Gaze Comes in Handy: Predicting and Preventing Erroneous Hand Actions in AR-Supported Manual Tasks

Julian Wolf
ETH Zurich

Quentin Lohmeyer
ETH Zurich

Christian Holz
ETH Zurich

Mirko Meboldt
ETH Zurich

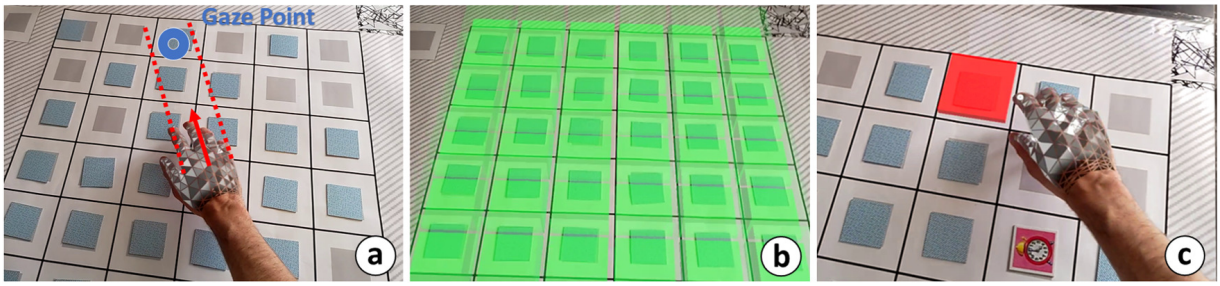

Figure 1: On the example of a memory card game, which requires hand-eye coordination, we demonstrate our closed-loop support system, which analyzes the user's gaze, hand velocity and hand trajectory in real-time to warn the user of predictably erroneous hand actions. (a) Our system records the user's gaze and hand movements projected into (b) the registered 3D environment to predict the next hand interaction. (c) Predictions are compared to a ground truth game layout to display either a green, yellow or red visual alert.

Abstract

Emerging Augmented Reality headsets incorporate gaze and hand tracking and can, thus, observe the user's behavior without interfering with ongoing activities. In this paper, we analyze hand-eye coordination in real-time to predict hand actions during target selection and warn users of potential errors before they occur. In our first user study, we recorded 10 participants playing a memory card game, which involves frequent hand-eye coordination with little task-relevant information. We found that participants' gaze locked onto target cards 350 ms before the hands touched them in 73.3 % of all cases, which coincided with the peak velocity of the hand moving to the target. Based on our findings, we then introduce a closed-loop support system that monitors the user's fingertip position to detect the first card turn and analyzes gaze, hand velocity and trajectory to predict the second card before it is turned by the user. In a second study with 12 participants, our support system correctly displayed color-coded visual alerts in a timely manner with an accuracy of 85.9%. The results indicate the high value of eye and hand tracking features for behavior prediction and provide a first step towards predictive real-time user support.

Index Terms: • Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Mixed / augmented reality

1 Introduction

Augmented reality head-mounted displays (AR HMDs) [24] are promising for industrial and clinical applications, providing operators with the information needed to perform manual tasks such as assembly [40], maintenance [32], or surgery [9]. Studies have shown that displaying contextual information can improve spatial understanding [8, 33] and reduce both time expenditures and the probability of errors [3, 16]. The same studies have also shown that users still perform errors while wearing AR glasses. In order to provide effective support during expert activities, recent work has used AR HMDs to capture and analyze user behavior by tracking visual markers on manipulated objects or by detecting certain steps of a procedure [27, 31]. Results showed that the relevant information can be adapted to provide the right instructions at the right time and place [42] or that real-time feedback on user actions can be provided [35]. So far, efforts have required processing footage from the integrated cameras while wearing AR glasses, limiting the depth of their processing stack for real-time purposes. Recent AR HMDs incorporate better hardware for computation and can thus provide eye gaze and hand tracking in real time, both of which have shown to be suitable for analyzing behavioral patterns outside AR contexts [4, 17, 21]. As gaze behavior is highly task-dependent [39], it provides deep insights into ongoing cognitive processes [10]. Hand tracking can be used to infer hand actions [12], which provide insights into the user's performance of manual tasks [21, 31].

Combining sensing modalities in recent HMDs creates a novel opportunity for capturing hand-eye coordination, which is the task-dependent relationship between hands and eyes [34]. Hand-eye coordination has been successfully tracked to automatically detect usability problems in eye tracking video recordings [29] or to predict user's target selection while reaching to a virtual object in a Virtual Reality (VR) space [6]. During hand-eye coordination, the eyes provide the necessary information to plan the motor system's movements [7, 37], making gaze a suitable indicator for predicting hand actions. This could be particularly useful in industrial and clinical applications, where real-time feedback to anticipated actions could combat the high cost of user errors.

In human-computer interaction, previous work on hand-eye coordination has investigated predicting target selection of virtual objects in VR [6], but no work has predicted target selection in *real-world handling tasks* that include physical object manipulation. Reaching for and picking up a physical object needs precise coordination that affects the time the gaze must arrive on the target for a seamless interaction [7].

In this paper, we investigate to what extent the real-time analysis of eye gaze and hand tracking lends itself to predicting hand actions in a *real-world task*. In a second step, we examine how effectively ongoing hand actions can be intercepted through visual

alerts before they are executed and how participants perceive this support. We introduce a method to analyze gaze patterns in real-time to predict target locations that users will reach next. Our method simultaneously tracks and analyzes hand movement to confirm the current gaze prediction and narrow the set of possible target locations. We illustrate our method on the example of a memory card game, which requires frequent hand-eye coordination during card turns with little task-relevant information and is thus representative of more general interaction. The memory game is particularly interesting because it is a fast, repetitive procedure where decisions are made on-the-fly and because it is characterized by a high frequency of target selections. It therefore supports the recording of high sample sizes in a well structured and controlled environment that is fully visible and accessible to the user (no obstacles or occlusions). A characteristic of memory games is that the correct choice of the second card depends on the first card choice. We therefore also investigate hand tracking features, i.e., tracked finger joints, that allow for the detection of the first card turn. Based on hand and gaze data recordings from a first user study, we derive a logic for closed-loop support that we then implement on an AR HMD to display color-coded visual alerts to the user. Our system monitors the user's fingertip position in proximity to card locations to detect the first card turn and then predicts the second card. Predictions are compared to a ground truth game layout stored on the device to display green, yellow or red visual alerts, depending on whether the predicted target is correct, incorrect but adjacent to the correct card, or neither correct nor adjacent. Our second user study investigated our method in real time with 12 more participants, showing that it predicted target locations online with 85.9% accuracy while being rated as supportive, well working and stimulating during qualitative interviews.

In summary, we make the following contributions in this paper:

1) a first study with 10 participants on the accuracy of hand motion prediction, showing that the gaze locked onto target cards 350 ms before touch in 73.3% of cases (averaged over both card turns), which coincided with the moment of hand movement deceleration. We further show that the set of possible targets can be significantly reduced based on the hand trajectory and that fingertip proximity to a card is a promising indicator for monitoring first card turns
2) a novel method for AR-supported manual real-world tasks that analyzes hand-eye coordination in real-time to predict hand actions during target selection. Our method extends previous work on predicting target selection in VR [6], i.e., using a velocity threshold and the gaze target, by combining gaze prediction with a hand trajectory and with a temporal coupling of gaze and hand features optimized for physical object manipulation
3) a second user study with 12 participants to evaluate the real-time effectiveness of our method to stop participants' motions in time (i.e., before they reach and start manipulating a target), showing correctly timed and placed visual alerts with an accuracy of 85.9% over 384 card pairs played.

