Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Blended Spaces: Perception and Interaction in Projection-Based Spatial Augmented Reality Environments

an der Universität Hamburg eingereichte Dissertation

vorgelegt von

Susanne Schmidt

Human-Computer Interaction
Fachbereich Informatik
Fakultät für Mathematik, Informatik und Naturwissenschaften

2020

Erstgutachter:	Prof. Dr. Frank Steinicke
Zweitgutachter:	Prof. Dr. Gregory F. Welch
Vorsitzende der Prüfungskommission:	Prof. Dr. Simone Frintrop
Stellvertretende Vorsitzende der Prüfungskommission:	Prof. Dr. Eva Bittner
Datum der Disputation:	14.08.2020

Für meinen Opa
Dr. Klaus Rose.

Abstract

The reality-virtuality continuum defined by Milgram [MTUK95] encompasses technological approaches to mix real and computer-generated content to create virtually enriched experiences. Most of today's applications can be assigned to a discrete stage within this continuum, but cannot take full advantage of the individual characteristics of each of the other stages. In this thesis, we introduce the concept of *Blended Spaces* to describe environments, in which (i) objects can transition between real, mixed, and virtual states, (ii) objects with different states can interact with each other, and (iii) users can experience the resulting real, *Augmented Reality* (AR), *Augmented Virtuality* (AV), and *Virtual Reality* (VR) conditions seamlessly. To implement such Blended Spaces, we exploit spatial projection-based AR technology. The resulting characteristics of Blended Spaces create new challenges in terms of their technical feasibility, their visual perception, and their support of collaboration; all of which will be investigated in detail in the scope of this thesis.

In Part II we discuss three hardware configurations in consideration of the needs and possibilities of different application fields. We also lay the technological foundation to link all interdependent system components, such as tracking cameras, projectors, and physical projection surfaces. In this context, customized calibration algorithms as well as rendering approaches are demonstrated. Part III investigates perceptual issues, which can be caused by conflicting depth cues and cannot be fully described by research in traditional hand-held and head-mounted AR due to differences in the setup. We consider a range of projected illusions that could affect the perceived depth of objects relative to each other and the environment. From the results of multiple user studies, we draw conclusions about the general consistency of depth perception in Blended Spaces and validate our hypotheses in several perceptual follow-up studies. The statistical results indicate that an increased parallax between multiple projections leads to higher depth misperceptions, however, they also suggest a strong dependency from the participant's experience with stereoscopic 3D. We conclude the part with a consideration of possible consequences of these observations on the usability of Blended Spaces in practical applications. Part IV focuses on the social characteristics of Blended Spaces and addresses the question of how to improve the interaction between a user and real as well as virtual cooperation partners, so-called intelligent virtual agents (IVAs), in such environments. To address multi-user support, we introduce two user interfaces that are well received by study participants in terms of usability as well as collaboration. Regarding the interaction with IVAs, we focus on their representation as well as their capabilities to interact with the physical surroundings. We collect subjective and objective data in a series of user studies, that indicates advantages of embodied, content-related agents over voice-only, generic agents in an exhibition scenario. Virtual guides with a visualized human body are perceived as more present and attractive than audio guides, whereas content-specific IVAs particularly achieve higher scores in the dimensions of attractiveness, stimulation, and novelty in comparison to generic IVAs. To investigate virtual-physical capabilities of IVAs, we demonstrate two exemplary manipulations that are affecting real-world objects either within or even outside of a Blended Space. Both manipulations receive positive feedback from participants of a user study, in terms of the subjectively perceived realism of IVAs as well as emotional responses such as surprise and enjoyment. Based on these considerations we finally envision future directions of Blended Spaces.

Zusammenfassung

Das Realitäts-Virtualitäts-Kontinuum nach Milgram [MTUK95] beschreibt technologische Ansätze zur Überlagerung realer und computergenerierter Inhalte. Heutige Anwendungen können typischerweise genau einer diskreten Stufe innerhalb dieses Kontinuums zugewiesen werden, sodass das Potential des Kontinuums nicht vollständig genutzt wird. Um diese Möglichkeiten besser auszuschöpfen, führen wir das Konzept der *Blended Spaces* ein. Dieses beschreibt Umgebungen, in denen (i) Objekte zwischen einem realen und gemischten sowie virtuellen Zuständen wechseln können, (ii) Objekte mit unterschiedlichen Zuständen interagieren können, und (iii) Nutzer nahtlos verschiedene Realitätsformen, einschließlich der unmittelbaren realen Umgebung sowie einer erweiterten Realität (AR) und einer vollständig virtuellen Realität (VR), erfahren können. Für die Implementierung solcher Blended Spaces bedienen wir uns der Technologie von *projektionsbasierter Spatial AR*. Aus den resultierenden Charakteristika ergeben sich neue Herausforderungen in Bezug auf die technische Realisierbarkeit von Blended Spaces sowie deren Wahrnehmung und die Unterstützung kollaborativer Aufgaben. Jeder dieser Aspekte wird im Rahmen der vorliegenden Dissertation näher beleuchtet.

Zunächst werden in Teil II alternative Hardware-Konfigurationen diskutiert, die auf die Anforderungen und Möglichkeiten verschiedener Anwendungsgebiete abgestimmt sind. Zudem werden technologische Grundlagen zur Verknüpfung aller involvierten Systemkomponenten wie Tracking-Kameras, Projektoren und Projektionsoberflächen geschaffen, insbesondere mit einem Fokus auf spezifische Kalibrierungsalgorithmen. Teil III untersucht perzeptuelle Fragestellungen, die sich aus der Existenz widersprüchlicher Tiefenhinweise ergeben und aufgrund konzeptioneller Unterschiede nicht vollständig durch Forschungsergebnisse aus verwandten AR-Gebieten abgedeckt werden können. Wir betrachten eine Palette von projektionsbasierten Illusionen, die die wahrgenommene Tiefe von Objekten beeinflussen können. Anhand mehrerer Nutzerstudien ziehen wir Schlüsse bezüglich der allgemeinen Konsistenz der Tiefenwahrnehmung in Blended Spaces und leiten daraus entsprechende Hypothesen ab. Deren Überprüfung legt nahe, dass eine wachsende Parallaxe zwischen Projektionen mit größeren Fehleinschätzungen der Tiefe korreliert, wobei letztere maßgeblich von der Erfahrung des Betrachters mit stereoskopischen 3D Displays abhängen. Inwiefern diese Beobachtungen Konsequenzen auf die Einsetzbarkeit von Blended Spaces in praktischen Anwendungen haben, ist Thema einer abschließenden Diskussion. Teil IV stellt soziale Faktoren in den Vordergrund und wirft die Frage auf, wie die Interaktion zwischen Nutzern und echten sowie virtuellen Kooperationspartnern, sogenannten intelligenten virtuellen Agenten (IVAs), verbessert werden kann. Zur Unterstützung mehrerer Nutzer werden zwei Interaktionskonzepte eingeführt, welche von Studienteilnehmern in Bezug auf Benutzbarkeit sowie Kollaboration positiv bewertet werden. Bei der Interaktion mit IVAs legen wir den Schwerpunkt auf deren audio-visuelle Repräsentation sowie die Fähigkeit mit der physikalischen Umgebung zu interagieren. Im Rahmen eines simulierten Museumsszenarios können Vorteile von kontextspezifischen IVAs mit humanoider visueller Repräsentation gegenüber generischer Audio-Guides festgestellt werden. Zudem werden zwei exemplarische Interaktionen vorgestellt, die IVAs eine Manipulation realer Objekte sowohl innerhalb als auch außerhalb eines Blended Spaces erlauben. Beide Umsetzungen werden durch Studienteilnehmer in Bezug auf den subjektiv empfundenen Realismus des IVA sowie hervorgerufene Emotionen positiv bewertet. Basierend auf diesen Ergebnissen diskutieren wir abschließend potentielle Entwicklungsrichtungen für Blended Spaces.

