758. [CrossRef] [PubMed]
75. Lum, M.J.H.; Rosen, J.; Lendvay, T.S.; Sinanan, M.N.; Hannaford, B. Effect of time delay on telesurgical performance. In Proceedings of the 2009 IEEE International Conference on Robotics and Automation, Kobe, Japan, 12–17 May 2009; pp. 4246–4252. [CrossRef]
76. Frank, L.H.; Casali, J.G.; Wierwille, W.W. Effects of Visual Display and Motion System Delays on Operator Performance and Uneasiness in a Driving Simulator. *Proc. Hum. Factors Soc. Annu. Meet.* **1987**, *31*, 492–496. [CrossRef]
77. Gnatzig, S.; Chucholowski, F.; Tang, T.; Lienkamp, M. A System Design for Teleoperated Road Vehicles. In *ICINCO (2)*; Ferrier, J.L., Gusikhin, O.Y., Madani, K., Sasiadek, J.Z., Eds.; SciTePress: Reykjavík, Iceland, 2013; pp. 231–238.
78. Hosseini, A.; Lienkamp, M. Enhancing telepresence during the teleoperation of road vehicles using HMD-based mixed reality. In Proceedings of the 2016 IEEE Intelligent Vehicles Symposium (IV), Gotenburg, Sweden, 19–22 June 2016; pp. 1366–1373. [CrossRef]
79. Monteiro, F.; Rocha, P.; Menezes, P.; Silva, A.; Dias, J. Teleoperating a mobile robot. A solution based on JAVA language. In Proceedings of the ISIE '97 Proceeding of the IEEE International Symposium on Industrial Electronics, Guimaraes, Portugal, 7–11 July 1997; Volume 1, pp. SS263–SS267. [CrossRef]
80. Aykut, T.; Zou, C.; Xu, J.; Van Opdenbosch, D.; Steinbach, E. A Delay Compensation Approach for Pan-Tilt-Unit-based Stereoscopic 360 Degree Telepresence Systems Using Head Motion Prediction. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 3323–3330. [CrossRef]
81. Oving, A.B.; van Erp, J.B.F. Driving with a Head-Slaved Camera System. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **2001**, *45*, 1372–1376. [CrossRef]
82. Pazuchanics, S.L. The Effects of Camera Perspective and Field of View on Performance in Teleoperated Navigation. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **2006**, *50*, 1528–1532. [CrossRef]
83. Arthur, K.W. Effects of Field of View on Performance with Head-Mounted Displays. Ph.D. Thesis, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, 2000.
84. Scribner, D.R.; Gombash, J.W. *The Effect of Stereoscopic and Wide Field of View Conditions on Teleoperator Performance*; Technical Report; Army Research Laboratory: Adelphi, MD, USA, 1998.
85. Smyth, C.; Gombash, J.W.; Burcham, P.M. *Indirect Vision Driving with Fixed Flat Panel Displays for Near-Unity, Wide, and Extended Fields of Camera View*; Technical Report; Army Research Laboratory: Adelphi, MD, USA, 2001.
86. Woods, D.D.; Tittle, J.; Feil, M.; Roesler, A. Envisioning human-robot coordination in future operations. *IEEE Trans. Syst. Man Cybern. Part C* **2004**, *34*, 210–218. [CrossRef]
87. Darken, R.P.; Kempster, K.; Kempster, M.K.; Peterson, B. Effects of Streaming Video Quality of Service on Spatial Comprehension in a Reconnaisance Task. In Proceedings of the Meeting of I/ITSEC, Orlando, FL, USA, 30 November–4 December 2001.