2 Related Work

Our work is related to hand-eye coordination, both in (1) real-world settings and in (2) human-computer interaction, to (3) predicting target selection and to (4) context-aware augmented reality.

2.1 Hand-Eye Coordination in Real-World Settings

Several studies have shown a task-dependent relationship between hands and eyes, namely, hand-eye coordination. Land et al. [25] investigated participants during "tea making" and found that each action is typically associated with four to six preceding fixations on task-relevant objects. Johansson et al. [19] extended the investigations to object manipulations and found similar behaviors on landmarks (e.g. objects and obstacles) relevant to the task. In a study conducted by Helsen et al. [14], participants had to move their hand as fast as possible from one physical button to another. They found that the gaze initiated 70 ms earlier than the hand movement, taking approximately two saccades to arrive on the target. The gaze stabilized on the target at about 50% of the total hand response time, which was also approximately the moment the hand started decelerating.

Similar to Garcia-Hernando et al. [12], we consider a hand action as an interaction between the hands and a physical 3D object (e.g., turning a screwdriver, pouring milk). The kinematics of hand actions can be divided into several phases, starting with the hands 'reaching towards an object' (target selection), 'grasping the object' to 'manipulating the object' [11, 19]. As we ultimately aim at supporting users in procedural tasks where gaze-behavior is highly task-dependent [39], we assume that 'target selection' can often be associated with the user's intent to perform a hand action with the respective object.

2.2 Hand-Eye Coordination in Human-Computer Interaction

Early work has dealt with analyzing mouse cursor trajectories and gaze behavior during interaction with graphical user interfaces [5,36] or web search [18]. While the gaze often led the mouse, researchers found several behavioral patterns compared to the more invariant patterns observed in real-world settings. Mutasim et al. [30] studied gaze movements in a VR hand-eye coordination training system that displayed a grid of virtual targets in front of a wall. They found the gaze arriving on target on average 250 ms before touch.

In a study setup similar to our work, Weill-Tessier and Gellersen [41] combined remote eye tracking with a Leap Motion hand tracking sensor to record the relation between gaze and hand movements while participants played a memory game on a tablet screen. They applied a velocity-based algorithm on the hand motion data to detect hovering states, i.e, when the hand was in a standby position, contrary to hand movement in our method. Their goal was to investigate whether the gaze behavior during hovering provided insights about the users' cognitive states in decision making (decisive, indecisive). Results showed that the number and duration of fixations during hover could not reveal indecision and that target selection was closely dependent on the target's location.

2.3 Predicting Target Selection

In user interfaces, target selection has a rich history in desktop environments. For example, Baudisch et al. [2] predicted possible targets during a drag-and-drop task on a large screen by analyzing cursor trajectories. Koochaki et al. [22] predicted user intent while participants were shown an image of a kitchen environment on a computer screen. Using a CNN to detect relevant objects and an LSTM to learn temporal features of the gaze transitioning between these objects, four different tasks were distinguished.

Target prediction also finds increasing use in VR. Marwecki et al. [28] analyzed eye gaze patterns to detect regions of interest in a virtual environment and covertly adapted the virtual scene, including the relocation of virtual elements to allow users to reach out and grasp physical props. Cheng et al. [6] predicted users' touch locations in VR by analyzing their gaze and hand motions to redirect the hand to a haptic prop. Using the gaze target and a velocity threshold of 3cm/s, their method achieved 97% accuracy. Contrary to our setup, hand movements were slow and participants were told which target to aim for. Our method is intended to work with very fast hand movements during real-world interaction and allows participants to make their own choice on-the-fly without restrictions.

2.4 Context-Aware Support in Augmented Reality

Context-aware augmented reality aims at automatically changing the content displayed in AR based on the current context (e.g., interpretation of the surrounding scene) to provide better support, mainly focusing on procedural applications such as surgery, assembly or maintenance.

Within surgical applications, research has primarily focused on robotic surgery or laparoscopy. Katić et al. [20] used different parameters during minimally invasive surgery (e.g., 'current instrument', 'distance to anatomical structures') to detect the current procedural step and to assess the current risk. They then combined this information to highlight specific anatomical structures. Gras et al. [13] calculated several Euclidean distance measurements between the tooltips, the gaze point, and the patient anatomy in simulated robotic surgery. Using these features, they trained a multi-Gaussian process model to automatically infer the desired AR display view at any point of the procedure.

In industrial applications such as maintenance, machine operation or assembly, much work on context-aware augmented reality has been done with AR Glasses. Henderson and Feiner [15] applied visual markers during AR-guided assembly to track the movement of handled objects and assess the user's current activity. Based on the relative position of these objects, they could automatically transition to the next step of the procedure or, if the user moved a wrong object, display an error message. Peterson and Stricker [35] proposed a system that compares video recording with a reference workflow to track the currently executed action at runtime. They used this awareness to adjust the displayed information for the user's needs. Ng et al. [31] detected the user's hands and particular task-relevant objects in video recordings. A real-time analysis of the spatial-temporal relation of the detected objects and hands then inferred the current step to provide contextual instructions in AR.

Taken together, previous work has explored means to automatically adapt AR support to the current context, but no work has investigated how hand and gaze features can be combined *online* to provide predictive AR support for potential errors before they occur.

3 Study 1: Patterns in Hand-Eye Coordination

In this study, two players played a memory game. The study's purpose was to record and analyze gaze and hand tracking data with a high level of task immersion to find a pattern that could be used to predict the next hand movement.

3.1 Apparatus

We implemented a Microsoft HoloLens 2 app using Unity's 3D game engine (2019.4.14f1) and the Mixed Reality Toolkit (MRTK 2.4.0). Our app positions a virtual playing field on the top of the real field, such that hand and eye gaze interactions with the real game cards resulted in measurable virtual interactions, as shown in Fig. 3. HoloLens 2 reports the wearer's gaze with an angular accuracy of 1.5° around the actual target and a recording rate of 30 fps [1]. Participants were standing in front of a table with an approximate distance between the head and memory card game of 60–130 cm, resulting in a measurement error of 1.50–3.25 cm. Through hand tracking, the 3D positions of 26 hand joints and the overall 3D velocity of the hand can be measured. We recorded the index fingertip, thumb tip, and hand velocity for our investigations. The recording rate varies from a low frame rate when the hand enters the field of view up to a maximum frame rate of 60 fps. Our app writes both gaze and hand tracking data into a buffer saved to a text file with a recording rate of 50 fps to synchronize all measurements. In this study, the AR HMDs did not display content and merely recorded hand tracking and gaze data next to a first-person video.

