# Assisting Manipulation and Grasping in Robot Teleoperation with Augmented Reality Visual Cues

Stephanie Arévalo Arboleda
Westphalian University of Applied Sciences
Gelsenkirchen, Germany
stephanie.arevalo@w-hs.de

Franziska Rücker
Westphalian University of Applied Sciences
Gelsenkirchen, Germany
franziska.ruecker@w-hs.de

Tim Dierks
Westphalian University of Applied Sciences
Gelsenkirchen, Germany
tim.dierks@w-hs.de

Jens Gerken
Westphalian University of Applied Sciences
Gelsenkirchen, Germany
jens.gerken@w-hs.de

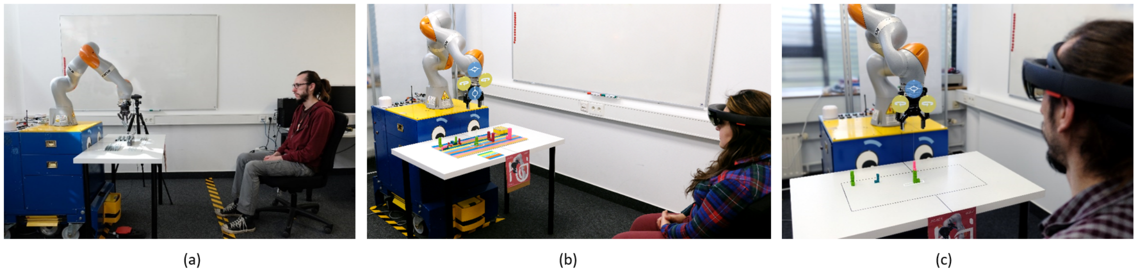


**Figure 1: (a) In this work, we explore how to improve performance in manipulation and grasping tasks by using augmented reality to present visual cues in teleoperation of a robotic arm in co-located spaces. To achieve that we designed two types of augmented visual cues (b) Basic Augmented Cues (c) Advanced Augmented Cues**

## ABSTRACT

Teleoperating industrial manipulators in co-located spaces can be challenging. Facilitating robot teleoperation by providing additional visual information about the environment and the robot affordances using augmented reality (AR), can improve task performance in manipulation and grasping. In this paper, we present two designs of augmented visual cues, that aim to enhance the visual space of the robot operator through hints about the position of the robot gripper in the workspace and in relation to the target. These visual cues aim to improve the distance perception and thus, the task performance. We evaluate both designs against a baseline in an experiment where participants teleoperate a robotic arm to perform pick-and-place tasks. Our results show performance improvements in different levels, reflecting in objective and subjective measures with trade-offs in terms of time, accuracy, and participants' views of teleoperation. These findings show the potential of AR not only in teleoperation, but in understanding the human-robot workspace.

## CCS CONCEPTS

• **Human-centered computing → Mixed / augmented reality**; *Empirical studies in interaction design.*

## KEYWORDS

human-robot interaction, visual cues, augmented reality, robot teleoperation

**ACM Reference Format:**
Stephanie Arévalo Arboleda, Franziska Rücker, Tim Dierks, and Jens Gerken. 2021. Assisting Manipulation and Grasping in Robot Teleoperation with Augmented Reality Visual Cues. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3411764.3445398

## 1 INTRODUCTION

Co-located teleoperation of robotic arms can be a challenging task that requires high level of attention, expertise and a good understanding of the environment, the robot, and the objects in it. In scenarios where the operator is not in direct proximity to the robot but at a certain distance (Figure 1a), some problems related to depth and distance perception may occur which in turn can affect performance. A prevailing problem in robot teleoperation is distance estimation in manipulation tasks, which is common in assembly environments. In addition, depth perception problems are part of the lack of situation awareness in teleoperation [15], which can result in lower performance in manipulation tasks. This also relates to the

*CHI '21, May 8–13, 2021, Yokohama, Japan*
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445398

ability to imagine how an object looks from a perspective different to the egocentric view, i.e., spatial orientation, which has been analyzed in remote teleoperation, showing how it affects teleoperation performance [33]. When analyzing co-located teleoperation (operator and robot located in the same physical space), the operator can experience problems related to depth perception and misjudge egocentric distances[43], similar to the problems experienced in remote teleoperation.

There have been a few efforts to improve the depth perception in remote teleoperation using cameras to capture different angles of a scene [23]. However, teleoperation has received fewer attention in co-located spaces. Problems in perception of distances and size can still occur when objects are located within the observer's personal space (2 m), i.e. overestimating distances [9]. The human eye naturally perceives the distance of objects around using visual cues processed by cognitive processes and can be divided into monocular and binocular cues [11]. Monocular cues aid to perceive distances in a near space (< 3m) [6] and comprise static and dynamic cues [22]. Dynamic cues are acquired by the motion of objects and/or observers and are crucial to perceive depth and object structure [46]. However, there are some scenarios where this motion is impossible or difficult (absence of dynamic cues) because the observer or the objects cannot move, e.g., people with mobility impairments, or scenarios where the visual angle cannot be changed due to workplace settings. Here, the observer can only count on natural static cues which might not suffice to get an accurate perception of distance and depth. Inspired by the static cues that humans naturally use to judge depth, we provide enhanced hints with Augmented Reality (AR) that consider elements from monocular cues such as lighting and interposition to provide a better understanding of the environment, the robot, and object affordances. As a use case, we considered a manufacturing environment with an industrial robotic arm where the operator is in a co-located space, teleoperating a robot to perform manipulation and grasping tasks.

The goal of this paper is to contribute to the research efforts of the use of AR for robot teleoperation from a human-centered robotics perspective. We present two designs of visual cues using AR, which we named, Augmented Visual Cues (AVC). These have the potential to improve task performance by providing better awareness of the objects' location and its depth within co-located spaces. In our first design, basic augmented visual cues, we augment the robot and the environment by providing a combination of real and virtual hints. In our second design, advanced augmented visual cues, we utilize sensor-based knowledge of the environment, i.e., using-pose object recognition, to present virtual hints that enhance the robot, environment and objects in it.

In the following sections, we present relevant work related to robot depth perception and the use of AR in robot teleoperation. Then, we introduce the reasoning behind AVC and present each design of basic and advanced augmented visual cues, followed by an experiment where we evaluated the user's performance in teleoperation using our designs of AVC against a baseline condition in a pick-and-place task. Next, we present our results, followed by a discussion section. Finally, we present the limitations of our findings and final conclusions.

## 2 RELATED WORK

### 2.1 Depth Perception

The human visual system captures environment information through the eye, transmits this information through the optic nerve to the brain, which in turn interprets the received information [2]. Among this information we find depth perception, which can be defined as the ability to see volume and the relative position of objects in a three-dimensional environment [52]. Further, size perception is closely related to depth perception since humans use depth cues to perceive size. Thus, in environments where the number of depth cues is reduced, humans' abilities to perceive size are also compromised [42].

Depth perception enables our interaction with surrounding objects and helps to determine size and distances, through monocular (provided by one eye) and binocular (provided by both eyes) cues [22]. Monocular cues provide information that allows to determine more or less accurately the distance and size of objects located at an egocentric distance of approximately 3m [6]. Monocular cues can be divided into static, i.e, do not need the observer to change perspective, and dynamic cues which are acquired by motion, e.g., the observer needs to move around the environment to change the egocentric perspective. Static cues comprise of perspective (linear and texture), interposition (occlusion and transparency), lighting (shading and shadow), aerial (optical haze and blur) and dynamic cues involving optical flow, motion parallax and accretion [22]. In general, humans combine different types of cues to understand distance, position, and the relation among objects in space.

From the cognition perspective, it is relevant to discuss cognitive maps. These are individual representations about "the spatial and environmental information of the geographical space" [26]. These maps help humans to locate objects in the environment, and are closely related to individual spatial abilities. In robot teleoperation, spatial orientation (own ability to imagine objects from different visual perspectives) and spatial relation (own ability to mentally manipulate objects) have proven to be relevant in performance [33].

### 2.2 AR and the Visual Space in Robot Teleoperation

AR explores different ways to superimpose virtual images into the real world seamlessly [18]. The latest widespread of the use of AR head-mounted displays (ARHMD), e.g. Microsoft HoloLens [34] and Magic Leap One [30], has brought along a growing interest in incorporating solutions grounded in the alignment of real and virtual objects in the user's line of sight [36]. Robotics and Human-Robot Interaction (HRI) can benefit greatly from the use of AR. In fact, there has been an increasing body of research on using AR in robotics [31], especially, using AR for robot control [13, 53]. One of the most explored areas of research is using AR to find ways to visualize robot motion, e.g., trajectory planning by previewing planned simulated paths for robot motion [12], focusing on path planning for collision avoidance [7, 12], and co-located teleoperation of aerial robots [21, 47, 48]. Combining robot control with computer vision techniques is a promising area of research, e.g. using deep learning for 3D object pose estimation combined with AR to create a robot system [38].

Augmenting the visual space in a human-robot context has gathered the interest of the HRI community. For instance, AR has been used to provide information that help robot industrial programmers, and it has shown promising results in reducing their cognitive load [41]. Previous work [17], presented a projection-based AR interface to provide instructions that will assist operators in robot programming pick-and-place tasks, and Gacem et al. [14] present a robotic arm that projects a spotlight to localize objects in unfamiliar and dense environments.

Research in AR for assembly environments has also gained popularity. Wang et al. [50] reviewed a number of papers published in 25 years that investigate AR in assembly tasks, which show the influence and potential of AR in assembly systems. Here, enhancing the visual space to facilitate robot control has become an area of interest in manufacturing environments, e.g., Clemente et al. [8] proposed showing visual feedback of the gripper force and closure. Aligned with this line of research, there has been an emphasis on the importance of visually showing the robots' intentions through ARHMD [49, 51], or using projections on the environment [5].

