

Diss. ETH No. 28704

Towards Advanced User Guidance and Context Awareness in Augmented Reality-guided Procedures

A thesis submitted to attain the degree of

Doctor of Sciences of ETH Zurich
(Dr. sc. ETH Zurich)

presented by

Julian Christian Wolf

M.Sc. Karlsruhe Institute of Technology (KIT)
born on 10.05.1990
citizen of Germany

Prof. Dr.-Ing. Mirko Meboldt, examiner
Prof. Dr. Philipp Fürnstahl, co-examiner

2022

Julian Wolf
julian.wolf@mavt.ethz.ch

©2022

ETH Zurich
Product Development Group Zurich
Leonhardstrasse 21
8092 Zurich
Switzerland

pd | **Z** Product Development Group Zurich
Produktentwicklungsgruppe Zürich

Acknowledgments

Time is the most precious commodity one can give. I am very grateful for all the people who have supported me during this PhD with their time, knowledge, and friendship. Many people have played an important role on my way and whom I would like to thank.

First, I would like to express my sincere gratitude to Prof. Mirko Meboldt for allowing me to pursue this PhD and for his continuous support and guidance. The opportunity to work on so many innovative industrial and clinical applications and to expand my skills in my field of interest is something for which I am very grateful.

I would also like to thank the group leader of the Human Behavior Group at pd|z, Dr. Quentin Lohmeyer, for the many long and exciting conversations about future research directions, for always advocating for me, and for giving me direct, honest, and actionable feedback. At the same time, I would like to thank all my colleagues at the pd|z for the discussions, team events, paper beers, and the unique spirit that makes the pd|z a special place. In particular, I would like to thank Stephan Hess, Felix Wang, Sophokles Ktistakis, and Tobias Stauffer, with whom I worked directly.

For my work on predictive AR support, I collaborated with Prof. Christian Holz, whom I would like to thank for his continuous support and guidance, especially during an extended re-submission period. I was fascinated by your positive, fast, and practical attitude and learned much about the computer science community and writing papers.

My sincere thanks also go to Prof. Philipp Frnstahl and Prof. Mazda Farshad for initiating the SURGENT project, whose vision has greatly inspired this work, and for the opportunity to conduct my research in a highly innovative and inspiring environment.

Since 2018, I have also been working with the competence center for extended reality at SBB (Swiss Federal Railways), where I would like to thank Gnhan Akarcay for all the inspiring meetings and co-creating a vision for expert operator guidance. At the same time, and more actively

since 2020, I have been working with Accenture Labs in Sophia-Antipolis to explore how augmented reality and collaborative robots can bring added value to industrial applications. Here I would like to thank Clément Rinaudo and Pierre Duffaut for the fruitful collaboration.

I would also like to thank all the students with whom I collaborated and who made valuable contributions to this work. Working with young, motivated minds eager to learn has been much fun and has given me invaluable experience as a supervisor.

To my family and friends, and my wonderful girlfriend Marianne, thank you for all your support.

Contents

Abstract.....	v
Zusammenfassung	viii
Nomenclature	xii
1 Introduction.....	1
2 Background.....	13
1.1 Augmented Reality Head-mounted Displays.....	13
1.2 Preliminary Work on Automating Semantic Eye Tracking Analysis	15
3 Goals and Contributions.....	18
3.1 Comparing AR against Conventional Instructions.....	21
3.2 Effective Visualization Strategies	22
3.3 Predicting Future Hand Actions	24
4 Study I: Comparing AR against Conventional Instructions	27
4.1 Introduction	28
4.2 Related Works	30
4.3 Methods.....	31
4.4 Results	38
4.5 Discussion.....	42
4.6 Conclusion.....	44
5 Study II: Effective Visualizations Strategies.....	47
5.1 Introduction	48

5.2	Related Work	50
5.3	Materials and Methods	52
5.4	Results	61
5.5	Discussion.....	63
5.6	Conclusion.....	66
6	Study III: Predicting Future Hand Actions.....	68
6.1	Introduction	68
6.2	Related Work	71
6.3	Study Part 1: Patterns in Hand-Eye Coordination	75
6.4	Results	80
6.5	Intermediate Discussion	84
6.6	Implementation	85
	Study Part 2: Validating Closed-loop User Support	89
6.7	Results	90
6.8	Discussion.....	93
6.9	Limitations.....	95
6.10	Conclusion	96
7	Conclusion and Outlook.....	99
7.1	Conclusion.....	99
7.2	Outlook	103
	References.....	108
	Curriculum vitae	120

Abstract

Procedural tasks are common in many professions, such as maintenance, assembly, or surgery, and are characterized by an operator performing a predefined sequence of steps to achieve a specific goal. Because these tasks often involve elaborated machines, devices, or even patients, they place the highest requirements on correct task execution.

Augmented reality (AR) head-mounted displays (HMDs) have been shown to provide effective support during procedural tasks. Compared to conventional information mediums, where information is often spread among multiple documents (e.g., maintenance) or external screens (e.g., surgery), AR HMDs display contextual information directly into the field of view of the operator without occupying the operators' hands. While with AR, displayed information is only changed based on manual user input, context-aware AR promises to further improve the support provided by automatically adapting displayed information to best address the operator's current needs and by providing feedback. Understanding the strengths and weaknesses of these two technologies is key to developing support systems that can improve the quality of task execution, making procedural tasks safer and improving outcomes. Previous studies on context-aware systems have focused primarily on manual execution without consideration of an important part of human interaction, the perception. Eye tracking allows to measure perception and provides deep insights into cognitive processes, and might therefore bring benefits to context-aware systems that are important to be investigated.

This work investigates different concepts of how AR and context-aware AR support systems can be designed, how they work, and how they affect operators' task performance. It further aims to advance context-aware AR support by integrating eye tracking and by deriving a suitable system model to describe the relationships between human behavior, AR, and context-aware AR. Three studies are presented in this work.

Study I investigates the benefits of contextual information in AR over traditional information mediums to provide training instructions. A study

Abstract

was conducted with 21 medical students performing an extracorporeal membrane oxygenation (ECMO) cannulation on a physical simulator setup. The evaluation comprised of a detailed error protocol with both a categorization into knowledge- and handling-related errors and an error severity ranking. The results showed clear benefits of AR over conventional instructions while pointing out certain limitations that might be improved by context-aware AR.

Study II investigates effective visualization strategies when real-time feedback is provided continuously. A study was conducted with 4 expert surgeons and 10 surgical residents performing surgical drilling on a physical simulator setup. The results show that continuous performance feedback generally levels task performance between novice and expert operators, reveal clear advantages and preferences of certain AR visualizations, and give insights into how AR visualizations guide visual attention. In particular, the peripheral field around the area of execution proved to be promising for displaying information as the operator can simultaneously perceive feedback and coordinate hand movement.

Study III investigates the suitability of eye and hand tracking features for predicting and preventing an operator's erroneous actions. A study was conducted on a memory card game to explore the potential and limitations of this approach. The first experiment, which involved 10 participants, recorded participants' eye and hand movement to derive a method for target prediction. The second experiment with 12 participants examined the timeliness and accuracy of the implemented method end-to-end and showed the method to be highly effective in preventing a user's erroneous hand actions.

One of the key conclusions of this work is that context-aware AR support can significantly improve procedural outcomes and even raise the task performance of less experienced operators to the level of experts. In addition, analyzing hand-eye coordination patterns in real-time allows for predictive AR support and error prevention, which might eventually provide a safety net for operators performing their first independent task executions. For future work, important research directions include integrating and advancing predictive AR support for more complex

procedures, investigating effective visualization strategies in environments with multiple dynamic visual stimuli, as well as effective feedback and support strategies while operators transition from their first training to independent execution and eventually become experts.

Zusammenfassung

Prozedurale Aufgaben sind in vielen Berufen, wie z. B. in der Wartung, Montage oder Chirurgie, üblich und zeichnen sich dadurch aus, dass ein Bediener eine vordefinierte Abfolge von Schritten ausführt, um ein bestimmtes Ziel zu erreichen. Da diese Aufgaben oft in Wechselwirkung mit komplexen Maschinen, Geräten oder sogar am Patienten durchgeführt werden, werden allerhöchste Anforderungen an die korrekte Ausführung gestellt.

Augmented Reality (AR) Head-Mounted Displays (HMDs) können eine wirksame Unterstützung bei prozeduralen Aufgaben bieten. Im Vergleich zu herkömmlichen Informationsmedien, bei denen die Informationen oft auf mehrere Dokumente (z.B. bei der Wartung) oder externe Bildschirme (z.B. in der Chirurgie) verteilt sind, zeigen AR HMDs kontextbezogene Informationen direkt im Sichtfeld des Bedieners an, ohne dabei die Hände zu blockieren. Während bei AR die angezeigten Informationen nur auf Grundlage manueller Benutzereingaben geändert werden, versprechen kontextbewusste AR-Systeme eine weitere Verbesserung der Unterstützung, indem sie die angezeigten Informationen automatisch an die aktuellen Bedürfnisse des Bedieners anpassen und den Bediener durch Feedback unterstützen. Ein Verständnis der Stärken und Schwächen dieser beiden Technologien ist der Schlüssel zur Entwicklung von Unterstützungssystemen, die die Qualität der Aufgabenausführung verbessern können, um prozedurale Aufgaben sicherer zu machen und die Endresultate zu verbessern. Frühere Studien über kontextbewusste AR-Systeme haben sich in erster Linie auf die manuelle Ausführung konzentriert und dabei einen wichtigen Teil der menschlichen Interaktion, nämlich die Wahrnehmung, außer Acht gelassen. Eye Tracking ermöglicht die Messung der Wahrnehmung und bietet tiefe Einblicke in kognitive Prozesse und könnte daher Vorteile für kontextbewusste Systeme bringen, die es zu untersuchen gilt.

Das Ziel dieser Dissertation ist es, zu untersuchen, wie sich AR und kontextbewusste AR Systeme gestalten lassen, wie sie funktionieren, und

wie sie die Ausführungsqualität des Bedieners beeinflussen. Dabei soll die kontextbewusste AR-Unterstützung insbesondere auch durch die Integration von Eye-Tracking und die Herleitung eines geeigneten Systemmodells vorangetrieben werden. In dieser Arbeit werden drei Studien vorgestellt.

Studie I untersucht die Vorteile kontextbezogener Informationen in AR zum Vermitteln von Ausführungsinstruktionen gegenüber herkömmlichen Informationsmedien. Eine Studie wurde mit 21 Medizinstudenten durchgeführt, die eine extrakorporale Membranoxygenierung (ECMO) an einem physikalischen Simulator durchführten. Die Auswertung umfasste ein detailliertes Fehlerprotokoll mit einer Kategorisierung in wissens- und handhabungsbezogene Fehler sowie eine Einstufung des Fehlerschweregrads. Die Ergebnisse zeigen deutliche Vorteile von AR gegenüber konventionellen Anleitungen und weisen gleichzeitig auf bestimmte Limitationen hin, die durch kontextbewusstes AR verbessert werden könnten.

Studie II untersucht effektive Visualisierungsstrategien für den Fall, dass Echtzeit-Feedback kontinuierlich bereitgestellt wird. Es wurde eine Studie mit vier erfahrenen Chirurgen und zehn Assistenzärzten durchgeführt, wobei Bohrungen an einem physikalischen Wirbelsäulenmodell gesetzt werden mussten. Die Ergebnisse zeigen, dass kontinuierliches Feedback im Allgemeinen die Aufgabenleistung zwischen Anfängern und Experten angleicht, zeigen klare Vorteile und Präferenzen bestimmter AR-Visualisierungen und geben Einblicke in die Art und Weise, wie AR-Visualisierungen die visuelle Aufmerksamkeit lenken. Insbesondere das periphere Feld um den Ausführungsbereich erwies sich als vielversprechend für die Anzeige von Informationen, da der Bediener gleichzeitig Feedback wahrnehmen und die Handbewegung koordinieren kann.

Studie III untersucht die Eignung von Blick- und Handbewegungsmessungen zur Vorhersage und Vermeidung von Fehlhandlungen des Bedieners bei prozeduralen Aufgaben. Um das Potenzial und die Grenzen dieses Ansatzes zu untersuchen, wurde eine Studie mit einem Memory-Kartenspiel durchgeführt. Im ersten

Zusammenfassung

Experiment mit 10 Teilnehmern wurden nur die Augen- und Handbewegungen der Teilnehmer aufgezeichnet, um eine Vorhersagemethode herzuleiten. Das zweite Experiment mit 12 Teilnehmern untersuchte die Geschwindigkeit und Genauigkeit der implementierten Methode mit Nutzern und zeigte, dass die Methode sehr effektiv darin ist fehlerhafte Handaktionen des Benutzers rechtzeitig zu verhindern.

Eine der wichtigsten Schlussfolgerungen dieser Arbeit ist, dass eine kontextbewusste AR-Unterstützung die Ergebnisse erheblich verbessern und sogar die Aufgabenleistung von weniger erfahrenen Bedienern auf das Niveau von Experten anheben kann. Darüber hinaus ermöglicht die Echtzeitanalyse von Hand-Augen-Koordinationsmustern eine vorausschauende AR-Unterstützung und Fehlervermeidung, die zukünftigen Bedienern ein Sicherheitsnetz bieten könnte, wenn diese zum ersten Mal eigenständig eine neue Aufgabe ausführen müssen. Wichtige Forschungsrichtungen für zukünftige Arbeiten sind die Integration und Weiterentwicklung von präventiver AR-Unterstützung für komplexere Prozeduren, die Untersuchung effektiver Visualisierungsstrategien in Umgebungen mit mehreren dynamischen visuellen Stimuli, sowie effektive Feedback- und Unterstützungsstrategien, die den Bediener vom ersten Training bis zur eigenständigen Ausführung bestmöglich unterstützen.

Nomenclature

Acronyms and abbreviations

2D	2-Dimensional
3D	3-Dimensional
AOI	Area of interest
AR	Augmented reality
cGOM	Computational gaze-object mapping
CNN	Convolutional neural network
DoF	Degrees of freedom
ECMO	Extracorporeal membrane oxygenation
fps	Frames per second
HMD	Head-mounted display
IMU	Inertia measurement unit
IR	Infrared
LSTM	Long short-term memory
MRTK	Mixed reality toolkit
OGD	Object-gaze distance
OR	Operation room
OST	Optical see-through
PCA	Perception-cognition-action
Post-op	After operation
RMS	Root mean square
SD	standard deviation
SLAM	Simultaneous location and mapping
SUS	System usability scale
TNR	True negative rate
TPR	True positive rate
UEQ	User experience questionnaire
USZ	University hospital Zurich
VR	Virtual reality

1 Introduction

Augmented reality (AR) is a technology that superimposes a computer-generated image on the real world to create a composite view [1]. Among the various digital devices that can create AR views (e.g., video projectors, tablets, screens), semi-transparent head-mounted displays (HMD) that are worn in front of the user's eyes (similar to glasses) have seen the biggest technological advancements over the last few years, with sophisticated hardware (e.g., Microsoft HoloLens 2) being commercially available, and have democratized the use of AR for a broader community.

Although AR has potential implications for many areas of application where digital information needs to be accessed and interacted with in the real world, it is particularly promising for procedural tasks and processes in professional disciplines such as maintenance and repair [2], assembly [3], or surgery [4]. The fundamental cognitive processes of operators in these domains, the support they need, and the procedural problems they face are quite similar [5]. Procedural tasks and processes can be defined as a series of actions conducted in a predefined order to achieve a goal or desired result [6]. By definition, procedures are highly similar to processes with the main difference that they are the official or accepted way of doing something [6]. Fig. 1 illustrates a process model at the example of drilling a hole.

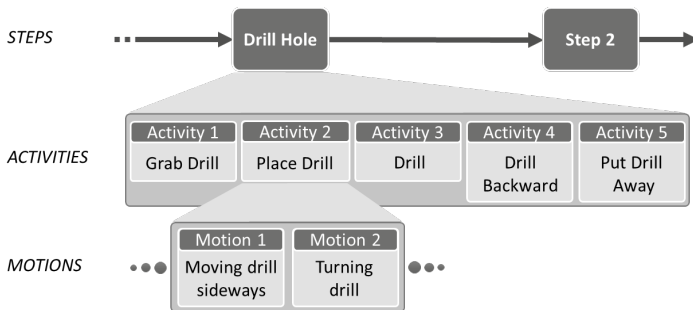


Fig. 1: Process model at the example of “drilling a hole”. Steps can consist of several activities, the smallest semantic entity of a physical task, and motions, which are simple trajectories without semantics (definitions adapted from [7-9]). (Note that the granularity level of these terms can differ between publications and depends on the task at hand.)

Training and execution of procedural tasks often require supplementary information that explain important details of the task, show images or schematic illustrations that support problem solving, or examples of ideal execution. In contrast to conventional information mediums where information is often spread among multiple documents (e.g., maintenance) or external screens (e.g., surgery), AR HMDs allow to display contextual information at the right place at the right time without occupying the operators’ hands. Studies have shown AR to potentially increase spatial understanding of anatomical structures [10] and spatial problem-solving skills [11] while causing fewer procedural errors [12, 13] and lower cognitive load [14].

Despite the potential benefits of AR for user guidance, there are many critical applications that still use conventional information mediums. For example, extracorporeal membrane oxygenation (ECMO) is a life-saving procedure for severe respiratory or cardiac failure, characterized by a high mortality rate of over 60% [15]. Regular training on physical simulators has been shown to reduce mortality rates [16], but training is time-consuming, and document-based training instructions, either printed or accessible via a desktop computer, require the trainee to periodically interrupt simulator training to consult training materials. AR step-by-step instructions could improve the effectiveness and efficiency of training by

making relevant information directly accessible during task execution (cf. Fig. 2), which could facilitate information retrieval and task comprehension, and could help clinicians be better prepared for emergency situations, even after single training runs. To date, few studies have examined how AR affects procedural errors, and these have mostly focused on assembly and maintenance procedures. Their results suggest that AR provides clearer task instructions that lead to fewer procedural errors. However, further work is needed to examine how these instructions affect errors in detail, such as the severity of errors or causes. A detailed error analysis would provide a better understanding of how and why errors occur and where AR can improve procedural outcomes.

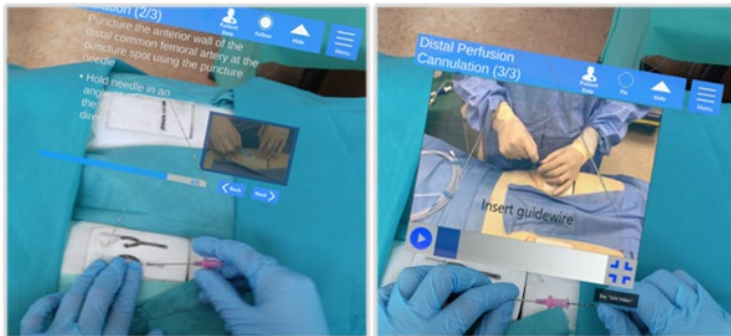


Fig. 2: Example of an ECMO cannulation training with AR step-by-step instructions displaying text, images, and videos next to the area of execution [17].

One of the key advantages of AR is that it gives a high degree of design freedom to display information in a way they best support the current context, often referred to as **contextual information**. Context can be considered to be, for example, the current step or the experience level of a user. A more general definition of context is provided by Anind K. Dey:

Context can be defined as “any information useful for characterizing the situation of an entity, where an entity can be a person, place, or object relevant to the interaction between user and application, including user and application themselves.” [18]

In addition to other devices that can provide digital guidance during a procedure (e.g., a screen), AR HMDs allow contextual information to be freely positioned in 3D space and with a high degree of design freedom, ranging from overlay to displaying information next to the area of execution, from 2D to 3D, and from abstract to geometric (or anatomical) representations.

Although previous studies have demonstrated that AR can improve procedural outcomes [12, 13], it is not possible to completely eliminate human error simply by displaying contextual information. As Dekker [19] noted, human error is systematically linked to a variety of external influences, such as the situation and environment people work in, the task, and the tools they use. Since external influences can change during a procedure, as can operator performance, it is questionable whether a static display of contextual information is sufficient for effective operator guidance. In addition, operators still need to manually navigate through the AR interface, which causes additional effort and can contribute to the stress level, especially for steps or procedures that are already cognitively demanding.

To overcome these limitations, the field of context-aware AR support has emerged. By computationally analyzing data from external sensors or sensors integrated in the AR device in real time, the current context of the operator within the procedural task is inferred. This allows the AR device to adjust contextual information to best address immediate user needs [20, 21], provide feedback on the user's current actions [22] and quality of execution [23], or transition to the next step when a step is complete to reduce manual user input [24]. Context-awareness can further improve collaboration in multi-operator tasks such as surgery, as it facilitates human-machine collaboration with robotic systems, tells staff about upcoming phases so tool changes can be prepared in time, and can estimate the remaining duration of the surgery to plan anesthetizations [21]. A general definition for context-awareness in computing systems is provided by Gartner:

Context-aware computing is a “style of computing in which situational and environmental information about people, places and things is used to anticipate immediate needs and proactively offer enriched, situation-aware and usable content, functions and experiences.” [25]

Fig. 3 shows the schematic relationship between AR and context-aware support in procedural tasks. By default, AR guidance, as any digital guidance, can be implemented in a way that it provides optimal support based on expectable contexts such as the current step or the expertise level of the user. Here, context is determined by manual user input into the system and otherwise remains static. In contrast, context-aware support continuously determines the context and can adapt contextual information dynamically without or with only little need for manual user input. In addition, by comparing current and expected execution in real time, the AR system can provide feedback on task execution.

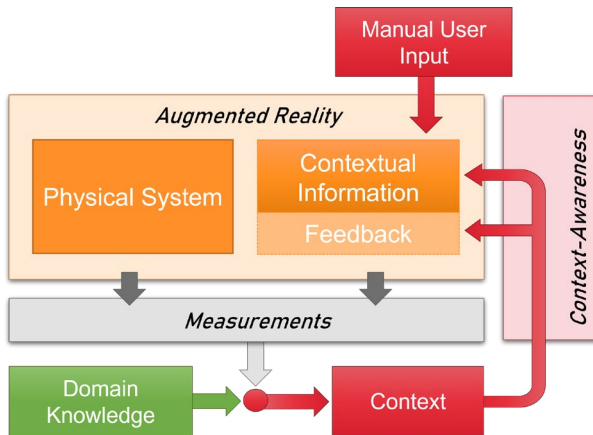


Fig. 3: System model showing the relationship between AR and context-aware support. AR can provide contextual information based on manual user input such as the current step or expertise level of the user. Context can be determined by processing measurement data and interpreting it based on domain knowledge (e.g., standard operating procedure, expected system parameters). Context-aware support automatically determines the current context and enhances AR by making contextual information adaptive and by providing feedback.