Contents

List of Figures

List of Tables

Part I.

Conceptualization of Blended Spaces for Interactive Shared Experiences

1 Chapter 1.

Definition of Blended Spaces

Digital transformation affected almost every aspect of our daily lives and will shape how we work, learn, communicate, and live in the future [SB19]. Two emerging technologies with the potential to have a long-term impact on both business and consumer markets are *Virtual Reality* (VR) and *Augmented Reality* (AR). The invention of VR, which denotes the generation of completely synthetic environments using computer technology [MTUK95], dates back to the 1960s when Morton Heilig patented one of the earliest examples of immersive computer-generated environments, the *Sensorama* [Hei62], and Ivan Sutherland envisioned the *Ultimate Display* as a "room within which the computer can control the existence of matter" [Sut65]. In 1968, it was Sutherland who created the first partially see-through head-mounted display (HMD) [Sut68] known as the *Sword of Damocles*, which is widely considered to be a precursor of current AR technology. Though having its origins in the same decade as VR, it took almost 30 years before AR emerged as an independent research field (for an extensive historical background of both VR and AR see [CGRR18]). In contrast to a full replacement of the real environment, AR aims at enhancing real-world views by embedding additional virtual content [BR05]. Despite fundamental differences between the two concepts, both became particularly popular in the gaming and entertainment sector [SH16]. According to the Gartner hype cycle 2015 [WB15], both technologies already passed the phases of technology breakthrough and the peak of inflated expectations, and are estimated to reach mainstream adoption by 2025. In the report of 2019, both VR and AR were even excluded from the list of emerging technologies because, according to Gartner, they already reached a mature state [SB19].

A more differentiated analysis of the strengths and weaknesses of both technologies led to a set of customized applications beyond the entertainment industry. VR can immerse users in environments, which represent a different place or time, while being isolated from the real world. This can be beneficial for applications such as therapy, military training, and design review [Jer15]. In contrast, AR technology is preferable to VR systems if users are required to see and interact with their real environment, for example, for navigation, maintenance tasks, and computer-assisted surgeries [SH16].

Regarding the virtualization of real-world objects, VR and AR are only two stages within a continuum, which was introduced by Paul Milgram in 1994 [MTUK95]. While purely real environments (REs) and completely virtual environments (VEs) mark the extremes of this continuum, a range of *Mixed Reality* (MR) environments, i.e. spaces that combine real and virtual elements, lies in between. Whitton et al. [WLIB05] describe fundamental challenges to enable such MR environments, including the mergence of real and virtual scene elements, the tracking of physical objects, as well as the simulation of plausible virtual-physical interactions. If the MR environment is predominantly real and only single physical objects are embedded, the overall state is denoted as AR. As opposed to this, *Augmented Virtuality* (AV) refers to a primarily VE to which some amount of real objects has been added [MTUK95].

Though Milgram's continuum provides a basis for the continuous virtualization of an environment, it is limited in the sense that it only considers augmentation, i.e. the addition of virtual content, as a possibility to increase the virtuality of objects. Related research projects, for example, by Broll et al. [HB10], demonstrate that REs can also be modified by (partially) removing real content, resulting in a *Diminished Reality* (DR) state. Analogous to the correlation of AV and AR, *Diminished Virtuality* (DV) is the equivalent to DR as it refers to a predominantly VE, from which some of the virtual elements are removed and replaced by visualizations of real-world objects. While augmentation and diminishment are opposite to each other on a conceptual level, they both increase the overall virtuality of the affected objects. This is because technologically the diminishment of real-world objects also requires the superposition of virtual content, which in this particular case represents the portion of the scene that was previously obscured by these objects.

Instead of adding virtual or removing real content, an alternative to achieve such a (partial or full) virtualization is the modulation of existing objects, for example, by using image filters in order to highlight specific features, or by virtually transforming visual properties of the physical object.

To integrate augmentation, diminishment, and modulation into a comprehensive taxonomy, Mann et al. [Man02] introduced an orthogonal axis to Milgram's continuum, called *Mediality*. However, we argue that modulation of REs also decreases the amount of unaltered real objects, and therefore can be understood as a third path to increase the virtuality of such environments. Consequently, we extend Milgram's continuum by two additional paths between reality and virtuality, as illustrated in Figure 1.1. In the following, we use the term *reality-virtuality (RV) continuum* to refer to a continuous scale between reality and virtuality, that results from the augmentation, diminishment, and modulation of physical and virtual elements within an environment. The three operations should not be understood to be mutually exclusive since they can be applied simultaneously to different parts of a physical object to increase its overall level of virtuality (see Fig. 1.1).