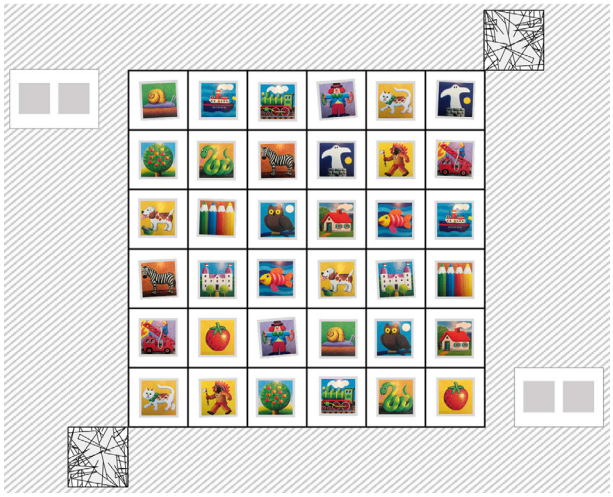

Figure 2: Paper sheet with imprinted 6 x 6 grid for memory cards and two Vuforia markers for 3D registration.

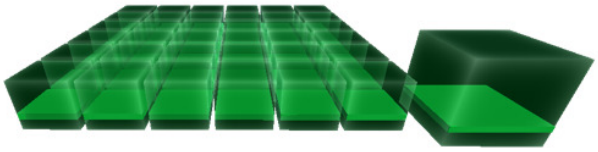

Figure 3: Front view of a two-layer virtual playing field with 36 fields of the same dimensions in the horizontal plane as the real playing field. The flat green cuboids register the user's eye gaze while the green transparent cuboids register when the fingertips are within proximity for a potential card turn. A touch of the large cuboid on the right-hand side is used during the second user study to reset the support system. The virtual playing field is only visible during calibration and is faded out before the game starts.

3.2 Task and Procedure

In each experiment, two players competed in a memory card game, where one player, i.e. the study participant, was wearing a Microsoft HoloLens 2. The players stood in front of a table with an imprinted 6 x 6 grid. Each field in the grid measured 10 cm x 10 cm and contained one memory card. The cards constituted a memory card game with 18 pairs of cards, i.e. 36 cards in total (Fig. 2). Players wearing the Hololens 2 were instructed to only play with one hand.

3.2.1 App Calibration

Before each game, participants calibrated the system. First, they were guided through the eye tracking system's calibration procedure, an automated routine available on Microsoft HoloLens 2. Second, participants were instructed to place a virtual grid over the physical grid by confirming the position of two Vuforia markers printed at two diagonally opposing corners of the physical grid (cf. Fig. 2). After confirmation of both marker positions via touch gestures, the virtual grid (cf. Fig. 3) was placed between both marker positions, inheriting the spatial orientation of the first marker. Participants could then either confirm correct placement and hide the virtual field or repeat the calibration process.

3.2.2 Game Structure

At the beginning of the game, all 36 cards were shuffled by the study moderator and placed on the table with their colored sides facing up (Fig. 2). Players then had one minute to memorize the location

of as many pairs of cards as possible. After the minute, the cards were flipped and the first player chose two cards to be turned over. If the cards belonged together, they were removed from the game and placed on the field on the right-hand side of the grid, the player scored a point and could turn over another pair of cards. If the cards did not match, the cards were turned face down again and the other player's turn started. The game finished when no more cards were left. The player with the most correctly identified pairs of cards won.

3.3 Participants

We recruited eleven participants (5 male, 6 female, mean age = 29.2 years, SD = 2.8 years) with normal or corrected-to-normal vision. All participants stated to be right-handed. One participant's records had to be excluded for insufficient tracking quality, resulting in a total number of ten participants.

3.4 Data Analysis

During the experiments, we recorded the gaze target, i.e., the card the participant was currently looking at, the 3D position of the index fingertip, the thumb tip, and the 3D velocity of the hand, with a fixed frame rate of 50 fps and saved all data to a text file. Simultaneously, we recorded a first-person video that displayed the current frame number in the bottom left corner. We observed and corrected a delay between video recording and displayed frame number of approximately 12 frames. All measurements were expressed in the coordinate system of the virtual playing field.

As a first postprocessing step, we defined the two events 'First Card Turn' (FCT) and 'Second Card Turn' (SCT) as the time the participant started turning the respective card. These events represent the start of a hand action we intend to predict with our method. Using the first-person video recordings for comparison, we manually labeled each of these events with the identification number (ID) of the turned cards, ranging from 1 to 36, in the output file recorded with HoloLens 2. Secondly, gaze behavior was then analyzed to find a predictor for target selection of future hand actions. Using a sliding window, we categorized 4 or more gaze measurements (80ms) on the same target as a 'fixation' and categorized remaining measurements as 'background'. This resulted in a time series with either 'fixation' or 'background' labels, where each data point of a fixation was associated with a card ID of the examined card. We then performed a retrospective analysis for each card event 'FCT' and 'SCT' and split the last 3 seconds of gaze behavior prior to the card events into windows of length 100 ms. For each FCT or SCT, we iterated through all windows and checked if the card ID of a fixation in a window matched with the card ID of the target card. If yes, this resulted in a value of '1' for the respective window. If not, it resulted in a value of '0'. For each window position, we summed up these results ('0'/'1') over all FCTs/SCTs and divided them by the total number of FCTs/SCTs. This resulted in the relative number of fixations on target cards for each window position, expressed in percent.

Hand movements were evaluated with a threefold objective. In a first step, we explored the hand velocity curve to investigate whether the hand movements 'card reach' and 'card turn' could be clearly distinguished. In this context, we investigated characteristic features in the hand velocity that occurred when the correct gaze prediction was made. Such a feature represents a trigger condition to confirm the current gaze prediction. As differences in hand tracking rate may occur, we interpolated missing data points with intermediate values.

Second, we investigated how the direction of the hand movement can be utilized as a boundary condition to limit possible targets. Based on the hand velocity vector in the horizontal plane, we calculated the shortest distances between all card locations and the current hand trajectory, i.e., the perpendicular distances d_{perp}, for each time step (cf. Equation 1). We tested different perpendicular and longitudinal distance thresholds to ensure the target card was

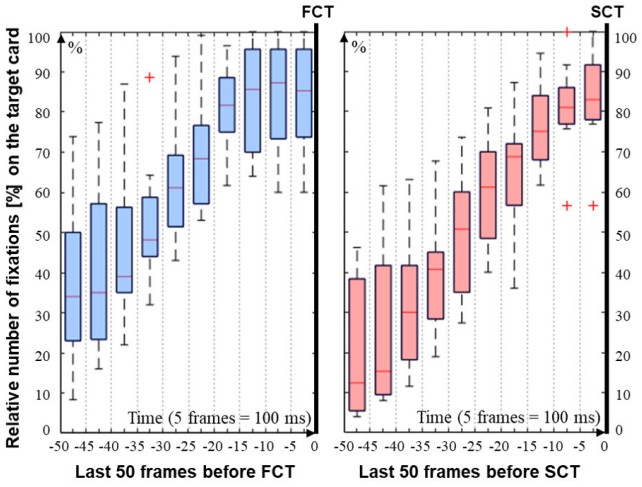

Figure 4: Last-second gaze behavior (50 fps) on the selected card before the first card turn (left) and second card turn (right) across all participants, divided into 100 ms time windows. Each value of a box plot represents the number of fixations on target cards for one participant, divided by all FCTs or SCTs played by the participant.

located within the trajectories bounds soon after the started card reach while excluding as many other cards as possible.

$$d_{perp} = \frac{\left\|(\overrightarrow{H_{Pos}C_{Pos}}) \times \vec{v}\right\|}{\|\vec{v}\|} \quad \begin{matrix} H_{Pos} = 3D\ HandPosition; \\ C_{Pos} = 3D\ CardPosition; \end{matrix} \qquad (1)$$

Last, we evaluated the positions of index fingertip and thumb tip for each card turn to investigate whether they could be used as an indicator for the first card turn. We defined cuboids above each card location that had the same horizontal dimensions as the fields and varied the height of these cuboids (similar to transparent cuboids in Fig. 3). We calculated the tracking rate, i.e., the amount of available hand tracking data points at a recording rate of 50fps, as well as the relative number of measured hits on the target card's cuboid for the index fingertip, thumb tip, and their center.