Context can be inferred from a variety of sensors that focus on the target system, e.g., condition monitoring of machines states [26] or vital parameters of a patient [27], or the operator, e.g., monitoring tool usage [28]. The latter one can be inferred by tracking objects and tools relevant to the task, the operators' hands, or other psychological measurements (e.g., heart rate, pupil dilation [29]) that indicate the mental or physical state of the operator.

Within the field of AR-guided procedures, previous work has primarily inferred context by tracking relative positions and movements of two or more relevant objects or the user's hands in the physical environment. For example, Ng et al. [24] applied a 2D convolutional neural network (CNN) during AR-guided maintenance to detect the user's hands and particular task-relevant objects in video recordings. Henderson and Feiner [22] applied visual markers during AR-guided assembly to track the movement of handled objects. Based on the relative position of these objects, they provided real-time feedback on the deviation between the current and desired final position of the assembly part. In addition, their system could automatically transition to the next step of the procedure or, if the user was moving the wrong object, display an error message. Tracking tool positions relatively to the registered patient anatomy is also common in surgical navigation systems. Liebmann et al. [23] tracked visual markers attached to a drill sleeve in a surgical simulator setup to provide real-time feedback on the current drill trajectory in relation to the planned (ideal) trajectory (cf. Fig. 4).

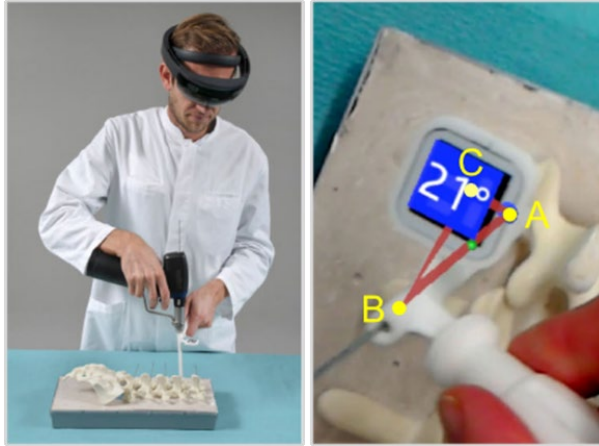


Fig. 4: Surgeon wearing an AR HMD that provides context-aware AR support while drilling into a lumbar spine model. The drill sleeve is enhanced with an ArUco image marker that is tracked by the AR HMDs cameras to display the current drill trajectory (\overline{AB}) next to the planned trajectory (\overline{AC}) as well as the angular deviation in radial degrees [23].

Depending on the purpose for which the **feedback** is provided, it can be divided into two categories: procedural and performance feedback (cf. Table 1). Procedural feedback is provided to ensure the operator does not deviate from the intended workflow. Because procedures can involve many steps, the operator may inadvertently execute a wrong step, skip a step, or only partially complete a step. Even highly trained and experienced surgeons can skip steps during complex procedures if no countermeasures are taken, as evidenced by the many recorded cases of gossypiboma (surgical sponge or a laparotomy pad left in the patient’s body) [30, 31]. Performance feedback focuses on the operator’s motions and is provided when a high accuracy of the task execution is required such as when drilling into bone surrounded by vital structures [23]. It involves continuous tracking of relevant objects and display of the target value (e.g., deviation from end position) in real time so that the operator can adjust his behavior accordingly.

Table 1: Feedback types of context-aware AR support systems

Feedback	Level of Granularity	Examples
Discrete procedural feedback	Step / Activity	<ul style="list-style-type: none"> - Is the right activity being executed? - Has the right step been accomplished?
Continuous performance feedback	Motions	<ul style="list-style-type: none"> - How much of the task (%) is completed? - How accurate is the execution?

Previous work has primarily determined context by detecting and relating two or more objects of interest (object-object context) or the user’s hands (hand-object). Neumann and Majoros [32] refer to this as the psychomotor phase, which represents manual task execution and stands in contrast to the informational or cognitive phase. During the cognitive phase, the user perceives information, understands and interprets that information, and derives how the task can be executed [32]. Other work describes this relationship in a perception, cognition, and action (PCA) model [33, 34]. In essence, these models encompass two key aspects of observable human interaction: the eyes perceiving task-relevant information (input) and the hands manipulating the physical environment (output), while cognitive states and processes are hidden and can only be estimated.

Similar to the input-output-interaction of the human operator and AR, context-aware support processes sensor measurements (input) and then controls the information displayed in AR (output). It is therefore important to derive a system model that can describe the relationships of these interacting systems. One way to model these relationships is to combine the previously presented system model (see Fig. 3) with the PCA model (cf. Fig. 5). The resulting model shows how the human operator perceives information from AR, processes this information, and then performs manipulations within the AR environment. Perception is influenced by visual cues from the virtually presented content, visual cues from the physical environment, and by the motor system when hand movement requires coordination [35]. Therefore, it is important to also integrate measurements into context-aware systems that provide insight into this interaction.

Recent AR HMDs are equipped with better hardware for computation and can therefore provide real-time eye-gaze and hand tracking, both of which have been shown to be suitable for behavior analysis outside of AR contexts [36-38]. Gaze behavior is highly task-dependent [39] and provides deep insights into perception and cognitive processes [40]. Hand tracking can be used to infer hand activities [41], providing insights into how users perform manual tasks [24, 38]. To date, there is limited work integrating eye tracking into context-aware AR systems. For example, Lindlbauer et al. proposed a context-aware system that measures pupil dilation as an indicator of cognitive load. They combined this input with information about task and environment to adapt when, where and how virtual content is displayed [29].

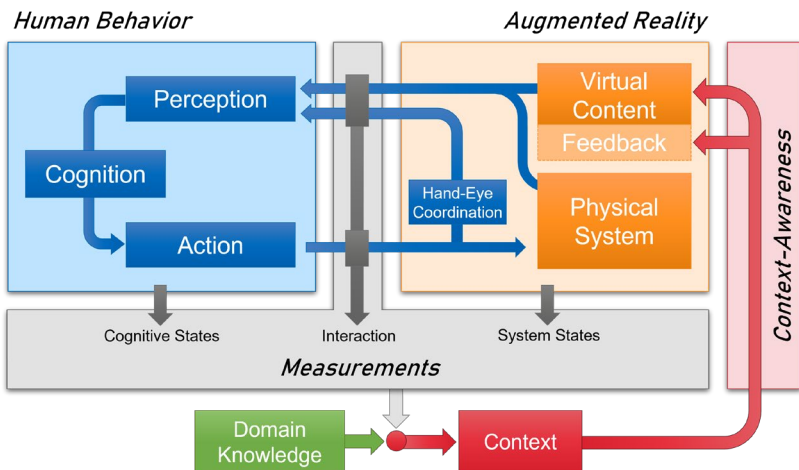


Fig. 5: System model extended with the perception-cognition-action model (adapted from [33, 34]). The model describes the relationship between human behavior, augmented reality, measurements, and context-aware support. Augmented reality adds virtual content to the physical system that is perceived by the operator's eyes, processed, and then executed in an action, i.e., a manipulation within the augmented reality. The measurements form the basis for understanding the interaction between operator and AR and well as states of operator and system. By comparing measurements with domain knowledge, the context can be inferred, which can then be processed by the context-aware support to adapt virtual content and to provide feedback.

Because the information displayed in AR adds visual cues to the real environment that affect operators' visual attention and can even distract them, eye tracking gives important insights into when the user perceived information from the real and when from the virtual world. It therefore helps to better understand how and where information should be displayed during context-aware AR support.

Combining sensing modalities of eye tracking and hand/object tracking further allows to analyze hand-eye coordination, which is the task-dependent relationship between hands and eyes [42]. Gras et al. [43] computed multiple Euclidean distance measurements between tooltips, viewpoint, and patient anatomy in a simulated AR-guided robotic surgery. They then trained a multi-Gaussian process model to automatically determine the desired AR view (overlay on/off) at each time point of the surgery. Such a system using features of the tooltip and the gaze point also learns patterns in hand-eye coordination, which are often distinct, reoccurring patterns of human interaction. One of these distinct patterns occurs during target selection, also referred to as a 'directing pattern' [35], and is usually present when a person moves their hand or a hand-held object towards a target position. Several studies have shown the eyes fixating on task-relevant objects and target locations before hand movement starts [44, 45]. It can therefore be expected that the eyes are not only an indicator for the current action, but also for future actions, and enable potentially new implications of real-time support. Such predictive support could be particularly useful in industrial and clinical applications to combat the high risk and cost of user errors.

In summary, the use of conventional information mediums comes with certain challenges that complicate the training and execution of procedural tasks in professional disciplines. AR can potentially improve procedural outcomes, but more work is needed to understand how contextual information in AR affects task performance. As a static display of contextual information is limited in providing procedural guidance, context-aware AR support promises to overcome these limitations by making contextual information adaptive and by providing feedback. AR offers a variety of possibilities to visualize and position information

during context-aware support and more work is needed to understand the effect of visualizations on task performance and user behavior. Here, eye tracking might be particularly interesting to also understand how displayed information affects visual attention and might be useful for optimizing AR support. In addition, a real-time analysis of hand and eye tracking features could enable predictive AR support and error prevention.

The next chapter gives a short overview of topics relevant to this work that are not covered within the main studies of this thesis (chapter 4-6), whereas chapter 3 explains the goals and contributions of this work in more detail.

2 Background

1.1 Augmented Reality Head-mounted Displays

Augmented reality (AR) is a technology that superimposes a computer-generated image on the real world to create a composite view [1]. One of the first prototypes was famously introduced by Ivan Sutherland in 1968 [46], and AR has been an active area of research ever since. AR is often defined by three characteristics: (1) It combines real and virtual world, (2) it is interactive in real time, and (3) virtual content is registered in 3D space [47]. Registration refers to the accurate alignment of virtual content on the physical environment so that virtual objects can, for example, remain fixed in 3D space even when the user is moving the device. While some work argues that AR is not limited to visual augmentations and should also address other human senses [48], the vast majority of AR experiences are based on visual augmentations. Visual augmentations can be achieved by video projectors, video see-through (VST) systems that first record the scene through cameras and then display the augmented video on a screen, or optical see-through (OST) systems that use semi-transparent displays [49]. Common devices include stationary screens, hand-held devices (e.g., smartphones), body-worn displays and in particular head-mounted displays (HMDs). OST-HMDs have seen the biggest technological advancements over the last few years with sophisticated hardware (e.g., Microsoft HoloLens 2) being commercially available, and have democratized the use of AR for a broader community.

Being one of the most sophisticated OST-HMDs on the market, a Microsoft HoloLens 2 (Microsoft, Redmond, Washington) was chosen for the investigations in this thesis. The device uses four head tracking cameras, an inertia measurement unit (IMU), a RGB camera and a near and far range depth camera for perceiving its environment [50]. These sensor inputs are processed by sophisticated SLAM (simultaneous location and mapping) algorithms to create a map of the environment and position the device within it. As a consequence, the device can place virtual objects stationary in 3D space even when the user is moving. The

device can further track the user's hand poses and eye movement. As the transformation between device and registered environment is known, as is the transformation between eye and hand tracking sensors and the device, these measurements can automatically be registered in the physical environment, providing a first semantic layer for behavior analysis. Fig. 6 shows the sensor inputs relevant to this work at the example of a surgical procedure. Besides eye and hand tracking, the presented system processes RGB camera input to predict ArUco marker positions in 3D space, and to provide continuous performance feedback on task execution.

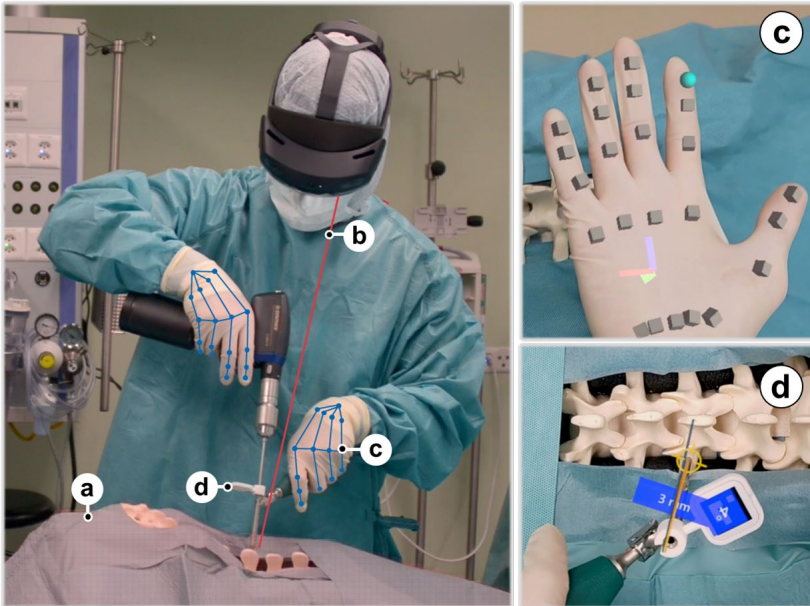


Fig. 6: Surgeon drilling into the spine while wearing a Microsoft HoloLens 2 with illustrated sensors inputs (left). The AR HMD tracks its position relatively to the (a) registered physical environment, tracks (b) the operator's eye-gaze direction, (c) the hand poses with 26 finger joints for each hand, and (d) the tool pose with 6 DoF. The two images on the right show the (c) hand tracking and (d) ArUco marker-based tool tracking with continuous performance feedback as displayed to the operator. (© raw images left and bottom right: Balgrist campus, ROCS group)

1.2 Preliminary Work on Automating Semantic Eye Tracking Analysis

Eye tracking has been shown to provide deep insights into perception and cognitive processes [40]. One of the most established ET methods for semantic interpretation of gaze behavior is the Area of Interest (AOI) analysis, where the stimuli is divided into areas (semantic entities) that are of interest to the evaluator. AOIs can be any visible part of the physical or virtual environment, including interface buttons, a part of a screen, or hand-held objects. Mapping the eye-gaze onto AOIs gives the eye tracking data semantic meaning and allows to calculate important metrics for depiction of visual attention, such as AOI dwell times, i.e., the total time spent looking on an AOI, AOI transitions, and AOI sequences. [51]

AR HMDs such as HoloLens 2 can measure the eye-gaze's direction in relation to the registered physical environment and can therefore automate the mapping of eye-gaze on the physical environment. By positioning invisible virtual objects in 3D space, eye-gaze interaction with AOIs in the physical space can be measured on the virtual level. However, this automated mapping does not work for tangible objects such as hand-held tools or assembly parts that are moved during the procedure. As of 2018, the gold standard for mapping eye-gaze to AOIs required the evaluator to iterate fixation by fixation through the whole eye tracking video recording and to manually assign the eye-gaze of each fixation to the respective AOI. This approach is highly tedious and inefficient, and is not applicable when eye tracking is to be processed for real-time support in context-aware systems.

To overcome the limitations of AOI analysis for dynamic tangible objects and for real-time purposes, we have proposed an automated gaze mapping approach, the computational gaze-object mapping (cGOM) [52]¹. The algorithm is based on a deep convolutional neural network (CNN) and trained to detect objects that are of interest. While active, it iterates

¹ J. Wolf, S. Hess, D. Bachmann, Q. Lohmeyer, and M. Meboldt. Automating areas of interest analysis in mobile eye tracking experiments based on machine learning. *Journal of Eye Movement Research*. 2018

through all video frames and makes inference with Mask R-CNN, a deep CNN, to predict (i) the pixel area that is expected to belong to the object, and then (ii) compares the pixel area to the pixel coordinate of the gaze point (cf. Fig. 7). If the gaze point lies within the pixel area, the fixation is assigned to the AOI. If no match is noted, the fixation is assigned to the background or whitespace (all remaining space). The algorithm’s performance was evaluated in a lab setup on eye tracking recordings of 10 participants with the two AOIs ‘syringe’ and ‘bottle’. Using only 264 labelled object representations for the syringe, it achieved a true positive rate (TPR) of 80% and a true negative rate (TNR) of 85% compared to manual AOI mapping.

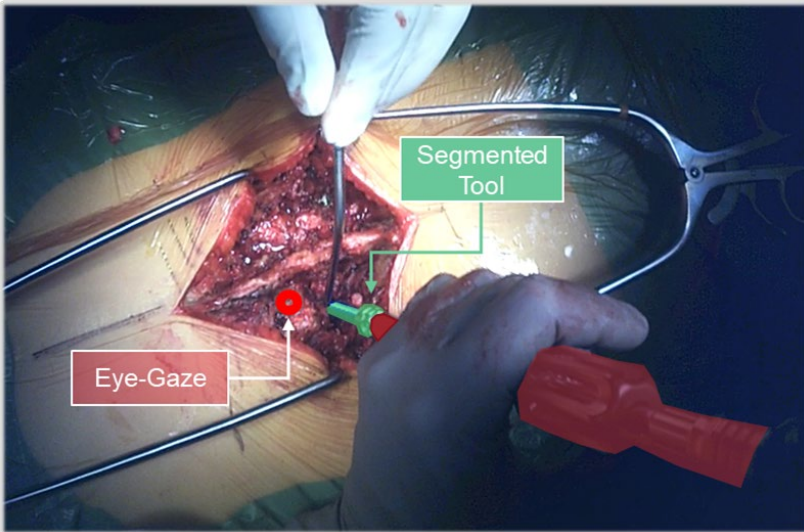


Fig. 7: Computational gaze-object mapping [52]. The two AOIs ‘screw’ (green) and ‘screwdriver’ (red) are predicted by Mask R-CNN and then compared to the coordinate of the eye-gaze point (red circle). While conventional AOI analysis only assigns the gaze point of a fixation to one AOI at a time, our work proposed an extension that calculates the 2D Euclidean distances between each AOI and the gaze point, also referred to as object-gaze distance (OGD), creating a timeseries that provides a better representation of visual attention [53].

To test the effectiveness of automated AOI analysis for more complex real-world applications, a second study was conducted in a real surgical environment (cf. Fig. 7). Two tools (a screw and a screwdriver) were selected as AOIs. Eye-tracking video recordings showed that the gaze point was often in the immediate periphery of the AOIs. Consequently, most fixations were assigned to white space, which has the same informative value as if the gaze had been far away from the AOIs. Moreover, if the gaze was located between the screw and the screwdriver, it could only be assigned to one AOI. It is therefore questionable whether such a binary assignment of fixations to a single AOI is sufficient to map the visual attention of the operator when working with multiple dynamic objects. In order to include the peripheral vision into AOI analysis and to make AOI analysis applicable to dynamic multi-object environments, we proposed an extension of the state-of-the-art AOI analysis, the object-gaze distance (OGD) [50]². The extended algorithm computes the 2D Euclidean distance between the gaze point and each AOI pixel region in the image plane, resulting in a multi-OGD time series in which 0px represents an AOI hit.

The analysis of the surgical data showed a significant increase in interpretable gaze data when near-peripheral vision was included, with fixation data increasing from 23.8% to 78.3% for the AOI ‘screw’ and from 4.5% to 67.2% for the AOI ‘screwdriver’. In addition, it has been shown that the combined evaluation of gaze distances to multiple objects reveals new gaze patterns and thus could provide a more accurate representation of operator gaze behavior. In addition, multi-OGDs are expected to be information-rich features for training time series models on the current step or activity of a procedural tasks.

² F.S. Wang, J. Wolf, M. Farshad, M. Meboldt, & Q. Lohmeyer. Quantifying near-peripheral gaze behavior in real-world applications. *Journal of Eye Movement Research*. 2018

3 Goals and Contributions

Procedural tasks in professional disciplines such as maintenance or surgery place the highest demands on correct execution, as human error can have serious consequences for elaborated machines, patients, or the operator himself. AR and context-aware AR promise to offer new opportunities to improve procedural outcomes. Understanding the strengths and weaknesses of these technologies is key to developing support systems that can improve the quality of task execution, making procedural tasks safer and increasing outcomes.

The first goal of this thesis is to investigate different concepts of how AR and context-aware AR support systems can be designed, how they work, and how they affect operators' task performance. As shown in the system model, perception is an important part of human interaction that needs to be understood when designing context-aware support. Therefore, as a second goal, this thesis focuses on understanding how context-aware AR support systems can benefit from eye tracking. The third goal is to validate the suitability of the proposed system model for explaining the study characteristics and underlying system relationships in this work.

The thesis is divided into three main studies. The first study aims at better understanding the advantages of contextual information displayed in AR over traditional information mediums. The second and third study investigate task performance of context-aware AR support with two different feedback types and eye tracking analysis both to understand visual behavior and for real-time support. The research questions and their motivation are explained below.

RQ 1: How do contextual information in AR affect execution errors?

One of the most common types of AR support is **AR step-by-step instructions**. Compared to conventional information mediums, AR instructions offer two chief advantages: operators' access to (1) contextual information, which reduces complexity to manageable increments, and (2) the proximity of information, which allows operators to continuously check their execution against the instructions in real time

and to adjust their behavior accordingly. Previous studies have shown that AR instructions can reduce the number of execution errors [12, 13]. A question that remains unanswered is whether these errors are caused by participants understanding of the task (perception and cognition), or by their ability to manually execute the task. To better understand error occurrences and causes, we conducted a user study on a complex medical procedure and manually evaluated a detailed error protocol with both a categorization into knowledge- and handling-related errors as well as an error severity ranking.

RQ 2: What are effective visualization strategies for continuous performance feedback?

When designing AR guidance, there are different visualization strategies in how to display information, ranging from overlay to displaying information next to the area of execution, from 2D to 3D, and from abstract to geometrical or anatomical representations. These strategies are expected to be particularly important when designing continuous performance feedback, as the operator's perception is confronted with two conflicting goals: checking the displayed target value while coordinating hand movement for accurate manual execution. It can therefore be assumed that eye tracking provides important insights into the visual behavior that is crucial for optimizing AR support. To investigate **effective visualization strategies during continuous performance feedback**, we conducted a user study focusing on a single step of a medical procedure and evaluated how different configurations of abstraction level (abstract or geometrical/anatomical), position (overlay or small offset), and dimensionality (2D or 3D) affect task performance, visual attention, and user experience.

