NeighboAR: Efficient Object Retrieval using Proximity- and Gaze-based Object Grouping with an AR System

ALEKSANDAR SLAVULJICA, University of St. Gallen, Switzerland
KENAN BEKTAŞ*, University of St. Gallen, Switzerland
JANNIS STRECKER, University of St. Gallen, Switzerland
SIMON MAYER, University of St. Gallen, Switzerland

Humans only recognize a few items in a scene at once and memorize three to seven items in the short term. Such limitations can be mitigated using cognitive offloading (e.g., sticky notes, digital reminders). We studied whether a gaze-enabled Augmented Reality (AR) system could facilitate cognitive offloading and improve object retrieval performance. To this end, we developed NeighboAR, which detects objects in a user's surroundings and generates a graph that stores object proximity relationships and user's gaze dwell times for each object. In a controlled experiment, we asked N=17 participants to inspect randomly distributed objects and later recall the position of a given target object. Our results show that displaying the target together with the proximity object with the longest user gaze dwell time helps recalling the position of the target. Specifically, NeighboAR significantly reduces the retrieval time by 33%, number of errors by 71%, and perceived workload by 10%.

CCS Concepts: • **Human-centered computing → Mixed / augmented reality**; **Ubiquitous and mobile computing systems and tools**; • **Computing methodologies → Perception**.

Additional Key Words and Phrases: augmented reality, cognitive offloading, eye tracking, object detection, human augmentation, mixed reality, working memory, visual search

ACM Reference Format:
Aleksandar Slavuljica, Kenan Bektaş, Jannis Strecker, and Simon Mayer. 2024. NeighboAR: Efficient Object Retrieval using Proximity- and Gaze-based Object Grouping with an AR System. *Proc. ACM Hum.-Comput. Interact.* 8, ETRA, Article 225 (May 2024), 19 pages. https://doi.org/10.1145/3655599

1 INTRODUCTION

The human visual system processes an immense amount of signals each second [34, 39], and many everyday activities such as object recognition or scene interpretation rely on the efficient processing of these signals [12]. The cognitive processing effort depends on the visual complexity of the given scene [31]; however, our perceptual and cognitive resources also have well-known limitations. For example, we can only accurately perceive a small part of our visual field of view [64] and the information we can actively hold in our memory is limited [2, 18, 48]. Because of such limitations, people may experience disorientation or confusion when working in complex environments with high degrees of visual clutter [3].

*Corresponding Author: kenan.bektas@unisg.ch

Authors' addresses: Aleksandar Slavuljica, aleksandar.slavuljica@student.unisg.ch, University of St. Gallen, St.Gallen, Switzerland; Kenan Bektaş, kenan.bektas@unisg.ch, University of St. Gallen, St.Gallen, Switzerland; Jannis Strecker, jannisrene.strecker@unisg.ch, University of St. Gallen, St.Gallen, Switzerland; Simon Mayer, simon.mayer@unisg.ch, University of St. Gallen, St.Gallen, Switzerland.

ACM 2573-0142/2024/5-ART225
https://doi.org/10.1145/3655599

Humans have developed a variety of strategies to cope with visually complex environments, for instance by performing specific actions (e.g., different forms of highlighting relevant elements in a cluttered environment) or by utilizing a variety of tools to make tasks mentally less demanding (e.g., sticky notes). Similarly, and specifically for fast retrieval, techniques such as sorting objects (e.g., on book shelves) or strategically placing or grouping objects according to their color, shape, or meaning can help to alleviate mental workload [10, 26]. In these ways, cognition can be offloaded onto our body or into the world to reduce information processing requirements of a task, and actions that afford such a reduction are collectively referred to as *cognitive offloading* [61]. Hence, also interactive visualizations [62] and strategic reminders [25] can store information outside of our body and reduce the demand for mental representations. Alternatively, previous studies have shown that fingers and eyes are being intuitively used for indexing the location of items and supporting working memory [4, 16]. Relying heavily on external resources (e.g., navigational aids and online tools) may however weaken cognitive abilities [22, 50, 75] or cause cognitive skill degradation [73]. Thus, technologies that aim to augment cognitive abilities of humans [58] need to better support the natural capabilities and resources of their users [79].

Augmented Reality (AR) devices, such as smartphones or head-mounted displays (HMDs), alter users' perception of the world with computer-generated overlays [3]. Lately, wearable AR devices are being used for sensory, motoric, and cognitive augmentation of humans [58]; specifically, they can be exploited to enable cognitive offloading [52, 67, 69]. We propose that the potential of AR as a cognitive offloading tool is even greater *when combined with eye tracking*. For instance, eye tracking permits taking into account the effective allocation of the user's attention to objects in their surroundings, in real-time. We explore this potential in the scope of a cognitive offloading solution for efficient retrieval of objects. Our proposed system, NeighboAR, combines proximity-based object grouping with information about overt user attention that we obtain from a gaze-enabled AR system. This permits us to determine *anchor objects* that receive the most gaze dwell time. When the user is tasked to retrieve an object later, NeighboAR uses associated nearby anchor object as a cue. Our experiments show that this considerably facilitates object retrieval tasks. From the perspective of cognitive offloading, our solution represents an approach that does not *substitute* a user's memory; rather, we use information about a user's previous attention allocation to guide their recall from memory. We have three main contributions:

- We present the NeighboAR concept for gaze-aware cognitive offloading. We discuss related research on visual search and eye tracking, visual object detection, and the use of AR technology for cognitive offloading, and point out what distinguishes NeighboAR as a novel approach to AR-based cognitive offloading.
- We introduce the NeighboAR prototype application. Our prototype is running on a Microsoft HoloLens 2 (HL2) headset that is paired with an external server.
- We present the results of a user experiment of the NeighboAR prototype. Our results show that significant improvements can be achieved in object retrieval performance and perceived workload when users are provided assistance with NeighboAR.

To the best of our knowledge, we present the first attempt to combine gaze data and object detection for cognitive offloading through an AR device, and to evaluate the benefits of this combination in object retrieval.

2 RELATED WORK

The conceptual foundation of NeighboAR is formed by related research on visual search and eye tracking, visual object detection, and the use of AR technology for cognitive offloading. In the

following, we discuss relevant research in these fields and point out what distinguishes NeighboAR as a novel approach to AR-based cognitive offloading.

2.1 Visual Search and Eye Tracking

Visual search is vital for humans (and animals), as it constitutes an essential part of many of our daily activities [12]. Thus, it is a topic of interest for researchers across diverse fields such as ophthalmology [76], psychology [19], and human-computer interaction [38]. Visual search may involve both overt and covert orientation of attention, where the former involves a deliberate movement of the head and *eyes* [57]. Previous research identified main factors that guide attention in visual search such as stimulus-, user-, and scene-driven attributes as well as the perceived value of and previous interactions with the target objects [77]. According to image interpretation experts (e.g., radiologists or analysts in satellite-based remote sensing settings), sufficient knowledge about the target is the key to a successful search [19]. In previous studies, several predictive and contextual cues were identified that reduce uncertainty of the target location where observers are informed about the cue-target relationships [19]. Typically, these cues guide observer's eye movements to locations where the target objects are expected because of the scene context (i.e., semantic consistency) or co-occurrence of cues and targets [19]. Castelhano and Heaven [14] show that co-occurrence facilitates search even in the presence of semantic *inconsistency*. In this work, we ask users to explore a scene while we measure their gaze dwell time on each object that they observed. We hypothesize that these previous interactions and co-occurrence determine a strong cue-target relationship that can inform users about the cue that they can intuitively recall when asked to retrieve a target object.

In eye tracking research, visual search is a commonly studied activity that can offer insights into the decision-making and problem-solving behavior of humans [19]. From eye-tracking data, conclusions about human attention can be drawn which have a wide range of applications—from the optimization of visual search interfaces [9] and product placements in vending machines [68] to the creation of more user-friendly software interfaces [36] and even the assessment of the complexity of declarative process models in software [1]. Previous research shows that users, when tasked with object retrieval in cluttered work environments, rapidly move their eyes from one object to another [24, 70]. Objects which have been gazed at for longer remain in a user's memory more accurately and for a longer time [17, 72], and in our work, we exploit this in the context of cognitive offloading. Yet, the frequency of eye movements is proportional to the complexity of the environment itself, meaning that users struggle with retaining object information when faced with an environment with many objects [33].

Recently, Peacock et al. [55] studied the link between multiple eye tracking features and working memory encoding. Their model shows that fixation duration is one of the predictive features for a successful encoding. The participants of their experiments were asked to navigate through the rooms of virtual apartments containing low clutter and find various targets (e.g., a table lamp) that were marked with an arrow. Then, the participants were asked to verbally recall these targets or their associated labels. However, in our experiment, participants freely viewed a set of physical objects without knowing about the target. Right before the retrieval, we showed them the target and an object that they gazed at most among other objects that were placed in the vicinity of the target. Thus, we do not disclose any hint about the target or its location beforehand and encourage users to rely on their own memory for retrieval.