## 3 AUGMENTED VISUAL CUES (AVC) FOR TELEOPERATION

In this paper, we build on related work by combining AR and robot teleoperation to address distance and depth perception problems, thus, improving task performance. Teleoperating a robot in a co-located assembly environment, requires the operator to process and understand the surroundings and spatial relations between objects and the robot gripper, especially if the visual perspective is fixed, leading to misestimations of the position of the robot gripper relative to target objects. These common misestimations negatively influence accuracy and efficiency during pick-and-place tasks, e.g., it leads to multiple attempts to fit an object in a specific area, repositioning the gripper several times, or failing to grasp a target object.

We present two types of AVC, designed for ARHMD, e.g. Microsoft HoloLens, to improve depth perception in pick-and-place tasks. Both approaches capitalize on the conceptual work of Walker et al. [47] who proposed a design framework considering the elements from AR for a user-centered HRI. They present three archetypes: "augmenting the environment", using visual cues embedded in the interaction environment; "augmenting the robot", where visual cues are directly connected to the robot in a way that they might change its real features; or "augmenting the user interface", where ego-centric imagery could provide system information to the user.

In particular, our AVC designs provide a set of visual cues inspired by certain monocular cues, namely lighting and interposition, which humans use to understand their spatial environment. We clarify that lighting relates to the shadows that an object projects over another object, providing information about relative depth [45].

Our AVC designs differ from each other concerning the knowledge of the environment they require. As a baseline, we propose Basic Augmented Visual Cues (Basic Cues) which only require information about the workspace where the robot is operating, such as the height of the robot above the tabletop and its dimensions. This type of cue has certain limitations in how it can assist the user, simply because there is no previous knowledge of the environment, for instance, the position of the objects on the workspace, or of the robot. Therefore, our second type of AVC, Advanced Augmented Visual Cues (Advanced Cues), investigates the use of object-pose recognition to build a spatial understanding of the objects in the scene and integrates this knowledge to further assist the operator.

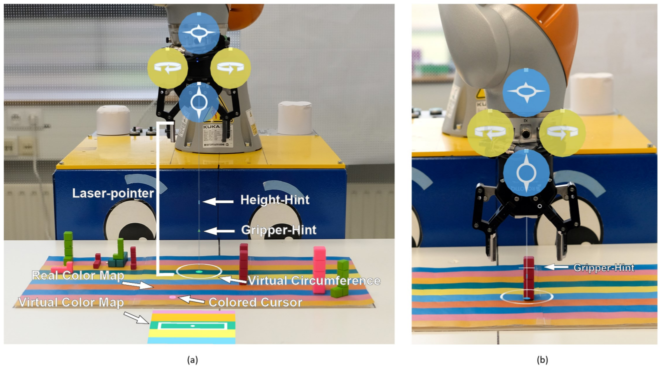


**Figure 2: (a) Visual representation of each basic augmented cue. (b) Gripper-hint showing adequate grasping position.**

Both AVC designs build on our general interaction principle for robot teleoperation. For this, we apply hands-free multimodal interaction by combining head-movements, using a head-gaze based cursor to point, and speech commands to commit certain actions. We also present a robot control interface that shows basic controls visually and spatially anchored to the gripper of the robot. These controls allow moving the gripper along the x-, y-, z-axis as well as rotating it. We will explain this interaction and interface design in detail in Section 4.7.

### 3.1 Basic Augmented Visual Cues (Basic Cues)

Our design of Basic Cues can be defined as indirect cues that suggest certain characteristics of the environments and the objects on it. These cues assist the operator to first determine the position of a target (depth perception) to then accurately positioning the gripper over a target. They aim to help teleoperation in unknown environments or when object-pose recognition is not possible.

In order to address distance and depth perception related problems, we designed a combination of physical and virtual hints. As the representative for real hints, we position a poster with a colormap upon the surface of the workspace. The colormap intends to provide a visual landmark that can help determine the exact or relative position of an object, e.g., the object in Figure 2b is located on the light blue stripe. However, it is still difficult to accurately determine the relative x,y position of the gripper above the workspace. Thus, we added a virtual colormap and other virtual hints that would indicate the position of the gripper on the workspace. Figure 2a shows a virtual semi-transparent line, which we named "Laser-pointer" reaching from the gripper towards the tabletop and crossing the dark green stripe on the physical colormap. Correspondingly, we provide a virtual colormap at the front of the table, which, with a rectangle, highlights the color or colors of the relative position of the gripper on the physical colormap.

Additionally, our virtual cues are inspired by the monocular cue of lighting, e.g., we present a virtual circumference under the gripper, emulating a shadow of the gripper on the tabletop. In the following section, we present the design and purpose of each basic cue showed in Figure 2 in detail:

**Laser-pointer.** This virtual cue consists of a semi-transparent straight vertical line. It starts at the center of the grip region until it reaches the tabletop. It has a small colored point at the end of the straight line on the tabletop. The color of that point matches the color of the physical colormap. The laser-pointer was designed to help to visualize the center position of the gripper and facilitate repositioning it in relation to the target object. Additionally, this semi-transparent line changes length dynamically when the gripper moves up or down along the y-axis (towards the tabletop). It also overlays objects located between the gripper and the tabletop, as it has no means of being aware of their presence or position.

**ColorMap.** This hint consists of a physical and virtual colormap. For the physical hint, we printed a poster with colored stripes —each 3 cm wide, considering the width of the gripper's fingers, which is of 2.2 cm, and covers the robot's workspace on the tabletop. It presents 7 different colors, which repeat sequentially to match the size of the workspace. The color combination was carefully selected to consider different types of color blindness following Wong's guidelines [55]. The virtual hint can be described as a window with colored stripes —showing the same colors presented in the poster. It is located at the front of the tabletop and is aligned to be in front of the current position of the gripper. Its function is to provide a visualization of the current position of the gripper on the tabletop, in the same way as a camera attached to the gripper facing the tabletop would. In addition, there is a white rectangle drawn over this window, symbolizing the grip region. It also has a white point in the middle of the rectangle that represents the laser-pointer. This cue intends to allow the operator to position the gripper right above a target object, reducing depth perception problems and assisting with partial occlusion.

**Colored cursor.** The cursor adopts the color of the physical stripe where it is currently pointing. This hint supports the operator by roughly pointing at a specific area on the tabletop, thereby assisting to visualize the position on the table the operator is pointing at.

**Colored-hints on the laser-pointer.** Along the laser-pointer we added two visual indicators, height-hint and gripper-hint. **The height-hint** highlights the mid-point between the gripper and the tabletop, Figure 2a. **The gripper-hint** is visible after the first movement along the robot's z-axis, and it indicates the future end-position of the gripper once its fingers are closed (grasping position) and it is presented as a white line that intersects the laser-pointer, Figure 2b. It indicates if the gripper is too close to the tabletop and it would collide with it while trying to grasp an object, or if the gripper is too far from the target and would fail to grasp it.

**Virtual Circumference.** Our design shows a virtual circumference with a diameter that matches the grip region, right below the position of the gripper on the tabletop. The main function of this virtual cue is to provide a better understanding of the diameter of the grip region over the tabletop. It was designed to assist the operator in determining if an object is within the grasping range or to determine if the gripper's fingers would collide with an object in the vicinity of the target object.

## 3.2 Advanced Augmented Visual Cues (Advanced Cues)

Advance Cues build upon the use of object-pose recognition of the objects in the scene. The resulting cues are therefore able to integrate this object-pose data to become environment-aware and provide explicit hints for interaction. This is achieved using the 3D model of the recognized objects and virtually overlaying them over the workspace. This virtual model is shaded to be invisible to operators and only occlude virtual elements behind it but not the real world. This allowed us to show a virtual representation of grasped objects that behave the same way as real objects. That is why we named the cues derived from object-pose recognition Advanced Cues, as we do not only present cues inspired in the monocular cue of lighting but also add occlusion to this design (Figure 3a, 3c). We propose and describe them as follows:

**Laser-pointer and virtual grip region.** We maintain this cue from the Basic Cues but added the occlusion capability, e.g., the laser-pointer projects from the center of the gripper down towards the tabletop until it reaches the nearest surface, which can be a target or neighboring objects, see Figure 3. Additionally, we provide a representation of the grip region on the tabletop in the form of a transparent rectangle with bolded borders. The virtual grip region is located just below the current position of the gripper and changes its diameter matching the current aperture of the gripper, e.g. when the fingers are opened or closed with an object in between (Figure 3a, 3b). This cue allows the operator to determine if an object is located within the grip region for picking and it presents the occlusion capability as seen in Figure 3c.

**Virtual extensions.** The design of this cue augments the robot's gripper with a virtual elongation of the gripper's fingers —similar to the laser-pointer. However, it is only visible before picking an object, e.g. when there is a danger of potential collision of one or both fingers with an object. It is presented by a virtual red semi-transparent plane that starts at the tip of the finger until it intersects with the area of the object that it would collide (Figure 3a). The purpose of this cue is to alert the operator of potential collisions and the need for repositioning the gripper.

**Ghost object and ghost gripper.** Figure 3b, 3c, and 3d show a virtual copy of the grasped object presented in a gray color. This ghost object appears once an object has been grasped and it follows the operator's gaze. It therefore allows to visualize the future position, pose, and space that an object occupies on a determined area (Figure 3b, 3c) without the need to actually conduct the interaction. The gaze-follow interaction can be activated/deactivated by a virtual button (green bottom button) located around the gripper. When this button is activated it shows the ghost object right under the current position of the gripper. The ghost gripper is related to the ghost object and it is also a virtual representation of the real gripper. It is showed in red, as shown in Figure 3d and appears when the operator points at a position on the tabletop where there would be a risk of some part of the gripper colliding with a neighboring object.