RQ 3: How suitable is a joint analysis of hand motions and eye movement for predicting and preventing erroneous hand actions?

Human error in industrial and clinical applications can be associated with high risk and cost. Previous work has proposed context-aware systems that provided **discrete procedural feedback** on the current action [22,

24] to tell operators when a wrong activity is being executed. One question that remains unanswered is whether wrong actions can also be predicted in advance to warn the user of potential errors before they occur. Eye tracking might be of particular importance as several studies have shown the gaze preceding hand movement during hand-object interactions, making eye-gaze a suitable indicator for **predicting future hand actions**. In this study, we investigated whether a joint analysis of eye and hand tracking features is suitable for predicting hand actions, and whether prediction happens early enough to stop the user’s hand movement before the erroneous hand action starts. We conducted a study in a lab environment on a fast and repetitive two-step procedure to test the timeliness and accuracy of this novel AR support.

Fig. 8 gives an overview of the study characteristics. The studies are explained in more detail in the following sections 3.1 – 3.3.





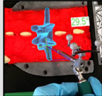



	Support Strategy		Primary Measurements		Application	
	Context	Feedback	Task Performance	Eye Tracking	Procedure	Case Study
Chapter 4 Study I (RQ1)	Contextual Information (non-adaptive)	✗	✓	✗	Linear, Multi-step 	ECMO Cannulation 
Chapter 5 Study II (RQ2)	Object-Object 	Continuous Performance Feedback	✓	Understand Visual Attention	Single Step 	Pedicle Drilling 
Chapter 6 Study III (RQ3)	Eye-Hand-Object 	Discrete Procedural Feedback	✓	Real-time Support	Two Steps (repetitive) 	Memory Game 

Fig. 8: Overview of the three main studies addressing research question (RQ) 1-3 and their system characteristics regarding support strategy, primary measurements, and application. All studies evaluate operators’ task performance. The first study investigates (non-adaptive) contextual information based on manual user input, whereas the second and third study investigate context-aware support with two different feedback types and eye tracking both to understand visual attention and for real-time support.

3.1 Comparing AR against Conventional Instructions

To better understand error occurrences and causes, we conducted a user study on a complex medical procedure and evaluated a detailed error protocol with both a categorization into knowledge- and handling-related errors as well as an error severity ranking. We chose the extracorporeal membrane oxygenation (ECMO) cannulation as a procedure as it is characterized by many steps that require detailed knowledge and fine-motoric execution. An AR step-by-step guide was developed for the Microsoft HoloLens 2 that combines the same text, images, and videos from the conventional training program with simple 3D models. Fig. 9 shows the system model applied to the first study setup.

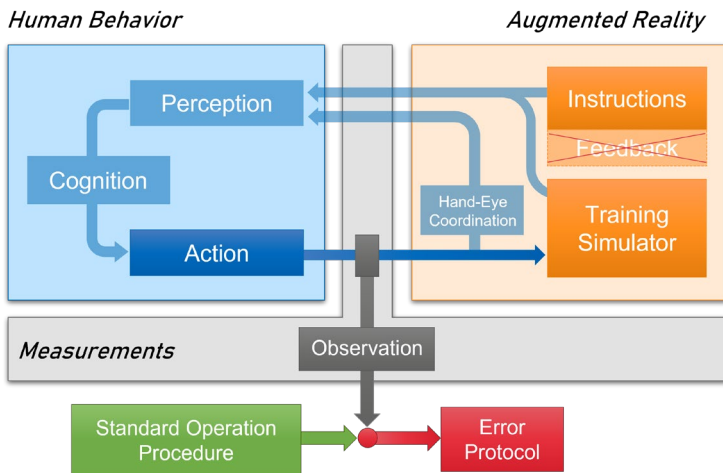


Fig. 9: System model applied to the first study setup. The study compared AR instructions to conventional instructions without context-aware support in a physical simulator setup and evaluated a detailed error protocol by comparing observed execution to the standard operating procedure. The study further evaluated 'completion time' and 'user experience' as secondary measurements that are not shown in the model.

These AR instructions were developed in an iterative process while regularly consulting experts from the Heart Center of the University

Hospital Zurich (USZ). A study was conducted with 21 medical students performing ECMO cannulation on a physical simulator. Training times, a detailed error protocol, and a standardized user experience questionnaire (UEQ) were evaluated.

The study is presented in Chapter 4. The content of this chapter has been published in the *International Journal of Computer Assisted Radiology and Surgery*. The AR application was developed as part of the ETH focus project ARORA and the study was greatly supported by the bachelor thesis of Viviane Wolfer.

- [17] J. Wolf, V. Wolfer, M. Halbe, F. Maisano, Q. Lohmeyer, and M. Meboldt. Comparing the effectiveness of augmented reality-based and conventional instructions during single ECMO cannulation training. *International Journal of Computer Assisted Radiology and Surgery*, 2021. doi: 10.1007/s11548-021-02408-y

3.2 Effective Visualization Strategies

To better understand how different visualization strategies in AR affect task performance, visual attention, and user experience, we conducted a user study focusing on a single step of a medical procedure that benefits from continuous performance feedback. Real-time performance feedback during tool handling is particularly important for surgical applications such as spinal fusion surgery. One particularly challenging step requires pre-drilling pedicle screw trajectories into the vertebrae with a surgical drill. Due to the proximity to vital structures, strong intraoperative bleeding and variability in morphology between patients [21,2,17,31,25], spinal fusion surgery is a very demanding procedure. Previous work has demonstrated the high accuracy of augmented reality (AR) head-mounted displays (HMD) for drilling pedicle trajectories. An important question that remains unanswered is how pedicle screw trajectories should be visualized in AR to best assist the surgeon.

In this study, we compared five AR visualizations displaying the drill trajectory via Microsoft HoloLens 2 with different configurations of

abstraction level (abstract or anatomical), position (overlay or small offset), and dimensionality (2D or 3D) against standard navigation on an external screen. The visualizations were derived and tested in an iterative process in close collaboration with experts from the Balgrist University Hospital Zurich. The final visualizations were tested in a study with 4 expert surgeons and 10 novices (residents in orthopedic surgery) on lumbar spine models covered by Plasticine. We assessed trajectory deviations ($^{\circ}$) from the preoperative plan, dwell times (%) on areas of interest (AOIs), and the user experience. Fig. 10 shows the system model applied to the second study setup.

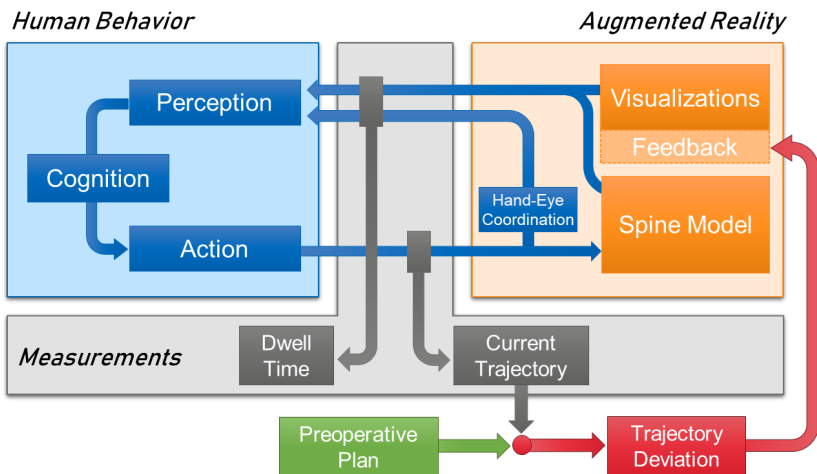


Fig. 10: System model applied to the second study setup. The study compared different AR visualizations to conventional navigation in a physical spine model setup while providing the trajectory deviation as continuous performance feedback. The primary measurements are 'dwell time' and 'trajectory deviation'. The study further evaluated the user experience with emphasis on 'visualization ranking', 'cognitive load', and 'ease of use', which are not presented in the model.

The study is presented in detail in Chapter 5. It was greatly supported by the master thesis of Dietmar Luchmann. The content of chapter 5 has been submitted to the *International Journal of Computer Assisted Radiology and Surgery* and is currently under review. This project is part of the

SURGENT project and was funded by University Medicine Zurich/ Hochschulmedizin Zürich.

J. Wolf, D. Luchmann, Q. Lohmeyer, M. Farshad, P. Fürnstahl, and M. Meboldt. How augmented reality visualizations for drilling affect trajectory deviation, visual attention, and user experience. *International Journal of Computer Assisted Radiology and Surgery (under review)*, 2022.

3.3 Predicting Future Hand Actions

To investigate the suitability of eye and hand tracking features for predicting and preventing erroneous hand actions, we conducted a user study in a simplified lab setup. We chose a memory card game for this study because it 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. The study consisted of two experiments. The first experiment, which involved 10 participants, was designed to record participants' eye and hand movement data in order to derive a method for target prediction. The second experiment included 12 participants and examined the timeliness and accuracy of the implemented method end-to-end. Fig. 11 shows the system model applied to the third study setup.

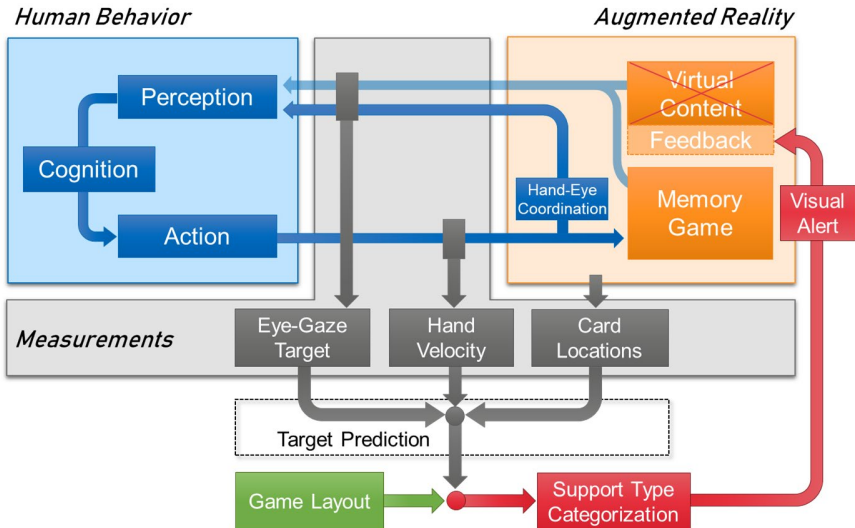


Fig. 11: System model applied to the third study setup. Participants played a memory card game without receiving task-relevant information. Eye-gaze and hand movements were recorded in relation to the card locations to predict the next card turn. Predictions were compared to a ground truth game layout to categorize the support type and provide participants with discrete procedural feedback, i.e., either green, yellow, or red visual alerts. The measurements comprised of ‘accuracy’ of the target prediction and the ‘timeliness’ of displayed visual alerts to stop the hand action before the card turn starts.

The study is presented in detail in Chapter 6. The content of this chapter has been published in *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. The project is part of the SURGENT project and was funded by University Medicine Zurich/ Hochschulmedizin Zürich.

- [54] J. Wolf, Q. Lohmeyer, C. Holz, and M. Meboldt. Gaze comes in handy: predicting and preventing erroneous hand actions in AR-supported manual tasks. *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2021.

4 Study I: Comparing AR against Conventional Instructions

The content of this chapter has been published in the *International Journal of Computer Assisted Radiology and Surgery* [17]. The study was greatly supported by the bachelor thesis of Viviane Wolfer. The AR application was developed as part of the ETH focus project ARORA.

Abstract

Effective training of extracorporeal membrane oxygenation (ECMO) cannulation is key to fighting the persistently high mortality rate of ECMO interventions. Though augmented reality (AR) is a promising technology for improving information display, only a small percentage of AR projects have addressed training procedures. The present study investigates the potential benefits of AR-based, contextual instructions for ECMO cannulation training as compared to instructions used during conventional training at a university hospital. An AR step-by-step guide was developed for the Microsoft HoloLens 2 that combines text, images, and videos from the conventional training program with simple 3D models. A study was conducted with 21 medical students performing two surgical procedures on a simulator. Participants were divided into two groups, with one group using the conventional instructions for the first procedure and AR instructions for the second and the other group using instructions in reverse order. Training times, a detailed error protocol, and a standardized user experience questionnaire (UEQ) were evaluated. AR-based execution was associated with slightly higher training times and with significantly fewer errors for the more complex second procedure ($p < 0.05$, Mann–Whitney U). These differences in errors were most present for knowledge-related errors, resulting in a 66% reduction in the number of errors. AR instructions also led to significantly better ratings on 5 out of the 6 scales used in the UEQ, pointing to higher perceived clarity of information, information acquisition speed, and stimulation. The results extend previous research on AR instructions to

ECMO cannulation training, indicating its high potential to improve training outcomes as a result of better information acquisition by participants during task execution. Future work should investigate how better performance in a single training session relates to better performance in the long run.

4.1 Introduction

Extracorporeal membrane oxygenation (ECMO) is a life-saving procedure for severe respiratory or cardiac failure that has evolved from a last-resort treatment to a more mainstream therapy over the past few years [55]. As ECMO cannulations gain more importance as an emergency treatment, the number of cases is increasing rapidly, and more and more hospitals are performing ECMO cannulations themselves [56]. While the mortality rate has decreased slightly as ECMO usage has increased, it remains high at over 60% [15]. Sufficient training for ECMO cannulation in general is shown to be linked with decreasing mortality rates [16]. Furthermore, simulation-based training shows significant improvements for ECMO cannulations [57, 58]. Frequent and realistic training therefore seems to be key when it comes to successful ECMO procedures. However, training is time-consuming, and there is no standardized certification or training process for ECMO cannulations [58]. Even though the Extracorporeal Life Support Organization (ELSO) has developed specific guidelines for safe ECMO practice, they are only used as a basic structure for ECMO centers to build varying institution-specific guidelines and programs around [58, 59]. Since ECMO cannulations are often emergency operations, a physician can go months or years without having to perform an ECMO cannulation before suddenly being confronted with a time-critical situation [60]. To be able to act precisely and quickly, physicians need to be provided with frequent and thorough training.

As a consequence, several approaches have been put forward, aiming to simplify and improve ECMO cannulation training options on the one hand and to work towards a standardized procedure on the other.

Simulation-based medical training in general is swiftly gaining ground, and high-fidelity simulators have been developed, with promising results, for ECMO cannulations as well [60]. Many of these involve mannequins and/or silicon-based tissue pads to make for realistic cannulation training [61]. Other simulators even involve the recreation of the ECMO- circuit and include applications where the instructor can manipulate ECMO data, among other factors, to simulate common problems [62]. Those, however, are mainly designed for post-cannulation problems.

In other areas of medicine, augmented reality (AR) has started to emerge as a training tool in recent years, offering fundamentally new possibilities for visualization and interaction with digital content. Modern devices (e.g., the Microsoft HoloLens 2) are affordable and easy to use and therefore widely accessible for training purposes. However, only a small percentage of AR projects have dealt with training procedures, while most of them have been applied to actual treatment scenarios [4]. Existing training applications based on AR depict, for example, the internal anatomy superimposed on a simulator [63] or overlay a CT scan to train methods for ultrasound [64]. Currently, there is no work on how Optical head-mounted displays (OST-HMD) can be utilized for ECMO cannulation. Furthermore, only limited research on step-by-step procedures and on the possible benefits of AR display options in medical training has been conducted. Azimi et al. [65] first trained, then assessed 20 participants in two emergency medical procedures to compare the effectiveness of AR-based instructions provided by an OST-HMD to conventional training. They found participants using the AR instructions spent more time training but were faster in completing the procedure in the assessment run. Participants further found the use of OST-HMDs more engaging and reported higher levels of confidence. In this article, we evaluate AR step-by-step instructions for ECMO cannulation training and compare them with the conventional training instructions regularly used at a university hospital. In addition to training times and user experience, we emphasize the quality of execution, i.e., the number of errors, observed in a single training run. Fewer errors during the training

are expected to lead to fewer necessary training iterations and less supervision needed to learn the procedure.

We see two chief advantages of AR step-by-step instructions: learners' access to contextual information, which reduces complexity to manageable increments, and (2) the proximity of information, which allows participants to continuously check their execution against the tutorial in real time, and to adjust their behavior accordingly. Consequently, we expect AR training instructions to result in shorter training times, fewer errors and better user experience, and thus lead to higher skill levels after a single training run.

4.2 Related Works

Even though AR applications for ECMO cannulation training have not yet been evaluated, previous work has investigated the benefits of ECMO cannulation training as well as AR for medical training.

4.2.1 Augmented Reality in Medical Training

AR is used and has been evaluated in different areas of medicine, including surgical environments [66], therapy [67] and training [68]. In [69], medical training with mobile AR was compared to textbook-based learning. Medical students were asked to study the given material (either mobile AR or textbook) for 45 minutes. No significance differences between the two groups were reported for either knowledge tests or experience questionnaires, but indicated that long-term retention of knowledge may be better with mobile AR. In contrast to the present study, the study only tested medical knowledge and no manual execution of steps.

Other work with Microsoft HoloLens 1 investigated the differences between AR- based and computer-based suture training for medical students [70]. Participants could choose to watch videos on either the HoloLens or a conventional computer and to execute a suturing pattern. The study showed that videos were watched more frequently on the

HoloLens, but there was no significant difference regarding the execution of or time spent on the manual task.

4.2.2 ECMO Cannulation Training

Several studies have shown that adequate cannulation training is crucial for minimizing complications and is therefore an important topic for further investigation and improvement [16, 71]. High fidelity cannulation simulators have been developed and various studies indicate that simulation-based training for ECMO implantations is highly effective for participants' knowledge and ability, and hence results in lower cannulation time and better execution [57, 58].

4.3 Methods

4.3.1 Study Design

To compare conventional and AR-based instructions for ECMO cannulation training, a study was conducted with 21 medical students. Participants had to perform two procedures, each using a different mode of instruction. Similar to the study performed by Azimi et al. [65], participants were split into two groups. One group performed P1 (procedure 1) with AR instructions and P2 with conventional instructions, while the other group performed the same training, but in reverse. Hence, each group was acting as the control group for one procedure. To minimize sequence effects of procedures, half of the participants of each group started with the second and the other half with the first procedure.

4.3.2 Task

Two surgical procedures were performed by each participant (cf. Fig. 12). The procedures were adapted to the study framework so that all steps could be completed on the simulator. Steps including ultrasound verification alone were skipped and steps including ultrasound guidance were adapted accordingly. More precisely, prior to the experiment, we

drew the bifurcations (venous and arterial) for the distal perfusion cannulation and marked the height of the puncture spot for the venous cannulation.

Distal perfusion cannulation included the identification of the left femoral artery branching, the puncture of the left distal common femoral artery (directed caudally), insertion of the limb perfusion cannula using the “Seldinger Technique”, removal of the guidewire, and an NaCl flush. Venous puncture comprised the identification of the right femoral vein, the incision of the skin at the puncture spot, puncture of the common femoral vein (directed cranially) and the measurement of the required cannula and wire length. Finally, venous cannulation included advancing the wire to the measured length, serial dilation of the vessel (using 3 dilators), the insertion of the venous cannula, and the removal of the wire and clamping of the cannula.

4.3.3 Experimental Procedure

Participants were first asked to fill out a questionnaire regarding their previous experiences with HoloLens and ECMO devices as well as a consent form. They then went through a ten-minute HoloLens 2 tutorial to familiarize themselves with the AR interface and the different navigation types (voice command, hand gesture). Participants then performed both procedures. Prior to starting a procedure, participants watched a video showing all steps to be performed, which aimed at imitating a live demonstration in real surgery prior to training. After each procedure, they filled out a user experience questionnaire (UEQ).

4 Study I: Comparing AR against Conventional Instructions

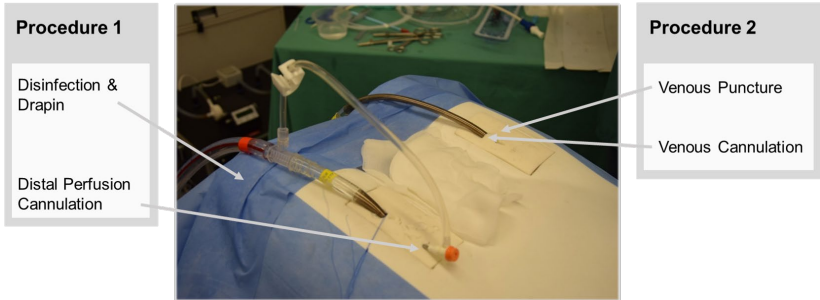


Fig. 12: The two procedures investigated during the study. Each procedure is performed on a different side of the simulator.

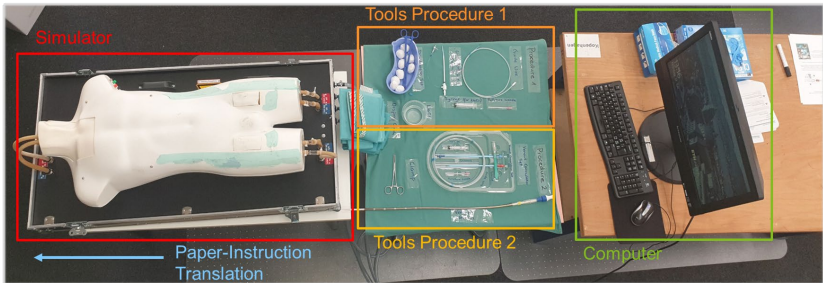


Fig. 13: Experimental setup consisting of a simulator, tools for procedure 1 and 2 and a desktop computer. Paper instructions are placed on the left-hand side of the simulator.

4.3.4 Experimental Setup

The setup consisted of a simulator, a table with a green tablecloth on which all tools were placed, and a stationary desktop computer (cf. Fig. 13). The simulator was placed in such a way that the hoses connected to it did not disturb the participant. All tools were labelled with the tool names used in the paper or AR instructions. They were divided into two sets, one for procedure 1 and the other one for procedure 2. A desktop computer was placed next to the tools, on which participants filled out the questionnaires and watched the initial videos. During the conventional

training, they could also use the computer to watch the video sequences presented in the AR instructions.

4.3.5 Participants

The study was conducted with 21 third- and fourth-year medical students (aged 22 - 30, 9 males, 12 females) who had never performed an ECMO cannulation before.

