

Received April 2, 2020, accepted April 13, 2020, date of publication April 23, 2020, date of current version May 8, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2990080

# Study of Multimodal Interfaces and the Improvements on Teleoperation

ELEFThERIOS TRIANTAFYLLIDIS<sup>1</sup>, CHRISTOPHER MCGREAVY, JIACHENG GU,  
AND ZHIBIN LI<sup>1</sup>, (Member, IEEE)

School of Informatics, The University of Edinburgh, Edinburgh EH8 9YL, U.K.

Corresponding author: Zhibin Li (zhibin.li@ed.ac.uk)

This work was supported by the EPSRC Future AI and Robotics for Space under Grant EP/R026092/1.

**ABSTRACT** Research in multimodal interfaces aims to provide immersive solutions and to increase overall human performance. A promising direction is to combine auditory, visual and haptic interaction between the user and the simulated environment. However, no extensive comparison exists to show how combining audiovisuohaptic interfaces would affect human perception and by extent reflected on task performance. Our paper explores this idea and presents a thorough, full-factorial comparison of how all combinations of audio, visual and haptic interfaces affect performance during manipulation. We evaluated how each combination affects the performance in a study ( $N = 25$ ) consisting of manipulation tasks with various difficulties. The overall performance was assessed using both subjective measures, by assessing cognitive workload and system usability, and objective measurements, by incorporating time and spatial accuracy-based metrics. The results showed that regardless of task complexity, the combination of stereoscopic-vision with the virtual reality headset increased performance across all measurements by 40%, compared to monocular-vision from a generic display monitor. Besides, using haptic feedback improved outcomes by 10% and auditory feedback accounted for approximately 5% improvement.

**INDEX TERMS** Audiovisuohaptic, auditory feedback, haptic feedback, immersive manipulation, immersive teleoperation, multimodal interface, multimodal interaction, virtual reality.

## I. INTRODUCTION

The growth of virtual reality, robotics and telecommunication technologies have spiked in recent years. This has led to an increase in teleoperation research – allowing humans the ability to remotely inhabit a foreign body, e.g. a robot as an avatar to complete a task [1]. With the recent outbreak of pandemics, remote robotic control and telepresence systems, have become more important than ever.

Teleoperation delegates the high-level control of a robot to a remote human operator, thus combining the human instinct as well as the computational and physical capabilities of robots. Humans are highly adaptable experts in motor control, constituting teleoperation a useful tool to help robots complete tasks in novel and dynamic environments. During a teleoperation task, the robot's performance is dictated by the controls being sent by the human. So how can we maximize human perception and by extent performance during task

supervision? The actions between an operator and a remote robotic system are physically detached from another, constituting the overall experience unnatural. This implies that policies which humans usually use to control their own bodies may not directly translate into effective control of a foreign body, which can lead to poor performance. To mitigate this, we can maximise feelings of immersion and by extent task performance in humans so that the foreign body feels more like their own. This can lead to improved performance when controlling another body in a remote environment [2], [3], and increasing immersion is known to increase the performance [4]–[6].

To increase the feeling of immersion and thus the performance, we can alter the way in which the human interacts with their avatar. In a primary setting, users can interact with their surrounding virtual setting, by using a monocular monitor to provide them with a visual representation of the environment, which may not necessarily lead to higher levels of immersion by itself. Using a virtual reality device could lead to increased performance because it offers richer visual

The associate editor coordinating the review of this manuscript and approving it for publication was Luigi De Russis<sup>1</sup>.

information, particularly attributed to stereoscopic depth [7], [8]. However, stimulating other senses may also affect performance, for example, using auditory and haptic feedback to superimpose information.

Previous work has compared the effect of combining some sensory interfaces. However, to the best of our knowledge, exhaustive comparisons have yet not been studied between visual, haptic and auditory sensory modalities, and how their combinations affect task performance of varying complexity. Our work aims to address this gap in the literature.

We use a pick and place task to compare the effects of these sensory interfaces on task performance. The setup for this task can be seen in Figure 4. The pick and place task is set in a virtual environment with different objects types, sizes and pick and place distances. We compare all combinations of visual (monocular or VR), auditory (presence or absence) and haptic (presence or absence) feedback. Changing these factors affects the difficulty of the task and we present a detailed analysis on how each combination of sensory inputs affects task performance.

Our study provides evidence to support a recommendation for the best performing combination of the sensory interface in manipulation tasks with varying complexity. By incorporating both subjective and objective measurements, we determine which combination offers the best performance for a given task. Throughout this paper, we present how we conduct, evaluate and analyse our experiments.

Contributions of our work include:

- A unique and reproducible interface which allows various combinations of sensory feedback for performing various tasks under different settings;
- A low-cost hardware and simple software approach in designing an effective vibrotactile haptic data glove;
- A virtual reality environment with high-fidelity physics simulation (friction, collision, contact forces) to closely resemble real-world interaction and make the best use of existing human motor skills;
- A concrete experimental design that can be used to test the effectiveness of new emerging technologies;
- To the best of our knowledge, this is the first exhaustive comparison of its kind between all combinations of visual, auditory and haptic interfaces for manipulation tasks of increasing difficulty.

## II. RELATED WORK

In this section, we discuss the previous work regarding the effectiveness of multisensory interfaces on immersion and performance, object interaction and manipulation. We group these studies in separate sensory modalities for clarity and identify gaps in current knowledge.

### A. MULTIMODAL INTERFACES

When operators embody a remote robot or are subjected to a virtual environment for training purposes, using only a visual monocular monitor, they can only experience that remote environment visually. By adding multiple modalities,

it was found that the workload of the visual cortex can be reduced, the awareness may be increased and thus the task performance can be improved [9], [10]. But when using multimodal interfaces, synchronisation is important. Otherwise, if signals of different modalities are out-of-synchronisation, overall spatial and temporal immersion is reduced, effectively nullifying the benefits of using multimodality [11], [12].

Furthermore, sensory feedback strategies need to be made prior to the implementation of a specific sensory channel. In most cases, the design decisions of one type of sensory feedback may be achieved via either a continuous manner, i.e. concurrent feedback, or after a desired event, i.e. terminal feedback [2].

This study focuses on audiovisuohaptic interfaces, since vision, hearing and touch are the highest developed and contributing the most towards embodiment [3], [11], [13] among all human senses. We present the previous work on visual, audio and haptic interfaces in the following sections.

#### 1) VISUAL CUES

Most research in this area has focused on the effect of visual interfaces between the human and the avatar. The dominance of vision in the sensory system is well supported [14]–[16], contributing to around 70% of overall human perception [13]. Thus, providing visual information in the best form is of vital importance.

The two primary sources of a visual interface include standard monocular monitors and virtual reality head-mounted displays (VRHMDs) that provide stereo vision. During the specific study of a target detection task in Unmanned Aerial Vehicles (UAVs), there were no significant differences in performance between the two [17], with the VRHMD even causing motion sickness potentially attributed to the illusion of self-motion of the vehicle. This is known as vection [18], which is a common complaint among VRHMD users in non-static situations. This is still an open problem, therefore, our study limits self-motion and only compares the effectiveness of both displays in static scenarios without the presence of navigation.

Though VRHMDs have drawbacks in some settings, they do offer many benefits over monocular screens. They offer better depth perception and environmental awareness than standard monitors [19]. This is of importance as studies have shown that humans overestimate their ability to perceive depth in virtual environments [8], [20], [21]. As such, increased depth information leads to reduced collisions with the surrounding environment and better performance during highly dexterous manipulation tasks [7], [22]. It is important, however, how the superimposition of information is delivered to the operator. One study showed that constantly providing feedback can be counterproductive both in user preference and time efficiency, compared to the feedback at the end of a task [23]. Providing a larger field of view can also result in increased performance and environmental awareness [22], [24], [25], but can decrease usability and increase perceived difficulty and workload demand [25].

## 2) AUDITORY CUES

Supplementing vision with auditory information can increase operator's awareness, especially during high visual load [2], [26]. Reducing as such overall mental workload is correlated to fewer accidents and better performance [27]. Audiovisual interfaces also improved intuitive control of a humanoid during manipulation tasks [28].

Though extra information, such as alarms and alerts, can be superimposed on a visual display, presenting them via an audio interface is a better approach that can decrease distraction [29]. Operators can also use auditory information to localise the sources of sounds, which is useful when FOV is limited [30]. Further studies in human walking also show that controlling auditory pitch may influence object clearance, with results indicating that participants indeed benefited from such sound sonification [31], [32]. This suggests that auditory information may provide a richer environmental experience and may be a valuable supplement to just relying on vision.

## 3) SOMATOSENSORY CUES

Tactile feedback can also augment visual information. Communicating spatial alerts via somatosensory stimulation can signal warnings without overloading visual pathways [33], [34]. In particular, manipulation can be improved by adding tactile feedback [1], [35], [36] that can result in better performance [37].

For diagnostic surgery simulators using virtual reality, complex and sophisticated tactile approaches of force feedback have been developed to reflect realistic reaction forces of deformable objects such as soft tissue [38]. Further research in kinesthetic force feedback has shown some advantages over lower-cost approaches [39], [40], particularly due to the ability to constrain the grasp motion of users hands, based on the virtual object they are holding [41]. However, providing high-resolution haptic feedback alone does not necessarily guarantee an increase in task performance [42]. Using only vibration feedback can increase spatial awareness for rigid, non-deformable objects [43]. Outputting vibrations proportional to the force applied by the robot also leads to improved performance [44], and therefore we used a similar approach.

## 4) AUDIOVISUOHAPTIC MULTIMODAL INTERFACES

A combination of all three modalities may also be effective in improving performance. One study hypothesises that audio-visuohaptic interfaces may increase task performance as the task gets gradually more difficult [2], but this is untested.

On one hand, an audiovisuohaptic interface did not significantly increase performance during a teleoperated navigation task [9], but operator's spatial ability and subjective performance did increase compared to using fewer interfaces. In another study, an audiovisuohaptic interface was implemented to assess performance in visual throwing tasks [45]. While not exhausting all comparisons of the interface or implementing varying task complexities, their results show

that point-based haptic devices and moreover auditory feedback did not significantly contribute to the improvement of task performance.

A meta-analysis of 45 studies showed that by supplementing visual information with either auditory or somatosensory (via vibrotactile cues) increased overall performance [46]. However, no extensive comparison has been conducted on how combining all three modalities affects immersion and by extent task performance reflected on higher levels of complexity.