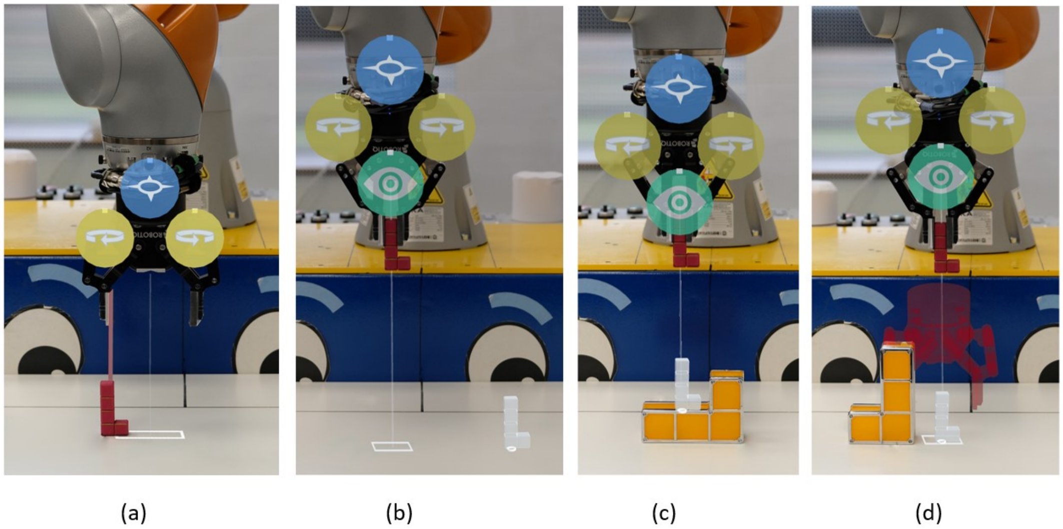


**Figure 3: (a) Virtual Extension and grip region showing occlusion. (b) Ghost object following the cursor. (c) Ghost object recognizes other object's pose and position. (d) Ghost object and ghost gripper.**

# 4 EXPERIMENT

We present an experiment where we test different presentations of AVC (Basic, Advanced, and No Cues) in co-located robot teleoperation to analyze their effect on task performance for manipulation and grasping tasks. While designing the different cues, we realized that Basic and Advanced Cues can affect task performance (time, accuracy, errors, subjective measures) on different levels. Additionally, our measure of accuracy could also hint at improvements in depth perception, as shown in previous experiments by Mather & Smith where depth cues improved accuracy and speed of performance [32].

## 4.1 Hypotheses

H1: We hypothesized that task performance in teleoperation would be the highest when Advanced Cues are used compared to Basic Cues or teleoperation without visual cues, as they explicitly show information of objects in the environment and assist the operator in sensing the environment. Specifically, high accuracy would be achieved with the lowest amount of time and errors across conditions.

H2: We hypothesized that task performance in teleoperation would be higher when Basic Cues are used compared to teleoperation without cues, as having hints that provide some information of the environment are necessary when located at a certain distance of the workspace. Specifically, Basic Cues would have higher accuracy with lesser time and a lower amount of errors than teleoperation without cues.

H3: H3.1. We hypothesized that teleoperation without AVC would cause a high cognitive load, while using Basic or Advanced Cues would have a low cognitive load, as showing visual hints aids to understand better the robot and the environment around it. This avoids the demand that operators must "imagine or guess" the position of the robot gripper and the target objects. H3.2. Additionally, subjective measures of performance would also reflect that Basic and Advanced Cues ease teleoperation with higher levels of usefulness, ease of use and learn, satisfaction, enjoyment and lower levels of concentration.

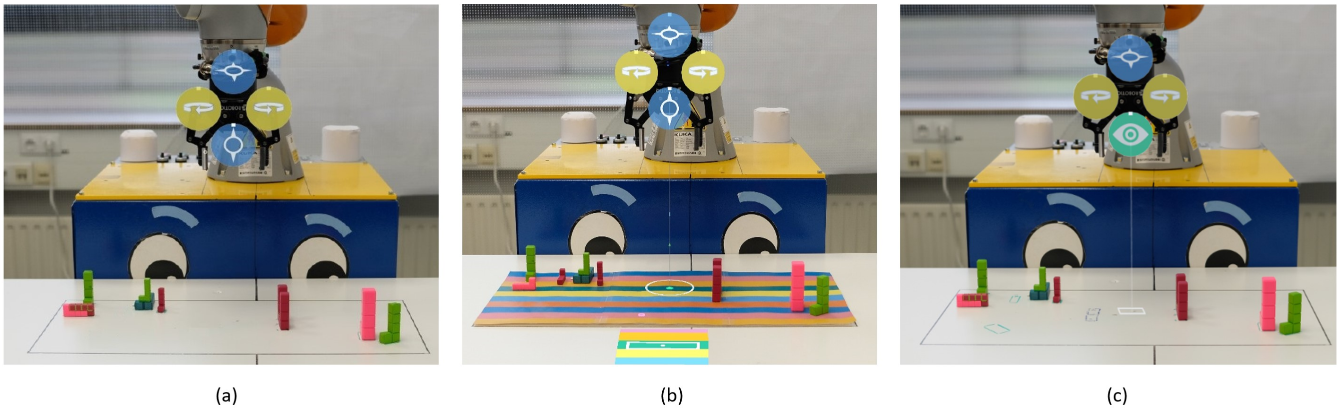
H4: We hypothesized that teleoperating a robotic arm would evoke different views on the presentation of AVC (Basic, Advanced, No Cues). Teleoperating a robot without cues would be perceived negatively, but Basic Cues would prompt neutral views of teleoperation and Advanced Cues would evoke positive views.

## 4.2 Participants

We recruited a total of 36 males and female participants, all of them students or university staff. They were randomly assigned to one of three experimental groups (No Cues, Basic Cues, Advanced Cues). Average age per condition was No Cues 29.42 (SD = 12.44), basic cues 29.25 (SD = 3.94), Advanced Cues 30.83 (SD = 3.38). The pre-test questionnaire revealed that 18 participants had some experience programming robots, and they were evenly distributed across conditions. Also, 15 participants reported to have some experience with the Microsoft HoloLens. Two participants reported to be colorblind —one took part in the Basic Cues condition which presented a color map and reported no problems in distinguishing the real and virtual colors. All participants received 7 euros for their participation.

## 4.3 Experimental Design

We conducted a 3 x 1 between-participants experiment to evaluate how our design of AVC affects task performance in co-located robot teleoperation. Our independent variable is the presentation of AVC (Basic, Advanced, and No Cues). No Cues, i.e., absence of visual cues, served as a baseline condition. In all three conditions, we used the same multimodal interaction (Section 4.7). However, the interface (virtual buttons) was slightly changed to accommodate each design of cues, see Figure 4. In particular, the bottom blue


(a)

(b)

(c)

**Figure 4: Experimental Conditions. (a) No Cues. (b) Basic Cues. (c) Advanced Cues.**

button used in the No Cues and Basic Cues conditions was replaced by a button that shows the ghost object under the current position of the gripper on the tabletop.

We collected the following objective measures to evaluate task performance: time (in seconds) spent positioning the gripper to pick and place an object after pointing to the target position, accuracy in picking and placing (distance in mm to achieve the perfect position, measured by recording the position where the participant grasps or places an object compared to the position where the object is located). Our measure of accuracy is also related to depth perception since it measures the positioning of the gripper above an object, where a higher accuracy would be related to an improvement of distance and depth perception. Also, we collected the number of errors in teleoperation, i.e., participants crashing the gripper against the table or other objects, failing to pick a target object, touching the neighboring objects differently with some part of the gripper, and dropping the target object while picking or placing.

Additionally, we collected subjective measures of task performance by using the NASA Raw-TLX (RTLX) [20] to evaluate workload. This version does not include the weighting of factors and pairwise comparisons. It is a common simplification and has proven to be robust and accurate [19].

In addition, we applied a questionnaire inspired by the USE [29] and Feeling of Flow [16]. We reformulated some items from the questionnaires to fit a human-robot collaborative scenario and evaluated their internal consistency using Cronbach alpha. The dimensions evaluated were enjoyment (4 items, Cronbach $\alpha$ = .729), concentration (4 items, Cronbach $\alpha$ = .762), usefulness (6 items, Cronbach $\alpha$ = .873), ease of use (8 items, Cronbach $\alpha$ = .895), ease of learning (4 items, Cronbach $\alpha$ = .92), and satisfaction (4 items, Cronbach $\alpha$ = .887). In order to understand the participants' views, we asked them to write three words to describe their feelings when controlling the robot and three words to describe the system (see sentiment analysis).

We used a mixed-methods approach to evaluate the data from the experiment. The objective data were analyzed using a one-way ANOVA considering the presentation of AVC as the between-subjects factor. We assessed the normal distribution of our variables using the Shapiro-Wilk test, homogeneity of variances using Levene's Test, and used Bonferroni correction to control Type I errors. Subjective data were analyzed using Kruska-Wallis test with Dunn's post-hoc test using Bonferroni correction. In order to evaluate the participants' views of teleoperation with the different presentations of AVC, we performed a sentiment analysis following a relatively simple approach. We identified the sentiment scores of the three words participants used to describe the system and their feelings. However, not every participant provided three words. The language the participants used was German, and therefore, we used SentiWs, a German-language dictionary for sentiment analysis [40]. If the word was not present in the corpus, the word was not included in the calculations. Once the polarity score had been determined, we calculated the sentiment per participant following a counting method, used in lexicon methods [25]. We reported the total number of sentiments per condition and not the aggregated results to avoid outweighing sentiments, e.g., a positive sentiment with a high score can outweigh two negatives with a low score. In order to look beyond these metrics and better understand the impressions of participants, we combined the sentiment analysis with thematic content analysis following Anderson's approach [3]. As criterion for relevancy, we established that a minimum of 3 different participants should have written the same term or a synonym of it. First, a researcher analyzed and grouped the terms into themes. This analysis was later rated by another researcher.

## 4.4 Task

The experiment consisted of a pick-and-place task with three subtasks and a similar training task to start the interaction. Each subtask (as well as the training task) involved picking and placing one particular green "L-shaped" object (O1, O2, O3), without crashing or touching the neighboring objects. Once the three objects were placed in their target positions the task was concluded, and the process was repeated 2 more times, resulting in 3 trials in total.