4.3.6 Information Material

Both conventional and AR instructions are based on a standard operating procedure for ECMO cannulation developed at the heart center of a university hospital. Some adjustments were made to the instructions to better suit the simulation setting. This included the replacement of ultrasound guidance with predefined marks on the simulator and some additional specifications to make the cannulation possible for participants with no previous ECMO experience.

4.3.7 Conventional Instructions

Conventional instructions consist of a printed version of the standard operating procedure that include text and supplementary images. In addition, participants could watch video sequences on a desktop computer (cf. Fig. 13).

4.3.8 Augmented Reality Instructions

The AR instructions were designed as a step-by-step guide that included the same text, pictures, and videos as the conventional instructions. In addition, simple 3D models were displayed in two steps (Fig. 14). It was developed for the Microsoft HoloLens 2 (Microsoft, Redmond, Washington) using the Unity 3D Game Engine (Unity Technologies, San Francisco, California). The application can be controlled both by hand gestures and voice commands. Using the outstretched index finger, the

4 Study I: Comparing AR against Conventional Instructions

user can interact with the interface simply by moving the finger "through" the projected button in the same way one would press a physical button. Audio as well as visual feedback indicate that a button was successfully pressed. Voice commands work either by reading out the name of a particular button or by focusing the eyes on the button and saying "select".

4.3.9 Simulator

A TF200 ECMO-Simulator by Erler Zimmer (AcuMax Med AG, Bad Zurzach, Switzerland) was used. It is specifically designed for ECMO cannulation training and regularly used during trainings at the university hospital. The simulator contains venous and arterial blood circulation through a hose system. Integrated pumps allow for individually adjustable blood flow and therefore realistic simulation. Common medical tools can be used on the simulator.

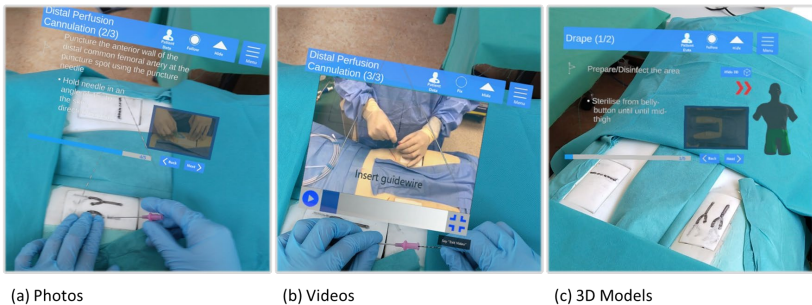


Fig. 14: Three information representation types complementing the text-based step-by-step instructions in AR.

4.3.10 Data Analysis

For the comparison of conventional and AR-based instructions, training time, number of errors and user experience were evaluated.

Error Analysis

We derived an error protocol that contains a list of possible errors and ranks each of them with a factor between 1 and 3 based according to their severity.

- 1-Point-Errors
 - Small errors
 - Errors without impact on further steps
 - Partially completed sub-steps
- 2-Point-Errors
 - Larger errors
 - Errors with impact on other steps or further progression
- 3-Point-Errors
 - Skipped steps
 - Maximum time for a step exceeded (leading to incompleteness)

Errors were further categorized as either handling errors, knowledge errors, or both. Handling errors are more related to participants' individual dexterity than to the clarity of the instructions, while knowledge errors are caused by the participants' lack of attention to the relevant, task-related information. The error protocol for procedure 2, including all errors and their respective categorization and severity ranking, is shown in Fig. 17.

User Experience Questionnaire

To assess the participant's personal experience, a User Experience Questionnaire (UEQ) [72] was handed out after each procedure. This standardized questionnaire consists of 26 questions. For each question, two contrasting adjectives were juxtaposed, and the participant was asked to decide where on the scale, from 1 (complete agreement with the left adjective) to 7 (complete agreement with the right adjective) their personal experience lay.

4 Study I: Comparing AR against Conventional Instructions

- Attractiveness: What is the overall impression?
- Perspicuity: Is it easy to get familiar with and is it easy to learn?
- Efficiency: Can the tasks be solved quickly and without unnecessary effort?
- Dependability: Is the system reliable and does the user feel in control of its handling?
- Stimulation: Is it exciting and motivating?
- Novelty: Is the product innovative and catchy?

All points awarded by the participants were rescaled for the evaluation so that the possible range of points lies between -3 (extremely bad) and +3 (extremely good). A neutral evaluation usually lies between -0.8 and 0.8, whereas values >0.8 signify a positive evaluation and values <-0.8 a negative one [73]. Fig. 15 shows all 26 UEQ questions for the six scales.

1		2	3	4	5	6	7				1		2	3	4	5	6	7			
Annoying	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Enjoyable	A	1	Unlikable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Pleasing	A	14
Not understandable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Understandable	P	2	Usual	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Leading edge	N	15
Creative	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Dull	N	3	Unpleasant	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Pleasant	A	16
Easy to learn	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Difficult to learn	P	4	Secure	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Not secure	D	17
Valuable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Inferior	S	5	Motivating	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Demotivating	S	18
Boring	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Exciting	S	6	Meets expectations	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Does not meet expectations	D	19
Not interesting	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Interesting	S	7	Inefficient	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Efficient	E	20
Unpredictable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Predictable	D	8	Clear	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Confusing	P	21
Fast	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Slow	E	9	Impractical	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Practical	E	22
Inventive	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Conventional	N	10	Organized	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Cluttered	E	23
Obstructive	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Supportive	D	11	Attractive	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unattractive	A	24
Good	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Bad	A	12	Friendly	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unfriendly	A	25
Complicated	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Easy	P	13	Conservative	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Innovative	N	26

Fig. 15: User Experience Questionnaire (UEQ) consisting of 26 questions for the six scales attractiveness (A), perspicuity (P), efficiency (E), dependability (D), stimulation (S) and novelty (N).

4.4 Results

4.4.1 Previous Experience

Among the 21 participants, 12 reported having completed their third year and 9 their fourth year of medical studies. No participant reported having any previous experience with ECMO cannulation, though 9 participants had previous experience with other cannulation procedures (e.g., venous cannulation or venous catheter into hand/arm). There was no significant difference in training time or error count for either procedure when comparing the group with cannulation experience to the group without. 5 participants claimed to have used AR glasses once before; 3 of them reported having used a Hololens 1. There was no significant difference between the group with previous AR experience and the group without any AR experience in terms of P1 training time or error counts in either P1 or P2. The training time of P2, however, was significantly lower for participants with previous AR experience ($p < 0.01$, t-test). Of these 5 experienced participants, 2 performed P2 with AR instructions and 3 with conventional instructions.

4.4.2 Overall Performance

Fig. 16 shows the training times and total error counts for procedure 1 and procedure 2 for both AR-based and conventional instructions. Training times of both procedures are subject to a normal distribution (Shapiro-Wilk-Test), which is not the case for the error count. For P1, AR instructions were associated with slightly higher training times and slightly lower error counts. P2 was characterized by significantly higher mean training times than P1. For P2, AR instructions were associated with slightly higher training times, but with only half the variance. AR instructions also resulted in significantly lower error counts than conventional instructions ($p < 0.05$, Mann-Whitney-U test). Error counts and training times were not significantly correlated (Spearman correlation).

4 Study I: Comparing AR against Conventional Instructions

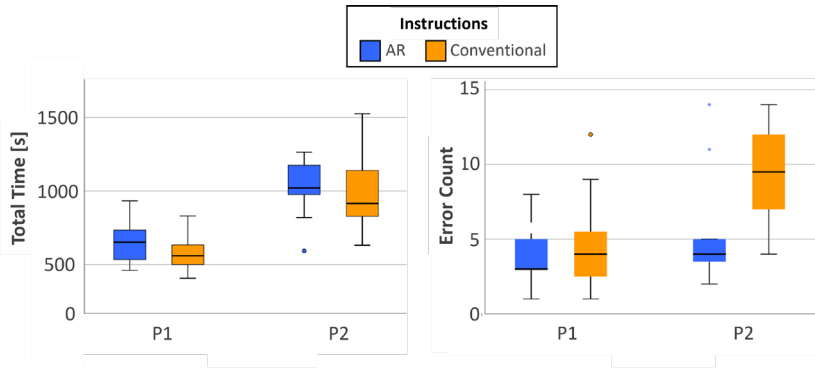


Fig. 16: Training times and error counts for procedure 1 (P1) and 2 (P2).

4.4.3 Detailed Error Analysis

Fig. 17 shows the full error protocol for P2. For the following comparisons, we rescaled errors associated with AR instructions to account for differences in participant numbers. Differences in errors occurred only for 1-point and 2-point errors. AR instructions resulted in 37% less 1-point errors, with 46 errors compared to 73 for conventional training. For 2-point errors, only one error was performed with AR, compared to 8 errors with conventional instructions. Both instruction types resulted in two 3-point errors.

Finally, we split the error counts according to their respective categories. Error counts for handling errors were similar for AR and conventional instructions, with a total of 21 each. The use of AR instructions resulted in a 66% decrease in error counts related to a lack of knowledge, reducing the error counts from 53 to 18. For those errors that could be related to both handling and knowledge, AR resulted in an error count of 15, compared to 21 for conventional instructions.

Error Description	Severity	Number of Errors for Each Participant																					Type			
		AR											Conventional													
		1	2	3	4	5	6	7	8	9	10	11	Σ	12	13	14	15	16	17	18	19	20		21	Σ	
Venous Puncture																										
Standing on the wrong side of the patient (correct: right)	2											0										1	1	K		
Identified artery instead of vein (vein: medial)	2											0												0	K	
Incision not done	3											0												0	K	
Incision done on wrong side of the patient	2											0										1	1	K		
Incision done wrong (not caudally or too far from spot)	1			1								1	1	1	1	1					1	1	5	K		
Puncture not done	3				1				1			2							1					1	H	
Puncture at the wrong spot	1	1										1										1	1	K		
Needle in wrong angle	1											0												0	K/H	
Puncture more than one attempt (# additional attempts)	1	2				2				2	1	7	1	2	1			3	1	1				9	H	
Needle pointing in wrong direction (caudally not cranially)	2											0				1	1							2	K	
Syringe not taken	2											0												0	K	
Syringe taken late	1											0												0	K	
Artery punctured instead of vein	2											0												0	K	
Distance not measured	2											0				1	1							2	K	
Distance not measured correctly	1					1						1						1						1	K/H	
Wrong wire taken	1											0												0	K	
Not holding on to measured wire	1			1		1		1	1			5	1					1					1	3	K/H	
Matching with cannula not done	1											0												0	K	
Venous Cannulation																										
Wire not inserted	3											0								1			1	1	K/H	
Wire not inserted to correct length	1							1		1		2	1		1	1							1	4	K/H	
More than one attempt to insert the wire	1		1			1			1			3			1						1			2	H	
Wire inserted but not through needle	2										1	1												0	K	
Needle not removed	2											0												0	K	
Wire pulled while needle removed	1											0												0	H	
Dilation not done	3											0												0	K	
Dilator inserted before needle removed	1							1				1												0	K	
Dilators in wrong order	2											0								1				1	K	
Dilation only partially done (not all dilators used)	1											0												0	K	
Dilator not completely inserted	1								1			1										3		5	K	
Dilator first inserted in the wrong way	1							1				1												1	0	K
Problems with wire handling	1	1										1									1	2	1	4	H	
Checking that wire is movable not done	1			2				3		3		8	2		1		3	3			3	3	15	8	K	
No pressure on groin when removing dilators	1							3		3		3				3	3		2					8	K	
Cannula insertion not done	3											0												0	K/H	
Cannula not measured	1							1				1				1			1					2	K	
Cannula not inserted to correct length	1		1				1	1		1	1	5				1	1			1				3	K/H	
Cannula inserted when dilator still in	1											0												0	K	
Insertor dislocated when inserting	1	1										1												1	0	H
Wire partially removed while inserting cannula	1							1				2												0	H	
Problems with wire handling	1											0				1								1	H	
More than one attempt to insert cannula	1											0												0	H	
No rotation while inserting	1											0												0	K/H	
Wire not removed	2											0												0	K	
Introducer not removed	2											0												0	K	
Introducer not completely removed	1									1		1												0	K/H	
Removal of wire and introducer not in right order	1											0			1									1	K/H	
More of cannula removed than only inserter	1						1					1				1								1	K/H	
Clamping not done	3											0												0	K/H	
Clamping not done immediately	1			1								1		1	1	1								4	K/H	
Problems with clamp handling	1				1	1				1		3												0	H	
Clamping at wrong spot	1											0			1							1		2	K	
Clamping while wire/introducer not removed	1										1	1					1							1	K	
A lot of blood (water) loss	2											0								1				1	H	
Total Error Counts: Σ [Severity × Errors]				5	2	4	5	4	5	2	14	3	11	4	59	4	9	6	14	14	10	12	8	11	7	95

Fig. 17: Full error protocol for P2 showing errors of each participant during AR-supported training (blue, n=11) and conventional training (brown, n=10). The right-most column for each training type shows the error total. Total error counts are calculated by multiplying the errors in each row with the error severity factor, ranging from 1 to 3. The last column shows the error categorization into handling errors (H), knowledge errors (K), or a combination of both (K/H).

4.4.4 User Experience

The point ratings for the six scales of the UEQ questionnaire are visualized in Fig. 18. It is evident that the AR version was evaluated positively on all six scales (points >0.8). The conventional version has a positive evaluation for Perspicuity and Dependability, a neutral one for Attractiveness, Efficiency and Stimulation (- 0.8 > points < 0.8) and a negative one for Novelty (points < -0.8). The best results for the AR version were obtained in the categories Attractiveness, Stimulation and Novelty. The AR version performed considerably better in five of the six categories and only shows similar results when it comes to the category Dependability.

Differences in scores between the AR and the conventional version are significant for the categories Attractiveness, Perspicuity, Efficiency, Stimulation and Novelty ($p < 0.05$, Mann-Whitney-U test). Clearly not significant is the difference in Dependability.

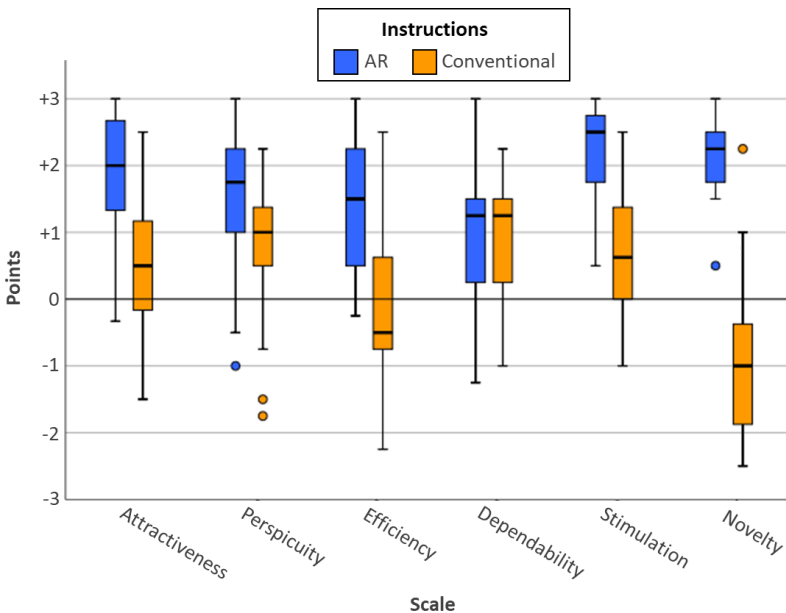


Fig. 18: Point means per UEQ scale.

4.5 Discussion

Previous experience in cannulation was not shown to have any significant effect on participants' performance. Although there was a significant difference in training times in P2 for participants with prior experience in AR, these participants were equally distributed between the group performing P2 with AR instructions and the one using conventional instructions. Therefore, we assume the influence of previous experiences to be negligible.

For the first procedure, no large differences in training times or error counts were found, suggesting that complexity was too low for differences in information presentation to have an impact on performance. Participants stated that they remembered relevant information from the initial video watched prior to the experiment and therefore often did not need to consult the instructions. For the more complex second procedure, instructions were consulted more frequently than in the first procedure but with important differences depending on the instruction type. Subjects using AR appeared to switch regularly between the displayed information and the point of execution, often evidenced by a brief pause in hand movement and sometimes by a brief but noticeable shift of the head. In contrast, subjects who used conventional instructions consulted the instructions less frequently but spent more time during each consultation. Differences in errors were mostly associated with knowledge errors, i.e., information that was missed during execution, which agrees well with our initial assumption about the benefits of having access to contextual information in close proximity.

Although we expected shorter training times for the procedures, this result is consistent with a previous study [65], which has attributed the differences to more engaging work with HMD. Similar to [70], we observed that videos were watched more frequently with the Hololens than on a computer. The current state of the application may have also had an effect on training times. As videos were rather short, at less than 15 seconds, we did not implement a functionality for jumping to a specific point in the video. To watch a part of the video in AR again, participants

had to replay it from the beginning, which was not the case for videos watched on the computer. Rather than watching the whole video in the AR instructions several times, participants only consulted long videos once or twice. As a result, error counts of the AR instructions were the highest for steps with the longest video sequences.

Deriving a detailed error protocol with error ranking and categorization has been shown to provide valuable insights on the effectiveness of training instructions. Error counts indicate that the significantly higher rankings in the UEQ were not only related to the novelty and excitement of using an OST-HMD but were also linked to the much more convenient presentation of information. The UEQ suggests that participants found AR instructions motivating and exciting to use (Stimulation) - both desirable characteristics for frequent training and long-term retention - and rated them highly in terms of clarity of information and ease of learning (Perspicuity) and information acquisition speed (Efficiency). Dependability was expected to be slightly higher for the conventional instructions, since the likelihood of encountering technical difficulties was higher for the novel AR technology. Both instructions were rated positively with a score of over 1, even if voice commands or hand gestures sometimes posed difficulties.

While the results in this paper indicate a high potential of AR instructions for surgical training, they are based on a study with only 21 medical students. Further experiments would strengthen the validity of these findings. As described earlier, participants could only replay the full video sequences when using the AR instructions and could not skip to the middle of a video when needed. Integrating this feature should further enhance the performance of AR instructions. To improve the realism of the training, currently missing steps like ultrasound verification or guidance should be integrated. Most importantly, this study only investigated the outcome of one training iteration and long-term training effects and the amount of information retained for future ECMO cannulations were not explored. A second iteration without any instructions could be of interest for future research, to investigate whether higher training times or lower error counts correlate with better

performance in the long run. However, with respect to the high differences in error counts in P2, we don't expect major changes in participants performances during a following assessment run, without feedback on previous performance. Future studies could therefore either perform a second training iteration with same instruction material and focus on how well performance improves during autodidactic training iterations. Alternatively, it could include a supervisor who provides feedback on participants' performance, and then relate performance to the time spent with the supervisor. We expect participants to require significantly less supervision before achieving error-free execution when using AR for training.

4.6 Conclusion

So far, there has been little research on how step-by-step instructions visualized by OST-HMD can be utilized to improve surgical training. In this paper, we demonstrated the potential of AR by taking the example of an ECMO cannulation. AR significantly reduced errors in the more complex second procedure and was clearly favored by the participants. Moreover, when comparing the variances in completion times and errors, AR instructions resulted in much more homogeneous performance levels. This is promising, as it helps to standardize training performance and makes training outcomes more predictable. We believe that these advantages of OST-HMD are generalizable to other surgical procedures. Further studies are needed to assess how AR training affects physicians' long-term knowledge and skill development. For this purpose, it would also be interesting to utilize the integrated eye-tracking capabilities of recent OST-HMD, which allow for a more fine-grained analysis of participants' behavior and their ongoing cognitive processes.

Acknowledgements: The AR application investigated in this study was developed during the focus project ARORA at ETH Zurich. We would like to thank all participants for their time and interest in our study.

4 Study I: Comparing AR against Conventional Instructions

Ethical approval: All studies have been approved and performed in accordance with ethical standards.

Informed consent: This article does not contain patient data. Informed consent was obtained from all participants included in the study.

5 Study II: Effective Visualizations Strategies

The content of chapter 5 has been submitted to the *International Journal of Computer Assisted Radiology and Surgery* and is currently under review. It was greatly supported by the master thesis of Dietmar Luchmann.

Abstract

Previous work has demonstrated the high accuracy of augmented reality (AR) head-mounted displays (HMD) for pedicle screw placement in spinal fusion surgery. An important question that remains unanswered is how pedicle screw trajectories should be visualized in AR to best assist the surgeon. We compared five AR visualizations displaying the drill trajectory via Microsoft HoloLens 2 with different configurations of abstraction level (abstract or anatomical), position (overlay or small offset), and dimensionality (2D or 3D) against standard navigation on an external screen. We tested these visualizations in a study with 4 expert surgeons and 10 novices (residents in orthopedic surgery) on lumbar spine models covered by Plasticine. We assessed trajectory deviations (°) from the preoperative plan, dwell times (%) on areas of interest (AOIs), and the user experience. Two AR visualizations resulted in significantly lower trajectory deviations (Mixed-Effects ANOVA, $p < 0.0001$ and $p < 0.05$) compared to standard navigation, whereas no significant differences were found between participant groups. The best user ratings for ease of use and cognitive load were obtained with an abstract visualization superimposed in the peripheral field around the entry point and a 3D anatomical visualization displayed with small offset from the entry point. While visual attention was predominantly guided to the visualizations, participants still spent 20% of their time examining the entry point area for visualizations displayed with a small offset. Our results show that navigation generally levels task performance, reveal

clear advantages and preferences of certain AR visualizations, and give insights into how AR visualizations guide visual attention.

5.1 Introduction

Spinal fusion surgery is indicated by severe spine disorders [74-76] and has seen an increase in performed surgeries of up to 200% in the last 30 years [77-82]. The procedure involves a superficial decortication of the entry point on dorsal side of the vertebral arch and insertion of a probe into the pedicle to make a channel for the screw. Verification of the trajectory and possible breaches are performed by intraoperative imaging and palpation. Subsequently the screws are inserted into the pre-drilled channels, and finally connected with metal rods and plates so that the corresponding vertebrae are firmly connected [83-85]. Due to the proximity of vital structures, strong intraoperative bleeding and variability in morphology between patients [86-88], spinal fusion surgery is a very demanding procedure, resulting in enormous health care costs [74, 89, 90].