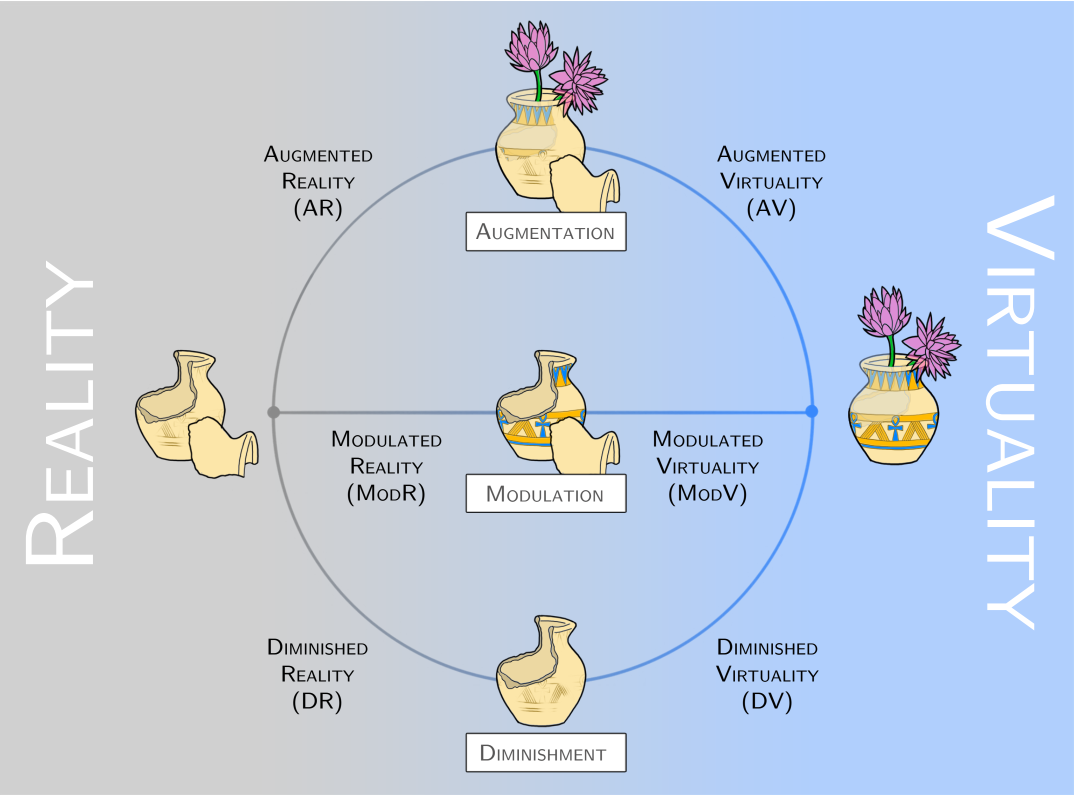


Figure 1.1.: Illustration of different states of a single physical object within a Blended Space. Parts of the object can be augmented (i.e., virtual elements are added), modulated (i.e., real elements are modified), or diminished (i.e., real elements are subtracted) to increase the overall virtuality along the RV continuum.

To illustrate the individual differences between stages of the RV continuum, we consider the example of a natural history museum with a focus on dinosaurs. In real conditions (without any virtual enhancements) the original skeleton of a dinosaur is visible. Using a combination of modulation and augmentation, selected bones can be highlighted and annotated to provide additional information on the anatomy of a dinosaur. To visualize the place of discovery, the real skeleton can be embedded in a VE which shows the archaeological site, therefore resulting in an AV state. Finally, by diminishing the real skeleton and augmenting the environment with animated dinosaurs, the user can explore an immersive virtual prehistoric environment. This example demonstrates that each stage of the RV continuum has individual advantages for the user. However, most of today's applications can be assigned to exactly one discrete stage within the continuum, and therefore do not exploit the full potential of computer-mediated reality. An environment that can transform into any stage along the RV continuum could overcome this limitation.

In the context of this thesis, we define the term *Blended Space* as an environment with one or more physical objects, which meets the following three requirements:

(R1) **RV Transformability**
By augmentation, modulation, and diminishment, all physical objects in a Blended Space can change their state along the RV continuum.

(R2) **RV Interactivity**
Within a Blended Space, real, mixed, and virtual objects can interact with each other in a physically plausible way.

(R3) **RV Seamlessness**
Transitions between RV states are seamless and can be experienced by users without requiring them to switch the display technology.

Hence, Blended Spaces can be imagined as environments, in which the virtuality of each physical object can be seamlessly increased and decreased using a slider, and in which both real and virtual objects interact naturally with each other. The overall state of a Blended Space depends on the RV states of all included objects. A Blended Space with 90% of virtual and 10% of real objects may overall be categorized to be in an AV state. If the last 10% of real objects also increase their virtuality using augmentation, diminishment, or modulation, the Blended Space transitions into a VR state.

Our notion of Blended Spaces is closely linked to the concept introduced by Benyon et al., who describe such spaces as MR environments that aim to create "a more harmonized and unified user experience" [BMA12]. To make a clear distinction to traditional MR, we impose the additional requirements (R1) to (R3) on Blended Spaces, regarding their temporal behavior as well as interactions between included objects. This leads to a slightly different concept in comparison to the proposed idea of Benyon et al. since Blended Spaces in the context of this thesis refer to environments that cover the entire RV continuum rather than just the subscale of MR environments.

To clarify the proposed requirements (R1) to (R3), we outline several related projects that have an overlap with the definition of Blended Spaces but do not satisfy at least one of the stated requirements.

Hybrid Spaces [RH17] suggested by Roo and Hachet aim at covering the entire RV continuum, including the real, AR, and VR stage. In the context of architectural visualization, users can explore a mock-up either with or without an HMD. In the AR condition, the physical 3D model can be augmented with 2D textures via projections, whereas the VR condition displays additional environmental information and allows for an egocentric 3D view of the scene (cf. R1). However, as transitions between the real, AR, and VR states require users to change the display technology (projectors vs. HMD), R3 of the definition of Blended Spaces is not satisfied.

Smart Terrain [Jac13] is a feature of the Vuforia SDK, which facilitates the development of AR applications on mobile devices. It allows users to scan their RE with a smartphone or tablet that is equipped with a depth camera. Based on the reconstructed environment, Smart Terrain computes matching virtual content that is attached to physical anchor points. The official example application showcases a tower defense game that demonstrates how the RV state of each object can be changed gradually through 3D animations, therefore supporting (R1) and partially (R3). In contrast to the previously presented *Hybrid Spaces*, users are not required to switch the output device in order to experience real, virtual, and mixed environments, however, abrupt changes of the RV state of an object can occur. This is because of the limited display size of smartphones, which only cover a small portion of the user's visual field when held at arm's length. By moving the smartphone, users can navigate through a Blended Space that is essentially larger than the display [BR05], however, this causes objects to abruptly change their RV state between virtual (when seen through the display), and real (when the display does not cover the object).

Other approaches bridge the gap between multiple stages of the RV continuum by separating a single room into distinct real, AR, AV, and VR regions. While this approach allows users to transition between RV stages just by moving around the room, it does not fulfill requirement R1 since not each physical scene object can pass through different changes of state along the RV continuum. The following two examples illustrate the concept and how it differs from Blended Spaces.

A Continuum of VE Experiences [DRHL$^+$03] was designed by Davis et al. to model multiple levels of immersion within the same environment. In a real-world room, selected physical objects and walls are covered with retroreflective material and therefore can serve as projection displays for virtual content. If users direct their gaze to an unaltered part of the real world, they experience a RE. By looking at a real physical table with a projected remote user, the application implicitly switches to an AR condition. Finally, if users are approaching one of the virtually enhanced walls, they immerse themselves in an AV or VR condition, depending on whether other physical objects are still visible in the field of view. In this manner, users can transition seamlessly between multiple stages of the RV continuum (R3) by navigating to different parts of their environment. This is in contradistinction to Blended Spaces, which are characterized by the RV transformability of all included objects (R1).

Traversable Interfaces [KSBG00] follow a similar concept since users dynamically relocate themselves along the RV continuum by moving through their environment. As in the previous example, each section of the environment is either constantly real or virtual over the entire experience. Unlike the above-mentioned project, all sections are separated through physical boundaries such as fabric curtains, water curtains, or a sliding door, which simultaneously serve as projection surfaces. By passing through these boundaries, users get the illusion of entering a different part of the local environment, which may also represent a new stage on the RV continuum. Again, scene objects cannot pass through different changes of state but instead are permanently assigned to a real, AR, AV, or VR section of the environment.