4 Results

On average, the ten recorded games took 5.2 min (SD=1.0 min) with a total of 141 card pairs played by the participants.

4.1 Analysis of Gaze Behavior

Eye gaze on the target card was generally low except for the last 1.5 seconds, where the fixations on the target card slowly started rising, and in particular for the last second, where this increase started climbing at a faster pace. Figure 4 shows how often participants were already examining the target card in the last second before the respective card turn divided into time windows with a duration of 100 ms.

Between 50 and 45 frames before FCT, participants were examining the target card on average in 35.4% of cases. This value rises steadily and starts stagnating approximately 20 frames before FCT with a mean of 81.1%, reaching its highest value just before the card turn with a mean value of 85.2%. We observe similar SCT behavior, though with an overall reduced percentage of fixations on the target card. Between 50 and 45 frames before SCT only 19.0% of fixations were registered on target cards. This value rises to 65.5% for the fourth-to-last window and reaches its maximum mean value of 83.3% just before SCT. Averaged over FCT and SCT, the gaze prediction reaches a value of 73.3% for the fourth-to-last time window, which corresponds to a prediction time of 350 ms.

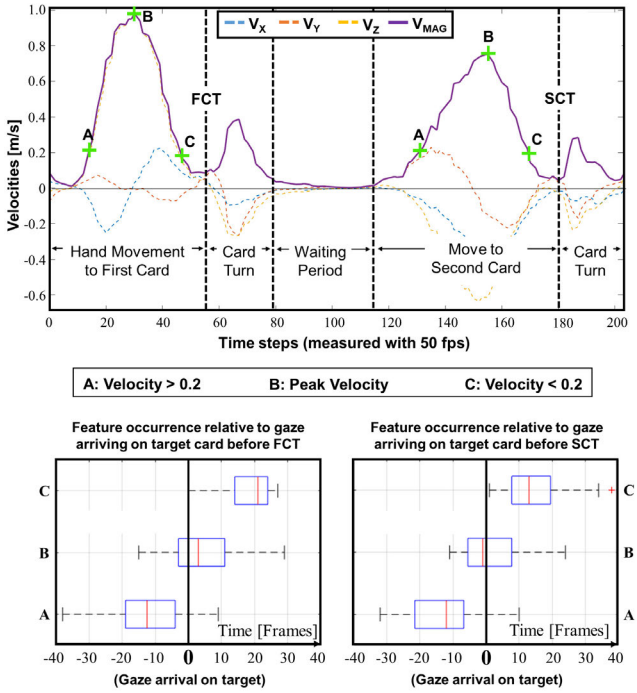

Figure 5: The top diagram shows the velocity components of an example hand sequence of one move (two card turns) and three hand velocity features that represent the start (A), peak velocity (B) and end (C) of a card reach. V_X represents the velocity in the lateral direction, V_Z represents the longitudinal direction, and V_Y represents the vertical direction. The bottom diagram shows the time interval between each hand velocity feature (A-C) and the gaze arriving on the target card before a card turn. A positive value indicates that the feature occurred after the gaze arrived on the target.

4.2 Analysis of Hand Movements

4.2.1 Hand Velocity

Figure 5 shows the hand velocity components and the resulting velocity magnitude for an example hand sequence. Each FCT and SCT consisted of two phases: (i) hand movement to a card (card reach) and the subsequent (ii) turning over of a card (card turn). Occasionally there were short periods during a move, in which the participant briefly interrupted their hand movement. These waiting periods occurred infrequently. We randomly selected and analyzed 30 (approximately 10% of all FCTs and SCTs) card reaches and card turns to differentiate the 'card reach' and 'card turn.' The average velocity during a card turn was 0.10 m/s (SD=0.02 m/s) with a duration of 0.38 s (SD=0.10 s). The average velocity during a card reach was 0.39 m/s (SD=0.11 m/s) with a duration of 0.92 s (SD=0.24 s). Both mean velocity and mean duration during card reach were significantly higher ($p<0.01$, Wilcoxon Signed-Rank Test) than when it was turned over. The two actions can thus be clearly distinguished from one another using these criteria.

4.2.2 Temporal Coupling of Eye Gaze and Hand Movement

Three features of each hand reach, i.e., the start, the peak, and the end of the movement, were extracted across all participants and related to the arrival of gaze on the target card (Fig. 5, Feature A-C) to derive a trigger condition for the current gaze prediction. For both FCT and SCT, the occurrence time of the peak velocity is, on average, very close to the time the gaze arrives on the target

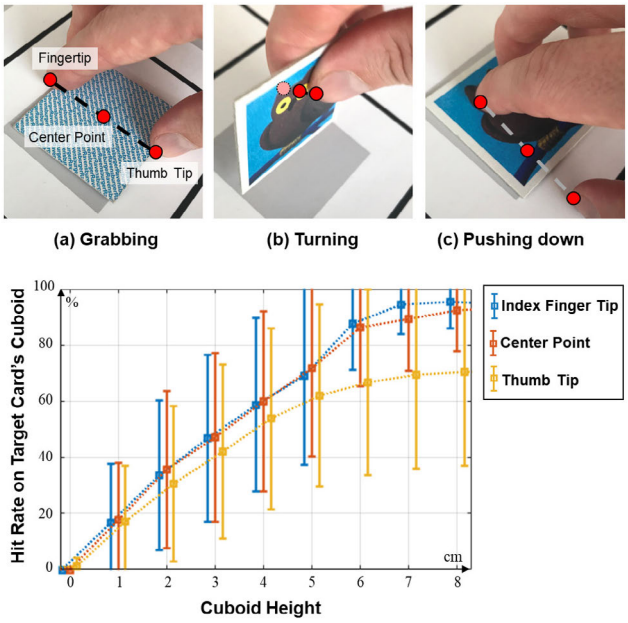

Figure 6: At the top, three characteristic scenes of a card turn are shown (a-c), with the two involved hand joints index fingertip and thumb tip as well as their center. At the bottom, over all first card turns the hit rate on the target card's cuboid is shown for different cuboid heights for the index fingertip, thumb tip and center. Error bars show the hit rates' standard deviation over all card turns.

card. The start of a hand movement represents an earlier but riskier prediction, while the end of a hand movement allows for a more conservative prediction.

4.2.3 Hand Trajectory Planning

Taking into account only cards located within a lateral distance of 6 cm (approx. half the size of a card field) around the current hand trajectory and 30 cm in the longitudinal hand direction, the 36 possible cards could be reduced to an average of 2.9 cards (SD=1.1). Approximately 470 ms (SD=220 ms) before SCT, the target card laid within the trajectories' tolerance field.