In previous research, gaze data has also been used for reducing cognitive load. For instance, Gazemarks [37] make use of a screen that displays visual reminders that take into account the user's last gaze position to help them reorient their attention after an attention switch or interruption. This leads to faster completion times for a resumed visual search task [37]. Similarly,

GeoGazemarkers provide users with visual orientation cues on a small hand-held display [23]. Specifically, GeoGazemarkers support map users in orientation tasks by providing them assistance based on where they have looked the most in a previous exploration phase. We propose that a solution that takes into account object gaze dwell times in retrieval tasks may be an effective way to facilitate cognitive offloading; hence, NeighboAR combines gaze tracking in an AR-based cognitive offloading system, and we evaluate its effect on users' object retrieval performance in a visually cluttered *physical* environment. To permit NeighboAR to assign gaze dwell times to objects, it needs to first *detect* relevant scene objects.

2.2 Visual Object Detection

The underlying mechanisms of human visual perception have inspired and informed the development of computer vision solutions [12], prominently in the field of object detection [80]. Visual object detection consists of locating objects of interest in an image and solving a classification problem. The output of an object detection algorithm is typically a label indicating a type of object and the location or bounding box of the object in a visual scene. Two-stage approaches, such as Region-based Convolutional Neural Network (R-CNN) [27], extract region proposals of candidate objects and compute features for each region through a large CNN in the first stage. Then, in the second stage, each region is classified using a Support Vector Machine (SVM) with a linear kernel. Although such two-stage approaches offer very good detection accuracy, they compromise on speed [81]. One-stage algorithms, in contrast, extract region proposals and classify them at the same time, thereby reducing computation time significantly. *You Only Look Once* (YOLO) [59] is the first algorithm that proposed an architecture in which a single neural network predicts bounding boxes and class probabilities at the same time. On top of YOLOv3 [60], which improves the algorithm's performance with small objects, and YOLOv4 [11], which provides an even faster and more accurate algorithm, YOLOv7 [74] shows further improvements by requiring 75% fewer parameters, 36% less computation time, while increasing average precision by 1.5% compared to YOLOv4. In NeighboAR, we chose to make use of YOLOv7 due to its high processing speed and its high level of flexibility and customizability. It also supports training for custom objects, which we require for conducting our experiments on the cognitive offloading features of NeighboAR.

Significant use of visual object detection is seen in fields such as security, surveillance, robotics, and autonomous vehicles [81]. Fang et al. [20], for instance, aimed to address the challenges of identifying and managing safety hazards in construction sites, which are dynamic environments with a high level of visual clutter. Their system can create contextual links between detected objects, derive further meaning, and ultimately detect whether certain objects pose a safety hazard. NeighboAR focuses instead on the proximity of objects to one another and combines this data with object gaze dwell times. Specifically, gaze dwell times are used to filter the proximity-based graph and thereby better assist users by taking into account which objects were attended most, hence, perhaps will be remembered best.

AR-capable hardware such as smartphones or head-mounted displays (e.g., HL2[1], Apple Vision Pro[2], or Magic Leap 2[3]) typically are equipped with RGB cameras, and therefore facilitate the combination of AR and object detection. Regarding the allocation of concerns in the object detection process across devices in an AR setting, Gammeter et al. [21] combine server-side object recognition with client-side object tracking, where an Android-platform AR device tracks objects on the client side and sends this data to the server-side component. The server-side component uses a deep

[1] https://www.microsoft.com/en-us/hololens. Last accessed February 10, 2024.
[2] https://www.apple.com/apple-vision-pro. Last accessed February 10, 2024.
[3] https://www.magicleap.com/magic-leap-2. Last accessed February 10, 2024.

learning model to classify tracked objects. Gammeter et al. found that this separated setup is faster and more accurate than performing all functions client-side. The HL2 that we use as part of the NeighboAR system is similarly limited regarding processing power. Therefore, we separate the NeighboAR application into HL2 and server functionalities.

2.3 Cognitive Offloading and AR

Cognitive offloading is a process in which we aid our mind in accomplishing its cognitive tasks by offloading some of the mental demand to the external environment. Risko and Gilbert [61] explain that the interactions of our body with the environment facilitate this process, often subconsciously partaking in this act. However, this process is not limited only to the use of our own bodies. People have historically used various tools for cognitive offloading, such as an abacus for calculations or knots and recording devices such as quipus for memorization [50]. Scaife and Rogers [62] reviewed the value of advances in graphical technologies for facilitating cognitive activities. Specifically, they discuss to what extent the construction of internal mental representations can be supported by external graphical representations such as diagrams, multimedia, or *interactive virtual reality* applications. According to Scaife and Rogers [62], such external representations can reduce the amount of cognitive effort (i.e., computational offloading) that is required in making inferences or solving problems.

Our interactions with other hardware and software artifacts (e.g., smartphones or search engines) can also serve for cognitive offloading [61]. With respect to the NeighboAR system, we specifically consider AR applications for cognitive offloading. This can range from simple textual overlays to highly interactive holographs, and use cases cover domains from entertainment to medicine and more [54]. Today, AR can be experienced through various devices such as HMDs, tablets, or smartphones. As such, most individuals in the world already have access to smartphone hardware which they could use for cognitive offloading in conjunction with some form of AR [65]. As AR devices are becoming affordable and widely available [30, 51, 56], we expect their potential with respect to cognitive offloading to grow further. Eye tracking and AR are two technologies which are often combined together, as AR head-mounted displays are frequently equipped with integrated eye tracking capabilities [66]. Eye-tracking data has already been used in conjunction with AR technology to help simplify various tasks, examples of which can be seen in AR surgical training solutions [43], as well as in AR-enabled human and robot intention communication [15]. However, these implementations do not measure cognitive offloading effects.

Strecker et al. [67] proposed an example cognitive offloading solution that combines *Semantically* integrated *Optical Character Recognition* (OCR) with *AR* (SOCRAR) to provide an efficient and user-friendly way to extract and interact with text in a user's physical environment. On top of the OCR, this system semantically lifts written text with respect to available ontologies, allowing the user to interact with the text in meaningful ways, such as for currency conversion, additional information search, etc. NeighboAR bears similarity to SOCRAR, as we likewise use an AR device for cognitive offloading. However, while Strecker et al. [67] apply semantic OCR to the camera feed of an HMD, our approach facilitates cognitive offloading with a visual object detector in combination with eye tracking and spatial clustering, and focuses on object retrieval tasks. Tang et al. [69] report that AR applications can be used to decrease cognitive demand during object assembly tasks. The participants of their study received instructions using an AR system on how to assemble an object. The findings of this study imply that AR guidance significantly improves performance across various metrics in comparison to traditional instruction methods. Other than decreasing errors and total assembly time, measurements of mental effort show that users benefited from cognitive offloading. We likewise use AR assistance to improve performance and decrease the cognitive demand of a task, but our system in addition uses eye tracking to measure user gaze

dwell times for individual objects, and then makes use of this data for providing assistance in object retrieval tasks.

AR applications often provide supplementary information to their user. For example, AR holograms can be used to point towards potential locations of an object in a shared space to avoid overwhelming the user. The size of the holograms can be adjusted based on how recent the registered locations were [52]. However, Tsurukawa et al. [71] employ AR to blur unnecessary peripheral information, thus reducing the complexity and clutter in an observed area. As a consequence, cognitive demand of the associated task is decreased. This is made possible by having a user define specific regions of significance in a 3D space. Then, the underlying algorithm determines which regions must become apparent to the user. Insignificant regions in the user's field of view are blurred to decrease visual clutter and, hence, cognitive demand. Our prototype does not use AR to blur the user's view. To assist users, NeighboAR provides them with task-relevant information based on the collected object gaze dwell times and the spatial proximity of target objects. This makes our system less obtrusive, while still enabling cognitive offloading.

AR has also been combined with eye tracking, albeit not in the context of supporting cognitive offloading: Panetta et al. [53] proposed an automated solution to a cognitively demanding and time-consuming manual annotation process (i.e., an essential part of eye-tracking data analysis). Their proposed software architecture uses a wearable eye tracker to gather gaze-overlaid video footage. This software then detects the specific object a user is focusing on and calculates the gaze dwell time. Barz et al. [6] similarly implemented such a solution for the HL2 based on the Unity game engine. Our system shares similarities with these approaches, as we likewise dynamically detect different objects in the user's field of view and calculate how long each object was gazed at. However, we go beyond annotation and use the combination of object detection and eye tracking to attempt to improve object retrieval performance of users in visually complex environments where objects are presented in arbitrary semantic and spatial constellations.