The participants wore the Microsoft HoloLens and were seated in front of the table at 160 cm from the workspace. They were instructed not to move or tilt their heads to the side to change their visual angle. This was done explicitly to evaluate the effect of the AVC in the operator's visual perspective.

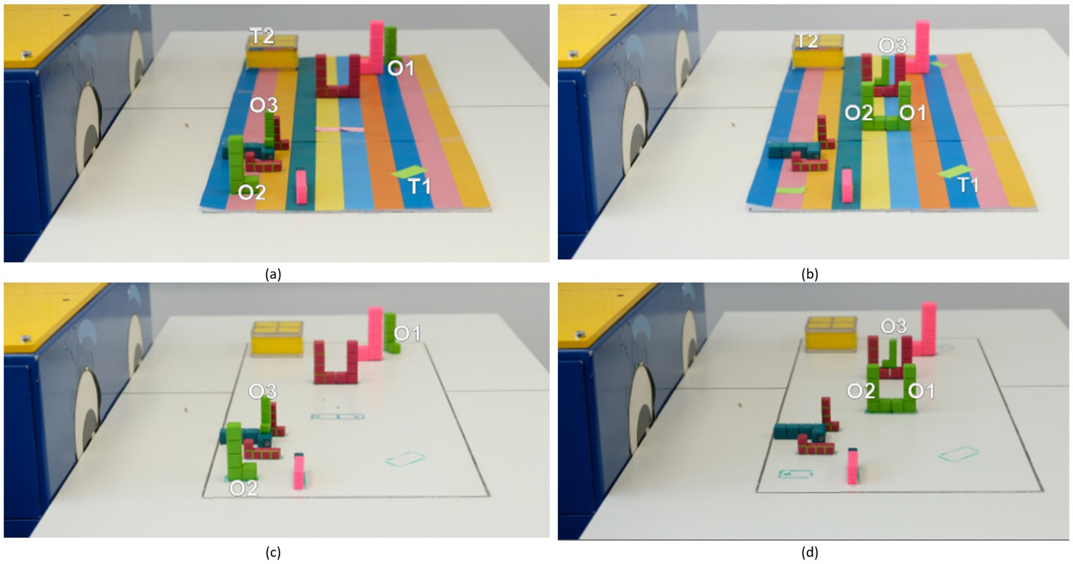


**Figure 5: Side view of the experimental task: the numbers represent the order in which each piece would be picked-and-placed. T1 and T2 relate to the positions for the training tasks. (a) Basic Cues workspace before the task. (b) Basic Cues workspace with the completed task. (c) Advanced Cues and No Cues workspace before the task. (d) Advanced Cues and No Cues workspace with the completed task.**

Figure 5 shows the workspace with "L-shaped" objects in different sizes and colors. Participants only manipulated the green objects and the numbers represent the order in which these needed to be picked and placed. Each object involved a different type of challenge and the target positions for placing were marked on the workspace, as shown in Figure 5b and 5d. O1 was slightly rotated to the side, requiring the operator to rotate the gripper and align its fingers accordingly. Additionally, participants needed to avoid touching or moving the large pink object next to it. The target position of O1 was in the center of the workspace with (at that time) no other objects in the vicinity. O1 still needed to be rotated further to match the marking. O2 was surrounded by a red and pink object of smaller size, posing some risk of collision on two sides but leaving space for movement around the other surrounding areas. On top of that, the pink object partially occluded O2. The target position of O2, from the operator's perspective, was behind the target position of O1 (Figure 5b, 5d) so O1 and O2 formed a U-shaped figure. O1 thereby occluded the target position and posed risk of crashing onto O1. O3 was smaller in size than O1 and O2, and it was located on top of a blue object (change of height during picking) with two red objects of smaller size in the vicinity (risk of collision). The target position of O3 was in between the U-shaped figure formed by two red objects (Figure 5b, 5d). In order to successfully place O3, the gripper needed to be precisely positioned and gave little space for error, so that it did not touch or crash onto the red objects.

## 4.5 Procedure

The experiment lasted approximately 60 minutes and consisted of 5 phases (1) introduction, (2) calibration, (3) familiarization and training phase (5 - 15 minutes), (4) task (7 - 30 minutes), (5) subjective questionnaires and interview.

(1) Participants were given an overview of the devices to be used, the goal of the experiment was explained and they were handed a consent form. (2) Then, they were shown an introductory video on how to wear the HoloLens and proceeded to calibrate it. (3) After that, they were shown a set of short clips regarding the interaction modalities, the interface, and the visual cues to be used. Participants on the No Cues condition were only shown the first two sets of videos. Then, they proceeded to perform the training task. During this phase, participants were instructed to use all the commands and virtual buttons to familiarize themselves with the interface and the interaction itself. Once they felt comfortable with the interface and interaction, we instructed them to take off the HoloLens and take a short break while filling out a background questionnaire. This was performed to avoid fatigue effects derived from extended periods of wearing the HoloLens. (4) Next, participants completed the task with three repetitions. (5) Finally, participants were given post-experiment questionnaires that evaluate their experience, and we conducted a short interview.

## 4.6 Implementation and Devices

We used a Kuka iiwa 7 R800 lightweight robotic arm, designed especially for HRI, with an attached Robotiq adaptive 2-Finger Gripper, both controlled via the Kuka iiwa's control unit. It has a reach of 80 cm, a payload of 7 kg and torque sensors in each joint for a safe collaboration with humans. For the user interface, we used the Microsoft HoloLens 1, which is equipped with inside-out tracking, an HD video camera and microphones that allow the use of head movement, and speech and gestures as input options. The field of view is $30^\circ \times 17.5^\circ$ with a resolution of $1268 \times 720$ px per eye. The application running on the HoloLens was programmed

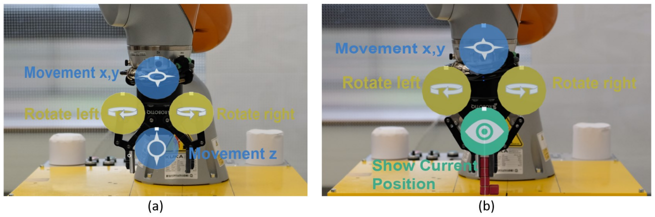


**Figure 6: (a) Robot controls for the basic cues: yellow buttons rotate the gripper, top blue button moves the gripper on x,y axis and bottom blue button moves the gripper on the z axis. (b) Robot controls for the Advanced Cues: yellow buttons rotate the gripper, top blue button moves the gripper on the x,y axis, and the green button shows the current position of the gripper on the tabletop.**

with Unity 2018.1.2 and the HoloToolkit Plugin [35], as well as, the Vuforia Plugin [1], using C#.

The Kuka iiwa and the HoloLens communicate via User Datagram Protocol (UDP) in a local network. The control unit of the Kuka iiwa runs a specially designed backend application that processes the received messages, moves the robot according to the receiving data, and returns to its current status. In order to display the augmentation at the right pose in the real world, we use the Vuforia plugin to detect a marker attached to the robot and align it by placing a virtual copy in the HoloLens' environment at the detected position.

To communicate pose data from the HoloLens to the control unit of the robot, we first converted the pose from Unity's left-handed coordinate system to the robot's right-handed coordinate system and use common length units (mm). Then, this cartesian position, along with rotation and velocity is sent via UDP. The backend recognizes the command and allows the robot to plan the movement via internal inverse kinematics to then moving the robot arm physically.

For a full object-pose recognition, first, a 2D object detection and the 6D pose recognition is necessary. In order to enable this we followed Sundermeyer et al.'s [44] approach. However, we did not actively scan and perform object-pose recognition for the experiment. After several trials, we had a failure of pose recognition at a mean of 7.5% overlap between objects, which was not suitable for our workspace with several objects that occlude each other. Thereby to keep the accuracy of object-pose recognition stable along the experiment, we decided to hardcode the object-pose from the workspace in the system. This avoided having different experiences in the use of Advanced Cues, which depend on the accuracy of object-pose recognition. Further, the analysis of the accuracy of object-pose recognition was not a subject of study in this paper but was the mean to develop our Advanced Cues.

## 4.7 Robot Control and Interaction

This research is part of a larger project in which we also investigate hands-free interaction. Therefore, in order to control the robotic arm for picking and placing tasks, we opted for multimodal hands-free interaction with a combination of head-movements for pointing and speech for committing an action. These two modalities have been proved to fit well together [28], are common interaction modalities in AR/Mixed-Reality (MR) which are natively supported by the HoloLens, and are a natural way to interact when giving commands. As our contribution of this paper lies within the Visual Cues, we kept the interaction modality stable to avoid confounding effects.

Our interface design shows virtual robot controls anchored to the gripper of the robot (Figure 6). This allows operators to have these controls not only within the user's line of sight but also precisely in the current operator's focus. Thereby, an interface anchored to the operator's current focus avoids unrequired attention shifts, which are common when using a control pad or an external screen for teleoperation.

In our scenario, this interaction involves pointing at a position on the tabletop where a target object is located, commanding the robotic arm to move the robot to that position by saying the command "Move", finding an adequate grasping position and adjusting the position of the gripper (using the top blue button, Figure 6a). Once the Operator has determined that the gripper is located at an adequate gripping position, the command "pick" commits the action of grasping, i.e. the gripper moves in the robot's z-axis towards the tabletop, closes its fingers, and moves to the initial position on the z-axis (height) again. Once the object has been successfully grasped, the operator points at another target position, commands the robotic arm to move to that position, adjusts the position, and then places the grasped object in the target position by saying the command "place".

If there is no previous knowledge of the environment, the operator needs to manually control and adjust the gripper's height (z-axis, up/down) to find the adequate grasping position. This is achieved using the blue bottom button as seen in Figure 6a. Also, the commands to pick and place objects function slightly differently when using object-pose recognition. When the workspace and the objects around are known by the system, the gripper can move automatically towards the target, after pointing to the object and repositioning on the x,y plane. The operator then says the command "pick" or "place" and the robot will move towards the target object, close the fingers and move to the initial height (z-axis).