## B. OBJECT INTERACTION AND MANIPULATION

To compare the effect of visual, auditory and haptic feedback on task performance, we must first define a task. We chose to measure the effect of these interfaces on manipulation tasks of different difficulties. Manipulation is a suitable choice, since it involves coarse and fine motor movements, depending on the object being grasped. The Southampton Hand Assessment Procedure (SHAP) [47], defines six clinically validated grasping classifications to test hand function. This comprises the entire range of human hand motion from fine to coarse manipulation. One study even addressed all the possible different grasping techniques a human can initiate with an object by implementing the SHAP in the physics engine MuJoCo. However, no comparison between the sensory modalities was drawn [48]. We were inspired by the aforementioned study and used a range of different objects and sizes during our experiments. By doing this we can examine the effect of combining sensory interfaces on the performance of different levels of human motor skills during object manipulation and interaction.

Our aim is to increase the task performance by improving the overall immersion. However, immersion is a complex phenomenon which can be negatively influenced by the so-called "Uncanny Valley" – a break in immersion when an artificial being appears *too* realistic, causing negative responses towards it [49]. More relevant to this study is the "Uncanny Valley of Haptics", which has a similar effect when haptic feedback does not coincide with other sensory feedback and reduces the perception of realism [42]. Neuroimaging studies support this concept, showing that visual and haptic activation overlaps in the occipital lobe [50]–[53]. We aim to investigate if the simultaneous presence of both modalities increases the performance.

## III. HYPOTHESES

The following hypotheses are formed from our review, while primarily hypothesizing that an audiovisuohaptic multimodal interface will prove to be significantly more effective when subjected to higher task complexity, compared to fewer modalities or the minimal representation of these.

*Hypothesis 1:* There will be lower perceived cognitive workload corresponding to higher performance with (a) the stereoscopic VRHMD than with the monocular display monitor, (b) the presence of somatosensory feedback than its

**TABLE 1.** The multimodal interface broken down into the  $2^3$  possible combinations of visual, auditory and haptic feedback.

	Vision		Audition		Haptics	
	Monitor	VRHMD	Absence	Presence	Absence	Presence
C1	X		X		X	
C2	X		X			X
C3	X			X	X	
C4	X			X		X
C5		X	X		X	
C6		X	X			X
C7		X		X	X	
C8		X		X		X

absence, and finally (c) the presence of auditory feedback than the absence of it.

*Hypothesis 2:* There will be higher perceived system usability corresponding to higher performance with (a) the stereoscopic VRHMD than with the monocular display monitor, (b) the presence of somatosensory feedback than its absence, and finally (c) the presence of auditory feedback than the absence of it.

*Hypothesis 3:* Faster performance corresponding to less placement and completion time will be observed with (a) the stereoscopic VRHMD than with the monocular display monitor, (b) the presence of somatosensory feedback than its absence, and finally (c) presence of auditory feedback than the absence of it.

*Hypothesis 4:* Better depth estimation with less distance error to target, will be measured in the order of interface conditions incorporating (a) the stereoscopic VRHMD than with the monocular display monitor, (b) the presence of somatosensory feedback than its absence, and finally (c) the presence of auditory feedback than absence.

*Hypothesis 5:* Higher placement precision, including higher spatial position and orientation accuracy, will be measured in the order of interface conditions incorporating (a) the stereoscopic VRHMD than with the monocular display monitor, (b) the presence of somatosensory feedback than its absence, and finally (c) the presence of auditory feedback than the absence of it.

#### IV. METHODOLOGY

This section describes the key hardware and software components in our study. First, to test our hypotheses, we designed a series of experiments, during which the participants performed a pick and place task under various conditions. All possible combinations of a visual, auditory and haptic interface are assessed. Each modality has two modes, as detailed in Table 1, providing a full factorial study.

We studied the presence and absence of audition and haptics, whereas for vision, we compared monocular view (the display monitor) with stereoscopic view (the VRHMD). All combinations of these modalities amount to  $2^3$  combinations. Assessment of performance is achieved via both objective and subjective metrics. Participants completed manipulation tasks under each of the above conditions.

#### A. PARTICIPANTS

A total of ( $N = 25$ ) participants were recruited in this study via an advertisement at the University of Edinburgh. Ages ranged from 21 to 44 ( $M = 26.36$ ,  $SD = 4.829$ ,  $Mdn = 26$ ), with 6 females and 19 males. Each had healthy hand control, normal hearing ability and normal/corrected vision. A 30-minute interactive experience using the VRHMD was given as compensation.

#### B. EQUIPMENT AND SYSTEM SETUP

For visual feedback, a computer monitor was used for the monocular condition and a VRHMD for stereoscopic vision. The monitor was a 27 inch HP Elite IPS display, with  $2560 \times 1440$  resolution and 60Hz refresh rate placed 75-100cm from the participant. The VRHMD was HTC Vive Pro with 3.5 inch AMOLED screen at  $2880 \times 1600$  resolution ( $1440 \times 1600$  pixels per eye), 90Hz refresh rate and  $110^\circ$  FOV. High-resolution displays were chosen to limit distance overestimation and a degraded longitudinal control [54]. An NVIDIA 2080 Ti was used to ensure consistent frame-rates.

Two stereo headsets provided the audio interface. One was integrated onto the HTC Vive VRHMD for stereo conditions. The other was separately attached during the display monitor monocular conditions. Audio quality was at 16 bit, 44100 Hz.

To provide haptic feedback, we constructed a custom haptic capable glove inspired by [55], which incorporated a vibration motor on the thumb and index finger of the glove. The vibration intensity in their study was accomplished and influenced by the proportion of the size of the virtual object the user was colliding and touching with. While their approach indeed shows a promising step towards immersive experiences in the branch of entertainment, we took their method a step forward by incorporating physical properties including kinetic energy and object penetration for manipulation scenarios detailed further along this paper, specifically in the methodology section. In the construction of our custom glove, 15 coin-vibration motors were used, with DC 3V 70mA 12000 RPM. Two motors were placed on each finger (proximal & distal phalanges). Five motors were placed on the palm. Wireless communication between the virtual environment and the glove ensured free movement. This was achieved using a Bluetooth transceiver for each glove.

We chose to use vibrotactile stimulation rather than force feedback for its lower cost and certain advantages over force feedback. Preliminary findings indicate that force feedback is only more beneficial than vibrotactile stimulation when presented at a high-resolution [42], [56]. However, this increases cost and size. Air jet-driven approaches exist for force feedback, while these show significant effectiveness, they nonetheless require large space and pose a substantially higher cost compared to vibration approaches [57]. Vibratory feedback, however, can be more beneficial than force feedback in direct manipulation tasks such as ours [58]. Overall vibrotactile stimulation is shown to be an effective substitute for force feedback according to another study [59].

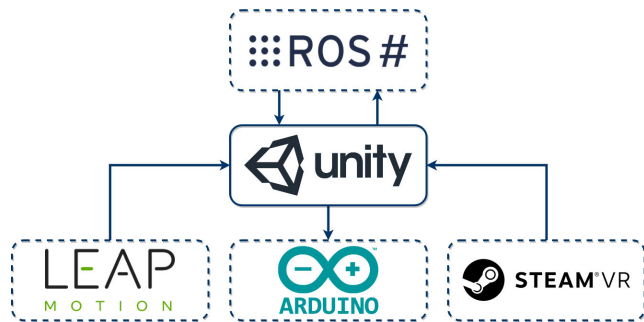


FIGURE 1. Diagram of the simulation setup with all the software plugins used.

The manipulation task for this study was performed by mapping the user's hand movements to an anthropologically human-robot hand in the simulation environment. To capture hand movements, we used the Leap Motion Hand Controller (LMHC). This uses a stereo camera system and infrared LEDs to capture hand motions. In all conditions, the device was fixed to the participant's forehead, either by a strap or on the front of the VRHMD. The LMHC was able to track the haptic gloves, as anthropomorphic features were retained.

### C. SOFTWARE AND SIMULATION ENVIRONMENT SETUP

In our experiments, the participant's conducted manipulation tasks in a virtual environment. As such, this study required a virtual environment which was connected to the hardware. The relationship of these components is shown in Figure 1.

The Unity3D engine was used as the core of our virtual environment. Two Shadow robotic hands acted as teleoperated manipulators. Physics simulations of the environment used the Unity3D engine, whereas robotic hand physics were handled by the ROS-Sharp physics engine. Unity obtained hand positions from the LMHC via the Leap Motion SDK. A plugin was developed to communicate between the Unity environment and haptic gloves via a Bluetooth module on the glove's Arduino controllers.

### D. HAND MANIPULATION AND CONTROL

The Leap Motion SDK outputs Cartesian joint positions in world frame, but joint angles are required to control the virtual hand. This translation was made by calculating the angle  $\theta$  between a joint  $\vec{b}_{i-1}$  and its parent  $\vec{b}_i$ :

$$\theta = \arccos \left( \frac{\vec{b}_i \cdot \vec{b}_{i-1}}{\|\vec{b}_i\| \|\vec{b}_{i-1}\|} \right). \quad (1)$$

A Proportional Derivative (PD) controller was used to control the joints. Each joint has one PD controller, formulated as follows for each timestep  $t$ :

$$u(t) = K_P \cdot e(t) + K_D \cdot \dot{e}(t), \quad (2)$$

where  $u(t)$  is the angular velocity control signal sent to the Shadow hand joints.  $e(t) = q_h(t) - q_r(t)$  is the current position error between the human joint and the robot joint and  $\dot{e}(t) = \dot{q}_{ref} - \dot{q}_r(t)$  is the velocity error between the robot and

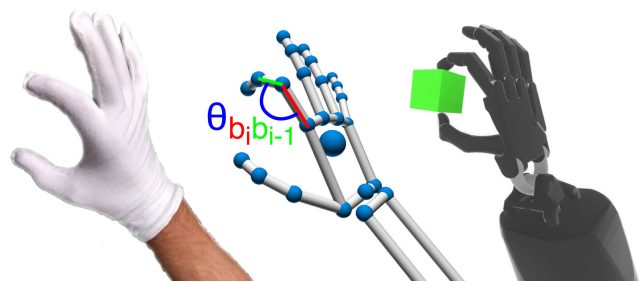


FIGURE 2. Hand control approach through direct joint angle re-targeting from our custom haptic glove to the final robotic hand.

the desired velocity, which here is set to zero.  $K_P$  and  $K_D$  are the gains which were tuned such that human and robot motion matched as accurately as possible. Depicted in Figure 2.