After showcasing several related projects that, to some extent, differ from the definition of Blended Spaces, we finally also want to illustrate an existing example that supports each of the three requirements.

The MagicBook [BKP01] was developed by Billinghurst et al. and specifically aimed at covering the entire RV continuum, including the real, AR, and VR stage. The real stage is represented by an ordinary book with text and illustrations that can be viewed without additional technology, or with a video see-through HMD (for details on the technology see Sec. 2.1.1). When specific patterns are recognized in the camera feed, matching virtual 3D models appear out of the pages. Without changing the output medium, the displayed virtual models can be explored from an egocentric view by switching to VR mode. Therefore, both RV transformability and seamlessness are supported by the MagicBook. The last requirement, RV interactivity, is satisfied in the sense that physical pages and virtual objects behave as in a traditional pop-up book. There are only a few interactions between objects, though, including the appearance, disappearance, and rotation of virtual scenes according to the currently visible book page.

As the MagicBook demonstrates, a Blended Space could be achieved using an AR display that covers the user's entire field of view and is able to overlay an arbitrary amount of virtual content. We pursue this approach in the next chapter by comparing a range of AR hardware by means of technical, usability, and social factors, intending to find the most suitable basic technology for a Blended Space.

2

Chapter 2.

Enabling Technology for Blended Spaces

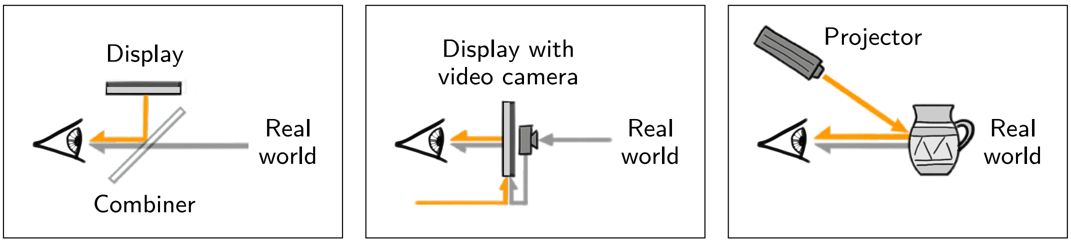
A basic technology to enable Blended Spaces has to be able to present all different stages within the RV continuum while supporting seamless transitions between them. Many of the conventional VR/AR displays turn out to be unsuitable, as they lack full coverage of the continuum. In the following, we take a glance at different display types and evaluate the most promising candidates regarding various technological, human, and economic factors. Finally, the selected basic technology will be presented in higher detail.

2.1. AR Display Types

To create a Blended Space, virtual and real content have to be combined using a single display. This functionality is already well-known from the field of AR, and therefore, established AR displays are a promising starting point for the process of selecting an appropriate basic technology for Blended Spaces. AR displays can be classified according to different criteria, including (i) the method of augmentation [SH16], and (ii) the display location [BR05]. As both classifications have a strong impact on the characteristics of the final application, we first want to take a look at the features of each display category before discussing the pros and cons for creating Blended Spaces.

2.1.1. Method of Augmentation

There are two basic methods to present virtual content to the user. First, a lens can be placed between the user's eyes and the environment; either in the form of an optical combiner or a video display. Second, the virtual content can be projected directly onto physical objects within the 3D space around the user. The three resulting techniques — optical see-through, video see-through, and projection-based — are described in the following sections (cf. Schmalstieg et al. [SH16]).



(a) Optical see-through **(b)** Video see-through **(c)** Projection-based

Figure 2.1.: Three technologies to merge real content (gray arrows) with virtual content (orange arrows).

Optical See-Through Displays The central part of most wearable optical see-through (OST) displays is an optical combiner with both transmissive and reflective characteristics, for example, a semi-reflective mirror or a diffractive waveguide [Auk16]. As the combiner allows a sufficient amount of light from the environment to pass, users can still see their surroundings directly through the lens. Virtual content is reflected from a projector unit that is placed on the sides or above the lens, and therefore appears superimposed to the view of the real world (see Fig. 2.1a). While this technology is suited for small screen sizes such as in glasses, larger displays can be implemented with transparent LCDs or OLEDs.

Video See-Through Displays For video see-through (VST) devices, the partly transmissive lens is replaced by some sort of video screen. Due to this modification, the virtual content does not have to be reflected to reach the user's eye, as it is displayed directly in the line of sight. However, as a direct view of the surroundings is blocked by the display, a video image of the real world has to be recorded to be overlayed by the virtual content. This is usually done by an RGB camera that is positioned close to the display to reduce the offset between its optical axis and the user's viewing direction to a minimum. The spatial arrangement of components is displayed in Figure 2.1b.

Projection-Based Technologies Unlike the previous two approaches, projection-based displays do not place a lens between users and their environment to combine virtual and real content. Instead, virtual objects are projected directly onto the surfaces of the physical surroundings, using single or multiple projectors. Apart from lightweight 3D glasses to perceive stereoscopic content, users do not have to wear any optics since the display is usually decoupled from the user's head (see Fig. 2.1c). There are some exceptions, however, including projection-based glasses such as *castAR* [Cas], and hand-held projectors such as *WALKABOUT Projection* [Wal] and *Lumen* [San].

2.1.2. Display Location

AR displays can also be classified according to their proximity to the user [BR05]. Displays can be placed directly in front of the user's eyes (i.e., *head-mounted displays*, for short *HMDs*), in arm-length distance (i.e., *hand-held displays*), or at an arbitrary position inside the room (i.e., *spatial displays*). Transitions between the classes are possible, for example, by installing smartphones, which are usually used as hand-held devices, in appropriate viewers such as *Samsung Gear VR* or *Google Cardboard*, and therefore turn them into HMDs.

2.2. Technological, Human, and Economic Factors for Building Blended Spaces

Based on the resulting nine categories of AR displays, we first consider the three requirements a technology for a Blended Space has to fulfill and discuss the displays' public availability afterwards.

Table 2.1.: Overview of nine categories of AR displays, each rated in terms of whether it (top left) complies with the definition of Blended Spaces, and (bottom right) is publicly available.

		Method of Augmentation		
		Optical see-through	Video see-through	Projection-based
Display Location	Head-mounted	currently do not meet (R3) publicly available	meet requirements (R1)–(R3) publicly available	meet requirements (R1)–(R3) still under research
	Hand-held	do not meet (R3) by design still under research	do not meet (R3) by design publicly available	will mostly not meet (R3) still under research
	Spatial	usually do not meet (R1) or (R3) available on request	usually do not meet (R1) or (R3) available on request	meet requirements (R1)–(R3) publicly available

As all AR displays are designed to compose real and virtual content, the RV transformation of a single physical object can be supported naturally by each of them. However, due to technological restrictions, some of the display types limit the amount of RV transformable objects within the scene. Regarding spatial OST and VST technology, each physical scene object that should be virtually extensible has to be covered by at least one display. Since spatial displays are typically arranged at fixed positions within the environment and hinder the direct interaction between users and the object, they are less suitable for the implementation of Blended Spaces with a large number of physical objects.