On top of the commands, we added two sets of virtual buttons anchored to the real gripper, which are activated by dwelling (Figure 6). The yellow buttons rotate the robot's gripper at a pace of 10 degrees per second and allow the operator to ensure that the fingers of the gripper match the affordances of the object to grasp it. The top blue button controls the gripper's movements on the robot's x,y axis (right, left, back, front). It follows the operator's head movements and is intended to be used for fine adjustment of the position of the gripper. It can move from 0 to 800 mm/s following the operator's head movements, and the operator controls the speed by directing the head gaze further from the gripper for faster movements, and closer to the gripper for slow movements, e.g., if the operator gazes at the corners of the tabletop, the gripper will move faster, and it will move slower when the operator gazes to the sides of the gripper. The bottom blue button is only present when there is no knowledge of the environment (Basic Cues), and it controls the gripper movements on the z-axis (up/down) to manually adjust the height when approaching an object on the tabletop. It behaves in a similar manner as the x,y movement button and follows the operator's head movements. The operator can also control the speed by directing the gaze further or closer to the gripper, e.g., the gripper moves faster if the operator gazes below the table or 30 cm above the gripper, and it moves slower if the operator gazes somewhere close to the center of the gripper. When a button is activated the rest of buttons are hidden to avoid the Midas-touch problems [24], especially when the robotic arm is following head movements. This bottom blue button is replaced by a green button in the Advanced Cues. Figure 6b shows the ghost object at the current position of the gripper.

## 4.8 Results

*4.8.1 Objective Measures.* We analyzed our data with a One-Way Anova with type of cues (No cues, Basic Cues, Advanced Cues) as the between-subjects factor and time, accuracy, and errors as dependent variables.

For time, the data were normally distributed, as assessed by the Shapiro-Wilk test ($\alpha$ = .05). Homogeneity of variances was asserted using Levene's Test which showed that equal variances could be assumed (p = .124). The mean task-time was statistically significant for the different levels of presentation of cues, Figure 7, F(2, 33) = 13.62, p < .001, $\eta p^2$ = .452. Post-hoc pairwise comparisons (Bonferroni adjusted) show that the time spent for picking and placing decreased significantly with Advanced Cues (M=29.35 s, SD=8.67s) when compared to Basic Cues (M=62.02 s, SD=22.70), (p < .001) and to the No Cues condition (M=57.86 s, SD=15.66), (p = .001) but no significant difference was found between the No Cues and Basic Cues condition.

For accuracy, the data were normally distributed, as assessed by the Shapiro-Wilk test ($\alpha$ = .05). Homogeneity of variances was asserted using Levene's Test, which showed that equal variances could be assumed (p = .414). The mean accuracy differed statistically for the different levels of type of cues, Figure 7, F(2,33) = 6.61, p = .004, $\eta p^2$ = .29. Post-hoc pairwise comparisons, applying Bonferroni correction, revealed a statistically significant improvement in accuracy when using Advanced Cues (M=10.97 mm, SD=1.84) compared to No Cues (M=13.54 mm, SD=2.70), (p = .021), and a significant improvement when using Basic Cues (M=10.54 mm, SD=1.89), (p = .006) compared to No Cues. No significant difference was found between Basic Cues and Advanced Cues.

Figure 7 shows a plot of the amount of time spent vs accuracy. Here, we can see the improvement in time and accuracy, especially for the Advanced Cues.

For the number of errors, the data was normally distributed for Basic Cues and Advanced Cues, but not for the No Cues condition, as assessed by the Shapiro-Wilk test ($\alpha$ = .05). As a result, we applied a non-parametric test statistic on the data, the Kruskal-Wallis, followed by pairwise comparisons with the Dunn's test. The Kruskal-Wallis test, which was corrected for tied ranks, was significant $\chi^2$ (2, N = 36) = 11.10, p = .004. For pairwise comparisons, the post-hoc Dunn's test showed significant differences for Advanced Cues (N=12, Md=3) compared to Basic Cues (N=12, Md=6) with p=0.008, and compared to No Cues (N=12, Md=6), with p=0.017 (Bonferroni corrected). However, no significant difference was found between No Cues and Basic Cues.

**Table 1: Results of Subjective Measures per AVC**

| | No Cues | | Basic Cues | | Adv. Cues | | | |
|---|---|---|---|---|---|---|---|---|
| | M | SD | M | SD | M | SD | $\chi^2(2)$ | p |
| Usefulness | 3.18 | 1.50 | 5.08 | 1.01 | 5.39 | 0.67 | 10.88 | .004 |
| Ease of Use | 4.38 | 1.51 | 4.85 | 1.09 | 5.47 | .92 | 3.63 | .16 |
| Ease of Learning | 4.94 | 1.78 | 6 | .94 | 6.33 | .54 | 4.47 | .11 |
| Satisfaction | 4.79 | 1.51 | 5.56 | .64 | 5.9 | .61 | 4.56 | .1 |
| Enjoyment | 5.96 | .82 | 6.06 | .69 | 6.33 | .59 | 1.68 | .43 |
| Concentration | 5.75 | .8 | 6.19 | .74 | 6.21 | .6 | 2.61 | .27 |

**Table 2: Sentiment Analysis Results**

| | System | | Feelings | |
|---|---|---|---|---|
| | No. Positive | No. Negative | No. Positive | No. Negative |
| No Cues | 14 | 6 | 9 | 13 |
| Basic Cues | 16 | 5 | 10 | 10 |
| Advanced Cues | 21 | 2 | 18 | 6 |

*4.8.2 Subjective Measures.* We did not find significant differences of our designs of AVC after performing the Kruskal-Wallis test in ease of use, ease of learning, satisfaction, enjoyment, and concentration, see Table 1. However, we found significant difference in usefulness ($\chi^2(2)$ = 10.88, p = .004). Post-hoc Dunn's test for pairwise comparisons, applying Bonferroni correction, showed a higher perception of usefulness when teleoperating the robot using Advanced Cues (p = .005) compared to teleoperation with No Cues, and teleoperation using Basic Cues (p = .048) compared to teleoperation with No Cues. No significant differences were found between Advanced and Basic Cues.

After evaluating the NASA RTLX using the Kruskal-Wallis test, we found significant differences between the different designs of AVC only in the performance factor ($\chi^2(2)$ = 7.93, p = .019); see Figure 8. Post-hoc Dunn's Test for pairwise comparisons, applying Bonferroni correction, revealed a significant difference only when comparing Basic Cues (p = .026) to the No Cues condition.

*4.8.3 Sentiment Analysis.* We used sentiment analysis to evaluate the participants' views. In order to enable that, we considered the words participants used to describe the system and their feelings with different presentations of AVC. We found that in general, the system evoked a positive sentiment across conditions. However, we identified nuances in their feelings while performing the experiment dependent on the conditions each group of participants experienced (Table 2). Here, the No Cues condition prompted a negative sentiment which is supported by some comments by the participants during the interview, e.g. P15, "Something with the control system was counter-intuitive;" P17, "There is a visual feedback missing;" P20, "You have certain spatial imagination but that was not enough to find the correct position." Participants who performed the experiment using Basic Cues had different sentiments regarding their feelings while teleoperating the robot; half of them expressed positive sentiments and the other half expressed rather negative

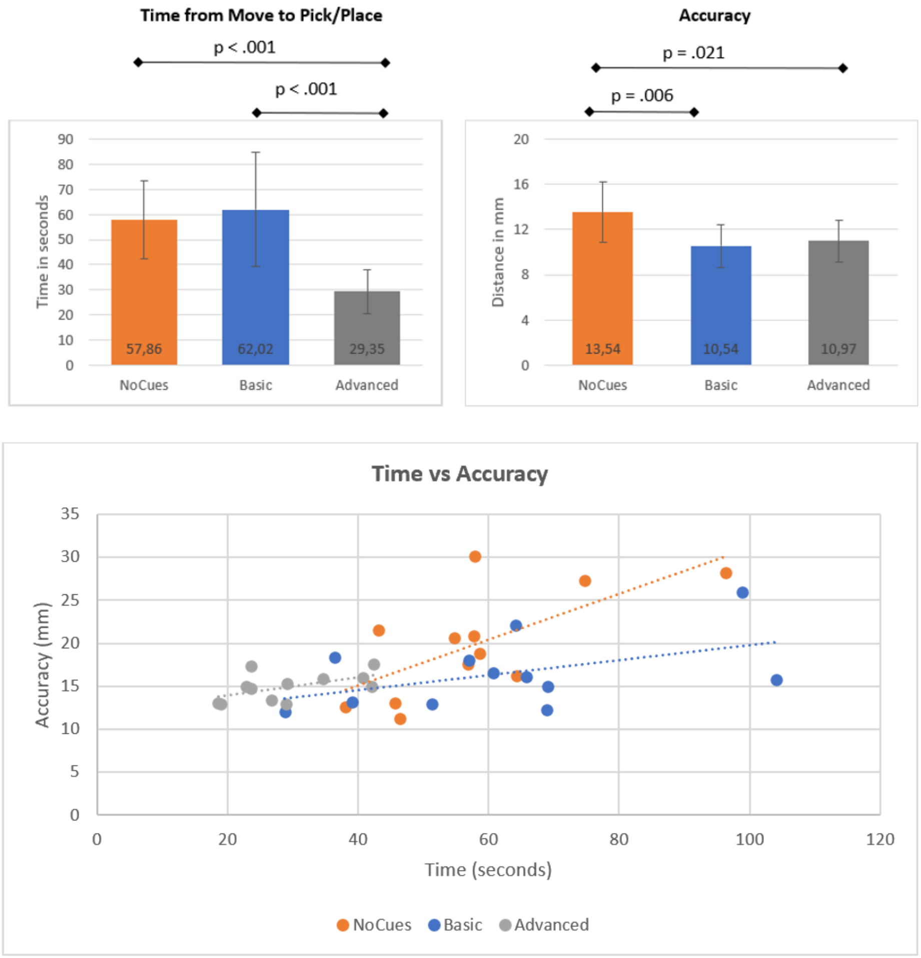


Figure 7: Top: Task performance measures of time and accuracy. Bottom: Plot chart of time vs accuracy