88. Witmer, B.G.; Sadowski, W.J. Nonvisually Guided Locomotion to a Previously Viewed Target in Real and Virtual Environments. *Hum. Factors* **1998**, *40*, 478–488. [CrossRef]
89. Livingstone, M.; Hubel, D. Segregation of form, color, movement, and depth: Anatomy, physiology, and perception. *Science* **1988**, *4853*, 740–749. [CrossRef]
90. Breedveld, P.; Stassen, H.; Meijer, D.; Stassen, L. Theoretical background and conceptual solution for depth perception and eye-hand coordination problems in laparoscopic surgery. *Minim. Invasive Ther. Allied Technol.* **2009**, *8*, 227–234. [CrossRef]
91. Bogdanova, R.; Boulanger, P.; Zheng, B. Depth Perception of Surgeons in Minimally Invasive Surgery. *Surg. Innov.* **2016**, *23*, 515–524. [CrossRef]

92. Avgousti, S.; Christoforou, E.; Panayides, A.; Voskarides, S.; Novales, C.; Nouaille, L.; Pattichis, C.; Vieyres, P. Medical telerobotic systems: Current status and future trends. *Biomed. Eng. Online* **2016**, *15*, 1–44. [CrossRef] [PubMed]
93. Masia, L.; Casadio, M.; Sandini, G.; Morasso, P. Eye-Hand Coordination during Dynamic Visuomotor Rotations. *PLoS ONE* **2009**, *4*, 1–11. [CrossRef] [PubMed]
94. DeJong, B.; Colgate, J.; Peshkin, M. Mental transformations in human-robot interaction. In *Mixed Reality and Human-Robot Interaction*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 35–51.
95. Macedo, J.A.; Kaber, D.B.; Endsley, M.R.; Powanusorn, P.; Myung, S. The Effect of Automated Compensation for Incongruent Axes on Teleoperator Performance. *Hum. Factors* **1998**, *40*, 541–553. [CrossRef]
96. Tittle, J.S.; Woods, D.D.; Roesler, A.; Howard, M.; Phillips, F. The Role of 2-D and 3-D Task Performance In the Design and Use of Visual Displays. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **2001**, *45*, 331–335. [CrossRef]
97. DeLucia, P.R.; Griswold, J.A. Effects of camera arrangement on perceptual-motor performance in minimally invasive surgery. *J. Exp. Psychol. Appl.* **2011**, *17*, 210–232. [CrossRef]
98. Almeida, L.; Menezes, P.; Dias, J. Improving robot teleoperation experience via immersive interfaces. In Proceedings of the 2017 4th Experiment@International Conference (exp.at'17), Faro, Portugal, 6–8 June 2017; pp. 87–92. [CrossRef]
99. Baddeley, A. Working memory. *Curr. Biol.* **2010**, *20*, R136–R140. [CrossRef]
100. Hanley, J.R.; Thomas, A. Maintenance rehearsal and the articulatory loop. *Br. J. Psychol.* **1984**, *75*, 521–527. [CrossRef]
101. Kessels, R.P.C.; van Zandvoort, M.J.E.; Postma, A.; Kappelle, L.J.; de Haan, E.H.F. The Corsi Block-Tapping Task: Standardization and Normative Data. *Appl. Neuropsychol.* **2000**, *7*, 252–258. [CrossRef]
102. Miller, G.A. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychol. Rev.* **1956**, *63*, 81–97. [CrossRef] [PubMed]
103. Lewis, J.R. *IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use*; Technical Report; IBM—Human Factors Group: Boca Raton, FL, USA, 1993.
104. Slater, M.; Usoh, M.; Steed, A. Depth of presence in virtual environments. *Presence-Teleoper. Virtual Environ.* **1994**, *3*, 130–144. [CrossRef]
105. Usoh, M.; Catena, E.; Arman, S.; Slater, M. Using Presence Questionnaires in Reality. *Presence Teleoper. Virtual Environ.* **2000**, *9*, 497–503. [CrossRef]
106. Watson, B.; Walker, N.; Woytiuk, P.; Ribarsky, W. Maintaining usability during 3D placement despite delay. In Proceedings of the IEEE Virtual Reality, Los Angeles, CA, USA, 22–26 March 2003; pp. 133–140. [CrossRef]
107. Watson, B.; Walker, N.; Ribarsky, W.; Johnson, V.A.S. Effects of Variation in System Responsiveness on User Performance in Virtual Environments. *Hum. Factors* **1998**, *40*, 403–414. [CrossRef]
108. Ellis, S.R.; Adelstein, B.D.; Baumeler, S.; Jense, G.; Jacoby, R.H. Sensor spatial distortion, visual latency, and update rate effects on 3D tracking in virtual environments. In Proceedings of the IEEE Virtual Reality (Cat. No. 99CB36316), Houston, TX, USA, 13–17 March 1999; pp. 218–221.
109. Lion, D.M. *Three Dimensional Manual Tracking Using A Head-Tracked Stereoscopic Display (Technical Report);* Human Interface Technology Lab.: Seattle, WA, USA, 1993.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).