All of the remaining AR displays are capable of transforming any number of physical objects in the environment but do not necessarily support seamless transitions between their RV states. As discussed for the example of *Smart Terrain* in Chapter 1, hand-held OST and VST devices such as smartphones or tablets have a limited screen size, and therefore only cover a small section of the physical environment when positioned at arm's length. While virtual elements can be added to the entire physical scene, only the subset in the display's field of view (FOV) is visible. When users change their viewpoint by moving the display device, virtual elements are clipped at the display's edges, and therefore (R3) is not satisfied. The same behavior can be observed for current implementations of OST-HMDs such as *Microsoft HoloLens* and *Meta* glasses. In contrast to hand-held devices, the narrow FOV of OST-HMDs is not a conceptual limitation but presumably will improve with future technological advances [BCL15]. Furthermore, smartphones could still be used to enable Blended Spaces by installing them in appropriate viewers such as *Samsung Gear VR*, or *Google Cardboard*. In this case, they would be categorized as head-worn devices due to the different display location.

In addition to these conceptual considerations of different AR displays, we also discuss their availability. All components that were used to build Blended Spaces as described in the following chapters are publicly available, and therefore can be used by any developer who plans to recreate a similar setup. We therefore excluded projected head-worn devices such as *castAR* [Cas] and hand-held projectors such as *WALKABOUT Projection* [Wal] from the list of potential basic technologies, since both are still in a prototype stage.

As a result, we focus on the three display technologies, for which the conceptual design complies best with the requirements of Blended Spaces, and which are publicly available nowadays:

- Optical see-through HMD (OST-HMD)
- Video See-through HMD (VST-HMD)
- Spatial projection-based AR (SAR)

To choose a final basic technology for Blended Spaces from the list of candidates, we consider a set of technological, human, and economic factors obtained from the literature (e.g., Schmalstieg et al. [SH16]). While the full comparison can be found in the appendix, we want to highlight the key arguments that finally led to the decision against both forms of HMDs in favor of SAR.

As we envision Blended Spaces being utilized in various domains, not necessarily by users with a technical background, aspects such as form factor, usability, and acceptance are of prime importance. SAR setups detach the display technology from the user and therefore get along with a minimum level of user instrumentation. This not only improves the system's ergonomics but also facilitates non-technical users to participate in the experience; a factor that particularly public applications could benefit from. SAR environments can be shared spaces since the same physical and virtual objects can be viewed by all users simultaneously. Nevertheless, multi-user support in SAR spaces is limited due to the view dependency of perspectively projected 3D content. Approaches that address this limitation are an essential part of this thesis and therefore are discussed in separate chapters (see Part IV). From a

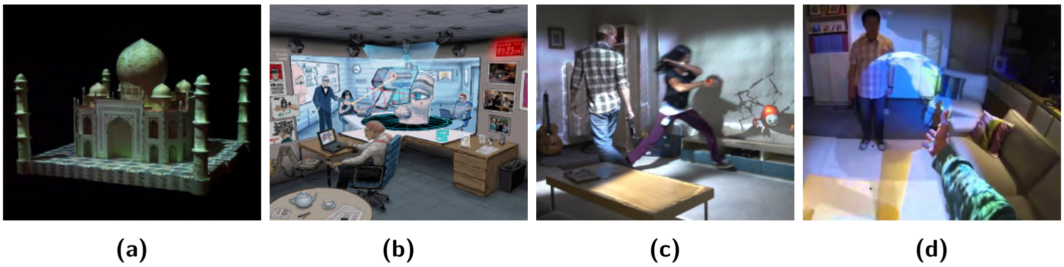


Figure 2.2.: Four examples of related SAR projects: (a) *Shader Lamps* [RWLB01], (b) *Office of the Future* [RWC+98], (c) *RoomAlive* [JSM+14], and (d) *Mano-a-Mano* [BWZ14].

technical viewpoint, SAR setups stand out for their capability to reproduce the human natural FOV. By installing multiple projectors, the users' entire surroundings can be augmented at once. Since the virtual content is projected onto physical surfaces in the environment rather than a display right in front of the user's eyes, the distance between virtual objects and display can be kept short in comparison to HMDs, and perceptual conflicts, as well as their negative effects such as eye fatigue or headache, can be reduced (see Sec. 8.4.2 for details).

Due to these advantages, we select SAR for realizing our first prototypes of Blended Spaces. Nevertheless, future improvements of the HMD ergonomics and optics have the potential to make head-worn or even retinal devices (i.e., smart contact lenses) a feasible solution for implementing Blended Spaces. In particular, we observe a trend towards on-body ubiquitous technology such as smartphones or smartwatches, which may apply to future AR displays as well. The mass distribution of HMDs would boost the development of corresponding applications, and concurrently reduce the need for shared AR setups that are available to the public.

2.3. Stereoscopic Projection-Based Spatial AR

Projection-based spatial AR (SAR), also known as *projection mapping*, combines the advantages of real-world augmentation and tangible user interfaces by projecting computer-generated images directly onto the surface of physical 3D objects [RWF98, BK03]. This technique allows users to perceive objects naturally, for example by walking around them or touching them [KSF10, Sch01], while the objects' appearance can be changed in various ways [Kau03]. Such integration of virtual content into our real-world environment found its way from projects in research and industry to our everyday life with applications including art, entertainment, education, and home automation (for a collection of examples, see [Jon]).

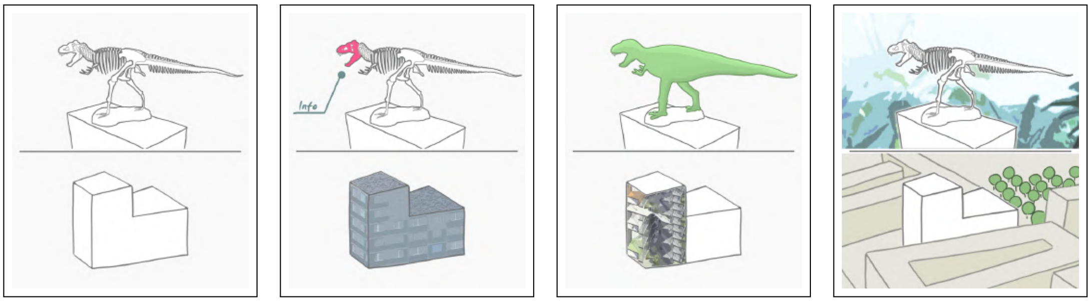
At the end of the last century, Raskar et al. [RCWS98] demonstrated a prototype implementation of SAR by overlaying non-planar surfaces with perspectively correct renderings of a virtual 3D scene using a multi-projector system. Since then, several SAR setups have been introduced and revised (see Fig. 2.2), for example, *Shader Lamps* [RWLB01], *Office of the Future* [RWC+98], and *Emancipated Pixels* [UUI99]. In most of these setups, 2D textures were projected onto a 3D geometry, i. e., the virtual content is aligned with the physical surfaces. More recently, Jones et al. and Benko et al. introduced the *RoomAlive* [JSM$^+$14], *IllumiRoom* [JBOW13] and *Mano-a-Mano* [BWZ14] setups, in which virtual objects can be displayed monoscopically at any arbitrary 3D location. For further research examples, we refer to an article by Marner et al. [MSWT14], that surveys promising research directions and applications of SAR.