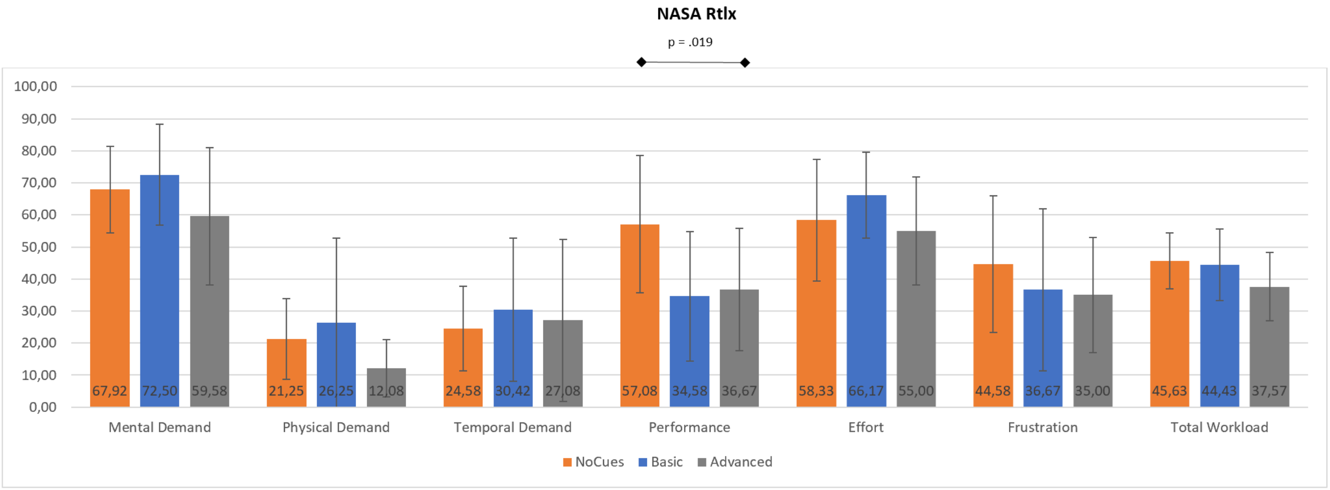


Figure 8: NASA Rtlx of Perceived Workload under the three conditions.

sentiments. Therefore, a further analysis is needed to understand the reasons behind this better (Section 4.8.4). The Advanced Cues aroused positive sentiments without any negative sentiment about teleoperation. Consequently, most participants expressed a positive sentiment about how they felt while teleoperating the robot using Advanced Cues.

*4.8.4 Thematic Content Analysis.* We considered it relevant to perform thematic content analysis since it could reveal underlying impressions of Basic and Advanced Cues. Hence, after the sentiment analysis, we looked for possible themes, and identified three that captured participants' impressions of both designs of AVC, and each design.

**Both designs of AVC were considered innovative, interesting, and intuitive.** The participant mentioned terms like "innovative" (3 times), "interesting" (3 times), and "intuitive" (4 times) on our Basic and Advanced Cues conditions. Teleoperating a robot using AR to augment the robot and its workspace was something novel to most of our participants, consequently evoking a sense of innovation. Despite having some participants with previous experiences with robots and AR/VR but not particularly robotic arms or teleoperation, it seemed that this combination (teleoperation with AR) was perceived as innovative. Related to innovation, participants referred to our system as "interesting" which might be also a consequence of this innovation and novelty factor, as presented in previous findings [10, 37]. Moreover, participants described our system as "intuitive". We relate this to our design of visual cues. P26 mentioned, "The ghost object was very easy to understand, as well as the red ghost gripper that appeared and showed collisions". However, P22, expressed, "It took me a while to trust the robot and the visuals. Once I did, it worked better for me." Also, we design our cues to be minimalistic and avoid overloading the operator's visual space.

**Basic Cues were referred to as futuristic and complex.** Under this condition participants used terms like "futuristic" (3 times) and complex (3 times). Most participants do not use robots in their day-to-day life and hence, teleoperating robotic arms with AR can still be perceived as something that is not quite in the present but rather forward-looking. Additionally, a complexity factor was mentioned; we attribute this to the design of Basic Cues since it had a rather large number of visual hints (6) visible at all times. This was also reflected when comparing Basic to Advanced Cues on our objective measures of task performance, e.g., a longer task time and higher number of errors. This hinted that for some participants, it was difficult to actively use all our Basic Cues, which could have led to ignoring some hints, e.g., P6 expressed, "You need to concentrate fully on moving the gripper, so you are looking at the gripper, and then also concentrate on the colors and that is hard." Participants' comments on Basic Cues were mostly directed towards the combination of virtual and real color maps. In fact, these were considered the most helpful hints by 9 out of 12 participants from the Basic Cues condition, e.g., as P3 said, "Without the color map it would be impossible to perceive depth." P22 stated, "The real color map and the virtual rectangle showing the color where the gripper was located were the most helpful."

**Advanced Cues were regarded as visionary.** Participants mentioned mostly a visionary aspect (5 times) while using Advanced Cues. We attribute this to the "ghost object and gripper" hint and the occlusion capability added to our cues. Being able to see different future states of an object and determine what would be the best position based on those states adds a visionary aspect to the interaction. Here, the ghost object was mentioned by half of the participants as the most helpful Advanced Cue, e.g., P10 said , "The ghost object was the most helpful." This was followed by the laser-pointer (5 participants), about which P29 expressed, "The visualization of the object (ghost object) and the laser-pointer were the most helpful." Finally, the grip region was also mentioned by 4 participants, e.g., P8 said, "The visualization of the gripping area was really helpful for picking and for placing, as well as the ghost object." Participants rarely referred to only one hint from the Advanced Cues as helpful; it was mostly to a combination of them. A possible explanation is that they showed only two different cues at a time.

## 5 DISCUSSION

In this section, we compare our hypotheses to the results obtained from the experiment. First, we will discuss in detail our hypotheses (H1, H2) related to the objective measures of task performance. Further, we will discuss our subjective measures (H3) together with the qualitative evaluation from our sentiment and thematic content analysis (H4).

In **H1**, we hypothesized that teleoperation with Advanced Cues will have the highest task performance. This should be reflected by the highest accuracy, the shortest amount of time, and the lowest number of errors across conditions. Our results partially support H1. These indeed show that teleoperation with Advanced Cues improved task performance compared to teleoperation without any cues. However, when comparing teleoperation with Advanced Cues to teleoperation with Basic Cues, we only find significant improvements in terms of time and number of errors (not in accuracy). These results suggest that both designs of AVC improve and support accuracy in teleoperation similarly. In fact, improvements in accuracy in task performance suggest having a close relation to depth perception [32]. Related to these results, we further question if this trade-off of time to favor accuracy affects the perception of the interaction. When looking closely at the performance factor of cognitive load, we found similar performance perception between Basic and Advanced Cues. This could indicate that having a high level of accuracy influences self-evaluation of performance. Still, when aiming to optimize pick-and-place tasks in terms of time and errors, Advanced Cues have shown to have a lower number of errors and a shorter task execution.

Regarding **H2**, we hypothesized that teleoperation with Basic Cues would improve task performance in terms of time, accuracy, and the number of errors, compared to teleoperation without cues. We can partially support this hypothesis. Our results show similar performance in amount of time and errors, but a significant improvement in terms of accuracy. This leads us to the conclusion that in the No Cues condition, people were able to trade-off time and errors for reduced accuracy. From the perspective of the Basic Cues design, we also believe that a certain amount of visual clutter due to the number of hints presented could have affected time and errors. We showed 6 different cues that were always visible. Nonetheless, during the interview, participants only referred to 3 of them as most helpful. It may be possible that participants might have tried to use all of them at the beginning but eventually ignored some of the hints as the experiment progressed. Our thematic content analysis also supported this line of thinking, by our Basic Cues being referred to as "complex" (Section 4.8.4). We highlight that in a real-world setting, a physical colormap requires changes in the workspace. We evaluated one type of landmark (combination of the virtual and physical colormap) that proved to increase the accuracy and usefulness in task performance. This can make this change on the physical workspace worthwhile. Yet, other types of landmarks that blend-in with the physical space might provide similar benefits.

Our results from H1 and H2 build upon similar results from Brizzi et al. [4] who showed how presenting visual cues through AR improved the accuracy and efficiency in task performance in pick-and-place tasks. Also, our findings align partially with those of [31] in 2D images, where a higher number of depth cues improved accuracy in a shorter time. We indeed found improvements in accuracy and time for the Advanced Cues but a time trade-off for the basic cues. This difference might be a result of the 3D environment and the design of our cues. Still, this time trade-off is in line with previous research about visual search in complex environments, where the number of items in the visual space causes extra effort (longer time) in discriminating targets [39].

**H3** regards our subjective measures. In H3.1., we hypothesized about a lower cognitive load of both designs of AVC. Our results partially support this hypothesis. We only found a significant difference in the perceived performance factor of the Nasa Rtlx when comparing Basic Cues to the No Cues condition. We believe that improvements in accuracy is also important for people's subjective assessment of their performance. H3.2 was also partially supported. Our results show only the significant differences in the perception of the usefulness of Basic and Advanced Cues compared to the No Cues condition. Our results did not show other significant differences in the rest of our subjective measures. A possible reasoning behind this can be the level of precision required for the task, added to the fact that most participants had never operated a robotic arm before. Expertise effects have been found to influence performance in teleoperation for pick-and-place tasks [4]. However, these can be leveled out when providing feedback through AR. Additionally, the metrics used could also have influenced these results. To better understand the subjective factor of our presentation of AVC, we explored the views of participants through sentiment and thematic content analysis.