4.2.4 Fingertip Proximity

Figure 6 shows three relevant hand features during a card turn (top) and the cuboid hit rates on target card's cuboids for each feature (bottom). The hit rates for the thumb tip are overall the lowest, indicating that the thumb was less often located over a field during card turn than the other two features. The hit rates of the index fingertip and the center point are very similar up to a height of 6 cm and then increase slightly more for the index fingertip.

The tracking rate, more precisely the number of available data points at a recording speed of 50 fps, reached a mean value of 29.4% (SD=26.6%) and a maximum of 60%. While running on-device video recordings, the recording rate is automatically reduced from 60 fps to 30 fps. Despite fluctuations in the tracking rate, the cuboid hit rates for the index fingertip and the center point were high during a card turn. Outliers occurred when the tracking rate was very low, and thus, registered hits on other cuboids had a more significant effect on the hit rate.

4.3 Intermediate Discussion

Gaze behavior on cards seemed to be random up to the last second before the card turn. In 73.3% of cases, the gaze arrived on target

card approximately 350 ms before card turn. The lower number of fixations on the target card during SCT than FCT is most likely related to the two-player setup. Participants who see a card whose counterpart they know during their opponent's move seemed to keep the position of that card in mind during their move. After revealing the expected matching card during their first card turn, they choose the second card without looking at it.

The peak velocity fits on average very well as a trigger condition for gaze prediction and errors due to the variance of peak velocity and gaze on target should be greatly reduced by only allowing targets on the hand trajectory. While the start of a hand movement can be well detected by a velocity threshold, the peak velocity can only be evaluated retrospectively. A possible alternative solution would be first to detect the start of a hand movement and then check for a negative acceleration of the hand.

The measurement of hit interaction of the index fingertip in the respective cuboids provides an excellent signal to detect the first card turn but is strongly affected by the hand tracking rate. For the best performance of our support system, it is advisable to test the system without first-person video recording and, thus, make full use of the device's capabilities to track hands with 60 fps. While we aimed for a high degree of task immersion during the behavioral analysis in the first study, we changed the setup to a single-player memory game to assess the support system's performance within the second user study.

5 Implementation: Closed-Loop User Support

Based on the results of the first study, we implemented our processing and analysis pipeline of gaze prediction, hand trigger, and hand trajectory on HoloLens 2 to display visual alerts to the user in real-time. In this section, we explain the functionality of the implemented closed-loop support system. As we aim to provide alerts for the second card turn based on selecting the first card, we first detect the first card turn by monitoring the fingertip position when near a respective card. Figure 7 shows the pseudo code of the closed-loop user support. We initialized the algorithm's thresholds based on the findings in our first study and refined them during a pilot study with three participants.

While the next card is set to the first card, all registered cuboid hits of the index fingertip are continuously written into a list of window size 20. We found that a cuboid height of 5.5 cm (Fig. 6) works well to detect card turns while avoiding false detections due to the hand moving across the field. Once the window size is reached, the tracking rate and cuboid hit rate are calculated. If at least 30% of data points are available and at least 60% of these data points register a hit on the same cuboid, the first card is selected. As a result, the respective field is outlined with green dashed lines (Fig. 8 (a)) and the next card is set to the second card.

Once the velocity of four consecutive frames is greater than 0.25 m/s, we detect the start of a new hand reach to a target. This Boolean allows us to filter out the majority of card turns and random hand movements (cf. Fig. 5). As missing data points can affect system performance, we interpolate single missing data points with an immediate value. Once the hand movement has started, the current gaze target is compared with the card located close to the current hand trajectory. Only cards within a maximum distance of 6 cm in the transverse direction and 30 cm in the longitudinal direction are considered. A color-coded visual alert is displayed above the examined card position when a match occurs between the gaze target and hand trajectory targets. If the predicted target matches the correct second card stored in the ground truth game layout, a green bounding box outlining the field is displayed (Fig. 8 (b)). A yellow alert is displayed in the event of a predicted incorrect target adjacent to the correct card. If neither the predicted target nor any adjacent fields are the correct card, a red warning sign is displayed (Fig. 8 (c-d)). At the beginning of our tests, we used a second Boolean

ALGORITHM 1: Closed-Loop User Support

```
Input: window size = 20 // equals 0.4s
Input: step counter = 0
Input: hand velocity, hand trajectory
Input: gaze target
// CF: number of consecutive frames //

while recording do
    if next card is first card then:
        increase step counter by 1;
        write latest cube touch event or default value into list;
        if step counter equals window size then
            calculate tracking rate of full window;
            calculate cuboid touch rate of full window;
            if tracking rate > 30% and at least 60% of touch
            measurements are within the same cube then
                first card has been selected;
                display confirmation;
                next card is second card;
            end
            step counter is set back to 0; clear list;
        end
    end
    if next card is second card then:
        velocity, trajectory ← get current hand data (4CF)
        gaze target ← get current gaze data (6CF)
        if hand velocity (4CF) > 0.25 then
            hand movement has started is true;
        end
        if hand movement has started and gaze target (6CF) is on
        hand trajectory (1CF) then
            if gaze target is not same as first card then
                second card has been selected;
                display warning / confirmation at target location;
            else
                reset target prediction;
            end
        end
    end
    if touch on reset cuboid is registered then
        reset move; // next expected card will be the first card
    end
end
```

Figure 7: Pseudo code for the implemented closed-loop user support.

condition after the detected start of a hand movement set true by three consecutive frames with negative acceleration (represents a feature slightly behind 'B', cf. Fig. 5). This implementation, however, proved to be generally too slow to issue visual alerts in-time and was dropped. The single velocity threshold used in our final implementation represents a feature between 'A' and 'B' (cf. Fig. 5). Further, we initialized the threshold for a card to be considered a gaze target with 4 consecutive frames on the same card. Increasing this value to 6 frames significantly reduced false positives during slower gaze transitions to the target.

The system was designed in a way that it would only provide one visual alert for each move. Issuing multiple warnings for wrong second card choices was expected to only result in a trial-and-error strategy instead of participants actually thinking about the card choice. Hence, after a visual alert was displayed, the system switched to standby until the cards were turned back and the next move started. To ensure that false detections were not propagated into future moves, participants had to reset the system once after each move. This was done by briefly moving their right hand over the single field on the right side of the grid (cf. Fig. 2), which was covered by an invisible cuboid (cf. Fig. 3). A touch with the cuboid resulted in the cuboid lighting up to confirm the reset.

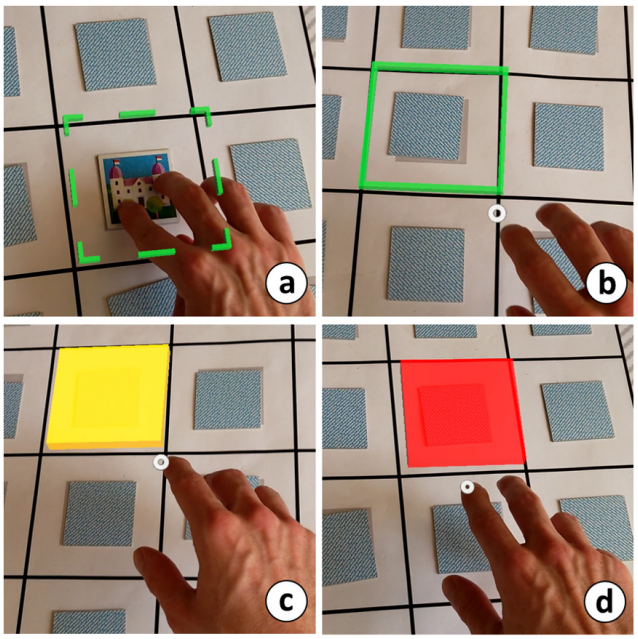

Figure 8: Confirmation of detected first card turn (a) and visual alerts for the second card prediction in case of correct target selection (b), a wrong target that is located directly next to the correct card (c), and a wrong target that is not located next to the correct card (d).