### E. SENSORY INTERFACE DESIGN

#### 1) VISUAL STIMULATION

We compare monocular and stereo feedback in our experiments using a generic display monitor and a VRMHD respectively. In addition to the visual disparity, the VRHMD also allows users to control the viewpoint in the virtual environment by moving their head. To conduct a fair experiment, we allow participants to change their viewpoint when using the monitor by using a computer keyboard using standard gaming keybindings, retain the optical hand controller consistently in a head-mounted state as well as using a monitor of similar resolution to the VRHMD. Acclimatization to these controls and technologies were allowed prior to commencing the experiments, detailed further along this work.

#### 2) AUDITORY STIMULATION

We hypothesize that auditory feedback will contribute to increased performance. Everyday sound effects "that make sense" were used to investigate how sound may compensate for the superimposition of visual information, without requiring prior context or explaining these to the participants, which would be inherently perceived as a substitute to text. Terminal auditory feedback is given in our case. More specifically, audio feedback is given in two situations.

Firstly, warnings and notifications were given via audio. A high-pitched alarm sound warned of imminent collisions between the robotic hands and the environment. A siren alarm sound, on the other hand, indicated time was running low. A successful "ding" indicated that at least part of an object had been placed inside the target volume irrespective of the placement accuracy.

Auditory feedback also relayed the sounds of interactions in the environment. Picking up, dropping or placing an object produced realistic bump and scrape sounds that one would expect when interacting with real objects.

#### 3) SOMATOSENSORY STIMULATION

Vibration feedback is applied to the gloves of the participants when the robotic arms collide with the environment. Here we describe how the vibration intensity is determined.

We are inspired by a similar study using appropriate “collision” signals to transmit variable frequency tactile feedback [1]. In a more previous study investigating vibrotactile approach, vibration intensity applied to users was proportional to the size of the virtual object being manipulated [55]. We adopt this approach where instead, vibration intensity is proportional to kinetic energy  $K_E$  and object penetration  $P$  of each finger segment in simulation. These are then combined to give the final intensity.

Kinetic energy  $K_E$  of the virtual collision is formulated as:

$$K_E = \frac{1}{2} \cdot m \cdot v^2, \quad (3)$$

where  $m$  is the body mass and  $v$  the velocity between the robot segment and the environment.

We use the relative penetration of  $P$  between the robot and the environment as a proxy for force. Since we operate in a simulation, we have access to the full state space of the environment. Penetration can then easily be defined by the relative distance between the robot segment  $v_r$  and virtual object  $v_o$  and the distance between the centre and surface of the object  $s_o$  as shown in Equation 4,

$$P = \left| 1 - \frac{\|v_r - v_o\|}{\frac{1}{2} \cdot s_o} \right|. \quad (4)$$

Equation 3 and Equation 4 can then be combined to calculate total vibration intensity shown in Equation 5,

$$V = V_{min} + a \cdot \frac{(V_{max} - V_{min})}{K_{E_{max}}} \cdot K_E + b \cdot \frac{(V_{max} - V_{min})}{P_{max}} \cdot P, \quad (5)$$

where  $V$  is the final vibration intensity transmitted to the vibration motors,  $V_{min}$  the minimum vibration intensity needed to distinguish vibrotactile stimulation when in contact. This is set to 25% based on a pilot study consisting of five participants. The second term calculates the vibration intensity based on the kinetic energy exerted and is controlled by a constant  $a$ .  $V_{max}$  is the maximum vibration intensity of the hardware,  $K_{E_{max}}$  is the maximum calculated kinetic energy in Joules with a velocity limit of 7 m/s set in the physics engine and  $K_E$  is the current kinetic energy exerted to the object. The kinetic energy is only applicable during the object acquisition and as masses are constant, it is only dependent upon the velocity of grabbing i.e. picking up. The final term calculates the vibration intensity based upon the penetration of the robotic hand with the object and is controlled by a constant  $b$ .  $P_{max}$  is the maximum penetration allowed which in our case is 100% and finally,  $P$  is the current penetration exerted to the object. Figure 3 illustrates our haptic glove in addition to its electronics, drive control board and motors exposed.

### F. MANIPULATION TASKS OF VARYING COMPLEXITY

All tasks required the participants to pick up an object from a set starting point and place it to a designated random target



**FIGURE 3.** Haptic glove (left module shown) in its final and first iteration with its electronics and motors exposed in the latter.

location illustrated with a slight-transparent shape. We integrated three basic types of three-dimensional object shapes to not only introduce the inherent different complexities that come with such objects, but also allowing us to assess different grasping techniques [47], [48]. While different shapes do indeed vary the task complexity, we also introduced different object sizes as well as placement distances.

#### 1) TASK A - CUBE MANIPULATION

The first task included manipulating a cube shape. A cube was used, as it does not flip or roll and we can assess both its position and rotation accuracy. Grabbing techniques employed included Precision Grasping via Palmar Pinch [47].

#### 2) TASK B - CYLINDER MANIPULATION

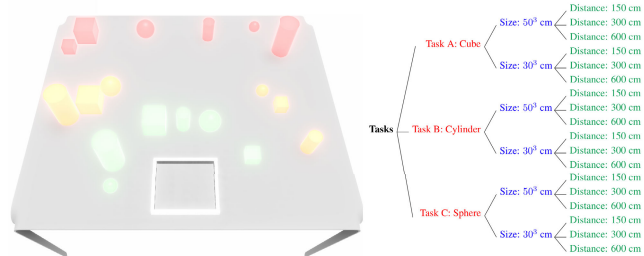
The second task included manipulating a cylinder shape. A cylinder can flip over and roll over a surface, making the task harder. We can also assess both the cylinder’s position and rotation. Grabbing techniques employed included Precision Grasping via Palmar Pinch, as well as Cylindrical Grasping, also known as Power Grasp [47].

#### 3) TASK C - SPHERE MANIPULATION

The third and final task was concerned with the manipulation of a sphere-shaped object. This was considered to be the hardest task due to the inherent ability of a sphere to roll over an even ideally horizontally placed surface if sufficient velocity would accumulate either from an inadequate precision velocity placement or release from a height offset. Grabbing techniques employed included Precision Grasping via Palmar Pinch as well as Spherical Grasping [47].

#### 4) OBJECT SCALE AND PLACEMENT DISTANCE

The aforementioned tasks are broken down into two sub-tasks assessing two object scales, large 50.0 x 50.0 x 50.0 (mm) (LxWxH) and small 30.0 x 30.0 x 30.0 (mm) (LxWxH). Furthermore, the aforementioned sub-tasks are broken down into sub-sub tasks assessing placement distances, defined as the absolute distance from the set starting point to a random



**FIGURE 4.** Image (left): All manipulation tasks illustrating the different three dimensional shapes, sizes as well as distances from 150, 300 and 600; green, yellow and red respectively. Tree (right): All 18 tasks broken down in type of object shapes (red), sizes (blue) and distances (green).

target location with distances ranging from 150.0, 300.0 and 600.0 (mm), making it progressively more difficult. A total of 144 trials were conducted per participant, stemming from our 8 interface conditions, two different object sizes, three different object shapes and distances. Across all participants, a total of 3600 trials were recorded. All of the manipulation tasks are visually depicted in Figure 4.

5) TASK PROGRESSION AND SUCCESSION

Progression to the next task is achieved when there is an intersection between the actual object and target position, regardless of the accuracy. When an overlap is achieved, the target placement slightly glows and a two-second progression timer is initiated which only pauses when the object does not retain its position. This countdown only pauses when the object is no longer colliding with the target placement volume i.e. indicating that the object has either been moved or has not remained stationary. Task progression is also achieved if the countdown timer, which has been set to 30 seconds for all tasks, reaches zero, however, in that case, the task is considered a fail rather than a success. Finally, for all tasks, an invisible collision wall was implemented to avoid objects falling out of physical bounds rendering a retrieval impossible.

V. EVALUATION

To evaluate each interface across all manipulation tasks, we implemented both subjective and objective measurements, since immersion and perception are highly subjective and our tasks are objective. We implemented both measurements to compensate for the inherent drawbacks of exclusively using questionnaires [60], [61]. Measurements are summarized in Table 2.

A. SUBJECTIVE MEASUREMENTS

We first measured cognitive workload for each interface condition through the use of the multidimensional assessment tool questionnaire NASA-Task Load Index, simply known as NASA-TLX [62]. Incorporating six sub-scales including mental demand, physical demand, temporal demand, effort, frustration, and performance.

In addition, we assessed overall system usability, through the use of the System Usability Scale questionnaire, just

**TABLE 2.** Summary of both objective and subjective measurements.

Measurement	Type	Metric
Cognitive Workload	Subjective	Questionnaire [Likert Scale]
System Usability	Subjective	Questionnaire [Likert Scale]
Task Succession	Objective	Percentage [%]
Placement Time	Objective	Seconds [s]
Completion Time	Objective	Seconds [s]
Target Error	Objective	Meters [m]
Position Accuracy	Objective	Percentage [%]
Rotation Accuracy	Objective	Percentage [%]

known as SUS [63]. Consisting of ten in total questions on a 5-point Likert scale, which range from “strongly disagree” to “strongly agree”, evaluating system complexity, consistency and cumbersome.

B. OBJECTIVE MEASUREMENTS

The overall task performance was measured by first comparing the total proportion of successful task completion, defined as placing the object to the target location within a time-countdown window of 30 seconds for each task regardless of accuracy, however, a minimum overlap with the target volume was required.

Time-based metrics were also incorporated, specifically placement and completion time to assess how fast performing each interface was. Placement time was defined as the time it took users to pick up the object and place it to the target location with potential accuracy corrections afterwards not being assessed, strictly the time stamp of the object and the target volume being in their very first collision. Completion time, on the other hand, was defined as the overall time it took users to complete successfully a task.

In addition to time, spatial-based metrics were also implemented to assess the accuracy of placing objects and how each interface may affect these, which is vital in remotely piloted systems concerned with fine manipulation. Target distance error was considered at the end of each task and defined as the distance between the center of the object and the target location, with higher values indicating worse performance. In addition, position accuracy was calculated by averaging all three axes from the euclidean space center of the object and the target (X,Y,Z) in one final percentage value. Orientation accuracy was similarly calculated but dependent upon the three-dimensional shape. For the cube and cylinder, a modulo operation of 45 and 90 degrees was performed respectively. Assessing the orientation of the sphere was disregarded due to its inherent shape.