In **H4**, we hypothesized about different views evoked by the presentation of our cues. A sentiment analysis supported this hypothesis in terms of the feelings expressed by the participants. Although participants were not able to compare the different presentations of AVC (because of a between-subjects experimental design), these were perceived differently and evoked contrasting views. The No Cues condition evoked a negative sentiment among participants, our Basic Cues aroused the same number of positive and negative sentiments, while the Advanced Cues gathered mostly positive sentiments. Moreover, our thematic content analysis tried to reveal hidden views of our designs of AVC. The positive sentiment is manifested with both designs of AVC as they were referred to as interesting, innovative, and intuitive. However, we are cautious with this result since the novelty factor might have played a definitive role. We consider the fact that participants referred to the Basic Cues as complex interesting. As discussed in H3, we attribute this to the number of cues presented to participants which could have hindered the positive sentiment that they could have evoked. Finally, our design of Advanced Cues was regarded as "visionary", which is also reflected by the high number of positive sentiments. This leads us to think that using previous knowledge of the environment to present explicit visual cues has a positive impact not only on objective performance but also on how the task and teleoperation is perceived.

## 6 LIMITATIONS

We acknowledge that our results might have limited generalizability due to our experimental design and our metrics. We decided to modify the existing subjective questionnaires (USE, Feeling of Flow) since they were constructed based on human-computer interaction systems and they did not completely transfer to HRI, e.g., teleoperating robotic arms and the use of AR. We recognize that different subjective questionnaires could have been used. However, we performed a sentiment and thematic content analysis that further explored the subjective aspect of the presentation of AVC.

Another limitation comes in terms of the type of interaction used. We recognize that multimodal hands-free interaction is a particular choice for robot teleoperation. However, we kept the interaction stable among conditions to minimize interaction effects and did not identify any major influence of the type of interaction in the evaluation of AVC. None of our participants reported problems with the general control of the robot that could be ascribed to the interaction modality. Head-gaze and speech are common interaction modalities in AR and MR and were natively supported by the technology we used, i.e., the Microsoft HoloLens 1. Additionally, this multimodal interaction modality is much simpler than learning to operate a robot control pad, which requires extensive training. On a different note, this work is particularly relevant for the growing body of research for hands-free interaction in HRI [28], [36].

The study is limited to the current AR technology, i.e., HoloLens 1. The tracking capabilities of the device are not precise enough to correctly allow for real-world objects to occlude the virtual ones. Besides, issues such as disparity planes and focal rivalry might arise prominently, as noted by Kruijff et al. [27], resulting in uncomfortable focus switches, although the location of virtual and real objects might be the same. Future or different AR technology might improve some of these issues. For example, see-through based AR (Varjo XR1) might reduce the visual clutter when virtual cues overlay or occlude real world objects, but it would need to be tested since it might arise other type of problems.

## 7 CONCLUSION

In this paper, we explored the design space of AR to enhance the operator's perception of the robot and its environment. We designed an interface with basic robot controls and used a multimodal interaction approach for teleoperation. This was used in combination with two designs of AVC (Basic and Advanced Cues) with the goal of providing a better understanding of depth and distances and thus improving task performance. In our first design, Basic Cues, we augmented the robot and the environment by providing a combination of real and virtual indirect hints. In our second design, Advanced Cues, we utilized previous knowledge of the environment —using object-pose recognition, to present explicit virtual hints that enhance the robot, the environment, and the objects in it. We evaluated our designs against a baseline (No Cues) in a user study where participants teleoperated a robotic arm to perform pick-and-place tasks that required some level of precision.

We hypothesized about the improvement of task performance at different levels in objective and subjective measures using both designs of AVC. Our findings show that both of our designs of AVC improve task performance in terms of accuracy, which is closely related to distance and depth perception compared to the baseline. However, our design of Advanced Cues also presents improvements in terms of time and number of errors. Our subjective measures hinted at nuances in participants' perception of self-performance and the usefulness of both designs of AVC. That is why we further explored the participants' views through sentiment analysis. Both designs of AVC gathered positive sentiments about the system. Nonetheless, we found differences in participants' feelings towards teleoperating a robot with Basic and Advanced Cues. In particular, our Advanced Cues evoked a positive sentiment while the Basic Cues aroused rather mixed sentiments. A possible reason behind this can be found in the results of our thematic content analysis. Here, we identified that our design of Basic Cues was regarded as complex, which we attribute to the number of hints presented to the operator.

All in all, our results show that both Basic and Advanced Cues offer a promising approach to enhance the visual space. They help to understand the relation between the environment and the robot in the workspace, thus, assisting the teleoperation of a robotic arm and increasing accuracy.

In future work, we will continue to explore the use of visual cues to enhance the operator's awareness of the workspace. We intend to go beyond improving depth perception and task performance and provide implicit visual cues that will help the operator to determine the changes and the potential problems derived from the interaction, e.g., objects falling, out-of-view objects, and partially occluded objects. A novel line of research has investigated robot deictic gestures used by social robots to provide instructions [54]. We consider it closely related to our augmented visual cues and we intend to extend this line of research in manufacturing environments with industrial robots.

## ACKNOWLEDGMENTS

We would like to thank in particular our co-authors Franziska Rücker and Tim Dierks. Their tremendous work during their master studies made this paper possible.

This research is supported by the German Federal Ministry of Education and Research (BMBF, FKZ: 13FH011IX6).

## REFERENCES

[1] 07/09/2020. SDK Download | Vuforia Developer Portal. https://developer.vuforia.com/downloads/sdk

[2] 2011. Appendix A: Human Visual Perception. In *Practical Image and Video Processing Using MATLAB®*, Oge Marques (Ed.). John Wiley & Sons, Inc, Hoboken, NJ, USA, 591–610. https://doi.org/10.1002/9781118093467.app1

[3] Rosemarie Anderson. 2007. Thematic content analysis (TCA). , 4 pages.

[4] Filippo Brizzi, Lorenzo Peppoloni, Alessandro Graziano, Erika Di Stefano, Carlo Alberto Avizzano, and Emanuele Ruffaldi. 2018. Effects of Augmented Reality on the Performance of Teleoperated Industrial Assembly Tasks in a Robotic Embodiment. *IEEE Transactions on Human-Machine Systems* 48, 2 (2018), 197–206. https://doi.org/10.1109/THMS.2017.2782490

[5] Ravi Teja Chadalavada, Henrik Andreasson, Robert Krug, and Achim J. Lilienthal. 2015. That's on my mind! robot to human intention communication through on-board projection on shared floor space. In *2015 European Conference on Mobile Robots*, European Conference on Mobile Robots and Tom Duckett (Eds.). IEEE, Piscataway, NJ, 1–6. https://doi.org/10.1109/ECMR.2015.7403771

[6] E. Laurence Chalmers. 1952. Monocular and Binocular Cues in the Perception of Size and Distance. *The American Journal of Psychology* 65, 3 (1952), 415. https://doi.org/10.2307/1418762

[7] J.W.S. Chong, S. K. Ong, A.Y.C. Nee, and K. Youcef-Youmi. 2009. Robot programming using augmented reality: An interactive method for planning collision-free paths. *Robotics and Computer-Integrated Manufacturing* 25, 3 (2009), 689–701. https://doi.org/10.1016/j.rcim.2008.05.002

[8] Francesco Clemente, Strahinja Dosen, Luca Lonini, Marko Markovic, Dario Farina, and Christian Cipriani. 2017. Humans Can Integrate Augmented Reality Feedback in Their Sensorimotor Control of a Robotic Hand. *IEEE Transactions on Human-Machine Systems* 47, 4 (2017), 583–589. https://doi.org/10.1109/THMS.2016.2611998

[9] James E. Cutting and Peter M. Vishton. 1995. Perceiving Layout and Knowing Distances. In *Perception of Space and Motion*. Elsevier, 69–117. https://doi.org/10.1016/B978-012240530-3/50005-5

[10] Andreas Dünser, Raphaël Grasset, Hartmut Seichter, and Mark Billinghurst. 2007. Applying HCI principles to AR systems design. (2007).

[11] Fatima El Jamiy and Ronald Marsh. 2019. Survey on depth perception in head mounted displays: distance estimation in virtual reality, augmented reality, and mixed reality. *IET Image Processing* 13, 5 (2019), 707–712. https://doi.org/10.1049/iet-ipr.2018.5920

[12] H. C. Fang, S. K. Ong, and A.Y.C. Nee. 2012. Interactive robot trajectory planning and simulation using Augmented Reality. *Robotics and Computer-Integrated Manufacturing* 28, 2 (2012), 227–237. https://doi.org/10.1016/j.rcim.2011.09.003

[13] Markus Funk, Andreas Bächler, Liane Bächler, Thomas Kosch, Thomas Heidenreich, and Albrecht Schmidt. 2017. Working with Augmented Reality?. In *PETRA 2017 (ICPS: ACM international conference proceeding series)*, Unknown (Ed.). ACM, New York, NY, USA, 222–229. https://doi.org/10.1145/3056540.3056548

[14] Hind Gacem, Gilles Bailly, James Eagan, and Eric Lecolinet. 2015. Finding Objects Faster in Dense Environments Using a Projection Augmented Robotic Arm. In *Human-computer interaction - INTERACT 2015*, Julio Abascal (Ed.). LNCS sublibrary. SL 3, Information systems and applications, incl. Internet/Web, and HCI, Vol. 9298. Springer, Cham, 221–238. https://doi.org/10.1007/978-3-319-22698-9{_}15

[15] Yiannis Gatsoulis, Gurvinder S. Virk, and Abbas A. Dehghani-Sanij. 2010. On the Measurement of Situation Awareness for Effective Human-Robot Interaction in Teleoperated Systems. *Journal of Cognitive Engineering and Decision Making* 4, 1 (2010), 69–98. https://doi.org/10.1518/155534310X495591

[16] Jawaid A. Ghani and Satish P. Deshpande. 1994. Task Characteristics and the Experience of Optimal Flow in Human—Computer Interaction. *The Journal of Psychology* 128, 4 (1994), 381–391. https://doi.org/10.1080/00223980.1994.9712742