3 NEIGHBOAR: COMBINING EYE TRACKING AND OBJECT DETECTION TO ENABLE AR-BASED COGNITIVE OFFLOADING

Similar to the GeoGazemarkers [23], for the design of the NeighboAR system we assume that users are able to better recall objects that they have previously looked at the most. Hence, we hypothesize that dwell time can serve as a cue to determine what visual elements should be displayed to users to support them in a given task. We focus on providing assistance to users for object retrieval tasks in cluttered environments where NeighboAR facilitates cognitive offloading. Specifically, NeighboAR achieves this by displaying an image of the target object (in AR) and supplementing this with the image of another object that received the longest gaze dwell time by the user among the objects in physical proximity to the target object.

NeighboAR combines AR, object detection, and eye tracking technologies. Our NeighboAR prototype runs on two devices, an HL2 and an external server, as shown in Figure 1. The application on the HL2 consists of four components: the *Eye Tracking Handler*, *3D Handler*, *UI Handler*, and *HTTP Listener*. We developed this application using the Unity Game Engine (2020.3.38) and the Mixed Reality Toolkit (MRTK; v2.8.2). The external server runs the *Object Detector* with the HL2's camera feed as input, and contains the *Administration* component to manage the user experiment.

Object Detector. NeighboAR's *Object Detector* is responsible for processing video frames coming from the HL2's front camera and detecting scene objects using YOLOv7 [74]. To prepare our experiments with the NeighboAR system, we trained our custom YOLOv7 model on an NVIDIA Tesla T4 GPU with 30 grocery products in the categories of chocolate, pasta, chips, and cereals. The test over the validation set resulted in an overall precision of 99.2% and a recall of 99.8%. The

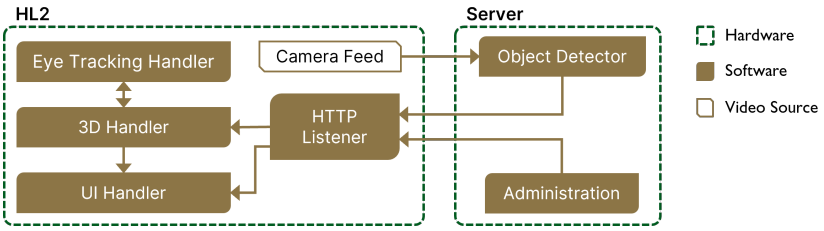


Fig. 1. The NeighboAR prototype consists of six software components (*Eye Tracking Handler*, *3D Handler*, *UI Handler*, *HTTP Listener*, *Object Detector*, and *Administration*) and two hardware devices (HL2 and an external server).

mean Average Precision (mAP) scores of 99.4% at an Intersection over Union (IoU) threshold of 0.5 and 97.6% across the range IoU=[0.5, 0.95] further indicate the robustness of our model. The *Object Detector* accesses the video feed of the HL2 front camera through the Mixed Reality Device Portal API [44] and processes the video frames with a frame rate of approximately 28 frames per second on an NVIDIA GeForce GTX 1080M mobile graphics chip. Through our testing, we concluded that the optimal confidence level for object detection is 0.88. Once an object is detected with at least this confidence level, an HTTP request is sent to the *HTTP Listener* component running on the HL2 that contains the object class and the 2D bounding box coordinates.

3D Handler. NeighboAR's *3D Handler* component has two main functions. First, it determines the 3D position of detected objects in real space. The coordinates of the 2D bounding box's central point (detected in and sent by the *Object Detector*) are converted into a ray, which is cast from the origin of the HL2's camera; we make use of the adjusted HL2 camera position provided by ARETT [35]. The point where this ray collides with the spatial mesh (i.e., a 3D geometrical representation of the user's environment created by MRTK's Spatial Awareness System [46]) is used as the preliminary 3D position of the object. To minimize inaccuracies due to rapid head movements, the *3D Handler* repeats this process, comparing the new 3D position to its previously detected position. If the 3D distance between these two consecutively detected positions is greater than 20 centimeters, both of these positions are deleted. For a given object, if five preliminary positions are successfully created, the average X, Y, and Z coordinates of these positions are used to determine the position of that object's 3D bounding box. The 3D bounding boxes have a fixed size which encloses all objects used in the experimental procedure, without causing overlapping. During pilot testing of our prototype, we found that these decisions allow the system to have an adequate detection performance. The bounding boxes remain invisible to the user and are used for the second main function of the *3D Handler*—determining what objects a user is gazing at. Concretely, when the *3D Handler* receives eye-tracking data from the *Eye Tracking Handler*, the 3D bounding boxes are used to find which object a user is gazing at. Given eye-tracking data, NeighboAR calculates a raycast from the user's eyes to their gaze target (i.e., the hit point with the 3D bounding box) to determine what object a user is gazing at. Every time the *3D Handler* determines that the user looks at an object, the object name, 3D position, and gaze dwell time are sent to the *Eye Tracking Handler*.

Eye Tracking Handler. This component is responsible for handling and logging eye-tracking data once the *3D Handler* has created at least one 3D bounding box. The HL2's eye tracker has a sampling rate of around 30Hz with an accuracy of approximately 1.5-3° [47]. NeighboAR uses MRTK's in-built functions to retrieve eye-tracking data from the HL2's eye tracker [45]. Whenever it receives a new gaze sample from the HL2's eye tracker, the *Eye Tracking Handler* sends it to

the *3D Handler*, which uses the sample to determine whether the user gazes at a particular object (as described above). If this is true, the *Eye Tracking Handler* receives the object's class name, 3D position, and gaze dwell time from the *3D Handler* and records this data.

UI Handler. In experimental object retrieval tasks, NeighboAR's *UI Handler* component provides users with visual cues and a button that is used in measuring the task duration. In the *unassisted mode*, only the target object is displayed, while in the *assisted mode*, both the target object and the object in proximity with the highest gaze dwell time are displayed (see Figure 2(e)). To determine the object in proximity of the target and that was gazed at the longest, the *UI Handler* constructs a proximity-based object grouping (i.e., the proximity graph) in the background, where the object positions are determined with the help of the *3D Handler*, as described earlier. Then, based on a maximum distance threshold of 30 centimeters between the detected objects (i.e., adjacent products), NeighboAR determines which objects are considered to be close to each other and thus clustered together into groups. In our experimental design, this grouping corresponds to the distribution of the objects in four separate cabinets. To determine the horizontal order of the clusters (i.e., which cluster corresponds to which cabinet; see Figure 2(a)), the X-coordinates of the first object in each cluster are sorted. Based on this ordering, the clusters are then labelled internally from the left-most (*C1*) to the right-most (*C4*) clusters. The clustered objects are then enriched with the current user's dwell time for each object, as determined by the *3D Handler* and *Eye Tracking Handler*. Within the cluster of the target object, the object with the highest dwell time is then selected to be displayed as the object in proximity.

HTTP Listener. NeighboAR's *HTTP Listener* component enables the HL2 and the server to communicate. The *HTTP Listener* runs on the HL2 and regularly checks for any incoming HTTP requests from the *Object Detector* and *Administration* components. After receiving an HTTP request, the contents of the request are forwarded to either the *3D Handler* or the *UI Handler*.

Administration. NeighboAR's *Administration* component is responsible for sending the user experiment-related commands to the *HTTP Listener* (see Section 4). Once received by the *HTTP Listener* component, these commands are executed by the *UI Handler*. These commands are used to set a target object during the experiment phase; to show, hide, reposition, or reset the display of the target and proximity object; to enable or disable the logging of eye-tracking data; to toggle the visibility of 3D bounding boxes for debugging; to start timers for data collection; to change between assisted and unassisted modes; and to initiate the user eye-tracking calibration process.

4 NEIGHBOAR USER EXPERIMENT

We conducted a controlled user experiment to evaluate NeighboAR and to verify our hypothesis that gaze dwell time together with physical proximity can serve as a cue to determine what visual elements should be displayed to users to support them in a given task. In this user experiment, participants were first tasked to remember the locations of 30 objects that were randomly arranged in four cabinets. The participants were tasked to retrieve a target object from the cabinets. During retrieval, either NeighboAR's *assisted* mode or its *unassisted* mode were active (see Figure 2(e)). In the *unassisted* mode, the users saw an image of the target object that they need to retrieve using their own memory (or their mind map). In the *assisted* mode, they additionally saw an object whose location is in proximity to the target object, and at which the user gazed the longest among all physically proximate objects while remembering the object locations.