Currently, free-hand surgery aided by 2D fluoroscopy is the most used approach in spinal fusion surgery, acting as the standard that more recent navigation systems are often compared with [91-93]. With technological advances and improved navigational methods being integrated into spinal fusion surgery [92], currently, two types (3D fluoroscopy, CT-guided navigation) of optical navigation systems are available. Both 3D fluoroscopy and CT-guided navigation systems were shown to be advantageous in accuracy and radiation exposure compared to free-hand execution or in combination with conventional visualizations [16, 86, 91-96].

As Härtl et al. have shown in 2013, only a minor part of surgeons utilize these new navigation technologies on a regular basis [97]. Factors stated are prolonged operating room (OR) times, a lack in ease of use and integration into the surgical workflow, and the high cost [97, 98]. Augmented reality (AR) head-mounted displays (HMDs) promise to offer a range of benefits compared to conventional navigation systems [23, 99],

such as increased anatomical understanding, execution speed and ease of use, and have seen growing interest over the last years [4]. By superimposing images into the field of view, the operator does not lose sight of the patient by gazing off at monitors [100]. Yoon et al. found no increase in operative time for AR navigation compared to freehand techniques [99]. Multiple studies found the accuracy of AR navigation to be comparable or better than conventional methods [101-103]. First studies on real patients have confirmed these results [101, 104].

While previous work on AR navigation has demonstrated the high accuracy of AR HMD for pedicle screw placement, an important question that remains unanswered is how pedicle screw trajectories should be visualized in AR to best assist the surgeon. Outside of AR, Brendle et al. [105] compared a hand-held navigation device with an integrated circular display showing different visualizations for pedicle screw trajectories against conventional navigation displayed on an external screen. They found a significant reduction ($p < 0.05$, Kruskal Wallis test) in cognitive load (NASA TLX survey) and a significantly better usability (SUS survey) when operating with the hand-held device. Using an AR HMD instead of a hand-held device, we are not limited to displaying information on the screen space of a display. Instead, we can anchor our AR interfaces in 3D space and, for example, display information in the peripheral area around the tool entry point. Moreover, augmented reality offers a variety of possibilities to display information, ranging from overlay to displaying information next to the patient, from 2D to 3D, and from abstract to anatomical representations. These configurations are expected to not only affect the surgical outcome, but also the user experience and the visual behavior, and are, thus, expected to greatly impact user acceptance.

In this paper, we compare five different augmented reality visualizations for pre-drilling pedicle screw trajectories with variations in abstraction level (abstract or anatomical), dimensionality (2D or 3D), and position (overlay or small offset) against conventional navigation on an external screen. We test these visualizations on a (L1-L5) lumbar spine model setup with 4 expert surgeons and 10 novices (residents in

orthopedic surgery). As it is difficult for a simulator setup to reproduce the same complexity that a real intervention has, we make two adjustments to our setup. First, we cover the spine model with a thin layer of red Plasticine to decrease visibility on the bone structure and thereby increase the need for navigation. Second, we instruct participants to pre-drill pedicle screw trajectories with less than 2° deviation from the preoperative plan to create a challenging and immersive task. We measure and evaluate the trajectory deviation between planned and realized trajectory ($^\circ$) as a metric for task performance, the dwell time (%) on areas of interest (AOIs) as a metric of visual attention, and the user experience with emphasis on ease of use and cognitive load. While our study is concerned with pedicle screw placement, we expect our findings to generalize to other AR-guided orthopedic interventions that are performed on partly occluded anatomy and that require a highly accurate execution.

Contrary to the outcome of free-hand execution, which has been shown to be significantly affected by the surgeons' experience ($p < 0.01$) [106, 107], we expect AR navigation to level the task performance between expert and novice groups. AR navigation should also result in lower trajectory deviations than standard navigation as the participants do not need to gaze off to a distant screen. Although visual attention should be primarily focused on the visualizations, we assume that surgeons will also need to look at the entry point area to coordinate the movement of the tool, which would indicate the benefits of having information close to or even superimposed on the entry point. Finally, we expect abstract visualizations to result in the lowest trajectory deviations and best user experience ratings due to the simplified information presentation.

5.2 Related Work

Several studies investigating AR navigation for pedicle screw placement have proposed visualizations for intraoperative guidance, such as overlaying CT slices [108] or 3D pedicle screw trajectories on the patient [109]. In our work, we are particularly interested in comparing different

configurations of AR visualizations and their impact on surgeons' behavior and preferences. The studies most related to our work are explained below.

Liebmann et al. [23] developed an AR navigation for pedicle screw placement using the stereo cameras of Microsoft HoloLens to detect and triangulate ArUco marker positions. Their AR interface superimposed the planned trajectory and the tool trajectory. The end of both trajectories were connected by a line and the numeric trajectory deviation was displayed on top of the ArUco marker. They evaluated their system on spine models and achieved a performance comparable to state-of-the-art navigation. We use the same combination of planned trajectory, tool trajectory and numeric trajectory deviation display for all our anatomical visualizations.

In a study similar to ours, Brendle et al. [105] compared a hand-held navigation device for pedicle screw placement with conventional navigation displayed on an external screen. Their hand-held device comprised of a drill sleeve with build-in circular display that showed the trajectory deviation in two different abstract visualizations. The first visualization, the 'circle display', shows several rings around the center of the display representing discrete trajectory deviations and the current tool trajectory as a point moving continuously across the underlying background. The target trajectory is achieved by moving the point into the central ring element. The 'grid display' divides the circular interface into 12 pie sections and four concentric circles of different radii. According to the relative orientation of the tool towards the planned trajectory, the respective grid field is highlighted in red, with the center field representing the target orientation. Both visualizations change the color of the center area if the target trajectory is achieved.

As part of our study, we also investigate two abstract visualizations, one with a continuous angle display like the 'circular display', and one using discrete ring segments like the 'grid display'. Contrary to Brendle et al. [105], we calculate and display the trajectory deviation projected into the sagittal and transverse plane instead of using tool-relative coordinates. Deviations along these planes can be adjusted independently

when holding the tool firmly and either leaning sideways (sagittal plane) or leaning forward/backward (transverse plane).

5.3 Materials and Methods

5.3.1 Apparatus

The AR navigation app was implemented for Microsoft HoloLens 2 using Unity 3D (2019.4.14f1) and the Mixed Reality Toolkit (MRTK 2.4.0). The standard navigation app was implemented as a desktop app using the same Unity 3D backend. The simulator setup (cf. Fig. 19) was based on a spine bed for a L1-L5 lumbar spine model (Synbone AG, Zizers, Switzerland) and had both a Vuforia image marker attached for initial registration with HoloLens 2 and a fixed infrared (IR) marker as a reference for tool tracking. The spine bed was reinforced with wooden pads so that no relative movement between spine and spine bed was possible. The spine was covered with red Plasticine to increase difficulty and thus the need for navigation. The tracking camera (Atracsys LLC, Puidoux, Switzerland; not visible in Fig. 19) was connected via cable to a desktop computer, running a server application that streamed incoming data points (i.e., transformation matrices of detected IR markers) to the client application (either to a Unity desktop app or to HoloLens 2). Three different tools were tracked by the external camera: a drill sleeve for navigation, a marker metal pin that can be inserted into the drilled pedicle channels to perform a post-op trajectory measurement, and a pointer marker for landmark registration of the spine model.

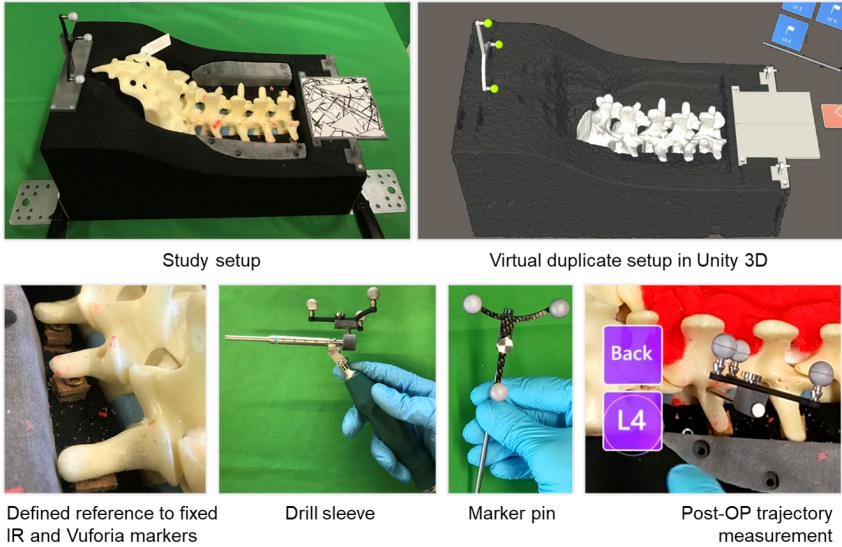


Fig. 19: Physical setup (left) with L1-L5 lumbar spine model in spine bed and virtual duplicate setup (right). The physical setup uses a fixed IR marker as a reference for tool tracking and a Vuforia image marker for initial registration with Microsoft HoloLens 2. The spine bed was reinforced with wooden pads to fully constrain relative movement between spine and spine bed. The setup uses a drill sleeve for navigation and a marker pin for post-op trajectory measurement.

5.3.2 System Calibration

The transformation between spine bed, fixed IR marker, and Vuforia image marker is known by design and remains constant. Prior to starting an experiment, we performed an 8-point landmark registration using the most distant points of the processus costalis to register the spine position to the spine bed. As the physical setup is attached to the underlying surface and remains static in 3D space, the Vuforia image marker only needs to be registered once before pre-drilling a side (5 pedicle screw trajectories) with HoloLens 2 to align the coordinate system of the virtual setup with the physical setup.

5.3.3 Outcome Parameters

We measured the absolute trajectory deviation (equation 1) between planned (\vec{A}) and executed trajectory (\vec{B}) and preoperative plan in radial degrees, dwell times on areas of interest in percent (cf. Section 3.7), and the user experience with emphasis on visualization ranking, ease of use, perceived accuracy, and cognitive load. Trajectory deviation measurements were conducted by inserting the marker metal pin into the drilled pedicle channels (cf. Fig. 19). This study was only concerned with trajectory alignment and did not investigate the accuracy of entry point placement.

$$TD = \cos^{-1} \left(\frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \|\vec{B}\|} \right) \quad (1)$$

5.3.4 Measurement System

Tracking System - An Atracsys fusionTrack250 (Atracsys LLC, Puidoux, Switzerland) with an accuracy of 0.09 mm (RMS) within 1.4 m distance and a measurement rate of 120 Hz was used. To determine the trajectory error between drill sleeve and marker pin, we fixated the drill sleeve in a vice, drilled ideal trajectories at low drilling speed, and subsequently measured the trajectory with the marker pin. We found a trajectory error between drill sleeve trajectory and marker pin trajectory of up to $\pm 0.17^\circ$.

Eye tracking - HoloLens 2 reports the wearer's eye-gaze with an angular accuracy of 1.5° around the actual target and a recording rate of 30 fps.

Questionnaire - The questionnaire consisted of an AR visualization ranking from 1 to 5, with the best visualization receiving 5 points and the worst 1 point, and point ratings for ease of use, perceived accuracy, and cognitive load ranging from 0 (not at all) to 10 (very much).

5.3.5 Visualizations

From the eight possible configurations of abstraction level, position, and dimensionality, five were considered most useful and implemented (cf. Table 2). Intuitively, we excluded a 2D anatomical overlay as we did not see benefits compared to a 3D overlay. We further excluded 3D abstract visualizations as trajectory deviations are only calculated along sagittal and transverse planes, which can be well displayed using two dimensions. Fig. 21 shows the implemented AR visualizations.

Table 2: Configurations of abstraction level (abstract or anatomical), position (overlay or small offset), and dimensionality (2D or 3D) used to derive the AR visualizations and the standard navigation.

Visualization Concept	Abstraction level	Position	Dimensionality	Device
'Standard Navigation'	Anatomical	Large offset	2D	External Screen
'3D Overlay'	Anatomical	Overlay	3D	HoloLens 2
'Virtual Twin'	Anatomical	Small offset	3D	HoloLens 2
'Sectional Views'	Anatomical	Small offset	2D	HoloLens 2
'Target Cross'	Abstract	Small offset	2D	HoloLens 2
'Peripheral Rings'	Abstract	Overlay	2D	HoloLens 2

The planned entry point position is highlighted by a green cross, while the current position of the drill sleeve tip, referred to as tool tip, is displayed as a purple cross. Participants can first position the tool tip on the entry point and then use the respective AR visualization to align the tool trajectory. All (AR and non-AR) visualizations are explained in the following.

Standard navigation - This non-AR navigation (cf. Fig. 20) displays sagittal and transverse slices of a segmented CT-model with planned (green) and current tool trajectories (purple) on an external screen and is used as the gold standard. It further displays a top view of the entire segmented 3D model and a cross-hair pointer for fine-positioning the tool tip at the entry point position.



Fig. 20: Standard navigation with user interface to connect to the external tracking camera, start recordings and select a vertebrae (left), a top view on the segmented spine model (bottom right), a magnifying view for fine-positioning of the tool tip on the entry point (red area in the middle), and 2D slices in sagittal and transversal direction (top right).

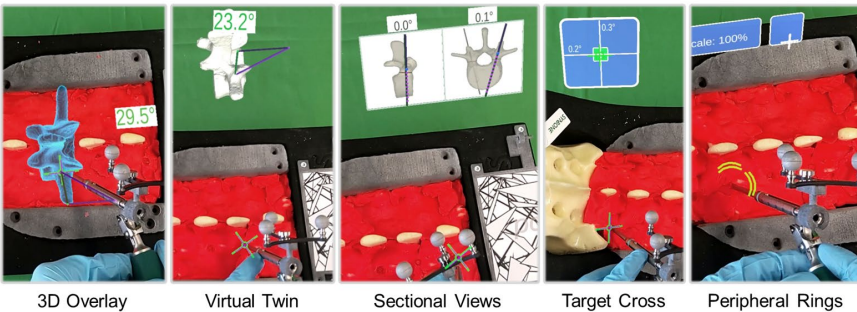


Fig. 21: All five AR visualizations shown from the participants' point of view.

3D overlay - This AR interface superimposes the whole virtual vertebrae on top of the real one. A green line represents the ideal trajectory from planning and a purple line represents the current tool trajectory. The absolute numeric deviation is displayed above the tool marker position.

Virtual twin - This AR interface displays the same information as the 3D overlay, yet next to the spine, and does, thus, not occlude the entry point

and the tool tip. As a consequence, the surgeon is not disturbed while operating but can countercheck the execution against the virtual model.

Sectional views - This AR interface aims to provide a more familiar navigational display that shows the same two sectional views used in the standard navigation, but closer to the entry point. By splitting the absolute angle along two axis, participants can optimize trajectory deviations independently for both directions.

Target cross - This AR interface aims at displaying directional trajectory deviations in a simpler way than the two sectional views by visualizing the trajectory as a red target cross moving continuously on a blue plane. The horizontal axis represents the deviations along the sagittal plane and the vertical axis represents the deviations along the transverse plane. The center area of the blue plane and the red target cross turn green while navigating within 2° deviation.

Peripheral rings - This AR interface aims at displaying the same abstract information as the target cross in the peripheral field of the entry point, thus allowing the participant to focus on the entry point without occluding it. Each ring segment represents an angular deviation of 1.41° in the respective direction with a maximum of four segments. If no ring segments are visible, the trajectory lies within the limits of $< 2^\circ$ deviation.

5.3.6 Study Design

The study was divided into an initial testing phase and a main study. During the initial testing, participants were introduced to the Microsoft HoloLens 2 and could test all visualizations without drilling. Within the main study, participants pre-drilled 35 pedicle screw trajectories (equals 3.5 L1-L5 lumbar spine models), of which 10 trajectories were drilled using standard navigation and 5 with each of the five AR visualizations. We derived a study protocol with configurations of visualization types and operation side so that all visualizations were performed an equal

number of times on the right and left side by each group. Participants were randomly assigned to these configurations. This study did not require the approval of the ethics committee.

5.3.7 Participants

14 participants were recruited and divided into an expert and a novice group. The expert group consisted of four expert surgeons (aged 35-42 years) specialized in spinal surgery with several years of work experience. The novice group was composed of ten residents in orthopedic surgery (aged 25-36 years). All experts and half of the novices stated previous experience with AR/VR, mainly with HoloLens 1.

5.3.8 Experimental Procedure

Fig. 22 shows the experimental setup used during the main study. The first spine model (both left and right side) was always navigated using standard navigation.



Fig. 22: Experimental setup showing the participant navigating with HoloLens 2, the spine bed, tracking camera and external screen.

Prior to starting with the second spine model, participants received a reintroduction to HoloLens 2 and were guided through the calibration procedure of the eye tracking system, which is an automated routine provided by HoloLens 2. Participants then calibrated HoloLens 2 to the experimental setup by confirming the position of the Vuforia image marker. After pre-drilling trajectories with each visualization, participants filled out a user experience questionnaire for the respective visualization. We used this time window to perform the post-op trajectory deviation measurement for each pedicle screw channel. Depending on the operating side specified in the study protocol, we then turned the setup by 180° and/or replaced the spine model, and performed a landmark registration to register the spine position to the spine bed. After the completion of the experiment, a general questionnaire was handed to the participants and complemented by an interview.

5.3.9 Eye Tracking Analysis

For quantification of visual attention, we divided the stimuli into five Areas of Interest (AOIs) (cf. Fig. 23). These AOIs comprise of the 'entry point' area, all visualizations displayed with offset to the entry point, and the 'background', which represents the remaining space. No separate AOIs were defined for the 3D overlay and the peripheral rings as these visualizations are displayed directly on top of the entry point. The time spent examining each AOI was summed up and divided by the total time spent on the task. This resulted in the relative dwell time for each AOI for the respective visualization, expressed in percent.

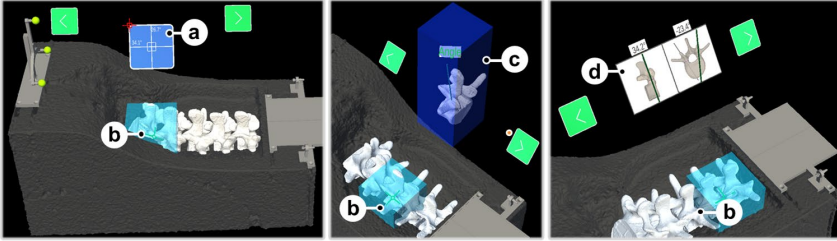


Fig. 23: Areas of interest used for eye tracking analysis, consisting of 'target cross' (a), 'entry point' (b), 'virtual twin' (c), 'sectional views' (d), and 'background' (all remaining space). The AOI 'entry point' is always placed above the currently operated vertebrae. The AOI boxes for (b) and (c) are not visible during navigation.

5.3.10 Statistical Analysis

Preliminary analysis revealed that trajectory deviations and dwell times were normally distributed, while questionnaire responses were undistributed. All statistical tests were conducted using the R environment. We modeled trajectory deviations in a mixed-effects ANOVA that included visualization, skill level, and operator side as fixed effects, and an operator random effect. A post-hoc analysis was performed with Benjamini-Hochberg adjusted pairwise t-tests. Each pedicle screw trajectory was considered as one measurement. Questionnaire responses were analyzed with a Friedman test and a post-hoc analysis with pairwise comparisons using the Wilcoxon signed rank test. Participant groups were pooled for statistical analysis of questionnaire responses to account for the small sample sizes. In a last step, we tested for correlation of trajectory deviation with either ease of use, perceived accuracy, or cognitive load using Spearman rank correlation.

5.4 Results

5.4.1 Trajectory Deviation

From the 490 pedicle screw trajectories drilled, 10 measurements (2%) were excluded due to problems in data recording with Hololens 2 and 20 measurements (4%) were excluded as the marker pin could not be fully inserted into the opened pedicle channel. Fig. 24 shows the post-op trajectory measurements for standard navigation and all AR visualizations separately for expert and novice groups. Mean trajectory deviations for both groups span from approximately 1° to 2° deviation. 73% of all measurements lied within the target deviation of 2° and 93% lied within 3° trajectory deviation. The mixed-effects ANOVA with Benjamini-Hochberg adjusted pairwise contrasts showed significant differences when comparing 'target cross' with 'standard navigation' and 'peripheral rings' (both $p < 0.0001$), 'overlay' ($p < 0.001$), 'virtual twin' and 'sectional views' (both $p < 0.05$). We also found significant differences when comparing 'virtual twin' against 'standard navigation' and 'peripheral rings' (both $p < 0.05$). We did not observe a significant effect of skill level or operating side on the trajectory deviation ($p > 0.05$).

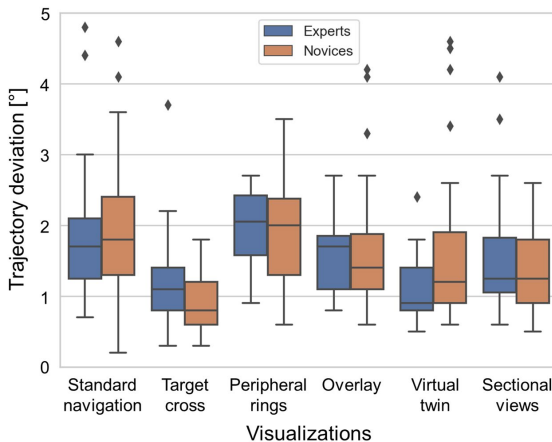


Fig. 24: Trajectory deviation measurements over all visualizations for expert (blue) and novice (brown) groups.

5.4.2 Visual Attention

Table 3 shows the relative dwell times on the AOI 'entry point' and all AR visualizations displayed with small offset while trajectory deviations were below 3°. Participants spent approximately 20% of their time examining the entry point area and the rest of their time on the respective visualization. Up to 10% of visual attention was usually registered on the AOI 'background'.

Table 3 Relative dwell times (means and stds) on AOI 'entry point' and the respective visualization for all time steps with trajectory deviation below 3°.