C. PROCEDURE

Prior to commencing the experiment, participants were briefed on the purpose of the experiment, gave formal written consent and were handed out the NASA-TLX as well as the SUS questionnaires to allow acquaintance with the scales. Once users got familiar with the questionnaires, their interpupillary distance (IPD), was measured for the VRHMD

and they were allowed for 10 minutes to get acquainted with the simulation environment. During this acclimatization procedure, the participants were able to familiarize themselves with the keyboard controls and the technologies implemented in the actual experiment, but not with the actual tasks. Furthermore, due to having eight different interface conditions, we also randomized the order of these multimodal interfaces for each participant, to counterbalance potential acclimatization or task adaption.

**VI. RESULTS**

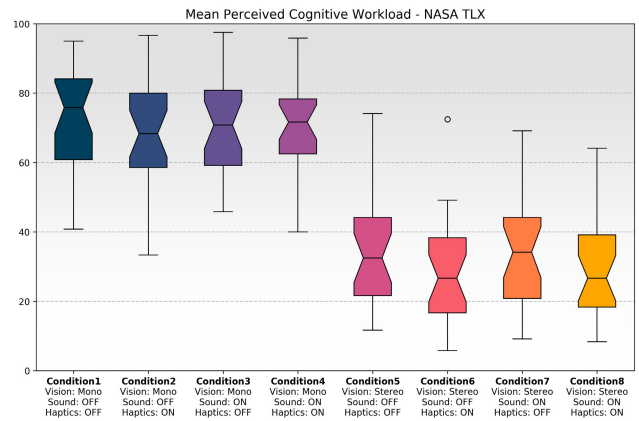
**A. ANALYSES TECHNIQUES AND METHODS**

A repeated-measures analysis of variance was used (RM-ANOVA), to analyze our parametric results that did not violate normality via Shapiro-Wilk Test. In addition, post-hoc analysis for the pairwise comparison of the eight different interface conditions was implemented. In cases where sphericity was not met, via Mauchly’s Test, a Greenhouse-Geisser correction was used to account for the violation and correct the degrees of freedom assuming a  $\epsilon < 0.75$ , otherwise a Huynh-Feld correction was used [64]. For non-parametric data, specifically for ordinal data i.e. likert scales, an Aligned-Rank Transform (ART) [65] was used to allow the use of parametric tests i.e. RM-ANOVA. For non-parametric continuous data, a Friedman’s test, similar to the RM-ANOVA, was used to test for significance across the eight interface conditions [66], and Wilcoxon signed-rank tests for post-hoc analysis for the pairwise comparison across the interface conditions. Samples that were classified as a Bernoulli distribution, the proportion of successful completion, a two times standard deviation from the mean was considered significant (95% CI) i.e. empirical rule [67]. Hereinafter, for all reported results, the significance levels are: \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$  and n.s not significant. Finally, in the Appendices, we summarize the overall results of each interface conditions across all measurements, thus giving new evidence to the hypothesized and untested effectiveness of each interface condition suggested by Sigrist et al. [2].

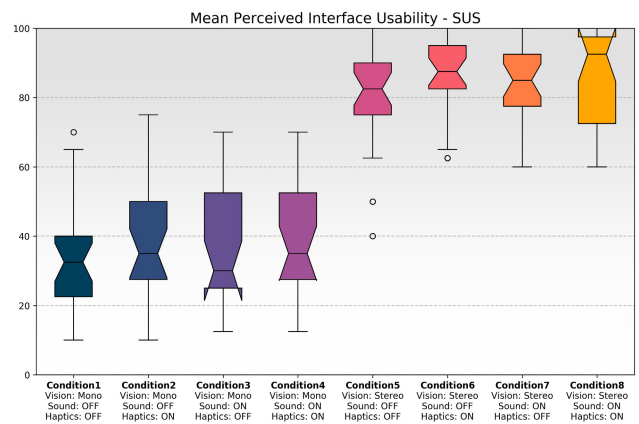
**B. SUBJECTIVE RESULTS**

**1) PERCEIVED WORKLOAD**

For the perceived workload, an ART was used to allow the use of parametric tests on ordinal data. A one way RM-ANOVA with a Greenhouse-Geisser correction ( $\epsilon = 0.342$ ) was used, yielding a highly significant difference across all eight interface conditions,  $(F(2.393, 57.437) = 70.473, p < 0.001, \eta_p^2 = 0.746)$ . Mean responses for perceived workload demand are shown in Figure 5 and Table 3. Post-hoc analysis showed partial support of hypothesis H1, specifically (a) that conditions incorporating monocular vision with the display monitor, C1,2,3 & 4, accounted to significantly higher perceived workload ( $p < 0.001$ ) than stereoscopic vision with the VRHMD, C5, C6, C7 & C8. Furthermore, (b) conditions incorporating somatosensory feedback only when paired with stereoscopic feedback. C6 & 8, showed



**FIGURE 5.** Box plot illustration across all eight interface conditions of the mean perceived workload, with higher scoring equal to worse performance. Dots represent outliers.



**FIGURE 6.** Box plot illustration across all eight interface conditions of the mean interface usability, with higher scoring equal to better performance. Dots represent outliers.

**TABLE 3.** Summary of all subjective results, reporting median and standard deviation across all eight interface conditions.

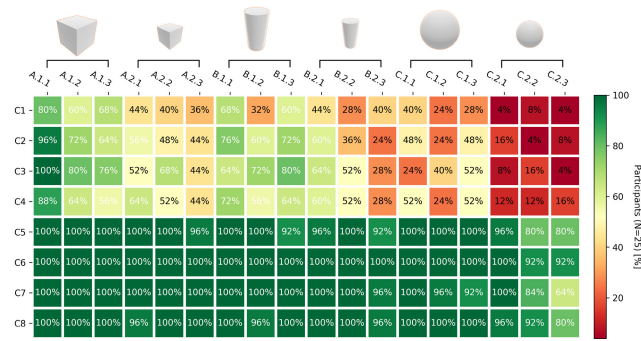
Subjective Measurements				NASA-TLX		SUS	
	Vision	Audio	Haptic	Med.	Std. D.	Med.	Std. D.
C1	Monitor	Off	Off	75.83	±13.82	32.50	±14.34
C2	Monitor	Off	On	68.33	±16.22	35.00	±17.15
C3	Monitor	On	Off	70.83	±15.00	30.00	±17.76
C4	Monitor	On	On	71.66	±14.93	35.00	±15.63
C5	VRHMD	Off	Off	32.50	±16.73	82.50	±15.82
C6	VRHMD	Off	On	26.66	±15.91	87.50	±11.33
C7	VRHMD	On	Off	34.16	±15.58	85.00	±11.38
C8	VRHMD	On	On	26.66	±14.52	92.50	±13.20

significantly lower perceived workload ( $p < 0.05$ ) than those who do not: C5 & 7 and when paired with monocular feedback, marginally lower workload was observed with somatosensory C2 ( $p = 0.056$ ), than only monocular C1. Finally, (c) conditions with audition only did not contribute to an observable difference in workload ( $p > 0.05$ ).

**2) INTERFACE USABILITY**

For the perceived system usability, an ART was used to allow the use of parametric tests on ordinal data. A one way RM-ANOVA with a Greenhouse-Geisser correction ( $\epsilon = 0.561$ )





**FIGURE 7.** Heat-map illustrating the proportion of task success rate going from lower to higher complexity, horizontal axis A.1.1 (left) to C.2.3 (right), across all the interface conditions C1 to C8, vertical axis.

was used, yielding a highly significant difference across all eight interface conditions, ( $F(3.930, 94.310) = 97.064$ ,  $p < 0.001$ ,  $\eta_p^2 = 0.802$ ). Average responses for interface usability are shown in Figure 6 and Table 3. Post-hoc analysis revealed that the same trend holds true for the system usability, as with the cognitive workload. Specifically, we again found partial support of our H2 hypothesis with (a) stereoscopic vision with the VRHMD C5, C6, C7 and C8 accounting to significantly higher interface usability than monocular vision with the display monitor C1, C2, C3 and C4, ( $p < 0.001$ ), (b) somatosensory feedback further increasing overall usability however again only when paired with stereoscopic visual feedback C6 and C8 ( $p < 0.05$ ), and finally (c) auditory feedback by itself making no significant difference ( $p > 0.05$ ).

## C. OBJECTIVE RESULTS

### 1) ERROR RATE

First, we analyzed the total proportion of successful task completion (%), across all interface conditions. Our sample was classified as a Bernoulli distribution and a two-times standard deviation from the mean, three-sigma rule, was used to test for significance. Results show that, interface conditions incorporating stereoscopic vision with the VRHMD (C5, C6, C7, C8) accounted to a significant observable difference, ( $p < 0.05$ ), in mean success rates 96.22% (SD=4.73%), 99.11% (SD=2.62%), 96.22% (SD=5.94%), 97.55% (SD=4.26%) respectively compared to the monocular display monitor (C1, C2, C3, C4) with rates 39.33% (SD=21.69%), 47.55% (SD=18.77%), 51.33% (SD=23.63%), 48.22% (SD=20.26%) respectively. No significant differences were observed between conditions incorporating haptic or auditory feedback ( $p > 0.05$ ). Results are depicted in Figure 7.