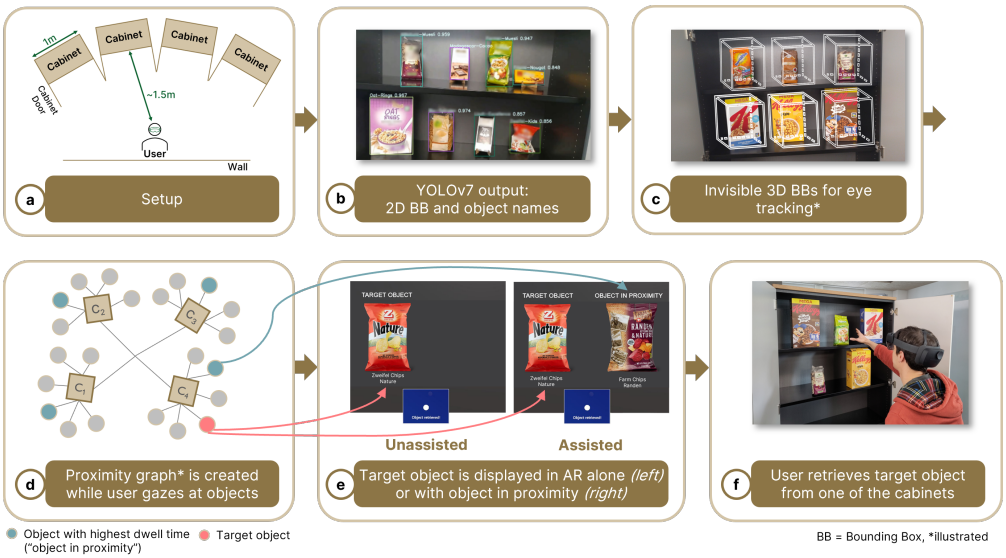


Fig. 2. The system flow of NeighboAR. (a) A user is wearing a HL2 and is sitting in front of four cabinets. (b) A screenshot taken from the output video of the Object Detector. The bounding boxes, class names and probabilities are not visible to the participants. (c) Based on the 2D bounding boxes, 3D bounding boxes are created invisibly to the user (see Section 3). (d) While the user inspects the objects, a proximity graph is created in the background (see Section 3). In each of the cabinets the object with the highest dwell time will be selected as an *object in proximity* if a target object of this cabinet is desired. (e) The left image shows the *unassisted* mode where users are presented with a picture and name of the target object they are asked to retrieve. The right image depicts the *assisted* mode where, additionally to the target object, the object in proximity to the target object which received the longest gaze dwell time from the user is also displayed. In both modes, the participant is expected to press the "Object retrieved!" button once they have found and retrieved the target object. (f) After being exposed to either the *assisted* or *unassisted* mode, the user is tasked to retrieve the target object from the right cabinet and press the button that is anchored to the wall.

4.1 Experiment Participants and Setup

Eighteen individuals (mean age 22 years; 4 females) participated in the user experiment and we excluded the data of one male participant due to data incompleteness. When asked to evaluate their familiarity with AR and Virtual Reality on a scale from 0 (not at all familiar) to 4 (very familiar), participants on average stated that their familiarity was 1.41 and 1.82, respectively.

For the experiment setup, we used an HL2 and connected it to an external server (laptop) over WiFi. We furthermore used four cabinets with shelves in which we placed 30 different grocery products (see Figure 2). Products were selected based on their size and ease of placing and stacking in a shelf, together with our requirement of stable object detection performance on our trained YOLOv7 detector. In each trial, all products were placed on the top two shelves of the cabinets. In this way, we ensured that the products were visible for the participants when they were sitting on an office chair with castors. While set in a context with grocery products, the experimental setup that we use is representative of a generic cluttered scene that includes multiple physical objects with different features (e.g., color, material, size). Such settings occur frequently in scenarios that range from industrial workshops (e.g., retrieval of a required machine tool) to private home environments. Thus, our findings are likely to generalize to many specific scenarios with visually

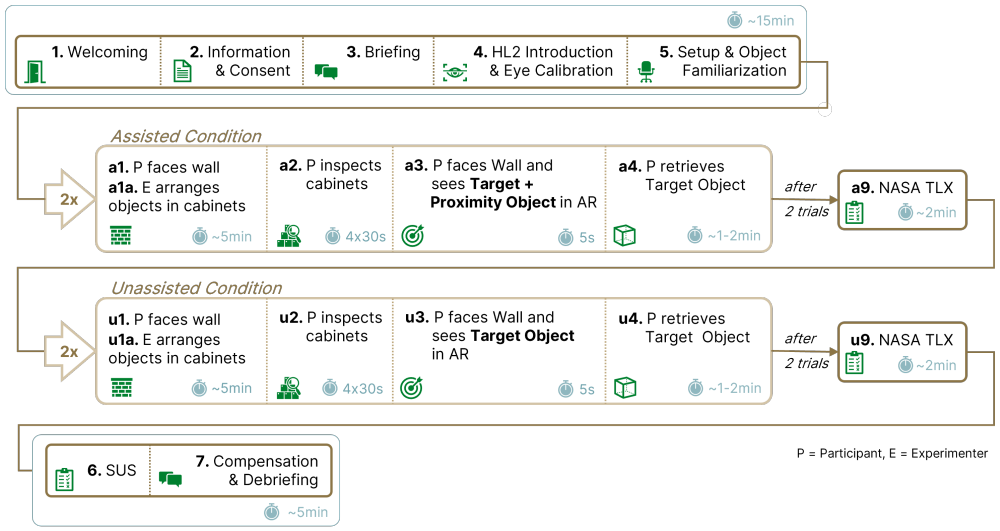


Fig. 3. A diagram showing the experimental procedure for one participant during the user experiment. Here, the participant first performs two trials in the *assisted* condition, followed by two trials in the *unassisted* condition. The order of conditions is randomized for every participant.

cluttered scenes. Our user experiment was exempt from a formal review by the ethics committee of our institution.

4.2 Experimental Procedure

We defined a within-subject design procedure that takes about 60 minutes and includes multiple predefined steps that we followed for all participants (see Figure 3). During pilot testing, we noticed that while some of these steps must have a fixed duration (e.g., inspection of the cabinets and object retrieval) others may require variable amount of time (e.g., eye calibration and familiarization of the participants and rearranging objects in the cabinets). In order to fix the total duration of the experiment and the amount of work across all participants we decided to have two trials for each experimental condition (i.e., *assisted* and *unassisted*), and thus four trials in total.

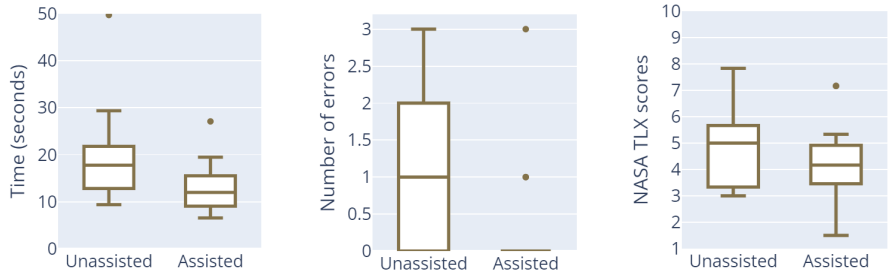
Before the start of the experiment, each participant read a general information sheet, thus was informed about the purpose of the experiment, as well as the tasks that are part of it. Each participant subsequently signed a consent form that informs them about how their data would be used and that they could withdraw their consent at any time and without any penalties. Participants confirmed they do not have any medical conditions (e.g., simulator sickness and visual disorders) which could affect them during the experiment. Before the first trial, each participant was assigned to one of the experimental conditions. We randomized the order of the conditions among the participants to avoid potential biases that would affect the results. The participant was then briefed on the experimental setup, put on the HL2 and started the in-built eye-calibration procedure. Each participant was assisted in placing and adjusting the HL2 on their head to avoid obstructing the visibility of virtual and physical content and causing discomfort to the user. To prevent any height biases which could systematically favor products on one of the shelves, the participant sat on a height-adjustable office chair with omnidirectional castors and the chair was adjusted so that the participant's eye level is at a comfortable position to perceive the two top-most shelves of the four cabinets. Finally, each participant was given the chance to familiarize themselves with the

products used in the experiment, so as to mitigate potential biases due to preexisting familiarity with products. Next, the participant was instructed to face a blank wall opposite of the cabinets while the products were placed into the four cabinets by the experimenter. The participant was then given two minutes, instructed to turn around and inspect all cabinets and to try to remember the locations of all products. Every 30 seconds, the HL2 played a bell sound to signify that the participant should move to the next cabinet. During inspection, participants were free to choose the order of inspection of the cabinets. In both *assisted* and *unassisted* conditions, NeighboAR generated a proximity-based graph of the objects during the inspection phase. It also tracked the participant eye movements while they inspect the cabinet to assess which object was gazed at for how long. The unassisted trials could have been performed without the HL2 by showing participants the target object using a printout. However, to avoid confounding effects from the hardware (e.g., physical demand or restrictions in peripheral vision), participants wore the HL2 in both experimental conditions.