Visualization	AOI 'Entry point'		AOI 'Visualization'	
	Exp	Nov	Exp	Nov
Target cross	19.5% ± 14.4%	19.2% ± 15.0%	78.3% ± 15.3%	74.9% ± 18.5%
Virtual twin	22.0% ± 16.6%	18.3% ± 14.3%	75.6% ± 17.3%	77.2% ± 16.6%
Sectional views	16.2% ± 10.7%	19.5% ± 16.1%	75.8% ± 15.9%	77.5% ± 17.7%

5.4.3 User Experience

Table 4 shows the summary statistics of the questionnaire results. While experts' preferences of visualizations varied greatly, resulting in overall similar point ratings between 2.8 and 3.3, novices' ratings were more determined, ranging from 3.8 for target cross to 2.2 for the sectional views. The Friedman test using joined participant groups showed significant differences between visualizations for ease of use and cognitive load ($p < 0.01$). Post-hoc analysis using pairwise Wilcoxon rank tests without p-value adjustment method showed significant differences in ease of use when comparing 'peripheral rings' against 'standard navigation', 'sectional views', and 'overlay' (all $p < 0.01$), and smaller differences when comparing 'target cross' and 'virtual twin' against 'overlay' ($p < 0.05$). For cognitive load, we found statistical differences when comparing 'peripheral rings' and 'virtual twin' against 'overlay' (both $p < 0.01$) and against 'standard navigation' (both $p < 0.05$). No significant differences were found when using the Benjamini-Hochberg

procedure for p-value adjustment. Finally, we found significant Spearman rank correlations between trajectory deviation and both perceived accuracy ($p < 0.05$, $\rho = -0.26$) and cognitive load ($p < 0.05$, $\rho = 0.23$).

Table 4: Questionnaire summary results: visualization point ranking [5=favorite visualization, 1=least liked] averaged over participants, means and stds for ease of use ([0,10] higher is better) and cognitive load (NASA-TLX, [0, 100] lower is better). The maximum values of each column are printed in bold.

Visualization	Ranking		Ease of Use		Cognitive Load	
	Exp	Nov	Exp	Nov	Exp	Nov
Standard navigation			8.0 ± 0.7	6.6 ± 1.5	25.8 ± 10.6	33.8 ± 8.9
Target cross	3.1	3.8	8.0 ± 1.6	7.4 ± 2.1	26.7 ± 11.3	28.7 ± 11.3
Peripheral rings	3.3	3.6	8.3 ± 1.1	8.8 ± 1.2	22.1 ± 2.2	22.5 ± 6.8
Overlay	2.8	2.4	6.8 ± 2.0	6.1 ± 2.0	32.5 ± 15.7	38.5 ± 17.5
Virtual twin	3.1	3.2	9.0 ± 0.7	7.5 ± 2.0	20.8 ± 6.7	23.2 ± 7.5
Sectional views	2.8	2.2	7.5 ± 1.5	6.7 ± 2.4	26.3 ± 9.2	32.7 ± 13.3

5.5 Discussion

The goal of this study was to better understand the effects of different AR visualizations for pedicle screw pre-drilling on task performance, visual attention and user experience both for expert surgeons and novice operators. Despite the fact that participants were not trained on the visualizations before the experiment and had previously only explored them, the majority of pedicle screw trajectories was successfully placed within the specified 2° trajectory deviation. The high similarity of outcomes for expert and novice groups indicates that navigation generally levels task performance.

Although we expected that abstract visualizations would receive the overall best ratings in terms of ease of use and cognitive load, only the peripheral rings were rated significantly better than standard navigation, whereas no such differences were found for the target cross. Participants appreciated that using the peripheral rings they could stay focused on the area of execution. As expected for abstract visualizations, the target cross achieved the best performance in terms of trajectory deviation. An

unexpected outcome was the significant difference in trajectory deviation between the target cross and the peripheral rings. Although both of these abstract visualizations show the same target angle of 2° , the target cross displays the angle continuously, whereas for the peripheral rings there is a blind spot within 2° range in which the participants do not receive feedback. We believe that by increasing the resolution of the peripheral rings to represent smaller angle increments (e.g., 0.5° increments instead of 1.41°), similarly low trajectory deviations are possible.

For anatomical visualizations, we expected the two sectional views to result in the lowest trajectory deviations as deviations can be optimized separately for both directions. Interestingly, participants executed the task more accurately using the virtual twin while both ease of use and cognitive load were much better rated. It seems that multiple anatomical views also increase the difficulty to incorporate this information during task execution.

The anatomical overlay was perceived as overall most distracting and resulted in higher trajectory deviations and worse user experience than the virtual twin. Participants found it irritating that they could not see the tool tip and entry point very well. This reduced visibility extends beyond the boundaries of the virtual object (likely due to the high brightness of the superimposed object) and only gradually decreases with lateral distance. In contrast, participants found the virtual twin particularly helpful in gaining an understanding of the anatomy and to locate the entry point. Participants further stated that it felt very intuitive that whenever the tool was in contact with the bone, they would also see this contact visually on the virtual twin.

In summary, our results support previous studies that have shown AR navigation to be comparable or better than standard navigation [101-103]. Our findings further support two key findings of a study conducted by Brendle et al. [105] comparing a hand-held navigation device with an external screen. First, we also found abstract visualizations to result in significantly lower cognitive load and higher usability ratings compared to standard navigation. Second, when comparing the same visualization either displayed in close proximity (sectional views) or on an external

screen (standard navigation) no significant differences were found in cognitive load and ease of use. We therefore conclude that cognitive load and ease of use are mostly affected by the design of a visualization and not its distance.

The trajectory deviations, however, were shown to be affected by the distance of the visualization. While navigating the tool, the eyes must provide the necessary information from the visualization to steer the hand movement. At the same time, the approximately 20% dwell time on the entry point area for visualizations displayed with offset indicate that it is not sufficient to only examine the visualization. Instead, participants also examine the entry point area to coordinate the hand movement, which, in the case of the external screen, requires the eye-gaze to travel much longer distances between screen and tool position. The dwell times further suggest that visualizations generally guide visual attention to where they are displayed. For visualizations displayed at a distance, this results in visual attention being pulled away from the patient, whereas with the peripheral rings the visual attention is actively guided to the area of execution.

For future studies, it would be interesting to investigate how a combination of the most promising anatomical and abstract visualizations, i.e., the virtual twin and the peripheral rings, affect the user experience and preferences. This was also suggested by several participants during the interviews. In addition, our study only examined AR visualizations for accurate alignment of the tool with the intended trajectory. Determining the position of the entry point is an important and challenging step that is also likely to benefit from different AR visualizations. The virtual twin could be particularly interesting, as participants already reported it as useful for understanding the anatomy and for locating the entry point.

Our study setup has several limitations. First, our results were generated in a simulator setup, so further studies are needed to verify these findings in real interventions. While the 490 pre-drilled pedicle screw trajectories were sufficient for statistical analysis of trajectory deviations, a higher number of participants would increase statistical power of

questionnaire results, especially as questionnaires are a subjective assessment metric. Furthermore, there may have been a learning curve in working with the physical setup that negatively affected standard navigation results. Allowing participants to operate two sides for standard navigation was aimed at compensating this effect. Despite the fact that the drill sleeve was jagged at the tip, participants sometimes slipped and had to reposition the tool at the entry point. While we do not expect an effect on the trajectory deviations, roughening the model around the entry point area would improve the tool handling for participants. Providing an additional tool for entry point preparation was expected to be unfeasible when used on 35 entry points, given the already complex study setup.

5.6 Conclusion

Using accurate and easy-to-use visualizations for navigation is important to effectively assist surgeons during a procedure. With our work, we contribute a study that investigates the advantages and disadvantages of different AR visualizations by jointly analyzing task performance, visual attention, and user experience. Taking the example of pedicle drilling, the design of AR visualizations has been shown to have a big impact on trajectory deviations, ease of use, and cognitive load. It is therefore important for future AR systems to consider different designs in the development process. When designing anatomical visualizations, it is not necessary to overlay anatomical information on the real anatomy to get a good spatial understanding during navigation. Our results suggest that it is actually easier to combine haptic and visual feedback when anatomical information is displayed with some offset, as the visibility on tool and entry point area are of high importance. In contrast to anatomical visualizations, the main advantage of abstract visualizations lies in the design freedom, such as the possibility to freely adjust the resolution or to fade in the information in the peripheral area of the tool entry point. The latter is particularly promising because visual attention is guided in such a way that information acquisition and coordination of hand movement can take place simultaneously, allowing surgeons to navigate without the need to take their eyes off the patient.

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

Compliance with ethical standards

Conflict of interest: The authors JW, DL, QL, PF, and MM declare no conflict of interest. MF is shareholder of a company developing mixed-reality application for orthopedics.

Ethical approval: All studies have been approved and performed in accordance with ethical standards.

Informed consent: This article does not contain patient data. Informed consent was obtained from all participants included in the study.

6 Study III: Predicting Future Hand Actions

The content of this chapter has been published in *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* [54]. (© 2021 IEEE)

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.

6.1 Introduction

Augmented reality head-mounted displays (AR HMDs) [110] are promising for industrial and clinical applications, providing operators with the information needed to perform manual tasks such as assembly[3], maintenance [2], or surgery [4]. Studies have shown that

displaying contextual information can improve spatial understanding [10, 11] and reduce both time expenditures and the probability of errors [12, 13]. 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 [23, 24]. Results showed that the relevant information can be adapted to provide the right instructions at the right time and place [111] or that real-time feedback on user actions can be provided [20]. 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 [36-38]. As gaze behavior is highly task-dependent [39], it provides deep insights into ongoing cognitive processes [40]. Hand tracking can be used to infer hand actions [41], which provide insights into the user's performance of manual tasks [24, 38].

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 [42]. Hand-eye coordination has been successfully tracked to automatically detect usability problems in eye tracking video recordings [112] or to predict user's target selection while reaching to a virtual object in a Virtual Reality (VR) space [113]. During hand-eye coordination, the eyes provide the necessary information to plan the motor system's movements [114, 115], 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 [113], 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 [114].

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 [113], 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.

6.2 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.

6.2.1 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. [44] 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. [45] 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. [116], 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. [41], 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’ [45, 117]. 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.

6.2.2 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 [118, 119] or web search [120]. 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. [121] 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 [122] 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.

6.2.3 Predicting Target Selection

In user interfaces, target selection has a rich history in desktop environments. For example, Baudisch et al. [123] predicted possible targets during a drag-and-drop task on a large screen by analyzing cursor trajectories. Koochaki et al. [124] 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. [125] 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. [113] 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.

6.2.4 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. Katic' et al. [21] 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. [43] 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 [22] 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 [20] 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. [24] 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.

6.3 Study Part 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.

6.3.1 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. 26. 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 [50]. 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.

6.3.2 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. 25). Players wearing the HoloLens 2 were instructed to only play with one hand.

6.3.3 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. 25). After confirmation of both marker positions via touch gestures, the virtual grid (cf. Fig. 26) 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.

6.3.4 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. 25). 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.

6.3.5 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.

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

$$d_{perp} = \frac{\|(\overrightarrow{H_{Pos}C_{Pos}}) \times \vec{v}\|}{\|\vec{v}\|} \quad \begin{array}{l} H_{Pos} = 3D \text{ HandPosition}; \\ C_{Pos} = 3D \text{ CardPosition}; \end{array} \quad (1)$$

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. 26). 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.

6.4 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.

6.4.1 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. Fig. 27 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.

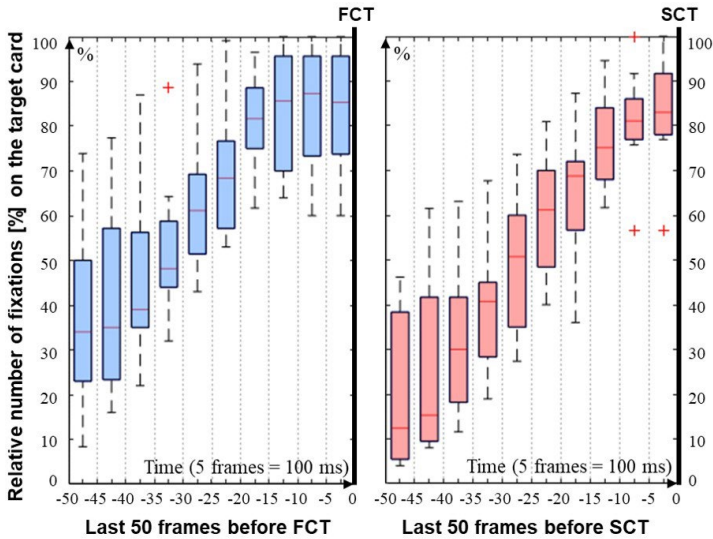


Fig. 27: Last-second gaze behavior (50 fps) on the selected card before the first card turn (left) and second card turn (right) across all participants, divided into 100 ms time windows. Each value of a box plot represents the number of fixations on target cards for one participant, divided by all FCTs or SCTs played by the participant.

6.4.2 Analysis of Hand Movements

Hand Velocity

Fig. 28 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.

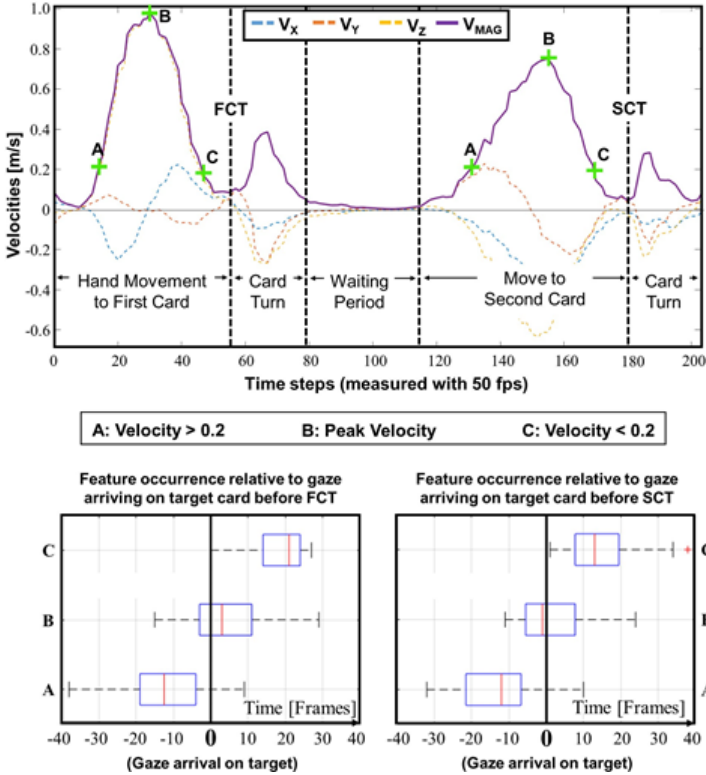


Fig. 28: The top diagram shows the velocity components of an example hand sequence of one move (two card turns) and three hand velocity features that represent the start (A), peak velocity (B) and end (C) of a card reach. V_x represents the velocity in the lateral direction, V_z represents the longitudinal 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.

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. 28, 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 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.

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.

6.4.3 Fingertip Proximity

Fig. 29 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.

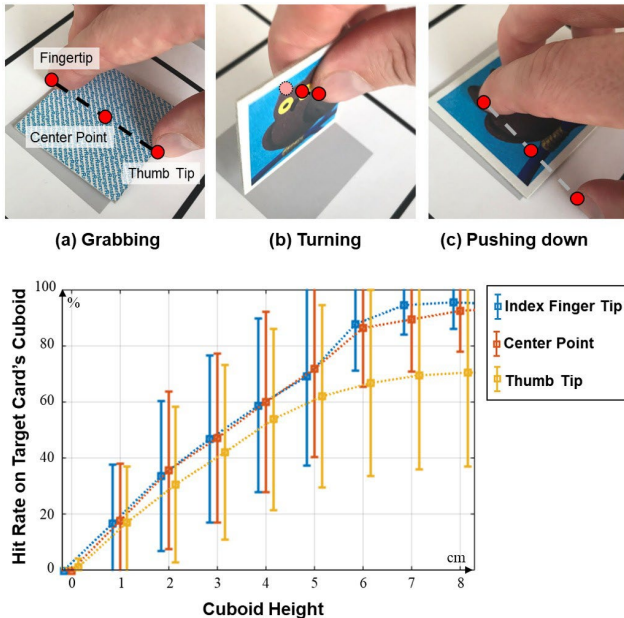


Fig. 29: At the top, three characteristic scenes of a card turn are shown (a-c), with the two involved hand joints index fingertip and thumb tip as well as their center. At the bottom, over all first card turns the hit rate on the target card's cuboid is shown for different cuboid heights for the index fingertip, thumb tip and center. Error bars show the hit rates' standard deviation over all card turns.

6.5 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.

6.6 Implementation

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. Fig. 30 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.