### 2) PLACEMENT AND COMPLETION TIME

For time-based metrics we considered only the successful instances. Transforming our data in a non-parametric state, Shapiro-Wilk Test for normality yielded ( $p < 0.05$ ) in both instances. Friedman's test was thus used, yielding a significant difference in mean placement as well as completion time across the eight interface conditions

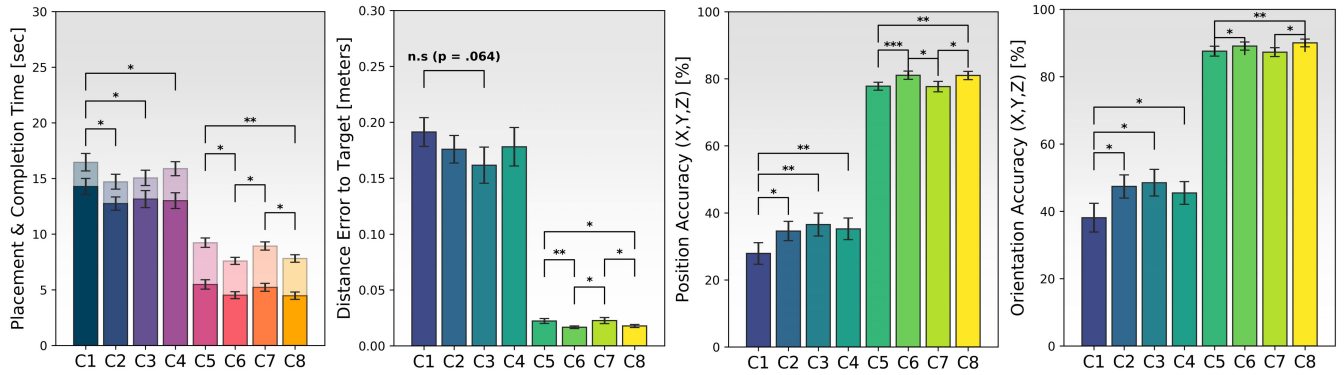
( $\chi^2(2) = 129.093$ ,  $p < 0.001$ ) and ( $\chi^2(2) = 131.093$ ,  $p < 0.001$ ) respectively. Placement and completion times are shown in Table 4 and Figure 8. Post-hoc analysis using the Wilcoxon Signed-Rank tests showed partial support of our H3 hypothesis, specifically (a) stereoscopic visual feedback with the VRHMD, C5, C6, C7 and C8 accounted to highly significantly less placement and completion time than with the monocular display monitor ( $p < 0.001$ ) C1, C2, C3 and C4, followed by (b) somatosensory feedback contributing additionally to significantly lesser placement and completion, however, only when paired with the VRHMD, C6 and C8 ( $p < 0.05$ ). Auditory feedback (c), did not contribute to an observable difference across all conditions ( $p > 0.05$ ).

### 3) DISTANCE ERROR

For distance-error to target, data was normally distributed, Shapiro-Wilk ( $p > 0.05$ ). As such, a one way RM-ANOVA with a Greenhouse-Geisser correction was used ( $\epsilon = 0.381$ ), yielding a highly statistical significance across the eight interface conditions, ( $F(2.664, 63.950) = 90.463$ ,  $p < 0.001$ ,  $\eta_p^2 = 0.790$ ). Distance error across all interfaces is shown in Table 4 and visually represented in Figure 8. Post-hoc analysis revealed partial support of our H4 hypothesis, specifically (a) conditions incorporating stereoscopic vision with the VRHMD, C5, C6, C7, C8 accounted to significantly lower distance error ( $p < 0.001$ ), compared to conditions incorporating monocular vision with the display monitor, C1, C2, C3, C4. Furthermore, (b) conditions incorporating somatosensory feedback, however only when paired with stereoscopic visual feedback, C6, C8, showed further significantly lower target error to the target placement ( $p < 0.05$ ), than conditions without C5, C7 respectively. Finally, auditory stimulation did not contribute to an observable difference in spatial accuracy ( $p > 0.05$ ).

### 4) POSITION AND ORIENTATION ACCURACY

Regarding spatial accuracy, specifically position and orientation accuracy, Shapiro-Wilk Test in both instances yielded ( $p > 0.05$ ) thus signifying normally distributed data. As such a one way RM-ANOVA with a Greenhouse-Geisser correction ( $\epsilon = 0.476$ ) and ( $\epsilon = 0.448$ ) respectively, yielded in both instances a highly significant difference ( $F(3.332, 79.962) = 174.488$ ,  $p < 0.001$ ,  $\eta_p^2 = 0.879$ ) and ( $F(3.139, 75.334) = 109.280$ ,  $p < 0.001$ ,  $\eta_p^2 = 0.820$ ) respectively. Position and orientation accuracy are shown in Table 4 and visually represented in Figure 8. Post-hoc analysis revealed full support of our H5 hypothesis, specifically (a) conditions incorporating stereoscopic vision with the VRHMD, C5, C6, C7, C8 accounted to significantly higher spatial accuracy both in position and orientation ( $p < 0.001$ ), than conditions incorporating monocular vision with the display monitor, C1, C2, C3, C4. Furthermore, (b) conditions incorporating somatosensory feedback C2, C4 and C6, C8, showed further significantly higher spatial accuracy ( $p < 0.05$ ), than those who do not C1, C3 and C5, C7 respectively. Finally,



**FIGURE 8.** Objective measurements represented in a bar graph in addition to standard error. From left to right, time-based metrics mean placement (opaque) and completion time (slightly transparent). Followed by spatial based metrics, specifically distance error, position and rotation accuracy. For all illustrated measurement results, all combinations between monocular and stereoscopic vision, i.e. C1, C2, C3, C4 with C5,C6,C7 and C8 are highly significant,  $p < 0.001$  and as a result significance bars are not visualised. For the remaining combinations, significance bars are shown, with each indicating \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$  and finally n.s not significant.

**TABLE 4.** Summary of all objective results including time-based and spatial-based metrics with mean and standard deviation across all eight interface conditions. In contrast with Figure 8, placement and completion times are shown here separately. Consult Figure 8 above for levels of significance.

Objective Measurements				Placement Time [s]		Completion Time [s]		Distance Error [cm]		Pos Accuracy (XYZ) [%]		Rot Accuracy (XYZ) [%]	
Condition	Vision	Audio	Haptic	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
C1	Monitor	Off	Off	14.27	±3.65	16.45	±3.91	19.12	±6.42	27.89%	±16.06	38.09%	±21.35
C2	Monitor	Off	On	12.73	±2.98	14.70	±3.32	17.58	±6.17	34.59%	±14.30	47.37%	±17.27
C3	Monitor	On	Off	13.14	±3.83	15.05	±3.42	16.15	±8.06	36.50%	±17.09	48.48%	±19.86
C4	Monitor	On	On	13.00	±3.55	15.88	±3.15	17.80	±8.61	35.22%	±16.11	45.44%	±16.97
C5	VRHMD	Off	Off	5.48	±2.10	9.22	±2.10	2.21	±1.08	77.75%	±5.93	87.55%	±7.30
C6	VRHMD	Off	On	4.51	±1.57	7.58	±1.60	1.65	±0.58	81.04%	±6.08	89.08%	±6.07
C7	VRHMD	On	Off	5.22	±1.77	8.92	±1.85	2.26	±1.37	77.65%	±7.75	87.27%	±6.65
C8	VRHMD	On	On	4.47	±1.64	7.80	±1.71	1.76	±0.70	80.97%	±6.16	90.01%	±5.67

(c) conditions incorporating only auditory stimulation C3 did also cause a greater increase in spatial accuracy than those without (C1) ( $p < 0.05$ ). Our findings here suggest that spatial accuracy increases significantly when stereo vision is used and furthermore when paired with either sound or somatosensory or even both, than just only relying on vision.

**VII. DISCUSSION**

Our results are summarised as follows: the overall performance of users increased by around 40% using stereoscopic vision with the VRHMD compared to monocular vision with the display monitor. Somatosensory feedback increased performance by a further 10% on top of all measurements. Auditory stimulation, however, had no significant effect on any measure, apart from an increase in spatial accuracy by less than 5%.

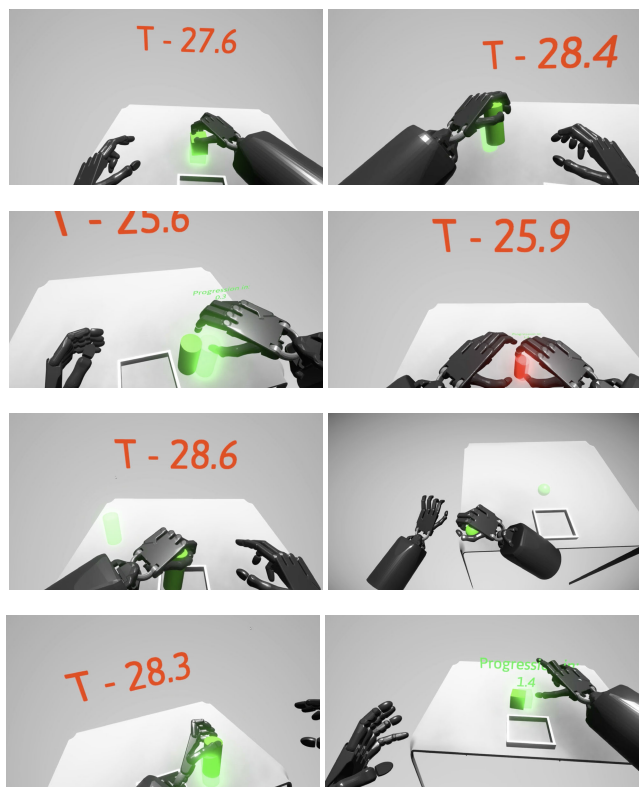
These results provide evidence to the untested hypothesis of [2]. More specifically, our results show that an audio-visuohaptic interface incorporating a stereoscopic VRHMD than a monocular monitor contributes to the highest task performance, followed by visuohaptic and less closely by audiovisual interfaces. See the Appendices for a cone-like illustration of each interface effectiveness that closely resemble the figures of [2].

Our results support existing research that vision is the dominant sense [14], [15], outperforming all other senses [13]. As depth information is important in manipulation tasks,

we can infer that better performance in VR may in part be due to the superior information available when using VRHMDs. This supports current literature [8], [20], [21].

Our results showed that less perceived cognitive workload was observed in the use of the VRHMD than in the monocular display. This contradicts previous work [17], but this may be attributed to significantly higher amounts of induced vection. Thus full conclusions cannot be drawn with our static scenario and further investigation is required to confirm. Our findings show that haptic feedback leads to better performance which is supported by some studies [37], but contradicts others [45]. The latter study found no significant effect of haptic feedback in a virtual throwing task. Since there are such a large number of options available for providing haptic feedback, findings may differ wildly simply by using a slightly different device. More research may be needed to investigate how small variations in the way haptic feedback is delivered, affects performance and a standardised device may be needed to compare the actual effect of haptics on humans.

The differences in the results for haptic devices may be partially explained by the “uncanny valley of haptics” [42]. This suggests that increasing the resolution of haptic feedback without the corresponding level of stimulation from other senses, will not contribute to a guaranteed increase in performance. Thus the resolution of all feedback interfaces has to be similar. Their study [42] used handheld controllers to



**FIGURE 9.** Example recordings of eight different participants during the manipulation experiment.

deliver haptic feedback. We used a custom vibrotactile glove which has a higher resolution than the handheld controllers, but this only increased performance when the resolution of visual stimulation was increased as well by switching from the monocular display monitor to the stereoscopic VRHMD, thus supporting [42].