After inspecting all cabinets, the participant was asked to face the wall opposite of the cabinets. In the *unassisted condition*, the NeighboAR application then displayed a picture of the target object along with the object's name (see Figure 2(e)). In the *assisted condition*, the NeighboAR application additionally displayed the object with longest gaze dwell time among all objects in the target object's proximity (see Figure 2(e)). The participant was given five seconds to recall which cabinet they believe contains the target object without turning around. After these five seconds, a bell sound played and NeighboAR started a timer to measure the participant's retrieval time. The participant now turned around and attempted to retrieve the object. To do this, they needed to approach the correct cabinet and take the object. Then they needed to return to their starting position and press a confirmation button that was displayed in the NeighboAR application (see Figure 2(e)) which stops the search timer. Note that both the target object and the confirmation button are spatially anchored to the wall shown in Figure 2(a). During each trial, the participant's number of errors was measured. Here, *errors* refer to the number of times that a participant stopped moving their chair and started inspecting the contents of the cabinet where the target object was not present. After completion, the trial was repeated with reshuffled object positions to prevent learning effects. Afterwards, the NASA Task Load Index (TLX) questionnaire [32] was answered by the participant on a 10-point Likert scale. Then, the trial condition (assisted/unassisted) was changed and all steps were repeated in two more trials. After a participant has completed all trials, they completed the System Usability Scale (SUS) questionnaire [13] and answered the additional question: "How useful was the displayed proximity object for finding the goal object?". Finally, each participant was debriefed and compensated for their time with 25 CHF per hour that our institution pays for participation in user experiments.

4.3 Hypotheses and Variables

In this user experiment, we tested three hypotheses. Our first null hypothesis (H_1) states that *retrieval times during unassisted trials are not significantly longer than during assisted trials.* To test H_1, we measured the *retrieval time* in seconds that is averaged across all participants and trials. Our second null hypothesis (H_2) states that *the number of errors made during unassisted trials is not significantly larger than during assisted trials.* To test H_2, we measured the average *number of errors* made during the object retrieval task across all participants and trials. Our third null hypothesis (H_3) states that *perceived workload during unassisted trials is not significantly larger than during assisted trials.* To test H_3, we measured the *perceived workload complexity* with the NASA TLX score that is averaged per condition. The independent variable of our user experiment is the mode of assistance provided in NeighboAR (*unassisted* vs. *assisted*).



(a) Mean Object Retrieval Times (b) Mean Number of Errors (c) Mean NASA TLX Scores

Fig. 4. Three box plots showing the results for different operation modes (*unassisted* or *assisted*) of the NeighboAR system.

5 RESULTS

In this section, we present the results of our user experiment and compare the differences in user performance and workload for the unassisted and assisted conditions (see Figure 2(e)). In addition, we present our findings about the system usability and perceived usefulness of NeighboAR.

Retrieval Speed (H_1). To test H_1, we compared the participants' *retrieval time* between the two conditions (*unassisted* and *assisted*), as shown in Figure 4(a). In the *unassisted* condition, the participants' mean object retrieval time is 19.38*s* ($\sigma = 9.6$), while it is 13.03*s* in the *assisted* condition ($\sigma = 5.37$). These two conditions are dependent and not all data follows a normal distribution according to the Shapiro-Wilk (SW) test (*unassisted*: $SW = 0.8$, $p = 0.002$; *assisted*: $SW = 0.915$, $p = 0.123$). As such, the Student's t-test cannot be used for statistical interference, due to the violation of the assumption of independence, as well as the assumption of normality. According to Scheff, equivalent non-parametric statistical interference methods can be used, if some of the aforementioned assumptions are violated [63]. Accordingly, Scheff suggests using the Wilcoxon signed-rank test. The results show that in the *assisted* condition, the participants needed significantly less time than in the unassisted condition ($W = 18.0, p = 0.002, r = 0.532, \alpha = 0.05$) and that we may hence reject H_1.

Number of Errors (H_2). Across all participants, the mean number of errors in the *unassisted* condition was 1 ($\sigma = 0.98$) while in the *assisted* condition it was 0.29 ($\sigma = 0.77$), as seen in Figure 4(b). We again used the Wilcoxon signed-rank test, because the data was not normally distributed according to the Shapiro-Wilk test (*unassisted*: $SW = 0.85$, $p = 0.011$; *assisted*: $SW = 0.45$, $p < 0.001$). The results show that the participants made significantly fewer errors in the assisted condition than in the unassisted condition ($W = 14.0, p = 0.024, r = 0.388, \alpha = 0.05$), and that we may hence reject H_2.

Perceived Workload (H_3). We used the NASA TLX questionnaire to measure participants' perceived workload. As this questionnaire consists of rating scales on a 10-point Likert scale, the data itself is ordinal in nature. However, Scheff explains that non-parametric statistical interference models can be used even with ordinal data, given that the data itself is first transformed into a ranked system [63]. In the unassisted condition, the average workload is rated as 4.79 ($\sigma = 1.33$), and in the assisted condition as 4.33 ($\sigma = 1.41$). The results of Wilcoxon signed-rank test show that

perceived workload for the assisted condition is significantly lower than the unassisted condition ($W = 25.5, p = 0.025, r = 0.385, \alpha = 0.05$), and we may hence reject H_3.

System Usability and Usefulness. The NeighboAR system received an average SUS score of 71.18 ($\sigma = 18.14$). According to Bangor et al. [5] this score is in the acceptable part (70 to 100 on a scale from 0 to 100) of the *Acceptability Range* and is classified as *GOOD* (between *OK* and *EXCELLENT*). Participants rated the assistance provided by NeighboAR as *useful* for the object retrieval (on average 3, $\sigma = 0.87$, on a scale from "0: Not useful at all" to "4: Very useful").

6 DISCUSSION

6.1 The Effect of NeighboAR on User Performance, Workload, and Comfort

The results depicted in Figure 4 demonstrate decreasing retrieval times, number of errors, and perceived workload when participants use NeighboAR in the *assisted* mode. Additionally, our statistical analysis show that the differences between *assisted* and *unassisted* conditions are statistically significant, thus we could reject all three null hypotheses. Our findings are consistent with previous work: *Informing individuals about their previous gaze-based interactions within an environment (e.g., a map interface [23] or a virtual apartment [55]) can be used to provide them with cues that allows them to efficiently use their cognitive resources and reduce their workload.*

Regarding our experimental design, it is worth noting that the participants reported about their workload after performing two trials per condition. Due to the time that we needed before each trial to redecorate the products, participants could have forgotten some details of how they felt during their first trial. The results hence include a training effect. However, this does not undermine our results, because the procedure was the same for all trials. In addition, the mandatory five-second thinking period before answering where the target object is placed could be obscuring our understanding of participant performance. In the *unassisted* condition, it is possible that some individuals would need all five seconds of thinking, but in the *assisted* condition, they could answer much faster. Three participants noted that the *assisted* condition caused some confusion, as they mixed up the target object with the proximity object in one of the trials. We have only sporadic oral feedback from the participants about these cases, but they might contribute to higher variance in assisted retrieval times and NASA TLX scores. Another limitation was our choice of objects in the experiment. Even though most objects were similar in packaging, some objects had more discernible features and thus higher visual saliency. This could understandably make it significantly easier to retrieve certain objects; however, this does not impact the expressivity of our results, because the target was randomly selected among 30 other objects.

The current version of NeighboAR was developed with an unobtrusive approach in mind. Our intention was to filter task-specific information in the background (i.e., on the graph), and to provide users with cues without cluttering their visual field. Further, our current implementation requires minimal interaction with the AR content, which makes it less challenging to train users, even if they had no previous experience with AR. As such, we could avoid causing discomfort or confusion resulting from unfamiliarity with AR systems. In addition, the SUS score of NeighboAR is relatively high for a research prototype, which is worthy of consideration as almost none of our participants worked with AR systems before. Thus, even when faced with a completely new technology, most participants deemed that it was sufficiently usable, in addition to significantly increasing their object retrieval performance.

6.2 Limitations, Future Applications, and Implications

Our current solution detects users' previous interactions with spatially co-located objects only from their gaze dwell time. NeighboAR can be extended with predictive models that can incorporate a

larger feature set of gaze dynamics. Then these models can be integrated into an AR system, as previously shown by others for detecting working memory encodings [55] or human activities [8], that can provide users with *gaze-contingent assistance* in real-time. This kind of assistance can potentially be beneficial for individuals with memory deficits, those who practice multitasking or must work in high cognitive workload conditions.