Beyond these research projects, SAR has become popular for public installations such as art exhibitions and theme parks, as well as for smaller-scale projects through a range of available consumer products. At CES 2014, *Whirlpool* was demonstrating a projected interactive cooktop that supports users with preparing a meal while connecting them to online services such as different social media feeds [Whi14]. Affordable projectors such as the *Sony Xperia Touch* transform any planar surface into an interactive touch screen and therefore open up new possibilities for using SAR, for example for tabletop games [Xpe]. The developers of *Lightform* [Lig19] provide an all-in-one solution with a projector, depth camera, and computer to turn any 3D geometry into a display surface, that can be used for decor, shop windows, or events among other applications. As we can only present a small proportion of available SAR examples, we refer to *Projection Mapping Central* [Jon], which collects current projects demonstrating the creative and functional potential of SAR for everyday applications.

Most of the previously mentioned setups do not provide stereoscopic display but rather rely on monoscopic cues such as view-dependent perspective to convey the sense of a spatial presence of the virtual object. Using stereoscopic projection, virtual 3D objects can be spatially embedded into the real-world surroundings of the user, creating a more realistic illusion of co-existing virtual and physical objects. In the course of this thesis, we will utilize both monoscopic and stereoscopic projections, depending on the requirements of the current use case. The research focus, however, lies on the latter technique, as it raises interesting questions regarding the perception of and interaction with the scene, that cannot be easily answered by applying results from conventional AR projects.

3 Chapter 3. Applications of Blended Spaces

Due to the advantages of the underlying basic technology of projection-based SAR, Blended Spaces may be used for a variety of applications. In particular, experiences in public spaces or, generally speaking, experiences that involve multiple users could benefit from the seamless, social and connective character of SAR technology. Figure 3.1 outlines two example cases: (i) the presentation of a real exhibit in a museum, and (ii) the discussion of a physical block model as part of the architectural design process. By using Blended Spaces, the physical objects can be augmented with additional virtual content to highlight specific details, change the objects' appearance or even show them in their natural context.



(a) original **(b)** 2D details **(c)** 3D details **(d)** 2D/3D context

Figure 3.1.: Illustration of four different states of the same (top) physical exhibit and (bottom) architectural block model along the RV continuum.

In the following two sections, we outline related projects in the context of both application fields and address the question, why a Blended Space may outperform traditional virtual and augmented environments for the given purposes. Apart from these examples, other applications could benefit from the presented setup as well, including but not restricted to training scenarios, rapid prototyping, tourism, and the real estate market.

3.1. Exhibitions

Though museums traditionally rely on physical 3D objects, they also benefit from the integration of virtual content (for an overview see [Rou01]). In the museum context, VR environments can be used in several ways, including the immersive display of an artifact's original environment, the exploration of places that are difficult to access or even non-existent, the possibility to encounter historical figures, or the visualization of and interaction with abstract concepts from natural sciences. For example, the *Kramer Museum* [Kre] is accessible exclusively in VR and therefore opens up to visitors from all over the globe without the requirement to be physically present.

Based on the variety of possible applications, the question arises whether museums still need real physical artifacts; an idea that was also debated in an essay of Tisdale [Tis11]. However, several surveys conducted by Reach Advisors revealed that meaningful experiences in museums are predominantly connected to original artifacts [Wil15]. An explanation of this outcome is the people's wish for authenticity as the digital revolution proceeds. While current technology allows the creation of continuously improving multisensory experiences by providing visual, auditory and even haptic feedback, the emotional connection between the visitor and an original exhibit cannot be reproduced in its entirety with a virtual substitute. For example, the *National Air and Space Museum* displays a touchable slice of real moon rock, which attracts thousands of visitors every day and evokes emotional responses that are unlikely for a virtual replica [Cra02]. Therefore, many experts agree that physical artifacts are still of great value to create emotionally engaging experiences in museums [Dud18, Tis11, Wil15].

Besides collections that either rely on original or virtual objects only, the combination of both types in a MR setup has the potential to overcome the individual limitations. For example, museums show objects that usually have their origin in a different time and a different place. By bringing the exhibit to the museum and its visitors, the link between the real object and its context gets lost. Virtual augmentations could (re-)establish those links by adding contextual information to the object. Example projects include restorations of the original hues at historical sites, such as the *Church Saint Climent de Taüll* [Pan] as well as the *Ara Pacis Altar* in Rome [Ara14]. Both applications are based on SAR to project original textures onto the historic walls. Another AR project that is linked to an actual historical site, is the *Archeoguide* [VKT$^+$01]. It allows visitors to overlay a ruined monument with a virtually reconstructed 3D model of its original state. The authors argue that by using AR the current physical state of the archaeological site is preserved (in contrast to a real reconstruction), and users can establish visual contact with the natural surroundings of the monument (in contrast to a VR display).

Another aspect of creating AR experiences is the rising demand for interactivity, that can partially be attributed to the increasing percentage of museum visitors that are digital natives, meaning that they grew up with digital technology such as computers or the Internet [Ste16]. While physical artifacts provide haptic feedback by nature, most museums prefer to conserve and protect them from inappropriate handling. One approach to address this issue is the usage of virtual content that complements a physical exhibit; in other words, the usage of an AR system. There is a variety of exhibitions with interactive AR installations, including *Exploring Pueblo Pottery* [Eti18], *Interactive Terracotta Warrior* [Cas15], and *The Revealing Flashlight* [RRL+14]. Each of these three examples is using SAR to project varying virtual content onto physical exhibits, while the users have control over the currently displayed state.

To summarize, most museum visitors still want to be able to experience the real exhibit with all its material qualities, however, they could also benefit from other views that show additional information or even overlay the physical object completely. Blended Spaces could cope with these demands as they allow the transition between different stages of the RV continuum.

3.2. Architecture

Architecture is another application field, in which physical objects have a long tradition and are still in use, despite a variety of available computer-aided design (CAD) tools. Since design proposals and revisions have to be coordinated between multiple stakeholders with different roles and objectives, collaboration is an essential aspect of architectural design processes. Physical block models are widely used to support collaborative work, as they are efficient tools for understanding design ideas and communicating them to involved parties such as building owners, investors, occupants, maintenance engineers, or other authorities. A rough model can be constructed quickly, using materials such as cardboard, wood, or foam. These models can be used in the early stages of a design process to obtain a sense of proportions, form, and the general structure of a building. However, despite the benefits of physical tangible 3D models, they are also subject to limitations regarding their flexibility. A large part of an architectural design process is characterized by the comparative consideration of alternative configurations, for example in terms of the building's layout, materials, and furnishing. An interactive switch-over between different options could support this process of decision-making, though it requires a great manual effort to be implemented in multiple versions of the same physical model. Moreover, questions about spatial configurations that are decisive for investors such as *"Is the street visible from the rooftop garden?"* cannot be answered easily using a rough block model.