6 Study 2: Validating Closed-Loop User Support

In our second study, 12 new participants were recruited to play a single-player memory game while our app now provided closed-loop user support (cf. Section 5, Algorithm 1). As observed in the first study, the use of first-person video recordings greatly reduces sensor performance. In particular, hand tracking is reduced from a possible 60 fps to approximately 30 fps. To test the support system at its best performance, we recorded participants' actions with an external camera while participants commented on their observations using the think aloud method.

6.1 Participants

We recruited 12 new participants from our institution (9 male, 3 female, mean ages = 27.3 years, SD = 2.9 years) with normal or corrected to normal vision. No participants were excluded.

6.2 Task

The goal of the game was to find all pairs of cards with as few card moves as possible during a single-player game. Participants were asked to select a different second card if a yellow or red alert was displayed in-time at the location of their initial card choice. Before each new move, participants once moved their right hand over the square to the right of the grid to reset the closed-loop support system.

6.3 Procedure

Participants were introduced to how the system worked and learned about the four visual aids (cf. Fig. 8) without addressing the underlying behavioral patterns. Participants then performed the app calibration and were able to test the system on three card pairs before starting the experiment. Participants were asked to think aloud and share their observations during the experiment. In the case of leaving out information, the experimenter asked questions. After the experiments, an interview was conducted.

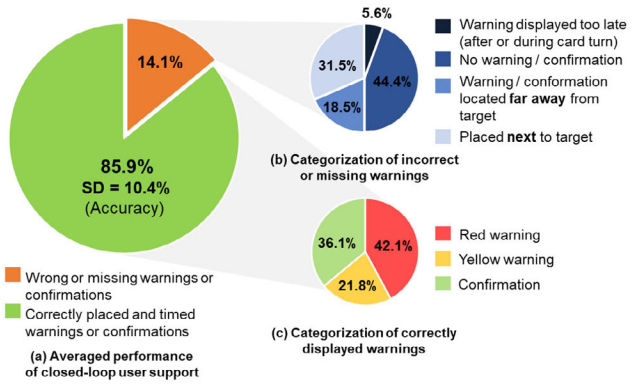

Figure 9: Warning system performance averaged over all participants (a) with categorization of incorrectly displayed warnings (b) and correctly displayed warnings (c).

6.4 Data Analysis

We analyzed the support system's performance and statements from interviews. In a first step, we examined the third-person video recordings and classified all warnings depending on time and place of occurrence as described by the participants during think aloud. A visual alert was considered timely when the participant recognized it before the card turn, resulting in an observable change in the target card after yellow and red warnings. The place of occurrence was categorized as either 'far away from target', if at least one field was separating the predicted and the actual target, as 'next to target' or as 'correct target'. We calculate the system accuracy by dividing the number of correctly timed and placed visual alerts by all second card turns. We did not quantify how often warnings subsequently led to a correct card choice, as this metric is highly affected by chance. Finally, during the interviews, we asked participants how they perceived the system's functionality, how they experienced the visual alerts, and whether they felt patronized, monitored or annoyed by the system at any point during the game.

7 Results

In total 384 card pairs were played by the participants. Only allowing one visual alert every move (turn of two cards), 330 (85.9%) hand actions in total correctly triggered a visual alert in-time, while 54 hand actions resulted in wrong, late or missing visual alerts. Figure 9 shows the mean performance across all participants and the breakdown of correct and incorrect warnings into subcategories.

Of the 54 card turns not resulting in a correct visual alert, only 5.6% were issued too late. A total of 31.5% of visual alerts were placed just next to the target, which, in the case of red and yellow alerts, still provided information about the actual target. A total of 18.5% of warnings placed far from the target usually occured when the participants moved their hand unconsciously over the field or when the hand movement and gaze overlapped while moving to a target. This was often an issue when moving the hand from the top left corner to the bottom right corner. In these cases, the gaze movement was slower, and the hand blocked sight on the cards while moving backward. Finally, 44.4% of second card turns were stated by the participants to have not issued any visual alert. Analysis of the event logs in the output file showed that for most of these cases, a visual alert was issued but was either not recognized by the participants or was placed outside of the field of view in AR. In approximately 7% of all first card turns recognition did not work properly. Either the green dashed lines appeared on the neighboring field or during the second card turn. In these cases, we recommended that participants simply reset and repeat the move.

We observed two fundamental strategies in dealing with the support system. Two-thirds of the participants found a natural pace from the beginning, where detection of the first card turn and prediction of the second card worked very well, reaching accuracies of the target prediction up to 97.0%. The other third of the participants initially performed random hand movements to test the system. After provoking false alerts, they quickly learned how the system worked. This group of participants then actively used hand-eye coordination to control the warning system, which became noticeable by the fixation on the target card and a short yet fast pointing gesture towards the target. Participants found it particularly helpful that visual feedback was shown for all card actions, including the first card, which allowed them to understand how the system worked and to collaborate better.

During the interview, all participants stated that the system worked very well and that it was helpful and supportive and stimulating to use. None of the participants felt patronized or monitored by the system. Two participants stated that the interpretation of visual aids and the effort for memorizing card pairs required an increased level of concentration. In contrast, two other participants stated that they had to think less during the task, using the support system as a tool, which they appreciated. While all visual warnings were perceived as helpful, preferences varied between participants. Perceptions of green warnings varied from participants experiencing them as positive and motivating feedback to participants having a rather neutral perspective. Yellow alerts were perceived as most useful, as they prevented incorrect hand action and gave hints about the correct target. This effect increased especially towards the end of the game when there were only a few cards left. Finally, red alerts were not perceived as negative by the participants. However, participants criticized that red alerts only pointed out a mistake without providing the user with additional task-relevant information. Two participants suggested displaying an arrow above the red warning that points in the approximate direction of the correct card to provide better support. Participants further stated that the reset cube was fast and easy to use but that they sometimes forgot about the reset, especially during their first moves, and thus needed to be reminded by the experimenter.

8 Discussion

Our goal was to investigate whether real-time analysis of hand-eye coordination is suitable for predicting hand actions during target selection.

Our investigations showed that the support of our implemented method was effective with a mean accuracy of 85.9%. While target prediction was lower for SCT than for FCT in the first user study, these differences were not present in the second user study. This could be a consequence of the change from a two-player to a single-player setup. Statements from the interviews suggest that the very robust predictions are also related, in part, to the fact that participants sometimes adjusted their behavior to interact with the support system in an optimal way. Despite the measured average prediction times of only 350 ms, most visual alerts were issued in time. This seems plausible, considering that simple reaction times range from 180 ms to 220 ms [23]. During hand-eye coordination, the eye continuously supplies information to control hand movement. If a warning sign obscures the target, the eyes cannot further guide the hand movement. In contrast, displaying green outlines did not interrupt the hand action in most of the cases.