The current prototype works under the assumption that the position of objects will not change after their detection. In the future, NeighboAR can be improved to support dynamic object tracking [42], which would expand the number of use cases where the prototype can be suitably employed. Furthermore, as this is not required for demonstrating the capabilities of the NeighboAR prototype in our experiments, it does not have an interactive user interface. Such an interface could be added, for instance to permit users to communicate to the system what objects they are currently looking for. Another extension of the prototype could allow a network of users to collaboratively build consistent proximity-based object groups and create shared interactive experiences. For example, when a user sees an object, its 3D position can be logged, as we described, and also simultaneously shared with selected users on the network. This would allow the system to automatically build object groupings and track object positions using several HMDs worn by different users (e.g., in [52]). As such, this implementation would allow the system to assist a user in retrieving objects which the user did not observe. This could result in improvements in (collaborative) retrieval tasks which could be particularly useful in large and complex environments such as shared workspaces or warehouses, where the chance of not observing all relevant objects is higher. Expanding NeighboAR's functionality to accommodate these extensions seems plausible, as existing AR headsets already support 3D object position tracking and wireless communication.

We argue that NeighboAR can enable efficient object retrieval and serves as a tool for cognitive offloading. However, our current findings rely on a user experiment that was executed in controlled experimental conditions and with limited number of participants and trials. We need evidence from followup studies that explore the longitudinal effects of NeighboAR on the cognitive abilities of diverse user groups and preferably in their daily social and professional routines. To enable other researchers to build on our work, we make the source code of our system openly available [4].

6.3 Privacy and Ethics Statement

In general, the usage of eye tracking systems and collection of eye-tracking data raise privacy-related concerns because sensitive insights about individuals (e.g., their personal preferences or mental condition) can be captured from their eye movements [40]. In modern gaze-enabled AR headsets, eye-tracking data can be further combined with other spatio-temporal (e.g., audio, visual, or biometric) recordings which can be used to predict (and extrapolate) more intimate details about the users and their personal life. Thus, the increasing availability of these technologies can be beneficial in addressing various problems of individuals or larger groups and, at the same time, they can be used for manipulative and harmful purposes [49]. Thus, features such as those we present in the NeighboAR approach and prototype should only be enabled after obtaining informed and deliberate consent from the user. Users should also be provided with the means to specify when and how the system will collect, store, and use their data [28, 41, 78]. As a technical mitigation measure, we suggest that eye-tracking data should be kept in personal data stores (e.g., in [7, 29]) which enable users to exert fine-grained control—for raw eye-tracking data as well as derived information—about what data is shared with third parties and what remains private.

[4] https://github.com/Interactions-HSG/NeighboAR

7 CONCLUSION

Object retrieval tasks that we frequently encounter in our personal and professional lives can be cognitively demanding especially in cluttered environments. To address this issue, we presented NeighboAR: a novel cognitive offloading solution that combines eye tracking and object detection through an AR device. We implemented a prototype and evaluated its benefits to support object retrieval tasks. Instead of directly pointing at the location of a target object that is detected among many other objects, NeighboAR provides its user with a cue that is informed by the previous allocation of user attention, which we measure through eye gaze. Specifically, in its *assisted* mode, NeighboAR displays the target together with an object that is co-located with the target and received the longest gaze dwell time by the user in the past. We showed that NeighboAR improves object retrieval performance and reduces the perceived workload while supporting rather than replacing a user's memory capabilities. This method of cognitive offloading can be potentially beneficial for individuals with memory deficits or it can be used for assisting those who work in highly cluttered and cognitively demanding environments.

ACKNOWLEDGMENTS

This work was supported by the Swiss Innovation Agency Innosuisse (Project #48342.1 IP-ICT) and the Basic Research Fund of the University of St.Gallen.

REFERENCES

[1] Amine Abbad-Andaloussi, Andrea Burattin, Tijs Slaats, Ekkart Kindler, and Barbara Weber. 2023. Complexity in declarative process models: Metrics and multi-modal assessment of cognitive load. *Expert Systems with Applications* 233 (2023), 120924. https://doi.org/10.1016/j.eswa.2023.120924

[2] G. A. Alvarez and P. Cavanagh. 2004. The Capacity of Visual Short-Term Memory Is Set Both by Visual Information Load and by Number of Objects. *Psychological Science* 15, 2 (2004), 106–111. https://www.jstor.org/stable/40063936

[3] Ronald T. Azuma. 1997. A Survey of Augmented Reality. *Presence: Teleoperators and Virtual Environments* 6, 4 (Aug. 1997), 355–385. https://doi.org/10.1162/pres.1997.6.4.355

[4] Dana H. Ballard, Mary M. Hayhoe, Polly K. Pook, and Rajesh P. N. Rao. 1997. Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences* 20, 4 (Dec. 1997), 723–742. https://doi.org/10.1017/S0140525X97001611

[5] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: adding an adjective rating scale. *J. Usability Studies* 4, 3 (May 2009), 114–123.

[6] Michael Barz, Sebastian Kapp, Jochen Kuhn, and Daniel Sonntag. 2021. Automatic Recognition and Augmentation of Attended Objects in Real-time using Eye Tracking and a Head-mounted Display. In *ACM Symposium on Eye Tracking Research and Applications* (Virtual Event, Germany) *(ETRA '21 Adjunct)*. ACM, New York, NY, USA, Article 3, 4 pages. https://doi.org/10.1145/3450341.3458766

[7] Kenan Bektaş, Jannis Strecker, Simon Mayer, and Kimberly Garcia. 2024. Gaze-enabled activity recognition for augmented reality feedback. *Computers & Graphics* 119 (2024), 103909. https://doi.org/10.1016/j.cag.2024.103909

[8] Kenan Bektaş, Jannis Strecker, Simon Mayer, Kimberly Garcia, Jonas Hermann, Kay Erik Jenß, Yasmine Sheila Antille, and Marc Solèr. 2023. GEAR: Gaze-enabled augmented reality for human activity recognition. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications (ETRA '23)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3588015.3588402

[9] Kenan Bektaş, Arzu Çöltekin, Jens Krüger, Andrew T. Duchowski, and Sara Irina Fabrikant. 2019. GeoGCD: improved visual search via gaze-contingent display. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*. ACM, Denver Colorado, 1–10. https://doi.org/10.1145/3317959.3321488

[10] Ed D. J. Berry, Richard J. Allen, Mark Mon-Williams, and Amanda H. Waterman. 2019. Cognitive Offloading: Structuring the Environment to Improve Children's Working Memory Task Performance. *Cognitive Science* 43, 8 (2019), e12770. https://doi.org/10.1111/cogs.12770

[11] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. *CoRR* abs/2004.10934 (2020), 17. arXiv:2004.10934 https://arxiv.org/abs/2004.10934

[12] Ali Borji and Laurent Itti. 2013. State-of-the-Art in Visual Attention Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1 (2013), 185–207. https://doi.org/10.1109/TPAMI.2012.89

[13] John Brooke. 1996. SUS: A 'Quick and Dirty' Usability Scale. In *Usability Evaluation In Industry*. CRC Press.

[14] Monica S. Castelhano and Chelsea Heaven. 2011. Scene Context Influences without Scene Gist: Eye Movements Guided by Spatial Associations in Visual Search. *Psychonomic Bulletin & Review* 18, 5 (Oct. 2011), 890–896. https://doi.org/10.3758/s13423-011-0107-8

[15] Ravi Teja Chadalavada, Henrik Andreasson, Maike Schindler, Rainer Palm, and Achim J. Lilienthal. 2020. Bi-directional navigation intent communication using spatial augmented reality and eye-tracking glasses for improved safety in human–robot interaction. *Robot. Comput.-Integr. Manuf.* 61, C (Feb. 2020), 15 pages. https://doi.org/10.1016/j.rcim.2019.101830

[16] Mingyuan Chu and Sotaro Kita. 2011. The Nature of Gestures' Beneficial Role in Spatial Problem Solving. *Journal of Experimental Psychology: General* 140, 1 (2011), 102–116. https://doi.org/10.1037/a0021790

[17] Marvin M. Chun and Nicholas B. Turk-Browne. 2007. Interactions between Attention and Memory. *Current Opinion in Neurobiology* 17, 2 (April 2007), 177–184. https://doi.org/10.1016/j.conb.2007.03.005

[18] Nelson Cowan. 2010. The Magical Mystery Four: How Is Working Memory Capacity Limited, and Why? *Current Directions in Psychological Science* 19, 1 (Feb. 2010), 51–57. https://doi.org/10.1177/0963721409359277

[19] M. P. Eckstein. 2011. Visual search: A retrospective. *Journal of Vision* 11, 5 (Dec. 2011), 14–14. https://doi.org/10.1167/11.5.14