We found little evidence to show that auditory feedback has a positive impact on performance, though spatial accuracy did increase in the audiovisual condition compared to the visual condition ( $p < 0.05$ ). Workload demand marginally decreased when auditory feedback was presented than just none at all, but not a significant level ( $p = 0.056$ ). However, a significant difference was found in previous work [27]. It is possible that this was a bi-product of the increase in performance when switching from mono to stereo vision, potentially overshadowing the contribution of audio in the subjective performance of participants.

In both objective and subjective measures, the combination of stereoscopic visual feedback i.e. the VRHMD with the addition of audio and haptic feedback, Condition 8, provided the best performance overall. This supports our primary hypothesis. This is in line with existing literature, that adding more modalities is correlated to improved performance in manipulation scenarios [46]. Though, there was no significant difference in performance when using only two modalities: stereoscopic visual i.e. VRHMD and haptic feedback, Condition 6. We did, nonetheless, see a marginal, but still significant drop in position and orientation accuracy

**TABLE 5.** Summary of Hypotheses support. Y: Yes, P: Partial, N: No.

Hypothesis	Support	Description
H1: Lower perceived workload	Partial	(a) Y (b) P (c) N
H2: Higher system usability	Partial	(a) Y (b) P (c) N
H3: Less task time	Partial	(a) Y (b) P (c) N
H4: Less distance error	Partial	(a) Y (b) P (c) N
H5: Higher placement precision	Full	(a) Y (b) Y (c) Y

(a) Vision with stereoscopic VR-HMD than monoscopic monitor  
 (b) Haptic feedback than without (c) Sound feedback than without  
 \*P: Partial; only effective when paired with stereo VR-HMD

in this condition, indicating that auditory did contribute to the effectiveness of spatial accuracy.

The main findings and design implications of our study include:

- Adding additional modalities increases performance.
- Relying on just one modality should be avoided.
- Vision dominates, making the highest contribution in performance when enhancing from mono to stereo vision.
- Effectiveness of multimodal interfaces is scenario-specific, this research explored it in the context of manipulation.
- Prioritization of visual, somatosensory and then auditory stimulation should be given for manipulation scenarios.
- Increasing task complexity lowers effectiveness as expected, but is not proportional for all multimodal interfaces.
- Vibrotactile feedback can be considered as a low-cost somatosensory approach while more focus can be given on the design of vibrotactile intensity to compensate for the inherent lack of force-feedback.

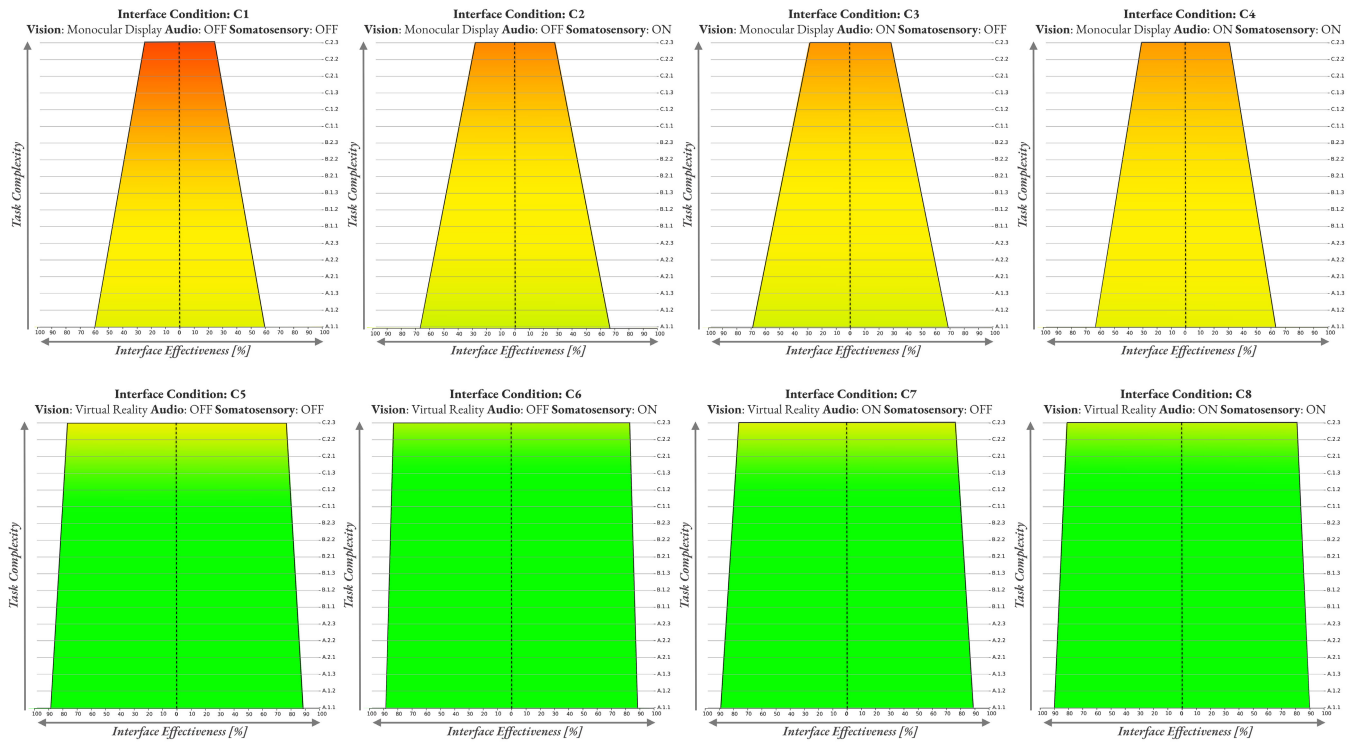
All of our hypotheses are summarized in Table 5 below, providing an overall overview of our findings.

### A. DESIGN AND RESEARCH IMPLICATIONS

Our low-cost haptic gloves show that expensive solutions are not required to achieve significant performance increases, in line with [42]. This may enable a wider range of research into haptic feedback and cost-effective multimodal interfaces.

We also show that adding haptic feedback to monocular feedback has no significant effect on performance. However, adding haptic feedback to a VRHMD does improve performance significantly. This seems to be in line with the “uncanny valley of haptics” [42], which supports that it is not enough to add extra sensory modalities, but the resolution of these modalities must be similar. This is highlighted in Conditions 2 (visuohaptic) & Condition 4 (audiovisuohaptic), where monocular vision is used. In this case, the additional sensory modalities did not contribute to an observable difference in performance apart from spatial accuracy, possibly due to a mismatch in resolution between monocular vision and other modalities.

Priorities should thus be given when designing multimodal interfaces for object manipulation. Our results support



**FIGURE 10.** Overall interface effectiveness through linear regression, across all measurements and across all tasks with an increasing task complexity from lower to higher. Width of the shapes represents the effectiveness, the wider the higher. Colouring also indicates the effectiveness increasing from red to green. The overall effectiveness is calculated linearly, specifically, the measurements are weighted  $(1 - 1/V_{max})$  where  $V_{max}$  is the maximum limit of the measurement. The data points from the scatter plot have been line fitted through linear regression to visualize a cone-like illustration. The width of the cone represents the effectiveness while the height of the cone the effectiveness of the interface at the specific task complexity. The specific task complexity is discussed in section “Manipulation Tasks of Varying Complexity”.

that researchers should aim to enhance visual stimuli before adding somatosensory feedback and lastly auditory.

Furthermore, based on our results, designers and researchers focusing on human performance in teleoperation, are encouraged to combine sensory interfaces as highlighted in this study. We observed that almost in all cases, bi-modal feedback i.e. visuohaptic and even more so audiovisuohaptic interfaces are significantly better performing than just relying on visual feedback. This may be even more the case for sensory channels that are already overloaded [9], [10], thus potentially opening more opportunities for researchers to investigate the effectiveness of such interfaces when channels are overloaded.

**B. LIMITATIONS AND FUTURE WORK**

The investigation of this research was focused on the contribution of the effectiveness of each sensory modality and combinations of these. However, we have not yet tested how auditory or somatosensory feedback would have compensated potentially overloaded visual information, which would have provided furthermore insight. Furthermore, we investigated the effectiveness of common visual feedback modalities i.e. the monitor display monitor and a VRHMD with their inherent capabilities. However, we did not explicitly and strictly investigated how monocular and stereoscopic visual feedback by themselves would influence performance.

Future road-map would include using the VRHMD with either monocular or stereoscopic rendering. In addition, multimodal design decisions are of paramount importance before implementing any kind of sensory feedback [2]. In our case, auditory feedback was implemented as the means of task indication and succession, instead of a continuous sonification i.e. concurrent type. Examples of concurrent auditory feedback would include controlling auditory pitch continuously based on target proximity, specific to manipulation tasks, which would potentially further enhance or supplement vision for depth perception. Thus, further evidence may be needed on how not only different types of sensory feedback may influence task succession, but also how the design decisions of each sensory channel affect task efficiency. We did assume zero to minimal latency during our experiments, knowing that time delays are correlated to simulator sickness. This is a real-world problem in teleoperation and further aggravated in wireless technologies. Latency in our experiments was <15ms and thus its effect was not studied. However, in real-world applications, latency can become a problem that causes simulator sickness and is also a challenge in teleoperation where communication bandwidth is limited [68]. Within this study, by thoroughly comparing an audiovisuohaptic multimodal interface, we have gained interesting insight on which modalities contribute to increased task performance, as long as time-delay is minimal.

## VIII. CONCLUSION

This paper explored how combining multiple sensory interfaces affects performance in manipulation tasks of varying complexity. Each combination of visual (monocular display monitor or a stereoscopic VRHMD), audio (with or without) and haptic (with or without) interface was tested. Task difficulty ranged from low to high by changing the size and shape of objects as well as the distance to the target placement.

The performance was measured both objectively and subjectively under experimental conditions. The results of these experiments showed a 40% increase in overall performance when using stereoscopic VRHMD visual feedback compared to a monocular display monitor. Somatosensory stimulation contributed a further 10% increase in performance, while auditory feedback only increased the spatial accuracy by an additional 5%.

Our evaluation found that by adding one more sensory modality in an interface is of a significant benefit than just relying on visual feedback. We thus conclude that task performance in teleoperation can be positively influenced by carefully selecting an appropriate combination of sensory feedback for a given task. As a result of this study, future researchers and designers should identify and prioritize certain modalities in order to design effective multimodal interfaces.