Based on our results and previous research on hand-eye coordination in target selection, there is strong evidence that our method is transferable to other cases. Several studies have shown the gaze preceding the hand during target selection [14, 19, 25], also referred to as a 'directing' pattern [26]. Our studies support these findings while demonstrating how hand and gaze features can be combined for target prediction. According to Crawford et al. [7], the object to be manipulated directly affects the time the gaze must arrive on the target. We therefore expect that some refinements of the thresholds used in our method will be necessary for optimal performance in other scenarios with other objects. We suggest that researchers record hand and gaze data for their specific scenario, following our implementation, and then fine-tune the parameters on their data to find a good compromise of prediction time and accuracy.

While the playing field used for our studies is two-dimensional, the invisible virtual objects for measuring user behavior, i.e., a thin layer for gaze interaction and a thicker layer for finger proximity (cf. Fig. 3), could be placed over any non-planar surface in 3D space. Both the velocity threshold and gaze target of our proposed method should be transferable to 3D settings. Only the trajectory is currently computed in the 2D plane and would have to also exclude targets along the vertical axis. Contrary to our setup, which was well structured and observable from different angles, more complex 3D setups might be characterized by occlusions and greater variations in target objects' distances and sizes, which might require case-specific extensions to our method.

We see our method in various procedural tasks where an operator follows a predefined sequence of actions such as, for example, during interaction with medical devices or machine interfaces, or while reaching for assembly parts. To integrate predictive support into more complex real-world tasks, however, the system needs a profound understanding of what the user is currently doing and how this is in alignment with a reference workflow. Such process monitoring has been studied in previous work [15, 31] and could be used as a basis for our system in the future. Hand tracking capabilities now also allow for direct monitoring of hand actions. In this work, we only monitored one hand joint, i.e., the index fingertip, in proximity to the cards to detect the first card turn, which was simple but very effective. Recent work has utilized all hand joints of a hand pose for training time series models (e.g., LSTM) on activity recognition of manual tasks, resulting in high accuracies [12, 38]. Training algorithms to recognize hand actions would allow future work to label them during data postprocessing automatically. Using the detected hand actions as output and the preceding hand and gaze behavior as input, supervised training pipelines can be implemented to learn more complex relations involving hand-eye coordination.

9 Limitations

The results are based on experiments with only 22 participants from a rather homogeneous sample population. Despite the small number of participants, the data set included 525 manually labelled first and second card selections (summed up over both studies), which we believe to be a solid basis to assess the performance of our method. Further studies would strengthen the validity of our findings and would be particularly interesting when conducted in other real-world settings.

While the heuristics derived in this paper work well on average, there is a distribution of temporal coupling between gaze and hand feature occurrence (cf. Fig.5), which can result in warnings sometimes being triggered at the wrong time, and thus at the wrong place. Such differences in temporal coupling cannot be fully accounted for by a system based on thresholds, but rather by jointly learning hand and gaze features from data. Combining the gaze prediction with a hand trajectory proved to be key to handling the variety in participants hand movements. During our initial investigations on target prediction, we found that simply using a velocity threshold and the gaze target (i.e., as proposed by Cheng et al. [6] for predefined targets) was not sufficiently robust when participants could make their own card choice on-the-fly. We suggest future work to also consider optimizing thresholds for hand movement direction, as hand movements from top left to bottom right corner were associated with a higher percentage of misplaced warnings. Finally, the thresholds were only optimized for the average target population. Customizing

thresholds to individual participants is expected to bring participants performances closer to those of participants who collaborated with the system and achieved accuracies of up to 97%.

The manual reset of the support system after each move might have had an effect on participants natural behavior. Playing the memory game without a reset cube would improve authenticity and could be achieved by integrating more pronounced process monitoring into the support system. There might have also been an effect of differences in participants spatial abilities. These differences, however, are expected to be rather small for the homogeneous group of young and healthy participants (mean age = 28 years) recruited for our studies.

In addition, as with any sensor, hand and gaze measurements are subject to certain measurement errors. The playing field dimensions were chosen to minimize error, particularly in measuring gaze behavior on cards. With state-of-the-art eye tracking glasses measuring gaze with 100 fps and angular accuracies between 0.5–1° (e.g., Tobii Pro Glasses 2), compared to HoloLens 2 with 30 fps and an accuracy of 1.5°, it is possible to analyze gaze behavior on more compact stimuli in the future, such as machine interfaces or surgical scenes, and with fine-grained analysis of eye movements. For hand tracking, data points were occasionally missing due to low tracking quality, which we also believe gradually improves with technological advancements. There may also be some errors due to the manual processing of the ground truth.

10 Conclusion

With the high cost that human error in industrial and clinical applications is associated with, error prevention is an important topic. In this paper, we presented a method that utilizes hand-eye coordination to predict hand actions during target selection. End-to-end testing of our method showed it to be highly effective in placing visual alerts over target locations and stop hand actions in a timely manner. Moreover, it showed that hand-eye coordination can be used as an intuitive way of interacting with a technical system and that transparent communication from the system to the user is key for effective collaboration.

To date, the field of context-aware augmented reality in manual tasks has primarily focused on providing feedback on current user behavior. With our work, we contribute a method that allows AR headsets to provide feedback at an earlier stage of a task. While the memory game proved to be an expedient case for this first investigation, future studies should investigate hand-eye coordination in industrial and clinical setups. It will be interesting to explore in the future what patterns exist during other real-world tasks, how they change in the course of a procedure and how they can be used for intelligent wearable support systems.

Acknowledgments

This work is part of the SURGENT project and was funded by University Medicine Zurich/ Hochschulmedizin Zürich.

References

[1] Eye tracking on hololens 2. `https://docs.microsoft.com`. Accessed: 2020-08-15.

[2] P. Baudisch, E. Cutrell, D. Robbins, M. Czerwinski, P. Tandler, B. Bederson, A. Zierlinger, et al. Drag-and-pop and drag-and-pick: Techniques for accessing remote screen content on touch-and pen-operated systems. In *Proceedings of INTERACT*, vol. 3, pp. 57–64, 2003.

[3] J. Blattgerste, B. Strenge, P. Renner, T. Pfeiffer, and K. Essig. Comparing conventional and augmented reality instructions for manual assembly tasks. In *Proceedings of the 10th international conference on pervasive technologies related to assistive environments*, pp. 75–82, 2017.

[4] A. Bulling, C. Weichel, and H. Gellersen. Eyecontext: recognition of high-level contextual cues from human visual behaviour. In *Proceedings of the sigchi conference on human factors in computing systems*, pp. 305–308, 2013.

[5] M. D. Byrne, J. R. Anderson, S. Douglass, and M. Matessa. Eye tracking the visual search of click-down menus. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, p. 402–409. Association for Computing Machinery, New York, NY, USA, 1999. doi: 10.1145/302979.303118

[6] L.-P. Cheng, E. Ofek, C. Holz, H. Benko, and A. D. Wilson. Sparse haptic proxy: Touch feedback in virtual environments using a general passive prop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, p. 3718–3728. Association for Computing Machinery, New York, NY, USA, 2017. doi: 10.1145/3025453.3025753

[7] J. D. Crawford, W. P. Medendorp, and J. J. Marotta. Spatial transformations for eye–hand coordination. *Journal of Neurophysiology*, 92(1):10–19, 2004. PMID: 15212434. doi: 10.1152/jn.00117.2004

[8] A. Deshpande and I. Kim. The effects of augmented reality on improving spatial problem solving for object assembly. *Advanced Engineering Informatics*, 38:760–775, 2018.