[20] Weili Fang, Ling Ma, Peter E. D. Love, Hanbin Luo, Lieyun Ding, and Ao Zhou. 2020. Knowledge Graph for Identifying Hazards on Construction Sites: Integrating Computer Vision with Ontology. *Automation in Construction* 119 (Nov. 2020), 103310. https://doi.org/10.1016/j.autcon.2020.103310

[21] Stephan Gammeter, Alexander Gassmann, Lukas Bossard, Till Quack, and Luc Van Gool. 2010. Server-side object recognition and client-side object tracking for mobile augmented reality. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. IEEE, San Francisco, CA, USA, 1–8. https://doi.org/10.1109/CVPRW.2010.5543248

[22] Aaron L. Gardony, Tad T. Brunyé, and Holly A. Taylor. 2015. Navigational Aids and Spatial Memory Impairment: The Role of Divided Attention. *Spatial Cognition & Computation* 15, 4 (Oct. 2015), 246–284. https://doi.org/10.1080/13875868.2015.1059432

[23] Ioannis Giannopoulos, Peter Kiefer, and Martin Raubal. 2012. GeoGazemarks: providing gaze history for the orientation on small display maps. In *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, Santa Monica California USA, 165–172. https://doi.org/10.1145/2388676.2388711

[24] Kerstin Gidlöf, Annika Wallin, Richard Dewhurst, and Kenneth Holmqvist. 2013. Using Eye Tracking to Trace a Cognitive Process: Gaze Behaviour During Decision Making in a Natural Environment. *Journal of Eye Movement Research* 6, 1 (Jan. 2013), 14. https://doi.org/10.16910/jemr.6.1.3

[25] Sam J. Gilbert. 2015. Strategic Use of Reminders: Influence of Both Domain-General and Task-Specific Metacognitive Confidence, Independent of Objective Memory Ability. *Consciousness and Cognition* 33 (May 2015), 245–260. https://doi.org/10.1016/j.concog.2015.01.006

[26] Sam J. Gilbert, Annika Boldt, Chhavi Sachdeva, Chiara Scarampi, and Pei-Chun Tsai. 2023. Outsourcing Memory to External Tools: A Review of 'Intention Offloading'. *Psychonomic Bulletin & Review* 30, 1 (Feb. 2023), 60–76. https://doi.org/10.3758/s13423-022-02139-4

[27] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*. IEEE Computer Society, USA, 580–587. https://doi.org/10.1109/CVPR.2014.81

[28] Fabian Göbel, Kuno Kurzhals, Martin Raubal, and Victor R. Schinazi. 2020. Gaze-Aware Mixed-Reality: Addressing Privacy Issues with Eye Tracking. In *Proceedings of the 1st Workshop on Exploring Potentially Abusive Ethical, Social and Political Implications of Mixed Reality Research in HCI at CHI'20*. ETH Zurich, Honolulu, HI, USA, 6. https://doi.org/10.3929/ETHZ-B-000409514

[29] Jan Grau, Simon Mayer, Jannis Strecker, Kimberly Garcia, and Kenan Bektaş. 2024. Gaze-based Opportunistic Privacy-preserving Human-Agent Collaboration. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, 6. https://doi.org/10.1145/3613905.3651066

[30] Jens Grubert, Tobias Langlotz, Stefanie Zollmann, and Holger Regenbrecht. 2017. Towards Pervasive Augmented Reality: Context-Awareness in Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics* 23, 6 (June 2017), 1706–1724. https://doi.org/10.1109/TVCG.2016.2543720

[31] Simon Harper, Eleni Michailidou, and Robert Stevens. 2009. Toward a definition of visual complexity as an implicit measure of cognitive load. *ACM Transactions on Applied Perception* 6, 2 (Feb. 2009), 1–18. https://doi.org/10.1145/1498700.1498704

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

[33] John M. Henderson, Myriam Chanceaux, and Tim J. Smith. 2009. The Influence of Clutter on Real-World Scene Search: Evidence from Search Efficiency and Eye Movements. *Journal of Vision* 9, 1 (Jan. 2009), 32. https://doi.org/10.1167/9.1.32

[34] Laurent Itti. 2000. *Models of bottom-up and top-down visual attention.* Ph. D. Dissertation. California Institute of Technology, USA. Advisor(s) Koch, Christof.

[35] Sebastian Kapp, Michael Barz, Sergey Mukhametov, Daniel Sonntag, and Jochen Kuhn. 2021. ARETT: Augmented Reality Eye Tracking Toolkit for Head Mounted Displays. *Sensors* 21, 6 (March 2021), 2234. https://doi.org/10.3390/s21062234

[36] Çağla Çığ Karaman and Tevfik Metin Sezgin. 2018. Gaze-based predictive user interfaces: Visualizing user intentions in the presence of uncertainty. *International Journal of Human-Computer Studies* 111 (2018), 78–91. https://doi.org/10.1016/j.ijhcs.2017.11.005

[37] Dagmar Kern, Paul Marshall, and Albrecht Schmidt. 2010. Gazemarks: gaze-based visual placeholders to ease attention switching. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, Atlanta Georgia USA, 2093–2102. https://doi.org/10.1145/1753326.1753646

[38] David E. Kieras and Anthony J. Hornof. 2014. Towards accurate and practical predictive models of active-vision-based visual search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 3875–3884. https://doi.org/10.1145/2556288.2557324

[39] Kristin Koch, Judith McLean, Ronen Segev, Michael A. Freed, Michael J. Berry, Vijay Balasubramanian, and Peter Sterling. 2006. How Much the Eye Tells the Brain. *Current Biology* 16, 14 (July 2006), 1428–1434. https://doi.org/10.1016/j.cub.2006.05.056

[40] Jacob Leon Kröger, Otto Hans-Martin Lutz, and Florian Müller. 2020. What Does Your Gaze Reveal About You? On the Privacy Implications of Eye Tracking. In *Privacy and Identity Management. Data for Better Living: AI and Privacy*, Michael Friedewald, Melek Önen, Eva Lievens, Stephan Krenn, and Samuel Fricker (Eds.). Vol. 576. Springer International Publishing, Cham, 226–241. https://doi.org/10.1007/978-3-030-42504-3_15

[41] Daniel J. Liebling and Sören Preibusch. 2014. Privacy Considerations for a Pervasive Eye Tracking World. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM, Seattle Washington, 1169–1177. https://doi.org/10.1145/2638728.2641688

[42] Luyang Liu, Hongyu Li, and Marco Gruteser. 2019. Edge Assisted Real-Time Object Detection for Mobile Augmented Reality. In *The 25th Annual International Conference on Mobile Computing and Networking* (Los Cabos, Mexico) *(MobiCom '19)*. ACM, New York, NY, USA, Article 25, 16 pages. https://doi.org/10.1145/3300061.3300116

[43] Shang Lu, Yerly Paola Sanchez Perdomo, Xianta Jiang, and Bin Zheng. 2020. Integrating Eye-Tracking to Augmented Reality System for Surgical Training. *Journal of Medical Systems* 44, 11 (Sept. 2020), 192. https://doi.org/10.1007/s10916-020-01656-w

[44] Microsoft. 2022. Device Portal API Reference - Mixed Reality. Retrieved February 10, 2024 from https://learn.microsoft.com/en-us/windows/mixed-reality/develop/advanced-concepts/device-portal-api-reference.

[45] Microsoft. 2022. Eye Tracking in Mixed Reality Toolkit — MRTK2. Retrieved February 10, 2024 from https://learn.microsoft.com/en-us/windows/mixed-reality/mrtk-unity/mrtk2/features/input/eye-tracking/eye-tracking-main?view=mrtkunity-2022-05.

[46] Microsoft. 2022. Spatial Awareness Getting Started - MRTK 2. Retrieved February 10, 2024 from https://learn.microsoft.com/en-us/windows/mixed-reality/mrtk-unity/mrtk2/features/spatial-awareness/spatial-awareness-getting-started.

[47] Microsoft. 2023. Eye Tracking on HoloLens 2. Retrieved February 10, 2024 from https://learn.microsoft.com/en-us/windows/mixed-reality/design/eye-tracking.