Immersive virtual environments (IVEs) have the potential to overcome these limitations since they allow for dynamic viewpoints, interactive choices, and a more realistic look in comparison to simplified block models. In [PS02] several trends using VR within the architectural domain are discussed. In particular, IVEs can provide architects and customers a spatial impression of not only a building's room layout but also its interior. The exploration can be implemented in the form of immersive walkthroughs, which present virtual 3D models at real scale from an ego-centric perspective. To allow users to explore a large virtual architectural model within a limited physical workspace, the *ArchExplore* project [BSH09] implemented a redirected walking strategy. Besides, the authors introduced portals to connect different virtual locations representing alternative design proposals. Furthermore, a lot of effort has been spent on the conceptual design and realization of IVEs that aim to support architects in designing and constructing 3D buildings [AEI03, AMR06, RFK+98]. Such approaches intend to provide users with the functionality of typical CAD tools within an IVE allowing a more natural and intuitive interaction.

Besides using traditional block models and IVEs, some tasks during the architectural design process can be facilitated by combining both physical and virtual objects. For example, Bruder et al. allowed users to see their own body as well as a variety of real-world tools, such as notepads and rulers, while immersed in the VE [BSVH10]. The developed AV studio for architectural exploration is based on a chroma keying approach to embed real objects in a VE. Furthermore, several AR setups that augment physical scale models with virtual content have been introduced [DSD+02, WD13]. The users can view the virtual augmentations by wearing VST- or OST-HMDs as described in Section 2.1.

As the previous examples show, each architectural design decision places special emphasis on different characteristics of the planned structure, for example, its 3D form, surroundings, floor plan, materials, or furniture. Using Blended Spaces, each of these aspects can be displayed in the best-suited mode of the RV continuum. Furthermore, by transitioning between views, multiple relevant factors can be involved without breaking the flow of a discussion.

Due to the positive aspects of Blended Spaces for both exhibitions and the architectural domain, we will focus on these application fields repeatedly throughout the following chapters.

4 Chapter 4. Research Questions

The conception of a Blended Space entails a variety of challenges, including technological, perceptual, and social factors. So far, only little research regarding fundamental differences between stereoscopic SAR and traditional VR/AR setups has been conducted. For that reason, this thesis focuses on the development of guidelines for the design and usage of Blended Spaces.

Next in this thesis, Part II discusses technical and mathematical fundamentals that are required to implement a Blended Space using SAR technology. In this context, we conceptualize an extended CAVE setup that is specifically adapted to work for all states a Blended Space can have. The main idea of Blended Spaces, as well as an implementation of a specific projection mapping approach, are introduced in the following publications:

[SS17] S. Schmidt and F. Steinicke. A Projection-Based Augmented Reality Setup for Blended Museum Experiences. In *Proceedings of the International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments (ICAT-EGVE) (Poster)*, pages 5–6, 2017

[SDBS15] S. Schmidt, S. Dähn, G. Bruder, and F. Steinicke. A Mobile Interactive Mapping Application for Spatial Augmented Reality On The Fly. In *Proceedings of the GI Workshop on Virtual and Augmented Reality (GI VR/AR)*, pages 1–9, 2015

In particular, we highlight related algorithms that are used to build a common basis for all involved system components and discuss modifications that are required to apply these algorithms to the specific projection-based environment. Based on the calibrated setup, we provide an overview of the rendering processes that are used in Blended Spaces.

From a technical perspective, we emphasize the coverage of the entire RV continuum, that can be addressed by a single Blended Space. Concerning the perception of the Blended Space, the monoscopic and stereoscopic projection on surfaces with different shapes, depths, and orientations opens up research questions that cannot be answered easily by applying results from studies in traditional AR or VR. Therefore, Part III of the dissertation addresses perceptual issues with a series of empirical studies. The results described in this thesis are based on our following publications:

[SBS16] S. Schmidt, G. Bruder, and F. Steinicke. Illusion of Depth in Spatial Augmented Reality. In *Proceedings of the IEEE VR Workshop on Perceptual and Cognitive Issues in AR (PERCAR)*, pages 1–6, 2016

[SBS17a] S. Schmidt, G. Bruder, and F. Steinicke. A Pilot Study of Altering Depth Perception with Projection-Based Illusions. In *Proceedings of the International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments (ICAT-EGVE) (Poster)*, pages 33–34, 2017

[SBS17b] S. Schmidt, G. Bruder, and F. Steinicke. Moving Towards Consistent Depth Perception in Stereoscopic Projection-Based Augmented Reality. In *Proceedings of the International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments (ICAT-EGVE)*, pages 161–168, 2017

[SBS20] S. Schmidt, G. Bruder, and F. Steinicke. Depth Perception and Manipulation in Projection-Based Spatial Augmented Reality. *Presence: Teleoperators & Virtual Environments*, 27(2):242–256, 2020

Part IV focuses on the social aspects of Blended Spaces. The proposed implementation of Blended Spaces with a separation between display and user opens up new possibilities to provide a shared space for multiple users and simultaneously changes the perception of and interaction with other virtual humans in the same space. In this context, we published research papers in the scope of HCI/VR conferences and journals. The chapters of Part IV are mainly based on the following publications, which address the interaction between users and real as well as virtual cooperation partners in Blended Spaces:

[SBS15] S. Schmidt, G. Bruder, and F. Steinicke. A Layer-based 3D Virtual Environment for Architectural Collaboration. In *Proceedings of the EuroVR International Conference*, pages 79–84, 2015

[SITS18] S. Schmidt, A. Irlitti, B. Thomas, and F. Steinicke. Floor-Projected Guidance Cues for Collaborative Exploration of Spatial Augmented Reality Setups. In *Proceedings of the ACM International Conference on Interactive Surfaces and Spaces (ISS)*, pages 279–289, 2018

[SBS18] S. Schmidt, G Bruder, and F. Steinicke. Effects of Embodiment on Generic and Content-Specific Intelligent Virtual Agents as Exhibition Guides. In *Proceedings of the International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments (ICAT-EGVE)*, pages 13–20, 2018

[NCSS18] D. Neves Coelho, S. Schmidt, and F. Steinicke. Kategorisierung und Evaluierung von Transitionen für CAVE Umgebungen. In *Proceedings of the GI Workshop on Virtual and Augmented Reality (GI VR/AR)*, pages 169–176, 2018

[SBS19] S. Schmidt, G. Bruder, and F. Steinicke. Effects of Virtual Agent and Object Representation on Experiencing Exhibited Artifacts. *Elsevier Computers & Graphics*, 83:1–10, 2019

[SAS19] S. Schmidt, O. Ariza, and F. Steinicke. Blended Agents: Manipulation of Physical Objects within Mixed Reality Environments and Beyond. In *Proceedings of the ACM Symposium on Spatial User Interaction (SUI)*, pages 1–10, 2019

In addition to the mentioned publications, the following papers were published during the PhD studies:

[Sch16] S. Schmidt. Interaction Techniques for Spatial Augmented Reality Setups. In *Proceedings of the IEEE Conference on Virtual Reality (VR Doctoral Consortium)*, 2016

[HSS17] F. Heinecke, S. Schmidt, and F. Steinicke. Präattentive Wahrnehmung von Farbe bei der Gestaltrichtung Flat Design. In *Proceedings of the Mensch und Computer (MuC)*, pages 391–394, 2017

They will be referenced at appropriate passages, without a detailed elaboration.