[17] L. L. Gong, S. K. Ong, and A. Y. C. Nee. 2019. Projection-based Augmented Reality Interface for Robot Grasping Tasks. In *Proceedings of the 2019 4th International Conference on Robotics, Control and Automation - ICRCA 2019*, Unknown (Ed.). ACM Press, New York, New York, USA, 100–104. https://doi.org/10.1145/3351180.3351204

[18] Scott A. Green, Mark Billinghurst, XiaoQi Chen, and J. Geoffrey Chase. 2008. Human-Robot Collaboration: A Literature Review and Augmented Reality Approach in Design. *International Journal of Advanced Robotic Systems* 5, 1 (2008), 1. https://doi.org/10.5772/5664

[19] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (2006), 904–908. https://doi.org/10.1177/154193120605000909

[20] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human mental workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, Amsterdam and Oxford, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[21] Hooman Hedayati, Michael Walker, and Daniel Szafir. 2018. Improving Collocated Robot Teleoperation with Augmented Reality. In *HRI'18*, Takayuki Kanda, Selma Ŝabanović, Guy Hoffman, and Adriana Tapus (Eds.). Association for Computing Machinery, New York, New York, 78–86. https://doi.org/10.1145/3171221.3171251

[22] I. P. Howard. 2012. *Perceiving in Depth, Volume 1: Basic Mechanisms*. Oxford University Press. https://books.google.de/books?id=A26JAgAAQBAJ

[23] J. A. Gomer, C. H. Dash, K. S. Moore, and C. C. Pagano. 2009. Using Radial Outflow to Provide Depth Information During Teleoperation. *Presence: Teleoperators and Virtual Environments* 18, 4 (2009), 304–320. https://doi.org/10.1162/pres.18.4.304

[24] Robert J. K. Jacob. 1995. Eye Tracking in Advanced Interface Design. In *Virtual Environments and Advanced Interface Design*. Oxford University Press, Inc, USA, 258–288.

[25] Alistair Kennedy and Diana Inkpen. 2006. SENTIMENT CLASSIFICATION of MOVIE REVIEWS USING CONTEXTUAL VALENCE SHIFTERS. *Computational Intelligence* 22, 2 (2006), 110–125. https://doi.org/10.1111/j.1467-8640.2006.00277.x

[26] R. Kitchin. 2001. Cognitive Maps. In *International encyclopedia of the social & behavioral sciences*, Neil J. edt Smelser and Paul B. edt Baltes (Eds.). Elsevier, Amsterdam and New York, 2120–2124. https://doi.org/10.1016/B0-08-043076-7/02531-6

[27] Ernst Kruijff, J. Edward Swan, and Steven Feiner. 2010. Perceptual issues in augmented reality revisited. In *9th IEEE International Symposium on Mixed and Augmented Reality (ISMAR), 2010*, Tobias Höllerer (Ed.). IEEE, Piscataway, NJ, 3–12. https://doi.org/10.1109/ISMAR.2010.5643530

[28] Dennis Krupke, Frank Steinicke, Paul Lubos, Yannick Jonetzko, Michael Gorner, and Jianwei Zhang. 10/1/2018 - 10/5/2018. Comparison of Multimodal Heading and Pointing Gestures for Co-Located Mixed Reality Human-Robot Interaction. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 1–9. https://doi.org/10.1109/IROS.2018.8594043

[29] Arnold M. Lund. 2001. Measuring usability with the use questionnaire12. *Usability Interface* 8, 2 (2001), 3–6.

[30] Magic Leap. [n.d.]. Magic Leap One. https://www.magicleap.com/

[31] Zhanat Makhataeva and Huseyin Varol. 2020. Augmented Reality for Robotics: A Review. *Robotics* 9, 2 (2020), 21. https://doi.org/10.3390/robotics9020021

[32] George Mather and David R. R. Smith. 2004. Combining depth cues: effects upon accuracy and speed of performance in a depth-ordering task. *Vision Research* 44, 6 (2004), 557–562. https://doi.org/10.1016/j.visres.2003.09.036

[33] M. Alejandra Menchaca-Brandan, Andrew M. Liu, Charles M. Oman, and Alan Natapoff. 2007. Influence of perspective-taking and mental rotation abilities in space teleoperation. In *Proceedings of the 2007 ACM*, Cynthia Breazeal, Alan C. Schultz, Terry Fong, and Sara Kiesler (Eds.). ACM, New York, 271. https://doi.org/10.1145/1228716.1228753

[34] Microsoft. [n.d.]. Microsoft Hololens. https://docs.microsoft.com/en-us/hololens/hololens1-hardware

[35] Microsoft‘. 2017. Mixed Reality Toolkit. https://github.com/microsoft/MixedRealityToolkit-Unity/releases

[36] Susanna Nilsson, Torbjörn Gustafsson, and Per Carleberg. 2007. Hands Free Interaction with Virtual Information in a Real Environment: Eye Gaze as an Interaction Tool in an Augmented Reality System. *Proceedings of COGAIN* (2007), 53–57.

[37] Thomas Olsson and Markus Salo. 2012. Narratives of satisfying and unsatisfying experiences of current mobile augmented reality applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (ACM Digital Library)*, Joseph A. Konstan (Ed.). ACM, New York, NY, 2779. https://doi.org/10.1145/2207676.2208677

[38] Yoon Jung Park, Hyocheol Ro, and Tack-Don Han. 2019. Deep-ChildAR bot. In *ACM SIGGRAPH 2019 Posters*. Association for Computing Machinery, [S.l.], 1–2. https://doi.org/10.1145/3306214.3338589

[39] Rajesh P.N. Rao, Gregory J. Zelinsky, Mary M. Hayhoe, and Dana H. Ballard. 2002. Eye movements in iconic visual search. *Vision Research* 42, 11 (2002), 1447–1463. https://doi.org/10.1016/S0042-6989(02)00040-8

[40] Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. SentiWS - A Publicly Available German-language Resource for Sentiment Analysis.

[41] S. Stadler, K. Kain, M. Giuliani, N. Mirnig, G. Stollnberger, and M. Tscheligi. 2016. Augmented reality for industrial robot programmers: Workload analysis for task-based, augmented reality-supported robot control. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 179–184. https://doi.org/10.1109/ROMAN.2016.7745108

[42] Bennett L. Schwartz and John H. Krantz. 2019. *Sensation & perception* (second edition ed.). SAGE, Los Angeles.

[43] Jose Aparecido Da Silva. 1985. Scales for Perceived Egocentric Distance in a Large Open Field: Comparison of Three Psychophysical Methods. *The American Journal of Psychology* 98, 1 (1985), 119. https://doi.org/10.2307/1422771

[44] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. [n.d.]. Implicit 3D Orientation Learning for 6D Object Detection from RGB Images. https://arxiv.org/pdf/1902.01275

[45] C. T. Swain. 1996. Integration of monocular cues to create depth effect. In *1997 IEEE international conference on acoustics, speech, and signal processing*. IEEE Comput. Soc. Press, 2745–2748. https://doi.org/10.1109/ICASSP.1997.595357

[46] Laurence P. Tidbury, Kevin R. Brooks, Anna R. O'Connor, and Sophie M. Wuerger. 2016. A Systematic Comparison of Static and Dynamic Cues for Depth Perception. *Investigative ophthalmology & visual science* 57, 8 (2016), 3545–3553. https://doi.org/10.1167/iovs.15-18104

[47] Michael Walker, Hooman Hedayati, Jennifer Lee, and Daniel Szafir. 2018. Communicating Robot Motion Intent with Augmented Reality. In *HRI'18*, Takayuki Kanda, Selma Ŝabanović, Guy Hoffman, and Adriana Tapus (Eds.). Association for Computing Machinery, New York, New York, 316–324. https://doi.org/10.1145/3171221.3171253

[48] Michael E. Walker, Hooman Hedayati, and Daniel Szafir. 3/11/2019 - 3/14/2019. Robot Teleoperation with Augmented Reality Virtual Surrogates. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 202–210. https://doi.org/10.1109/HRI.2019.8673306

[49] Stuart Walker and Ed Dorsa. 2001. Making Design Work: Sustainability, Product Design and Social Equity. *The Journal of Sustainable Product Design* 1, 1 (2001), 41–48. https://doi.org/10.1023/A:1014412307092

[50] X. Wang, S. K. Ong, and A. Y. C. Nee. 2016. A comprehensive survey of augmented reality assembly research. *Advances in Manufacturing* 4, 1 (2016), 1–22. https://doi.org/10.1007/s40436-015-0131-4

[51] Atsushi Watanabe, Tetsushi Ikeda, Yoichi Morales, Kazuhiko Shinozawa, Takahiro Miyashita, and Norihiro Hagita. 2015. Communicating robotic navigational intentions. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Wolfram Burgard (Ed.). IEEE, Piscataway, NJ, 5763–5769. https://doi.org/10.1109/IROS.2015.7354195

[52] Marcus R. Watson and James T. Enns. 2016. Depth Perception. In *Reference module in neuroscience and biobehavioral psychology*, John Stein (Ed.). Elsevier, [Place of publication not identified]. https://doi.org/10.1016/B978-0-12-809324-5.06398-7

[53] Rong Wen, Wei-Liang Tay, Binh P. Nguyen, Chin-Boon Chng, and Chee-Kong Chui. 2014. Hand gesture guided robot-assisted surgery based on a direct augmented reality interface. *Computer methods and programs in biomedicine* 116, 2 (2014), 68–80. https://doi.org/10.1016/j.cmpb.2013.12.018

[54] Tom Williams, Matthew Bussing, Sebastian Cabrol, Elizabeth Boyle, and Nhan Tran. 3/11/2019 - 3/14/2019. Mixed Reality Deictic Gesture for Multi-Modal Robot Communication. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 191–201. https://doi.org/10.1109/HRI.2019.8673275

[55] Bang Wong. 2011. Color blindness. *Nature methods* 8, 6 (2011), 441. https://doi.org/10.1038/nmeth.1618