[48] George A. Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63, 2 (March 1956), 81–97. https://doi.org/10.1037/h0043158

[49] Rainer Mühlhoff. 2023. Predictive Privacy: Collective Data Protection in the Context of Artificial Intelligence and Big Data. *Big Data & Society* 10, 1 (Jan. 2023), 205395172311668. https://doi.org/10.1177/20539517231166886

[50] John F. Nestojko, Jason R. Finley, and Henry L. Roediger. 2013. Extending Cognition to External Agents. *Psychological Inquiry* 24, 4 (Oct. 2013), 321–325. https://doi.org/10.1080/1047840X.2013.844056

[51] Jason Orlosky, Misha Sra, Kenan Bektaş, Huaishu Peng, Jeeeun Kim, Nataliya Kos'myna, Tobias Höllerer, Anthony Steed, Kiyoshi Kiyokawa, and Kaan Akşit. 2021. Telelife: The Future of Remote Living. *Frontiers in Virtual Reality* 2 (Nov. 2021), 763340. https://doi.org/10.3389/frvir.2021.763340

[52] Hiroto Oshimi, Monica Perusquía-Hernández, Naoya Isoyama, Hideaki Uchiyama, and Kiyoshi Kiyokawa. 2023. LocatAR: An AR Object Search Assistance System for a Shared Space. In *Proceedings of the Augmented Humans International Conference 2023* (Glasgow, United Kingdom) *(AHs '23)*. ACM, New York, NY, USA, 66–76. https://doi.org/10.1145/3582700.3582712

[53] Karen Panetta, Qianwen Wan, Aleksandra Kaszowska, Holly A. Taylor, and Sos Agaian. 2019. Software Architecture for Automating Cognitive Science Eye-Tracking Data Analysis and Object Annotation. *IEEE Transactions on Human-Machine Systems* 49, 3 (2019), 268–277. https://doi.org/10.1109/THMS.2019.2892919

[54] Pranav Parekh, Shireen Patel, Nivedita Patel, and Manan Shah. 2020. Systematic Review and Meta-Analysis of Augmented Reality in Medicine, Retail, and Games. *Visual Computing for Industry, Biomedicine, and Art* 3, 1 (Sept. 2020), 21. https://doi.org/10.1186/s42492-020-00057-7

[55] Candace E. Peacock, Ting Zhang, Brendan David-John, T. Scott Murdison, Matthew J. Boring, Hrvoje Benko, and Tanya R. Jonker. 2022. Gaze dynamics are sensitive to target orienting for working memory encoding in virtual reality. *Journal of Vision* 22, 1 (Jan. 2022), 2. https://doi.org/10.1167/jov.22.1.2

[56] Alexander Plopski, Teresa Hirzle, Nahal Norouzi, Long Qian, Gerd Bruder, and Tobias Langlotz. 2023. The Eye in Extended Reality: A Survey on Gaze Interaction and Eye Tracking in Head-worn Extended Reality. *Comput. Surveys* 55, 3 (March 2023), 1–39. https://doi.org/10.1145/3491207

[57] Michael I. Posner. 1980. Orienting of Attention. *Quarterly Journal of Experimental Psychology* 32, 1 (Feb. 1980), 3–25. https://doi.org/10.1080/00335558008248231

[58] Roope Raisamo, Ismo Rakkolainen, Päivi Majaranta, Katri Salminen, Jussi Rantala, and Ahmed Farooq. 2019. Human augmentation: Past, present and future. *International Journal of Human-Computer Studies* 131 (Nov. 2019), 131–143. https://doi.org/10.1016/j.ijhcs.2019.05.008

[59] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, NV, USA, 779–788. https://doi.org/10.1109/CVPR.2016.91

[60] Joseph Redmon and Ali Farhadi. 2018. YOLOv3: An Incremental Improvement. *CoRR* abs/1804.02767 (2018), 6. arXiv:1804.02767 http://arxiv.org/abs/1804.02767

[61] Evan F. Risko and Sam J. Gilbert. 2016. Cognitive Offloading. *Trends in Cognitive Sciences* 20, 9 (Sept. 2016), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

[62] Mike Scaife and Yvonne Rogers. 1996. External cognition: how do graphical representations work? *International Journal of Human-Computer Studies* 45, 2 (Aug. 1996), 185–213. https://doi.org/10.1006/ijhc.1996.0048

[63] Stephen W. Scheff. 2016. *Fundamental Statistical Principles for the Neurobiologist: A Survival Guide*. Academic Press is an imprint of Elsevier, Amsterdam Boston. https://doi.org/10.1016/C2015-0-02471-6

[64] D. J. Simons and D. T. Levin. 1997. Change Blindness. *Trends in Cognitive Sciences* 1, 7 (Oct. 1997), 261–267. https://doi.org/10.1016/S1364-6613(97)01080-2

[65] Statista. 2023. Number of smartphone users worldwide 2013-2028. Retrieved May 5, 2023 from https://www.statista.com/forecasts/1143723/smartphone-users-in-the-world.

[66] Michael Stengel and Marcus Magnor. 2016. Gaze-Contingent Computational Displays: Boosting perceptual fidelity. *IEEE Signal Processing Magazine* 33, 5 (2016), 139–148. https://doi.org/10.1109/MSP.2016.2580913

[67] Jannis Strecker, Kimberly García, Kenan Bektaş, Simon Mayer, and Ganesh Ramanathan. 2023. SOCRAR: Semantic OCR through Augmented Reality. In *Proceedings of the 12th International Conference on the Internet of Things* (Delft, Netherlands) *(IoT '22)*. ACM, New York, NY, USA, 25–32. https://doi.org/10.1145/3567445.3567453

[68] Shigeo Takahashi, Akane Uchita, Kazuho Watanabe, and Masatoshi Arikawa. 2022. Gaze-Driven Placement of Items for Proactive Visual Exploration. *Journal of Visualization* 25, 3 (2022), 613–633. https://doi.org/10.1007/s12650-021-00808-5

[69] Arthur Tang, Charles Owen, Frank Biocca, and Weimin Mou. 2003. Comparative effectiveness of augmented reality in object assembly. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. ACM, New York, NY, USA, 73–80. https://doi.org/10.1145/642611.642626

[70] Vu Tuan Tran and Norbert Fuhr. 2012. Using eye-tracking with dynamic areas of interest for analyzing interactive information retrieval. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Portland, Oregon, USA) *(SIGIR '12)*. ACM, New York, NY, USA, 1165–1166. https://doi.org/10.1145/2348283.2348521

[71] Junichi Tsurukawa, Mohammed Al-Sada, and Tatsuo Nakajima. 2015. Filtering visual information for reducing visual cognitive load. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers* (Osaka, Japan) *(UbiComp/ISWC'15 Adjunct)*. ACM, New York, NY, USA, 33–36. https://doi.org/10.1145/2800835.2800852

[72] Freek van Ede, Sammi R. Chekroud, and Anna C. Nobre. 2019. Human gaze tracks attentional focusing in memorized visual space. *Nature Human Behaviour* 3, 5 (May 2019), 462–470. https://doi.org/10.1038/s41562-019-0549-y

[73] Katherine Volz, Euijung Yang, Rachel Dudley, Elizabeth Lynch, Maria Dropps, and Michael C. Dorneich. 2016. An Evaluation of Cognitive Skill Degradation in Information Automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 60, 1 (Sept. 2016), 191–195. https://doi.org/10.1177/1541931213601043

[74] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. 2023. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Vancouver, BC, Canada, 7464–7475. https://doi.org/10.1109/CVPR52729.2023.00721

[75] Adrian F. Ward, Kristen Duke, Ayelet Gneezy, and Maarten W. Bos. 2017. Brain Drain: The Mere Presence of One's Own Smartphone Reduces Available Cognitive Capacity. *Journal of the Association for Consumer Research* 2, 2 (April

2017), 140–154. https://doi.org/10.1086/691462

[76] Jeremy M. Wolfe. 2021. Guided Search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review* 28, 4 (Aug. 2021), 1060–1092. https://doi.org/10.3758/s13423-020-01859-9

[77] Jeremy M. Wolfe and Todd S. Horowitz. 2017. Five factors that guide attention in visual search. *Nature Human Behaviour* 1, 3 (March 2017), 0058. https://doi.org/10.1038/s41562-017-0058

[78] XRSI. 2020. *The XRSI Privacy and Safety Framework*. Technical Report. XR Safety Initiative. https://xrsi.org/publication/the-xrsi-privacy-framework

[79] Shengdong Zhao, Felicia Tan, and Katherine Fennedy. 2023. Heads-Up Computing Moving Beyond the Device-Centered Paradigm. *Commun. ACM* 66, 9 (Sept. 2023), 56–63. https://doi.org/10.1145/3571722

[80] Zhong-Qiu Zhao, Peng Zheng, Shou-Tao Xu, and Xindong Wu. 2019. Object Detection With Deep Learning: A Review. *IEEE Transactions on Neural Networks and Learning Systems* 30, 11 (Nov. 2019), 3212–3232. https://doi.org/10.1109/TNNLS.2018.2876865

[81] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. 2023. Object Detection in 20 Years: A Survey. *Proc. IEEE* 111, 3 (2023), 257–276. https://doi.org/10.1109/JPROC.2023.3238524

Received November 2023; revised January 2024; accepted March 2024