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
  if next card is second card then:
    velocity, trajectory  $\leftarrow$  get current hand data (4CF)
    gaze target  $\leftarrow$  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
```

Fig. 30: Pseudo code for the implemented closed-loop user support.

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. 29) 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. 31 (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. 28). 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. 31 (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. 31 (c-d)). At the beginning of our tests, we used a second Boolean 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. 28). 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. 28). 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. 25), which was covered by an invisible cuboid (cf. Fig. 26). A touch with the cuboid resulted in the cuboid lighting up to confirm the reset.

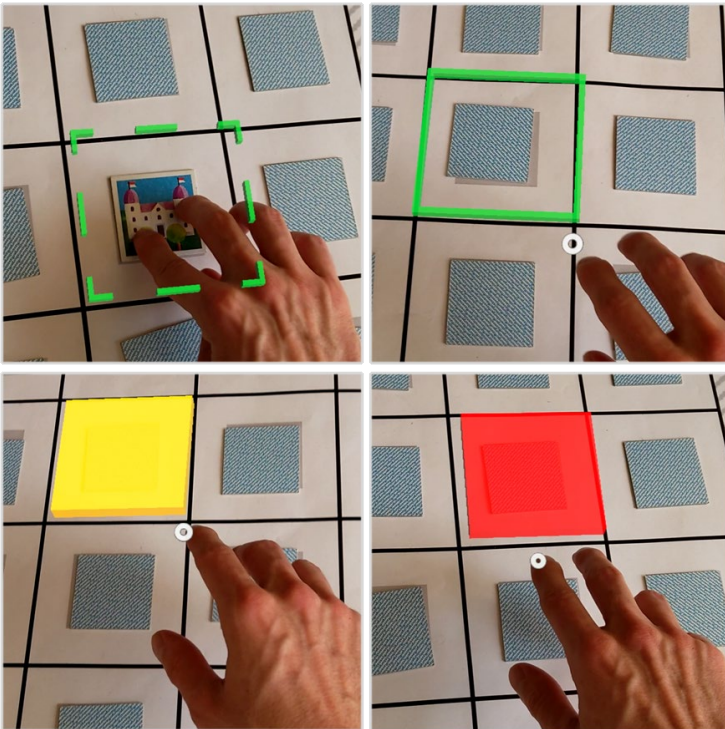


Fig. 31: Confirmation of detected first card turn (a) and visual alerts for the second card prediction in case of correct target selection (b), a wrong target that is located directly next to the correct card (c), and a wrong target that is not located next to the correct card (d).

Study Part 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.

6.6.1 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.

6.6.2 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.

6.6.3 Procedure

Participants were introduced to how the system worked and learned about the four visual aids (cf. Fig. 31) 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.

6.6.4 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.

6.7 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. Fig. 32 shows the mean performance across all participants and the breakdown of correct and incorrect warnings into subcategories.

6 Study III: Predicting Future Hand Actions

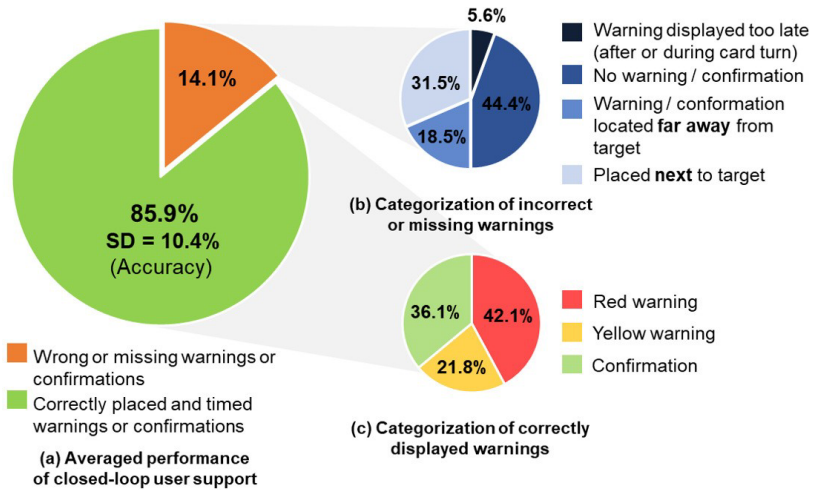


Fig. 32: Warning system performance averaged over all participants (a) with categorization of incorrectly displayed warnings (b) and correctly displayed warnings (c).

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 occurred 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.

6.8 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 [126]. 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 [44, 45, 116], also referred to as a ‘directing’ pattern [35]. Our studies support these findings while demonstrating how hand and gaze features can be combined for target prediction. According to Crawford et al. [114], 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. 26), 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 [22, 24] 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 [41, 127]. 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.

6.9 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. 28), 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. [113] 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.

6.10 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.

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

7 Conclusion and Outlook

AR and context-aware AR support show great potential in improving procedural outcomes. This work aims at providing a better understanding of how these technologies affect task performance, and to investigate the benefits that eye tracking can bring to context-aware AR support. A system model was proposed to explain the study characteristics and underlying system relationships in this work, which will also be discussed in regard to its suitability to model context-aware systems. This chapter first concludes the findings of the main studies and the specified goals and then outlines potential research directions for future work.

7.1 Conclusion

Three studies were conducted to address the research questions in this thesis. The first study was aimed at better understanding the advantages of contextual information displayed in AR over traditional information mediums. The second and third study investigated task performance of context-aware support with two different feedback types and eye tracking analysis both to understand visual behavior and for real-time support. The following paragraphs conclude the study results presented in chapter 4-6 and then lead to a final conclusion.

RQ 1: How do contextual information in AR affect execution errors?

The first study examined the benefits of contextual information in AR over traditional information mediums taking the example of ECMO cannulation. The evaluation of the detailed error protocol showed a 66% reduction in knowledge-related errors in the more complex second procedure. This is consistent with our expectation that contextual information in AR can be better incorporated during task execution, reducing procedural errors such as missed steps and partially completed steps. Despite the improved outcomes, a considerable amount of information was still missed or misinterpreted during execution, and

handling errors were not impacted, being even slightly higher with AR than with the conventional instructions. These results suggest that for complex procedures, a one-directed feed of information from AR to the operator is not sufficient for an error-free execution. Instead, the operator must be made aware of how his execution differs from the intended workflow, which requires a feedback loop. This feedback can either be provided by a supervisor or by integrating context-awareness, which was the scope of the following studies.

RQ 2: What are effective visualization strategies for continuous performance feedback?

The second study investigated visualizations for continuous performance feedback with different configurations of abstraction level, dimensionality, and position, taking the example of pedicle screw placement in spinal fusion surgery. It showed that the design of a visualization has a significant effect on ease of use, cognitive load, and task performance. It further showed that continuous performance feedback levels task performance, independently of the visualization. Both abstract and geometric visualizations resulted in very high execution accuracy and user experience ratings, although significant differences were found between designs. Abstract visualizations are particularly interesting as they allow to freely adjust resolution of the displayed target value, whereas geometrical visualizations are bound to the dimensions of the physical 3D space.

As displaying continuous performance feedback causes virtual information to change continuously, the eyes must perceive this information while coordinating hand movement. Eye tracking showed the distribution of visual attention between these conflicting goals, suggesting that virtual information should be displayed as close as possible to the area of execution to reduce the distance the eye-gaze must travel between information intake and hand coordination. While information proximity is beneficial, it is of great importance to not overlay virtual information on the area of execution or other visual stimuli that the operator must perceive during task execution. Here, the peripheral

field of vision has been shown to be particularly interesting as visual attention is guided in a way that information acquisition and coordination of hand movement can take place simultaneously without occluding relevant parts of the stimuli.

RQ 3: How suitable is a joint analysis of hand motions and eye movement for predicting and preventing erroneous hand actions?

The third study investigated the suitability of a real-time analysis of eye tracking and hand tracking data for predicting future hand actions, taking the example of a memory game. A novel method was presented that utilizes hand-eye coordination to predict hand actions during target selection. End-to-end testing of the method showed it to be highly effective in placing visual alerts over target locations and stop hand actions in a timely manner. The results suggest that the combination of eye-gaze target, hand velocity, and direction of hand motion are suitable indicators for future hand actions during target selection. More work is needed to test the transferability of the findings to other, more complex real-world applications.

While previous work had primarily focused on providing feedback on past or current execution, this work contributes a method that provides feedback at an earlier stage of a task, offering fundamentally new opportunities for user support.

Benefits of eye tracking for context-aware AR support systems

The second and third study investigated eye tracking both to understand visual attention and for predictive AR support. Eye tracking provided deep insights into how visual behavior is guided when continuous performance feedback is provided and helped to explain why information proximity to the area of execution is important. In contrast to our study setup where only two conflicting goals for operators' perception were observed, other setups might be characterized by multiple dynamic visual cues in the virtual and physical world that conflict with each other and the coordination of hand movement. Eye tracking holds promise for understanding these conflicting goals and optimizing where and when

information is displayed. In addition, the combination of eye tracking and hand tracking has been shown to be highly effective in providing predictive AR support, a novel type of support and an area of research that is currently largely unexplored, which offers great potential for improving procedural outcomes and enhancing safety. As mentioned in the introduction, previous work on context-aware systems has primarily focused on manual execution. Our work demonstrates why perception is an important aspect of human interaction and strongly argues for integrating eye tracking into future context-aware systems.

Suitability of the proposed system model

This work has proposed a system model to describe the relationships between human behavior, augmented reality, and context-aware support. The model proved effective in explaining the three studies presented in this thesis and their unique specifications. It also helped explain the conflicting goals during the perceptual tasks and why eye tracking is a crucial component for understanding the interaction between human behavior and contextual AR support. As seen in chapter 3.1 – 3.3, context-aware systems vary primarily in what they measure, what domain knowledge they use to determine the context, and what kind of contextual support they provide. The model can serve future work as a foundation for designing and explaining scientific studies at the interface of human behavior and context-aware support. A current shortcoming of the model is the ability to describe temporal relationships in the course of a procedure (e.g., temporal relation of eye and hand features). Here, it will be interesting to either add a third dimension to the proposed model or derive a procedural/temporal model optimized for this specific goal.

Final conclusion

Assistance systems are already part of our everyday lives, whether it is navigation using Google Maps or the lane departure warning system in cars. These systems also measure human behavior in a context-aware manner and provide real-time assistance tailored to the user's current task. This work has demonstrated the great potential of context-aware support

in AR-guided procedural tasks and the benefits of integrating eye tracking into context-aware systems. Although research is still at an early stage, the results suggest that context-aware AR support will make task execution safer in the future, reducing risks and potential harm to both humans and machines.

In the last decades, a great amount of time and effort has been invested in the development of robotic systems that can take over tasks of human operators. Context-aware AR support systems promise to improve operators' performance and strengthen their position in future work environments, and will enable better collaboration between humans and technical systems.

7.2 Outlook

This work has demonstrated the potential of AR and context-aware AR support to improve task performance in procedural tasks. From initial training to independent execution, procedural tasks differ only in how much support and knowledge is required by the operator. For future context-aware AR systems, we envision a fully integrated system that guides the operator from initial training to independent execution, gradually adjusting contextual information and feedback as the participant's skill level increases. This requires (1-2) gradually integrating predictive AR support into real-world applications to provide a safety net for operators performing their first independent task executions, (3) understanding effective visualization strategies in task environments with multiple dynamic visual stimuli (e.g., with robotic systems), and (4) investigating feedback and support strategies that are tailored to the operator's skill level.

Apply predictive AR support to real-world applications.

With our work on predicting target selection, we have demonstrated the suitability of gaze and hand tracking for predicting and preventing erroneous hand actions in a memory game. Further investigations are needed to refine the method to be applicable in real-world applications.

While both the velocity threshold and the gaze target should be applicable to 3D settings, the trajectory is currently only computed within the 2D plane and would also need to include the vertical axis. As our game setup was well observable from different viewpoints, interaction in real-world applications might encounter more occlusions and larger variations in object sizes and distances, which might require case-specific extensions to the method. With these adjustments, the prediction method is expected to already work for spacious systems such as industrial machines, where the operator reaches over several centimeters distance.

Surgical applications are more complex and characterized by smaller dimensions of the area of execution, which might require an AR HMD with more accurate eye tracking sensors. In addition, the method currently predicts discrete target locations of AOIs. Although AOI's for target prediction can also be defined in the virtual model superimposed on the patient (e.g., entry point area, vertebrae), predicting the tool position continuously and integrating this information into a statistical process and risk model [21] is expected to be more relevant for surgical applications. Future work should investigate whether regression-based machine learning models can also learn a mapping function from eye-gaze and tool movement to future tool positions.

Learn more advanced patterns in hand-eye coordination.

This work has only investigated a single hand-eye coordination pattern that occurs while the operator reaches for a target location, also referred to as a 'directing' pattern. According to Land and Hayhoe [35], other patterns in hand-eye coordination exist that might be used for predicting future hand actions. Moreover, this work derived a prediction method based on simple heuristics of hand motions and eye movement while treating the hand as a rigid object. To investigate more complex patterns in hand-eye coordination, it is crucial to have a precise computational understanding of the hand while interacting with physical objects including the finger movement. State-of-the-art hand tracking sensors (e.g., HoloLens 2) predict 3D hand poses with all hand and finger joints, but they are not robust during hand-object interactions, and finger

movements are often arbitrary when fingers are occluded. Predicting 3D hand poses and hand-object poses from ego-centric cameras is an active field of research in computer vision, and several models have been proposed over the last years [41, 127] that are now finding their way into procedural tasks. For example, Doughty and Ghugre [128] proposed a system that efficiently streams camera footage from HoloLens 2 to a workstation, where images are passed through an EgoPose network that jointly predicts 6-degree-of-freedom (6DoF) hand and object poses and sends this information back to the HMD. They used this system to provide surgical guidance during pre-drilling of pedicle screw trajectories. This approach already works well for interactions that are characterized by static hand-object poses, i.e., when the relative position of hand and object does not change during interaction. For dynamic hand-object interactions where fingers move in relation to the object, as it is the case during ECMO cannulation, and fingers can be occluded during interaction, it might be beneficial to integrate other modalities that support the inference of hand-object interactions. For example, Meier et al. [129] used the IMU from a wrist-worn smartwatch to detect finger taps on a physical surface. Such a measurement also works when fingers are occluded and can therefore complement and refine vision-based models for hand-object pose prediction and activity recognition.

Visual strategies in AR environments with multiple dynamic visual stimuli

In our second study, participants' perception was confronted with two conflicting objectives: a dynamic virtual cue (continuous performance feedback) and the coordination of hand movement. As shown in the system model, perception can additionally be confronted with visual cues from the real world that require regular checking. Such dynamic visual cues could be coming from the target system (e.g., check patient bleeding) or from a robotic arm that moves dynamically in the scene. These areas should also not be occluded, and the gaze path should be kept as short as possible when information is displayed. It will be interesting to understand how information must be visualized during such scenarios

with multiple dynamic visual stimuli. For example, information could be displayed in between two areas of interest, move with the operators' head or eye movement, or even be duplicated at several AOI locations to reduce the number of conflicting objectives. Eye tracking will be crucial for understanding where information is needed and the effect of different visualization strategies. Moreover, in this work, we only investigated the spatial distribution of visual attention based on dwell times. Future work should also investigate temporal gaze behavior, such as AOI sequences, to clarify where and when information should be displayed.

Feedback and support strategies while transitioning from training to independent execution.

The results of the second study have shown that context-aware AR support based on continuous performance feedback can raise novice operators' performance to expert level, indicating that the systems takes over responsibilities that are otherwise represented by expert knowledge. While such support greatly boosts operators' performance, it is questionable whether a novice will ever learn executing the task independently and with the same accuracy as an expert when always being guided with real-time feedback. Therefore, future work should investigate feedback strategies from the first training to executing the task independently with little or no guidance (depending on the gold standard). For example, instead of always providing operators with direct feedback, their performance can be recorded during the training and only shown as a summary statistic after completing the task, allowing them to reflect on their results after completion. Investigating the right way to provide and visualize feedback is important to efficiently develop operators' skills.

In addition to feedback, it will also be important to adapt virtual content between training cycles based on the current operator skill level. Too much information can result in low cognitive load and thus low knowledge retention. With too little information, the trainee can lose interest or shift to a trial-and-error strategy. Ideally, the operator is confronted with information that is sufficiently complex to keep his interest but not so complex as to discourage the operator. This state of

neither being under nor overchallenged is referred to as the ‘flow’ state, where learners’ efficiency is high [130]. A key requirement for changing virtual content over training cycles is finding behavior metrics that allow to infer the operators’ skill level based on recorded measurement data. Eye-gaze patterns have been shown to provide insights into participants learning curve [131]. An analysis of manual execution by tracking object-object or hand-object contexts as well as hand-eye coordination patterns are also promising measurements of expertise and skill development that should be explored in future work.

References

- [1] "Oxford Dictionary" oxforddictionaries.com (accessed August 1st, 2022).
- [2] R. Palmarini, J. A. Erkoyuncu, R. Roy, and H. Torabmostaedi, "A systematic review of augmented reality applications in maintenance" *Robotics and Computer-Integrated Manufacturing*, vol. 49, pp. 215-228, 2018, doi: 10.1016/j.rcim.2017.06.002.
- [3] X. Wang, S. K. Ong, and A. Y. C. Nee, "A comprehensive survey of augmented reality assembly research" (in English), *Advances in Manufacturing*, vol. 4, no. 1, pp. 1-22, Mar 2016, doi: 10.1007/s40436-015-0131-4.
- [4] M. Eckert, J. S. Volmerg, and C. M. Friedrich, "Augmented Reality in Medicine: Systematic and Bibliographic Review" *JMIR Mhealth Uhealth*, vol. 7, no. 4, p. e10967, Apr 26 2019, doi: 10.2196/10967.
- [5] R. I. Cook and D. D. Woods, "Operating at the sharp end: the complexity of human error" in *Human error in medicine: CRC Press*, 2018, pp. 255-310.
- [6] "Cambridge Dictionary" <https://dictionary.cambridge.org> (accessed August 1st, 2022).
- [7] F. Lalys and P. Jannin, "Surgical process modelling: a review" *Int J Comput Assist Radiol Surg*, vol. 9, no. 3, pp. 495-511, May 2014, doi: 10.1007/s11548-013-0940-5.
- [8] H. C. Lin, I. Shafran, D. Yuh, and G. D. Hager, "Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions" *Computer Aided Surgery*, vol. 11, no. 5, pp. 220-230, 2006.
- [9] D. Kragic and G. D. Hager, "Task modeling and specification for modular sensory based human-machine cooperative systems [includes Notice of Correction]" in *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2003, vol. 4, pp. 3192-3197.
- [10] E. Pelanis *et al.*, "Use of mixed reality for improved spatial understanding of liver anatomy" *Minimally Invasive Therapy & Allied Technologies*, vol. 29, no. 3, pp. 154-160, 2020.
- [11] A. Deshpande and I. Kim, "The effects of augmented reality on improving spatial problem solving for object assembly" (in English), *Advanced Engineering Informatics*, vol. 38, pp. 760-775, Oct 2018, doi: 10.1016/j.aei.2018.10.004.

- [12] 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*, 2017.
- [13] 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" (in English), *Journal of Computing and Information Science in Engineering*, vol. 20, no. 6, Dec 1 2020, doi: 10.1115/1.4046006.
- [14] A. Tang, C. Owen, F. Biocca, and W. Mou, "Comparative effectiveness of augmented reality in object assembly" in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2003, pp. 73-80.
- [15] C. Karagiannidis *et al.*, "Extracorporeal membrane oxygenation: evolving epidemiology and mortality" *Intensive Care Med*, vol. 42, no. 5, pp. 889-896, May 2016, doi: 10.1007/s00134-016-4273-z.
- [16] G. W. Kim *et al.*, "The effect of an improvement of experience and training in extracorporeal membrane oxygenation management on clinical outcomes" *Korean J Intern Med*, vol. 33, no. 1, pp. 121-129, Jan 2018, doi: 10.3904/kjim.2015.027.
- [17] J. Wolf, V. Wolfer, M. Halbe, F. Maisano, Q. Lohmeyer, and M. Meboldt, "Comparing the effectiveness of augmented reality-based and conventional instructions during single ECMO cannulation training" *Int J Comput Assist Radiol Surg*, vol. 16, no. 7, pp. 1171-1180, Jul 2021, doi: 10.1007/s11548-021-02408-y.
- [18] A. K. Dey, "Understanding and Using Context" (in English), *Personal and Ubiquitous Computing*, vol. 5, no. 1, pp. 4-7, Feb 2001, doi: 10.1007/s007790170019.
- [19] S. Dekker, *The field guide to understanding 'human error'*. CRC press, 2017, doi: 10.1201/9781317031833.
- [20] N. Petersen and D. Stricker, "Cognitive Augmented Reality" (in English), *Computers & Graphics-Uk*, vol. 53, pp. 82-91, Dec 2015, doi: 10.1016/j.cag.2015.08.009.
- [21] D. Katic *et al.*, "Context-aware Augmented Reality in laparoscopic surgery" *Comput Med Imaging Graph*, vol. 37, no. 2, pp. 174-82, Mar 2013, doi: 10.1016/j.compmedimag.2013.03.003.