## APPENDIX OVERVIEW OF OVERALL RESULTS

In this appendix section, we summarize the overall interface effectiveness from our experiments and visualise the overall findings of our results in Figure 10. In these figures, we visualise the overall effectiveness of each individual interface condition across all measurements and all tasks for the final overview of our entire experimental results.

## ACKNOWLEDGMENT

We would like to thank Prof. Taku Komura and Prof. Robert Fisher for their support during this study.

## REFERENCES

- [1] R. J. Stone, "Haptic feedback: A brief history from telepresence to virtual reality," in *Haptic Human-Computer Interaction*, S. Brewster and R. Murray-Smith, Eds. Berlin, Germany: Springer, 2001, pp. 1–16.
- [2] R. Sigrist, G. Rauter, R. Riener, and P. Wolf, "Augmented visual, auditory, haptic, and multimodal feedback in motor learning: A review," *Psychonomic Bull. Rev.*, vol. 20, no. 1, pp. 21–53, Feb. 2013, doi: 10.3758/s13423-012-0333-8.
- [3] J. Y. C. Chen, E. C. Haas, and M. J. Barnes, "Human performance issues and user interface design for teleoperated robots," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 37, no. 6, pp. 1231–1245, Nov. 2007.
- [4] H. Yanco and J. Drury, "Where am i? Acquiring situation awareness using a remote robot platform," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, Dec. 2004, pp. 2835–2840.
- [5] B. P. DeJong, J. E. Colgate, and M. A. Peshkin, "Improving teleoperation: Reducing mental rotations and translations," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Apr. 2004, pp. 3708–3714.
- [6] C. Jennett, A. L. Cox, P. Cairns, S. Dhoparee, A. Epps, T. Tijss, and A. Walton, "Measuring and defining the experience of immersion in games," *Int. J. Hum.-Comput. Stud.*, vol. 66, no. 9, pp. 641–661, Sep. 2008, doi: 10.1016/j.ijhcs.2008.04.004.
- [7] H. Martins and R. Ventura, "Immersive 3-D teleoperation of a search and rescue robot using a head-mounted display," in *Proc. IEEE Conf. Emerg. Technol. Factory Autom.*, Sep. 2009, pp. 1–8.
- [8] D. R. Lampton, D. P. McDonald, M. Singer, and J. P. Bliss, "Distance estimation in virtual environments," *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, vol. 39, no. 20, pp. 1268–1272, Oct. 1995, doi: 10.1177/154193129503902006.
- [9] C. E. Lathan and M. Tracey, "The effects of operator spatial perception and sensory feedback on human-robot teleoperation performance," *Presence, Teleoperators Virtual Environ.*, vol. 11, no. 4, pp. 368–377, Aug. 2002, doi: 10.1162/105474602760204282.
- [10] G. Burdea, P. Richard, and P. Coiffet, "Multimodal virtual reality: Input output devices, system integration, and human factors," *Int. J. Hum.-Comput. Interact.*, vol. 8, no. 1, pp. 5–24, Jan. 1996, doi: 10.1080/10447319609526138.
- [11] G. V. Popescu, G. C. Burdea, and H. Trefftz, "Multimodal interaction modeling," in *Handbook of Virtual Environments: Design, Implementation, and Applications* (Human Factors and Ergonomics), K. M. Stanney, Ed. Mahwah, NJ, USA: Erlbaum, 2002, pp. 435–454.
- [12] P. Richard, G. Burdea, D. Gomez, and P. Coiffet, "A comparison of haptic, visual and auditive force feedback for deformable virtual objects," in *Proc. International Conf. Autom. Technol. (ICAT)*, vol. 49, 1994, p. 62.
- [13] M. L. Heilig, "EL cine del futuro: The cinema of the future," *Presence, Teleoperators Virtual Environments*, vol. 1, no. 3, pp. 279–294, Jan. 1992.
- [14] I. Rock and J. Victor, "Vision and touch: An experimentally created conflict between the two senses," *Science*, vol. 143, no. 3606, pp. 594–596, Feb. 1964. [Online]. Available: <https://science.sciencemag.org/content/143/3606/594>
- [15] R. L. Klatzky, J. M. Loomis, A. C. Beall, S. S. Chance, and R. G. Golledge, "Spatial updating of self-position and orientation during real, imagined, and virtual locomotion," *Psychol. Sci.*, vol. 9, no. 4, pp. 293–298, Jul. 1998. [Online]. Available: <http://www.jstor.org/stable/40063340>
- [16] J. P. McIntire, P. R. Havig, and E. E. Geiselman, "Stereoscopic 3D displays and human performance: A comprehensive review," *Displays*, vol. 35, no. 1, pp. 18–26, Jan. 2014. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0141938213000929>
- [17] J. Brooks, R. Lodge, and D. White, "Comparison of a head-mounted display and flat screen display during a micro-UAV target detection task," *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, vol. 61, no. 1, pp. 1514–1518, Sep. 2017, doi: 10.1177/154193121601863.
- [18] B. Keshavarz, B. E. Riecke, L. J. Hettinger, and J. L. Campos, "Vection and visually induced motion sickness: How are they related?" *Frontiers Psychol.*, vol. 6, p. 472, Apr. 2015.
- [19] L. B. Rosenberg, "The effect of interocular distance upon operator performance using stereoscopic displays to perform virtual depth tasks," in *Proc. IEEE Virtual Reality Annu. Int. Symp.*, Sep. 1993, pp. 27–32.
- [20] B. G. Witmer and P. B. Kline, "Judging perceived and traversed distance in virtual environments," *Presence, Teleoperators Virtual Environ.*, vol. 7, no. 2, pp. 144–167, Apr. 1998, doi: 10.1162/105474698565640.
- [21] J. E. Swan, G. Singh, and S. R. Ellis, "Matching and reaching depth judgments with real and augmented reality targets," *IEEE Trans. Vis. Comput. Graphics*, vol. 21, no. 11, pp. 1289–1298, Nov. 2015.
- [22] D. R. Scribner and J. W. Gombash, "The effect of stereoscopic and wide field of view conditions on teleoperator performance," *Hum. Res. Eng., Army Res. Lab. Aberdeen Proving Ground, MD, USA, Tech. Rep. AD-a341 218*, 1998. [Online]. Available: <https://apps.dtic.mil/dtic/tr/fulltext/u2/a341218.pdf>
- [23] X. Shang, M. Kallmann, and A. S. Arif, "Effects of correctness and suggestive feedback on learning with an autonomous virtual trainer," in *Proc. 24th Int. Conf. Intell. User Interface Computing (IUI)*, New York, NY, USA, 2019, pp. 93–94, doi: 10.1145/3308557.3308675.
- [24] C. C. Smyth, "Indirect vision driving with fixed flat panel displays for near unity, wide, and extended fields of camera view," *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, vol. 44, no. 36, pp. 541–544, Jul. 2000, doi: 10.1177/154193120004403624.
- [25] S. Johnson, I. Rae, B. Mutlu, and L. Takayama, "Can you see me now?: How field of view affects collaboration in robotic telepresence," in *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, 2015, pp. 2397–2406, doi: 10.1145/2702123.2702526.
- [26] R. D. Shilling and B. Shinn-Cunningham, "Virtual auditory displays," in *Handbook Virtual Environments*. Boca Raton, FL, USA: CRC Press, 2002, pp. 105–132.