[9] M. Eckert, J. S. Volmerg, and C. M. Friedrich. Augmented reality in medicine: systematic and bibliographic review. *JMIR mHealth and uHealth*, 7(4):e10967, 2019.

[10] M. K. Eckstein, B. Guerra-Carrillo, A. T. M. Singley, and S. A. Bunge. Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Developmental cognitive neuroscience*, 25:69–91, 2017.

[11] C. Fermüller, F. Wang, Y. Yang, K. Zampogiannis, Y. Zhang, F. Barranco, and M. Pfeiffer. Prediction of manipulation actions. *International Journal of Computer Vision*, 126(2-4):358–374, Apr. 2018. doi: 10.1007/s11263-017-0992-z

[12] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 409–419, 2018.

[13] G. Gras and G. Yang. Context-aware modeling for augmented reality display behaviour. *IEEE Robotics and Automation Letters*, 4(2):562–569, 2019.

[14] W. F. Helsen, D. Elliott, J. L. Starkes, and K. L. Ricker. Temporal and spatial coupling of point of gaze and hand movements in aiming. *Journal of Motor Behavior*, 30(3):249–259, 1998. doi: 10.1080/00222899809601340

[15] S. J. Henderson and S. K. Feiner. Augmented reality in the psychomotor phase of a procedural task. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pp. 191–200, 2011.

[16] M. Hoover, J. Miller, S. Gilbert, and E. Winer. Measuring the performance impact of using the microsoft hololens 1 to provide guided assembly work instructions. *Journal of Computing and Information Science in Engineering*, 20(6), 2020.

[17] C.-M. Huang, S. Andrist, A. Sauppé, and B. Mutlu. Using gaze patterns to predict task intent in collaboration. *Frontiers in psychology*, 6:1049, 2015.

[18] J. Huang, R. White, and G. Buscher. User see, user point: Gaze and cursor alignment in web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, p. 1341–1350. Association for Computing Machinery, New York, NY, USA, 2012. doi: 10.1145/2207676.2208591

[19] R. Johansson, G. Westling, A. Bäckström, and J. R. Flanagan. Eye–hand coordination in object manipulation. *The Journal of Neuroscience*, 21:6917 – 6932, 2001.

[20] D. Katić, A.-L. Wekerle, J. Görtler, P. Spengler, S. Bodenstedt, S. Röhl, S. Suwelack, H. G. Kenngott, M. Wagner, B. P. Müller-Stich, R. Dillmann, and S. Speidel. Context-aware augmented reality in laparoscopic surgery. *Computerized Medical Imaging and Graphics*, 37(2):174 – 182, 2013. Special Issue on Mixed Reality Guidance of Therapy - Towards Clinical Implementation. doi: 10.1016/j.compmedimag.2013.03.003

[21] R. C. King, L. Atallah, B. P. Lo, and G.-Z. Yang. Development of a wireless sensor glove for surgical skills assessment. *IEEE Transactions on Information Technology in Biomedicine*, 13(5):673–679, 2009.

[22] F. Koochaki and L. Najafizadeh. Eye gaze-based early intent prediction

utilizing cnn-lstm. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1310–1313, 2019.

[23] R. J. Kosinski. A literature review on reaction time. *Clemson University*, 10(1), 2008.

[24] G. A. Koulieris, K. Akşit, M. Stengel, R. K. Mantiuk, K. Mania, and C. Richardt. Near-eye display and tracking technologies for virtual and augmented reality. In *Computer Graphics Forum*, vol. 38, pp. 493–519. Wiley Online Library, 2019.

[25] M. Land, N. Mennie, and J. Rusted. The roles of vision and eye movements in the control of activities of daily living. *Perception*, 28(11):1311–1328, 1999. doi: 10.1068/p2935

[26] M. F. Land and M. Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision Research*, 41(25):3559 – 3565, 2001. doi: 10.1016/S0042-6989(01)00102-X

[27] F. Liebmann, S. Roner, M. von Atzigen, D. Scaramuzza, R. Sutter, J. Snedeker, M. Farshad, and P. Fürnstahl. Pedicle screw navigation using surface digitization on the microsoft hololens. *International journal of computer assisted radiology and surgery*, 14(7):1157–1165, 2019.

[28] S. Marwecki, A. D. Wilson, E. Ofek, M. Gonzalez Franco, and C. Holz. Mise-unseen: Using eye tracking to hide virtual reality scene changes in plain sight. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, UIST '19, p. 777–789. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3332165.3347919

[29] M. Mussgnug, D. Singer, Q. Lohmeyer, and M. Meboldt. Automated interpretation of eye–hand coordination in mobile eye tracking recordings. *KI-Künstliche Intelligenz*, 31(4):331–337, August 2017.

[30] A. K. Mutasim, W. Stuerzlinger, and A. U. Batmaz. Gaze tracking for eye-hand coordination training systems in virtual reality. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, p. 1–9. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3334480.3382924

[31] L. X. Ng, J. Ng, K. T. Tang, L. Li, M. Rice, and M. Wan. Using visual intelligence to automate maintenance task guidance and monitoring on a head-mounted display. In *Proceedings of the 2019 5th International Conference on Robotics and Artificial Intelligence*, pp. 70–75, 2019.

[32] R. Palmarini, J. A. Erkoyuncu, R. Roy, and H. Torabmostaedi. A systematic review of augmented reality applications in maintenance. *Robotics and Computer-Integrated Manufacturing*, 49:215 – 228, 2018. doi: 10.1016/j.rcim.2017.06.002

[33] E. Pelanis, R. P. Kumar, D. L. Aghayan, R. Palomar, Å. A. Fretland, H. Brun, O. J. Elle, and B. Edwin. Use of mixed reality for improved spatial understanding of liver anatomy. *Minimally Invasive Therapy & Allied Technologies*, 29(3):154–160, 2020.

[34] J. Pelz, M. Hayhoe, and R. Loeber. The coordination of eye, head, and hand movements in a natural task. *Experimental brain research*, 139(3):266—277, August 2001. doi: 10.1007/s002210100745

[35] N. Petersen and D. Stricker. Cognitive augmented reality. *Computers & Graphics*, 53:82–91, 2015.

[36] B. A. Smith, J. Ho, W. Ark, and S. Zhai. Hand eye coordination patterns in target selection. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, pp. 117–122, 2000.

[37] B. W. Tatler and M. F. Land. Everyday visual attention. *The handbook of attention*, pp. 391–421, 2015.

[38] B. Tekin, F. Bogo, and M. Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4511–4520, 2019.

[39] J. Triesch, D. H. Ballard, M. M. Hayhoe, and B. T. Sullivan. What you see is what you need. *Journal of vision*, 3(1):9–9, 2003.

[40] X. Wang, S. K. Ong, and A. Y. Nee. A comprehensive survey of augmented reality assembly research. *Advances in Manufacturing*, 4(1):1–22, 2016.

[41] P. Weill-Tessier and H. Gellersen. Correlation between gaze and hovers during decision-making interaction. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, ETRA '18. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3204493.3204567

[42] J. Zhu, S.-K. Ong, and A. Y. Nee. A context-aware augmented reality assisted maintenance system. *International Journal of Computer Integrated Manufacturing*, 28(2):213–225, 2015.