- [22] S. Henderson and S. Feiner, "Augmented reality in the psychomotor phase of a procedural task" in *Proceedings of the 10th IEEE International Symposium on Mixed and Augmented Reality*, 2011, doi: 10.1109/ISMAR.2011.6092386.

References

- [23] F. Liebmann *et al.*, "Pedicle screw navigation using surface digitization on the Microsoft HoloLens" (in English), *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 7, pp. 1157-1165, Jul 2019, doi: 10.1007/s11548-019-01973-7.
- [24] L. X. Ng, J. Ng, K. T. W. 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 5th International Conference on Robotics and Artificial Intelligence 2019*, doi: 10.1145/3373724.3373727
- [25] "Gartner" <https://www.gartner.com/en/information-technology/glossary/context-aware-computing-2> (accessed August 1st, 2022).
- [26] R. K. Mobley, *An introduction to predictive maintenance*. Elsevier, 2002.
- [27] H. Lee, S. Y. Shin, M. Seo, G. B. Nam, and S. Joo, "Prediction of Ventricular Tachycardia One Hour before Occurrence Using Artificial Neural Networks" (in English), *Scientific Reports*, vol. 6, no. 1, pp. 1-7, Aug 26 2016, doi: 10.1038/srep32390.
- [28] R. Stauder *et al.*, "Random forests for phase detection in surgical workflow analysis" in *International Conference on Information Processing in Computer-Assisted Interventions*, 2014: Springer, pp. 148-157.
- [29] D. Lindlbauer, A. M. Feit, and O. Hilliges, "Context-Aware Online Adaptation of Mixed Reality Interfaces" in *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, 2019, doi: 10.1145/3332165.3347945.
- [30] A. Rajagopal and J. Martin, "Gossypiboma - "A surgeon's legacy" - Report of a case and review of the literature" (in English), *Diseases of the Colon & Rectum*, vol. 45, no. 1, pp. 119-120, Jan 2002, doi: 10.1007/s10350-004-6124-1.
- [31] K. E. Bani-Hani, K. A. Gharaibeh, and R. J. Yaghan, "Retained surgical sponges (gossypiboma)" *Asian J Surg*, vol. 28, no. 2, pp. 109-15, Apr 2005, doi: 10.1016/s1015-9584(09)60273-6.
- [32] U. Neumann and A. Majoros, "Cognitive, performance, and systems issues for augmented reality applications in manufacturing and maintenance" in *Proceedings. IEEE 1998 Virtual Reality Annual International Symposium (Cat. No. 98CB36180)*, 1998: IEEE, pp. 4-11.
- [33] C. Lenz *et al.*, "Joint-action for humans and industrial robots for assembly tasks" in *RO-MAN 2008-The 17th IEEE International*

Symposium on Robot and Human Interactive Communication, 2008: IEEE, pp. 130-135.

- [34] R. Felix and J. Rajan, "Human Factors in Safety-Critical Systems" ed: Butterworth Heinemann, 1997.
- [35] M. F. Land and M. Hayhoe, "In what ways do eye movements contribute to everyday activities?" *Vision Research*, vol. 41, no. 25-26, pp. 3559-3565, 2001, doi: 10.1016/s0042-6989(01)00102-x.
- [36] 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*, 2013, doi: 10.1145/2470654.2470697
- [37] C. M. Huang, S. Andrist, A. Sauppe, and B. Mutlu, "Using gaze patterns to predict task intent in collaboration" (in English), *Frontiers in Psychology*, vol. 6, p. 1049, Jul 24 2015, doi: 10.3389/fpsyg.2015.01049.
- [38] R. C. King, L. Atallah, B. P. Lo, and G. Z. Yang, "Development of a wireless sensor glove for surgical skills assessment" *IEEE Trans Inf Technol Biomed*, vol. 13, no. 5, pp. 673-9, Sep 2009, doi: 10.1109/TITB.2009.2029614.
- [39] J. Triesch, D. H. Ballard, M. M. Hayhoe, and B. T. Sullivan, "What you see is what you need" (in English), *Journal of Vision*, vol. 3, no. 1, pp. 86-94, 2003, doi: 10.1167/3.1.9.
- [40] 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*, vol. 25, pp. 69-91, 2017.
- [41] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 409-419, 2018.
- [42] J. Pelz, M. Hayhoe, and R. Loeber, "The coordination of eye, head, and hand movements in a natural task" *Exp Brain Res*, vol. 139, no. 3, pp. 266-77, Aug 2001, doi: 10.1007/s002210100745.
- [43] G. Gras and G. Z. Yang, "Context-Aware Modeling for Augmented Reality Display Behaviour" (in English), *Ieee Robotics and Automation Letters*, vol. 4, no. 2, pp. 562-569, Apr 2019, doi: 10.1109/Lra.2019.2890852.
- [44] M. Land, N. Mennie, and J. Rusted, "The roles of vision and eye movements in the control of activities of daily living" *Perception*, vol. 28, no. 11, pp. 1311-28, 1999, doi: 10.1068/p2935.

References

- [45] R. S. Johansson, G. Westling, A. Backstrom, and J. R. Flanagan, "Eye-hand coordination in object manipulation" *J Neurosci*, vol. 21, no. 17, pp. 6917-32, Sep 1 2001, doi: 10.1523/JNEUROSCI.21-17-06917.2001
- [46] I. E. Sutherland, "A head-mounted three dimensional display" in *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*, 1968, pp. 757-764.
- [47] R. T. Azuma, "A survey of augmented reality" (in English), *Presence-Virtual and Augmented Reality*, vol. 6, no. 4, pp. 355-385, Aug 1997, doi: 10.1162/pres.1997.6.4.355.
- [48] J. Carmigniani, B. Furht, M. Anisetti, P. Ceravolo, E. Damiani, and M. Ivkovic, "Augmented reality technologies, systems and applications" (in English), *Multimedia Tools and Applications*, vol. 51, no. 1, pp. 341-377, Jan 2011, doi: 10.1007/s11042-010-0660-6.
- [49] F. Heinrich, L. Schwenderling, F. Joeres, K. Lawonn, and C. Hansen, "Comparison of Augmented Reality Display Techniques to Support Medical Needle Insertion" *IEEE Trans Vis Comput Graph*, vol. 26, no. 12, pp. 3568-3575, Dec 2020, doi: 10.1109/TVCG.2020.3023637.
- [50] "Microsoft Docs" <https://docs.microsoft.com/> (accessed August 1st, 2022).
- [51] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer, *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford, 2011.
- [52] J. Wolf, S. Hess, D. Bachmann, Q. Lohmeyer, and M. Meboldt, "Automating Areas of Interest Analysis in Mobile Eye Tracking Experiments based on Machine Learning" *J Eye Mov Res*, vol. 11, no. 6, Dec 10 2018, doi: 10.16910/jemr.11.6.6.
- [53] F. S. Wang, J. Wolf, M. Farshad, M. Meboldt, and Q. Lohmeyer, "Object-Gaze Distance: Quantifying Near- Peripheral Gaze Behavior in Real-World Applications" *J Eye Mov Res*, vol. 14, no. 1, May 19 2021, doi: 10.16910/jemr.14.1.5.
- [54] J. Wolf, Q. Lohmeyer, C. Holz, and M. Meboldt, "Gaze Comes in Handy: Predicting and Preventing Erroneous Hand Actions in AR-Supported Manual Tasks" in *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2021, pp. 166-175. doi: 10.1109/ISMAR52148.2021.00031
- [55] M. S. Firstenberg, "Introductory Chapter: Evolution of ECMO from Salvage to Mainstream Supportive and Resuscitative Therapy" in *Extracorporeal Membrane Oxygenation: Advances in Therapy*, 2016, Chapter 1.

- [56] A. Combes *et al.*, "Position paper for the organization of extracorporeal membrane oxygenation programs for acute respiratory failure in adult patients" *Am J Respir Crit Care Med*, vol. 190, no. 5, pp. 488-96, Sep 1 2014, doi: 10.1164/rccm.201404-0630CP.
- [57] C. K. Allan *et al.*, "An extracorporeal membrane oxygenation cannulation curriculum featuring a novel integrated skills trainer leads to improved performance among pediatric cardiac surgery trainees" *Simul Healthc*, vol. 8, no. 4, pp. 221-8, Aug 2013, doi: 10.1097/SIH.0b013e31828b4179.
- [58] S. Y. Chan, M. Figueroa, T. Spentzas, A. Powell, R. Holloway, and S. Shah, "Prospective assessment of novice learners in a simulation-based extracorporeal membrane oxygenation (ECMO) education program" *Pediatr Cardiol*, vol. 34, no. 3, pp. 543-52, Mar 2013, doi: 10.1007/s00246-012-0490-6.
- [59] D. Dirnberger *et al.*, "Extracorporeal Life Support Organization (ELSO)" *Guidelines for ECMO Transport*, 2015.
- [60] D. Palmer *et al.*, "A High-Fidelity Surgical Model and Perfusion Simulator Used to Demonstrate ECMO Cannulation, Initiation, and Stabilization" *The journal of extra-corporeal technology*, vol. 51, no. 2, p. 94, 2019.
- [61] D. M. McMullan, "Novel ECMO surgical cannulation simulators" *Qatar Medical Journal*, vol. 2017, no. 1, p. 61, 2017, doi: 10.5339/qmj.2017.swacelso.61.
- [62] M. Aldisi *et al.*, "Design and implementation of a modular ECMO simulator" *Qatar Medical Journal*, vol. 2017, no. 1, p. 62, 2017, doi: 10.5339/qmj.2017.swacelso.62.
- [63] F. G. Hamza-Lup, J. P. Rolland, and C. Hughes, "A distributed augmented reality system for medical training and simulation" *arXiv preprint arXiv:1811.12815*, 2018.
- [64] T. Blum, S. M. Heining, O. Kutter, and N. Navab, "Advanced training methods using an augmented reality ultrasound simulator" in *Proceedings of the 8th IEEE International Symposium on Mixed and Augmented Reality*, 2009, doi: 10.1109/ISMAR.2009.5336476
- [65] E. Azimi *et al.*, "Evaluation of Optical See-Through Head-Mounted Displays in Training for Critical Care and Trauma" (in English), *25th 2018 Ieee Conference on Virtual Reality and 3d User Interfaces (Vr)*, pp. 511-512, 2018, doi: 10.1109/VR.2018.8446583.
- [66] P. Vavra *et al.*, "Recent Development of Augmented Reality in Surgery: A Review" *J Healthc Eng*, vol. 2017, p. 4574172, 2017, doi: 10.1155/2017/4574172.

References

- [67] C. Gorman and L. Gustafsson, "The use of augmented reality for rehabilitation after stroke: a narrative review" *Disabil Rehabil Assist Technol*, vol. 17, no. 4, pp. 409-417, May 2022, doi: 10.1080/17483107.2020.1791264.
- [68] E. Z. Barsom, M. Graafland, and M. P. Schijven, "Systematic review on the effectiveness of augmented reality applications in medical training" *Surg Endosc*, vol. 30, no. 10, pp. 4174-83, Oct 2016, doi: 10.1007/s00464-016-4800-6.
- [69] C. Noll, U. von Jan, U. Raap, and U. V. Albrecht, "Mobile Augmented Reality as a Feature for Self-Oriented, Blended Learning in Medicine: Randomized Controlled Trial" *JMIR Mhealth Uhealth*, vol. 5, no. 9, Sep 14 2017, doi: 10.2196/mhealth.7943.
- [70] H. Lia *et al.*, "HoloLens in suturing training" in *Proceedings Medical Imaging*, 2018 doi: 10.1117/12.22939342018.
- [71] L. Rupprecht, D. Lunz, A. Philipp, M. Lubnow, and C. Schmid, "Pitfalls in percutaneous ECMO cannulation" *Heart Lung Vessel*, vol. 7, no. 4, pp. 320-6, 2015. PMID: 26811838; PMCID: PMC4712035.
- [72] B. Laugwitz, T. Held, and M. Schrepp, "Construction and evaluation of a user experience questionnaire" In: *Holzinger, A. (eds) HCI and Usability for Education and Work. USAB 2008. Lecture Notes in Computer Science*, vol 5298, 2008, doi: 10.1007/978-3-540-89350-9_62008.
- [73] M. Schrepp, "User experience questionnaire handbook" *All you need to know to apply the UEQ successfully in your project*, 2015.
- [74] B. I. Martin, S. K. Mirza, N. Spina, W. R. Spiker, B. Lawrence, and D. S. Brodke, "Trends in Lumbar Fusion Procedure Rates and Associated Hospital Costs for Degenerative Spinal Diseases in the United States, 2004 to 2015" *Spine (Phila Pa 1976)*, vol. 44, no. 5, pp. 369-376, Mar 1 2019, doi: 10.1097/BRS.0000000000002822.
- [75] T. Maruyama and K. Takeshita, "Surgical treatment of scoliosis: a review of techniques currently applied" *Scoliosis*, vol. 3, no. 1, p. 6, Apr 18 2008, doi: 10.1186/1748-7161-3-6.
- [76] J. J. Verlaan *et al.*, "Surgical treatment of traumatic fractures of the thoracic and lumbar spine: a systematic review of the literature on techniques, complications, and outcome" *Spine (Phila Pa 1976)*, vol. 29, no. 7, pp. 803-14, Apr 1 2004, doi: 10.1097/01.brs.0000116990.31984.a9.
- [77] T. P. Ville, T. H. Tuomas, H. N. Marko, P. Liisa, P. R. Jussi, and M. M. Ville, "National trends in lumbar spine decompression and

- fusion surgery in Finland, 1997–2018" *Acta Orthopaedica*, vol. 92, no. 2, pp. 199-203, 2021, doi: 10.1080/17453674.2020.1839244
- [78] I. A. Harris and A. T. Dao, "Trends of spinal fusion surgery in Australia: 1997 to 2006" *ANZ J Surg*, vol. 79, no. 11, pp. 783-8, Nov 2009, doi: 10.1111/j.1445-2197.2009.05095.x.
- [79] K. Kobayashi, K. Ando, Y. Nishida, N. Ishiguro, and S. Imagama, "Epidemiological trends in spine surgery over 10 years in a multicenter database" *Eur Spine J*, vol. 27, no. 8, pp. 1698-1703, Aug 2018, doi: 10.1007/s00586-018-5513-4.
- [80] R. A. Deyo and S. K. Mirza, "Trends and variations in the use of spine surgery" (in English), *Clinical Orthopaedics and Related Research*, vol. 443, no. 443, pp. 139-146, Feb 2006, doi: 10.1097/01.bl.0000198726.62514.75.
- [81] W. C. Pannell, D. D. Savin, T. P. Scott, J. C. Wang, and M. D. Daubs, "Trends in the surgical treatment of lumbar spine disease in the United States" *Spine J*, vol. 15, no. 8, pp. 1719-27, Aug 1 2015, doi: 10.1016/j.spinee.2013.10.014.
- [82] R. A. Deyo, D. T. Gray, W. Kreuter, S. Mirza, and B. I. Martin, "United States trends in lumbar fusion surgery for degenerative conditions" *Spine (Phila Pa 1976)*, vol. 30, no. 12, pp. 1441-5; discussion, Jun 15 2005, doi: 10.1097/01.brs.0000166503.37969.8a.
- [83] E. Van de Kelft, *Surgery of the spine and spinal cord: A neurosurgical approach*. 2016.
- [84] A. Manbachi, *Towards ultrasound-guided spinal fusion surgery*. 2016.
- [85] K. H. Bridwell and M. Gupta, *Bridwell and DeWald's Textbook of Spinal Surgery*. 2019.
- [86] T. Laine, T. Lund, M. Ylikoski, J. Lohikoski, and D. Schlenzka, "Accuracy of pedicle screw insertion with and without computer assistance: a randomised controlled clinical study in 100 consecutive patients" (in English), *European Spine Journal*, vol. 9, no. 3, pp. 235-240, Jun 2000, doi: 10.1007/s005860000146.
- [87] M. M. Panjabi, J. D. O'Holleran, J. J. Crisco, 3rd, and R. Kothe, "Complexity of the thoracic spine pedicle anatomy" *Eur Spine J*, vol. 6, no. 1, pp. 19-24, 1997, doi: 10.1007/BF01676570.
- [88] M. Merc, I. Drstvensek, M. Vogrin, T. Brajljih, and G. Recnik, "A multi-level rapid prototyping drill guide template reduces the perforation risk of pedicle screw placement in the lumbar and sacral spine" *Arch Orthop Trauma Surg*, vol. 133, no. 7, pp. 893-9, Jul 2013, doi: 10.1007/s00402-013-1755-0.

References

- [89] M. Oglesby, S. J. Fineberg, A. A. Patel, M. A. Pelton, and K. Singh, "Epidemiological Trends in Cervical Spine Surgery for Degenerative Diseases Between 2002 and 2009" (in English), *Spine*, vol. 38, no. 14, pp. 1226-1232, Jun 15 2013, doi: 10.1097/BRS.0b013e31828be75d.
- [90] F. Raciborski, R. Gasik, and A. Klak, "Disorders of the spine. A major health and social problem" *Reumatologia*, vol. 54, no. 4, pp. 196-200, 2016, doi: 10.5114/reum.2016.62474.
- [91] P. Merloz *et al.*, "Fluoroscopy-based navigation system in spine surgery" *Proc Inst Mech Eng H*, vol. 221, no. 7, pp. 813-20, Oct 2007, doi: 10.1243/09544119JEIM268.
- [92] L. T. Holly and K. T. Foley, "Intraoperative spinal navigation" *Spine (Phila Pa 1976)*, vol. 28, no. 15 Suppl, pp. S54-61, Aug 1 2003, doi: 10.1097/01.BRS.0000076899.78522.D9.
- [93] L. P. Nolte *et al.*, "A new approach to computer-aided spine surgery: fluoroscopy-based surgical navigation" *Eur Spine J*, vol. 9 Suppl 1, no. 1, pp. S78-88, Feb 2000, doi: 10.1007/pl00010026.
- [94] L. T. Holly and K. T. Foley, "Image guidance in spine surgery" *Orthop Clin North Am*, vol. 38, no. 3, pp. 451-61; Jul 2007, doi: 10.1016/j.ocl.2007.04.001.
- [95] S. C. Thakkar, R. S. Thakkar, N. Sirisreetreerux, J. A. Carrino, B. Shafiq, and E. A. Hasenboehler, "2D versus 3D fluoroscopy-based navigation in posterior pelvic fixation: review of the literature on current technology" *Int J Comput Assist Radiol Surg*, vol. 12, no. 1, pp. 69-76, Jan 2017, doi: 10.1007/s11548-016-1465-5.
- [96] E. Van de Kelft, F. Costa, D. Van der Planken, and F. Schils, "A prospective multicenter registry on the accuracy of pedicle screw placement in the thoracic, lumbar, and sacral levels with the use of the O-arm imaging system and StealthStation Navigation" *Spine*, vol. 37, no. 25, pp. E1580-E1587, 2012.
- [97] R. Hrtl, K. S. Lam, J. Wang, A. Korge, F. Kandziora, and Audig, "Worldwide survey on the use of navigation in spine surgery" *World neurosurgery*, vol. 79, no. 1, pp. 162-172, 2013.
- [98] M. Nadeau, J. Batke, C. Fisher, and J. Street, "A Qualitative Web-Based Expert Opinion Analysis on the Adoption of Intraoperative CT and Navigation Systems in Spine Surgery" *Global Spine Journal*, vol. 5, no. 1, 2017, doi: 10.1055/s-0035-1554209.
- [99] J. W. Yoon, R. E. Chen, P. K. Han, P. Si, W. D. Freeman, and S. M. Pirris, "Technical feasibility and safety of an intraoperative head-up display device during spine instrumentation" *Int J Med Robot*, vol. 13, no. 3, p. e1770, Sep 2017, doi: 10.1002/rcs.1770.

- [100] E. Leger, S. Drouin, D. L. Collins, T. Popa, and M. Kersten-Oertel, "Quantifying attention shifts in augmented reality image-guided neurosurgery," *Healthc Technol Lett*, vol. 4, no. 5, pp. 188-192, Oct 2017, doi: 10.1049/htl.2017.0062.
- [101] L. Jud *et al.*, "Applicability of augmented reality in orthopedic surgery - A systematic review," *BMC Musculoskelet Disord*, vol. 21, no. 1, p. 103, Feb 15 2020, doi: 10.1186/s12891-020-3110-2.
- [102] G. Vadalà, S. De Salvatore, L. Ambrosio, F. Russo, R. Papalia, and V. Denaro, "Robotic spine surgery and augmented reality systems: a state of the art" *Neurospine*, vol. 17, no. 1, p. 88, 2020, doi: <https://doi.org/10.14245%2Fns.2040060.030>.
- [103] F. Muller, S. Roner, F. Liebmann, J. M. Spirig, P. Furnstahl, and M. Farshad, "Augmented reality navigation for spinal pedicle screw instrumentation using intraoperative 3D imaging" *Spine J*, vol. 20, no. 4, pp. 621-628, Apr 2020, doi: 10.1016/j.spinee.2019.10.012.
- [104] M. Farshad, P. Furnstahl, and J. M. Spirig, "First in man in-situ augmented reality pedicle screw navigation" *N Am Spine Soc J*, vol. 6, Jun 2021, doi: 10.1016/j.xnsj.2021.100065.
- [105] C. Brendle, L. Schütz, J. Esteban, S. M. Krieg, U. Eck, and N. Navab, "Can a Hand-Held Navigation Device Reduce Cognitive Load? A User-Centered Approach Evaluated by 18 Surgeons" In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2020, doi: 10.1007/978-3-030-59716-0_38.
- [106] A. F. Samdani *et al.*, "Accuracy of free-hand placement of thoracic pedicle screws in adolescent idiopathic scoliosis: how much of a difference does surgeon experience make?" *European Spine Journal*, vol. 19, no. 1, pp. 91-95, 2010.
- [107] R. A. Lehman Jr, L. G. Lenke, K. A. Keeler, Y. J. Kim, and G. Cheh, "Computed tomography evaluation of pedicle screws placed in the pediatric deformed spine over an 8-year period" *Spine*, vol. 32, no. 24, pp. 2679-2684, 2007.
- [108] A. Elmi-Terander *et al.*, "Surgical Navigation Technology Based on Augmented Reality and Integrated 3D Intraoperative Imaging: A Spine Cadaveric Feasibility and Accuracy Study" *Spine (Phila Pa 1976)*, vol. 41, no. 21, pp. E1303-E1311, Nov 1 2016, doi: 10.1097/BRS.0000000000001830.
- [109] Y. Abe *et al.*, "A novel 3D guidance system using augmented reality for percutaneous vertebroplasty" (in English), *J Neurosurg-Spine*, vol. 19, no. 4, pp. 492-501, Oct 2013, doi: 10.3171/2013.7.Spine12917.

References

- [110] G. A. Koulieris, K. Akit, M. Stengel, R. K. Mantiuk, K. Mania, and C. Richardt, "Near-Eye Display and Tracking Technologies for Virtual and Augmented Reality", *Computer Graphics Forum*, vol. 38, no. 2, pp. 493-519, 2019, doi: 10.1111/cgf.13654.
- [111] J. Zhu, S. K. Ong, and A. Y. C. Nee, "A context-aware augmented reality assisted maintenance system" (in English), *International Journal of Computer Integrated Manufacturing*, vol. 28, no. 2, pp. 213-225, Feb 1 2015, doi: 10.1080/0951192x.2013.874589.
- [112] M. Mussgnug, D. Singer, Q. Lohmeyer, and M. Meboldt, "Automated interpretation of eye-hand coordination in mobile eye tracking recordings Identifying demanding phases in human-machine interactions" (in English), *Kunstliche Intelligenz*, vol. 31, no. 4, pp. 331-337, Nov 2017, doi: 10.1007/s13218-017-0503-y.
- [113] 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" presented at the *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017, doi: 10.1145/3025453.3025753.
- [114] J. D. Crawford, W. P. Medendorp, and J. J. Marotta, "Spatial transformations for eye-hand coordination" *J Neurophysiol*, vol. 92, no. 1, pp. 10-9, Jul 2004, doi: 10.1152/jn.00117.2004.
- [115] B. W. Tatler and M. F. Land, "Everyday Visual Attention" (in English), *Handbook of Attention*, pp. 391-421, 2015.
- [116] 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" *J Mot Behav*, vol. 30, no. 3, pp. 249-59, Sep 1998, doi: 10.1080/00222899809601340.
- [117] C. Fermüller *et al.*, "Prediction of Manipulation Actions" (in English), *International Journal of Computer Vision*, vol. 126, no. 2-4, pp. 358-374, 2017, doi: 10.1007/s11263-017-0992-z.
- [118] M. D. Byrne, J. R. Anderson, S. Douglass, and M. Matessa, "Eye tracking the visual search of click-down menus" presented at the *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1999. doi: 10.1145/302979.303118.
- [119] 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*, 2000, doi: 10.1145/355017.355041
- [120] J. Huang, R. White, and G. Buscher, "User see, user point" presented at the *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012, doi: 10.1145/2207676.2208591.

- [121] 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*, doi: 10.1145/3334480.3382924.
- [122] 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*, 2018, doi: 10.1145/3204493.3204567.
- [123] P. Baudisch *et al.*, "Drag-and-pop and drag-and-pick: Techniques for accessing remote screen content on touch-and pen-operated systems" *Interact*, vol. 3, 2003.
- [124] F. Koochaki and L. Najafizadeh, "Eye Gaze-based Early Intent Prediction Utilizing CNN-LSTM", *2019 41st Annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, 2019, doi: 0.1109/EMBC.2019.8857054.
- [125] 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" presented at the *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, 2019, doi: 10.1145/3332165.3347919.
- [126] R. J. Kosinski, "A literature review on reaction time" *Clemson University*, vol. 10, no. 1, 2008.
- [127] 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.
- [128] M. Doughty and N. R. Ghugre, "HMD-EgoPose: head-mounted display-based egocentric marker-less tool and hand pose estimation for augmented surgical guidance" *Int J Comput Assist Radiol Surg*, Jun 14 2022, doi: 10.1007/s11548-022-02688-y.
- [129] M. Meier, P. Streli, A. Fender, and C. Holz, "TapID: Rapid touch interaction in virtual reality using wearable sensing" in *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, 2021, pp. 519-528, doi: 10.1109/VR50410.2021.00076
- [130] M. Csikszentmihalyi and M. Csikszentmihalyi, *Flow: The psychology of optimal experience*. Harper & Row New York, 1990.
- [131] F. S. Wang, C. Gianduzzo, M. Meboldt, and Q. Lohmeyer, "An algorithmic approach to determine expertise development using object-related gaze pattern sequences" *Behavior Research Methods*, vol. 54, no. 1, pp. 493-507, 2022, doi: 10.3758/s13428-021-01652-z

Curriculum vitae

Personal data:

Name: Julian Wolf
Place of birth: Ludwigshafen am Rhein, Germany
Nationality: German

Education:

01/2018 – 12/2022 **Doctor of Science (Dr. sc. ETH Zurich)** in Mechanical Engineering, Product Development Group Zurich, ETH Zurich (ETHZ), Switzerland

10/2014 – 11/2017 **Master of Science (M.Sc.)** in Mechanical Engineering, Karlsruhe Institute of Technology (KIT), Germany

03/2016 – 12/2016 **Exchange Year** at the University of Technology Sydney (UTS), Engineering Management and Business, Sydney

04/2012 – 09/2014 **Bachelor of Science (B.Sc.)** in Mechanical Engineering, Karlsruhe Institute of Technology (KIT), Germany

10/2010 – 03/2012 **Diploma Program** in Mechanical Engineering, Technical University Kaiserslautern, Germany

08/2000 – 03/2009 **High School Education**, Werner-Heisenberg-Gymnasium, Germany

List of publications

1. J. Wolf, D. Luchmann, Q. Lohmeyer, M. Farshad, P. Fürnstahl, M. Meboldt (2022). How augmented reality visualizations for drilling affect trajectory deviation, visual attention, and user experience. *International Journal of Computer Assisted Radiology and Surgery*. (Under review)
2. J. Wolf, Q. Lohmeyer, C. Holz, M. Meboldt (2021). Gaze Comes in Handy: Predicting and Preventing Erroneous Hand Actions in AR-Supported Manual Tasks. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (pp. 166-175), doi: 10.1109/ISMAR52148.2021.00031.
3. J. Wolf, V. Wolfer, M. Halbe, F. Maisano, Q. Lohmeyer, M. Meboldt (2021). Comparing the effectiveness of augmented reality-based and conventional instructions during single ECMO cannulation training. *International Journal of Computer Assisted Radiology and Surgery*, 16(7), 1171-1180, doi: 10.1007/s11548-021-02408-y.
4. F. Wang, J. Wolf, M. Farshad, M. Meboldt, Q. Lohmeyer (2021). Object-gaze distance: Quantifying near-peripheral gaze behavior in real-world applications. *Journal of Eye Movement Research*, 14(1), doi: 10.16910/jemr.14.1.5.
5. J. Wolf, S. Hess, D. Bachmann, Q. Lohmeyer, M. Meboldt (2018). Automating areas of interest analysis in mobile eye tracking experiments based on machine learning. *Journal of Eye Movement Research*, 11(6), doi: 10.16910/jemr.11.6.6.