- [27] Y. Nagai, S. Kimura, S. Tsuchiya, and T. Iida, "Audio feedback system for engineering test satellite vii," *Proc. SPIE*, vol. 3840, pp. 144–152, Nov. 1999. [Online]. Available: <https://doi.org/10.1117/12.369274>
- [28] S. Tachi, K. Komoriya, K. Sawada, T. Nishiyama, T. Itoko, M. Kobayashi, and K. Inoue, "Telexistence cockpit for humanoid robot control," *Adv. Robot.*, vol. 17, no. 3, pp. 199–217, Jan. 2003, doi: [10.1163/156855303764018468](https://doi.org/10.1163/156855303764018468).
- [29] R. Secoli, M.-H. Milot, G. Rosati, and D. J. Reinkensmeyer, "Effect of visual distraction and auditory feedback on patient effort during robot-assisted movement training after stroke," *J. Neuroeng. Rehabil.*, vol. 8, no. 1, p. 21, 2011.
- [30] B. D. Simpson, R. S. Bolia, and M. H. Draper, "Spatial audio display concepts supporting situation awareness for operators of unmanned aerial vehicles," *Hum. Perform., Situation Awareness, Automat., Current Res. Trends HPSAA II*, vol. 2, p. 61, Dec. 2013.
- [31] T. Erni and V. Dietz, "Obstacle avoidance during human walking: Learning rate and cross-modal transfer," *J. Physiol.*, vol. 534, no. 1, pp. 303–312, Jul. 2001.
- [32] M. Wellner, A. Schaufelberger, J. V. Zitzewitz, and R. Rieni, "Evaluation of visual and auditory feedback in virtual obstacle walking," *Presence, Teleoperators Virtual Environments*, vol. 17, no. 5, pp. 512–524, Oct. 2008.
- [33] R. W. Cholewiak and A. A. Collins, "The generation of vibrotactile patterns on a linear array: Influences of body site, time, and presentation mode," *Perception Psychophys.*, vol. 62, no. 6, pp. 1220–1235, Sep. 2000, doi: [10.3758/BF03212124](https://doi.org/10.3758/BF03212124).
- [34] J. L. Rochlis and D. J. Newman, "A tactile display for international space station (ISS) extravehicular activity (EVA)," *Aviation, space, Environ. Med.*, vol. 71, no. 6, pp. 571–578, 2000.
- [35] F. Gemperle, N. Ota, and D. Siewiorek, "Design of a wearable tactile display," in *Proc. 5th Int. Symp. Wearable Comput.*, 2001, pp. 5–12.
- [36] V. Yem, K. Vu, Y. Kon, and H. Kajimoto, "Effect of electrical stimulation haptic feedback on perceptions of softness-hardness and stickiness while touching a virtual object," in *Proc. IEEE Conf. Virtual Reality 3D User Interface (VR)*, Mar. 2018, pp. 89–96.
- [37] D. Brickler, S. V. Babu, J. Bertrand, and A. Bhargava, "Towards evaluating the effects of stereoscopic viewing and haptic interaction on perception-action coordination," in *Proc. IEEE Conf. Virtual Reality 3D User Interface (VR)*, Mar. 2018, pp. 1–516.
- [38] V. Vuskovic, M. Kauer, G. Szekeley, and M. Reidy, "Realistic force feedback for virtual reality based diagnostic surgery simulators," in *Proc. IEEE Int. Conf. Robot. Automat. Sympo. Process.*, Apr. 2000, pp. 1592–1598.
- [39] M. Sinclair, E. Ofek, M. Gonzalez-Franco, and C. Holz, "CapstanCrunch: A haptic VR controller with user-supplied force feedback," in *Proc. 32nd Annu. ACM Symp. User Interface Softw. Technol.*, New York, NY, USA, Oct. 2019, p. 815, doi: [10.1145/3332165.3347891](https://doi.org/10.1145/3332165.3347891).
- [40] X. Gu, Y. Zhang, W. Sun, Y. Bian, D. Zhou, and P. O. Kristensson, "Dexmo: An inexpensive and lightweight mechanical exoskeleton for motion capture and force feedback in VR," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, May 2016, p. 1991, doi: [10.1145/2858036.2858487](https://doi.org/10.1145/2858036.2858487).
- [41] I. Choi, E. W. Hawkes, D. L. Christensen, C. J. Ploch, and S. Follmer, "Wolverine: A wearable haptic interface for grasping in virtual reality," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 986–993.
- [42] C. C. Berger, M. Gonzalez-Franco, E. Ofek, and K. Hinckley, "The uncanny valley of haptics," *Sci. Robot.*, vol. 3, no. 17, Apr. 2018, Art. no. eaar7010.
- [43] J. Aleotti, S. Bottazzi, and M. Reggiani, (Oct. 2002). *A Multimodal User Interface for Remote Object Exploration in Teleoperation Systems*. [Online]. Available: <https://pdfs.semanticscholar.org/98bd/ce82196ceb7e2f8f0f1a6a480f0a2a00e80d.pdf>
- [44] A. M. Murray, R. L. Klatzky, and P. K. Khosla, "Psychophysical characterization and testbed validation of a wearable vibrotactile glove for telemanipulation," *Presence, Teleoperators Virtual Environ.*, vol. 12, no. 2, pp. 156–182, Apr. 2003, doi: [10.1162/105474603321640923](https://doi.org/10.1162/105474603321640923).
- [45] E. Frid, J. Moll, R. Bresin, and E.-L. Sallnäs Pysander, "Haptic feedback combined with movement sonification using a friction sound improves task performance in a virtual throwing task," *J. Multimodal User Interface*, vol. 13, no. 4, pp. 279–290, Dec. 2019, doi: [10.1007/s12193-018-0264-4](https://doi.org/10.1007/s12193-018-0264-4).
- [46] J. L. Burke, M. S. Prewett, A. A. Gray, L. Yang, F. R. B. Stilson, M. D. Coovert, L. R. Elliot, and E. Redden, "Comparing the effects of visual-auditory and visual-tactile feedback on user performance: A meta-analysis," in *Proc. 8th Int. Conf. Multimodal Interface*, New York, NY, USA, 2006, pp. 108–117, doi: [10.1145/1180995.1181017](https://doi.org/10.1145/1180995.1181017).
- [47] C. M. Light, P. H. Chappell, and P. J. Kyberd, "Establishing a standardized clinical assessment tool of pathologic and prosthetic hand function: Normative data, reliability, and validity," *Arch. Phys. Med. Rehabil.*, vol. 83, no. 6, pp. 776–783, Jun. 2002.
- [48] V. Kumar and E. Todorov, "MuJoCo HAPTIX: A virtual reality system for hand manipulation," in *Proc. IEEE-RAS 15th Int. Conf. Hum. Robots*, Nov. 2015, pp. 657–663.
- [49] M. Mori, K. MacDorman, and N. Kageki, "The uncanny valley [From the Field]," *IEEE Robot. Autom. Mag.*, vol. 19, no. 2, pp. 98–100, Jun. 2012.
- [50] A. Ghazanfar and C. Schroeder, "Is neocortex essentially multisensory?" *Trends Cognit. Sci.*, vol. 10, no. 6, pp. 278–285, Jun. 2006.
- [51] A. Amedi, R. Malach, T. Hendler, S. Peled, and E. Zohary, "Visuo-haptic object-related activation in the ventral visual pathway," *Nature Neurosci.*, vol. 4, no. 3, pp. 324–330, Mar. 2001.
- [52] T. W. James, G. K. Humphrey, J. S. Gati, P. Servos, R. S. Menon, and M. A. Goodale, "Haptic study of three-dimensional objects activates extrastriate visual areas," *Neuropsychologia*, vol. 40, no. 10, pp. 1706–1714, Jan. 2002.
- [53] K. Sathian and A. Zangaladze, "Feeling with the mind's eye: Contribution of visual cortex to tactile perception," *Behavioural Brain Res.*, vol. 135, nos. 1–2, pp. 127–132, Sep. 2002.
- [54] J. B. Van Erp and P. Padmos, "Image parameters for driving with indirect viewing systems," *Ergonomics*, vol. 46, no. 15, pp. 1471–1499, Dec. 2003.
- [55] M. Kim, C. Jeon, and J. Kim, "A study on immersion and presence of a portable hand haptic system for immersive virtual reality," *Sensors*, vol. 17, no. 5, p. 1141, May 2017, doi: [10.3390/s17051141](https://doi.org/10.3390/s17051141).
- [56] B. Weber, M. Sagardia, T. Hulín, and C. Preusche, "Visual, vibrotactile, and force feedback of collisions in virtual environments: Effects on performance, mental workload and spatial orientation," in *Virtual Augmented Mixed Reality. Designing Developing Augmented Virtual Environments*, R. Shumaker, Ed. Berlin, Germany: Springer, 2013, pp. 241–250.
- [57] Y. Suzuki and M. Kobayashi, "Air jet driven force feedback in virtual reality," *IEEE Comput. Graph. Appl.*, vol. 25, no. 1, pp. 44–47, Jan. 2005.
- [58] D. A. Kontarinis and R. D. Howe, "Tactile display of vibratory information in teleoperation and virtual environments," *Presence, Teleoperators Virtual Environ.*, vol. 4, no. 4, pp. 387–402, Jan. 1995, doi: [10.1162/pres.1995.4.4.387](https://doi.org/10.1162/pres.1995.4.4.387).
- [59] M. J. Massimino and T. B. Sheridan, "Sensory substitution for force feedback in teleoperation," *Presence: Teleoperators Virtual Environ.*, vol. 2, no. 4, pp. 344–352, Jan. 1993, doi: [10.1162/pres.1993.2.4.344](https://doi.org/10.1162/pres.1993.2.4.344).
- [60] M. Slater, M. Usoh, and A. Steed, "Depth of presence in virtual environments," *Presence: Teleoper. Virtual Environ.*, vol. 3, no. 2, pp. 130–144, Jan. 1994. [Online]. Available: <http://dx.doi.org/10.1162/pres.1994.3.2.130>
- [61] W. Ijsselstein, H. Ridder, de, J. Freeman, and S. Avons, "Presence: Concept, determinants and measurement," in *Human Vision and Electronic Imaging V*, B. Rogowitz and T. Pappas, Eds. Bellingham, WA, USA: SPIE, 2000, pp. 520–529.
- [62] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," *Adv. Psychol. Hum. Mental Workload*, vol. 15, p. 139–183, Dec. 1988.
- [63] J. Brooke, "SUS-A quick and dirty usability scale," *Usability Eval. Ind.*, vol. 189, no. 194, pp. 4–7, 1996.
- [64] H. Abdi, "The greenhouse-geisser correction," *Encyclopedia Res. Des.*, vol. 1, pp. 544–548, May 2010.
- [65] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins, "The aligned rank transform for nonparametric factorial analyses using only anova procedures," in *Proc. Annu. Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, 2011, pp. 143–146, doi: [10.1145/1978942.1978963](https://doi.org/10.1145/1978942.1978963).
- [66] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *J. Amer. Stat. Assoc.*, vol. 32, no. 200, pp. 675–701, Dec. 1937, doi: [10.1080/01621459.1937.10503522](https://doi.org/10.1080/01621459.1937.10503522).
- [67] F. Pukelsheim, "The three sigma rule," *Amer. Stat.*, vol. 48, no. 2, p. 88, May 1994. [Online]. Available: <http://www.jstor.org/stable/2684253>
- [68] S. A. McGlynn and W. A. Rogers, "Considerations for presence in teleoperation," in *Proc. Companion ACM/IEEE Int. Conf. Human-Robot Interact. (HRI)*, New York, NY, USA, 2017, p. 203, doi: [10.1145/3029798.3038369](https://doi.org/10.1145/3029798.3038369).



**ELEFThERIOS TRIANTAFYLLIDIS** received the B.S. degree in computer engineering from the Eastern Macedonia and Thrace Institute of Technology, Kavala, Greece, in 2016, and the M.S. (by Research) degree (Hons.) in robotics and autonomous systems from The University of Edinburgh, U.K., in 2019, where he is currently pursuing the Ph.D. degree in robotics and autonomous systems.

From 2015 to 2016, he was an Intern at Audi AG, Germany, working closely with other researchers and professionals on state-of-the-art human–computer interaction techniques. His research interests include human–computer interaction, multimodal interfaces, multisensory integration, mixed reality technologies, and robotics.



**CHRISTOPHER MCGREAVY** received the B.S. degree in human psychology from the University of Leicester, in 2010, the M.Sc. degree (Hons.) in computational neuroscience and cognitive robotics from the University of Birmingham, in 2016, and the M.Sc. (by Research) degree (Hons.) in robotics and autonomous systems from The University of Edinburgh, in 2018, where he is currently pursuing the Ph.D. degree in robotics and autonomous systems.



**JIACHENG GU** received the B.S. degree in electrical & electronic engineering and the M.S. in artificial intelligence from The University of Edinburgh, U.K., in 2017 and 2018, respectively, where he is currently pursuing the Ph.D. degree in robotics and autonomous systems.



**ZHIBIN LI** (Member, IEEE) received the joint Ph.D. degree in robotics from the University of Genova and also from the Istituto Italiano di Tecnologia, Italy, in 2012. From 2012 to 2016, he was a Postdoctoral Researcher with the Istituto Italiano di Tecnologia, Genova, Italy.

He is currently a Lecturer with the Institute of Perception, Action and Behaviour (IPAB), School of Informatics, The University of Edinburgh. His research interests include a variety of technologies in robot control, optimization and machine learning for solving challenging problems in dynamic motion control (manipulation, grasping, and locomotion), and complex behaviours of mobile and legged robots (wheeled & tracked, quadruped & humanoid).

...