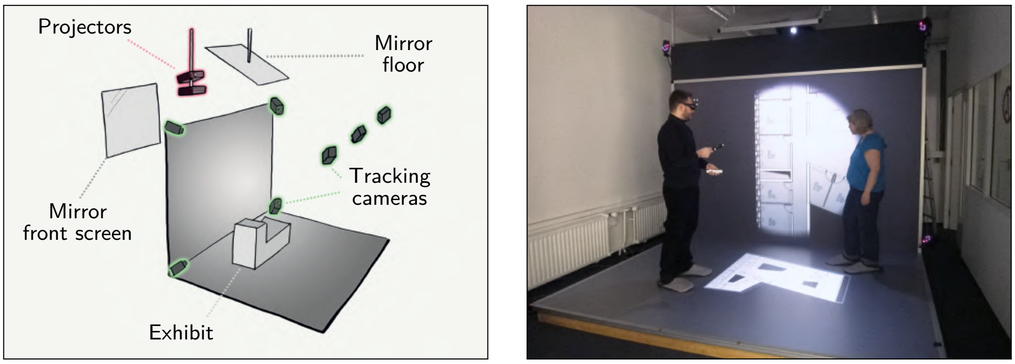
Part II.

Technological Fundamentals for Blended Spaces

5 Chapter 5. Hardware Setup

The implementation of SAR-based Blended Spaces does require simultaneous projection on individual objects and their surroundings. Thus, the user can be immersed either in a partially augmented or a purely virtual environment. The *Cave automatic virtual environment* (CAVE), introduced by Cruz-Neira et al. at the Showroom of ACM SIGGRAPH in 1992, displays a general approach for projection-based virtualization of an entire room [CNSD93]. In its original definition, a CAVE is described as a “nonintrusive easy-to-learn high-resolution virtual reality interface” [CNSD+92]. The first prototype involved two walls and was later extended by a third wall and the floor [CNSD+92]. In 1998, *COSMOS*, the first six-sided version of the CAVE, was demonstrated by Yamada [Yam98]. As an extension of the traditional CAVE, Jones et al. presented *RoomAlive* [JSM+14], which is a prototype that uses projectors to transform arbitrary rooms into virtually augmented environments, in particular, for entertainment experiences.

In the scope of this PhD thesis, three different types of CAVEs were implemented as illustrated in Figures 5.1 to 5.3. Each of them was built with regard to individual structural conditions of the surrounding environment and therefore features a different projection technology, layout, and system architecture. In the following chapter, we will discuss the setup of each CAVE to explain different design options for Blended Spaces. These options involve the number of projection walls (≥ 1), the positioning of projectors in relation to the walls (front vs. rear), the stereoscopy (monoscopic vs. stereoscopic) and view dependency (head tracking vs. no head tracking) of the projected content, as well as the projector light source (lamp vs. laser). An overview of the used software and hardware for each of the three case studies can be found in the appendix.



(a)

(b)

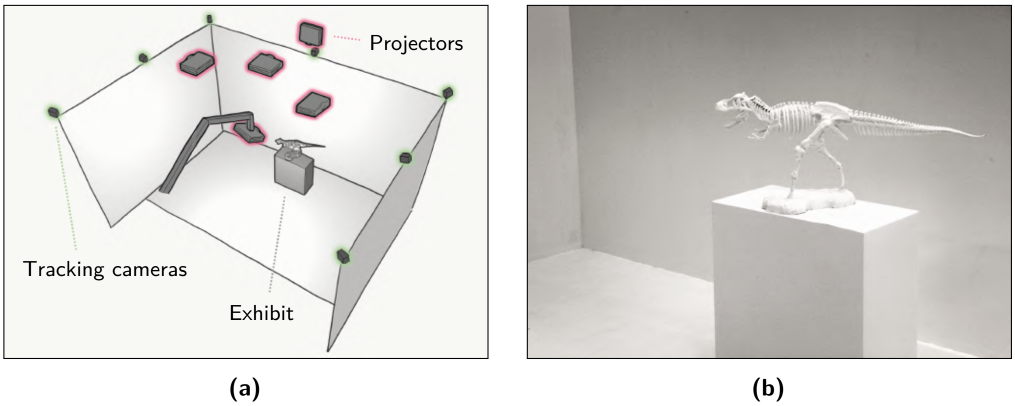
Figure 5.1.: The L-Shape setup with (a) schematic illustration of different components, and (b) photo of an example architectural application.

5.1. Case Study: L-Shape

Our first setup consists of two orthogonal projection screens, forming the shape of a capital ‘L’ (see Fig. 5.1a). While this layout is not fully immersive by design, it allows the augmentation of a physical object as well as its immediate vicinity. For example, this could be used for staging a physical exhibit in a museum with limited space. The L-Shape illustrated in Figure 5.1b consists of a floor screen that measures $2.24 \times 3.0m$ and a front screen with $2.02 \times 3.0m$. Two ProjectionDesign projectors with SXGA+, one for the front and one for the floor screen, are arranged behind the L-Shape at a height of about 2 meters.

The choice of front or rear projection depends on several factors, including the available space behind, below or above the projection screens. In our L-Shape setup, 2 meters clearance behind the CAVE allowed for a rear projection of the front screen, while the floor is projected from above. To reduce the required distance between projectors and screens, two mirrors are used that reflect the images to the back of the front screen and downward to the floor, as shown in Figure 5.1. The main benefit of this setup is that users do not cast any shadows on the walls, even when they get close to the projections.

As we intended to use the L-Shape both for stereoscopically projected objects that appear in front of or behind the projection screen (see [SBS15]) and for monoscopic augmentations of physical object surfaces (see [SDBS15]), the used projectors support both modes. For stereoscopic display, users wear active shutter glasses that synchronize with the projectors using DLP link technology. For head tracking, five retroreflective markers are attached to the glasses. The markers are arranged in a certain geometry to build a unique target, whose position and orientation can be tracked by an optical tracking system. For good tracking coverage and robust tracking under occlusions, we use seven infrared (IR) cameras, which are mounted to the ceiling and corners of the L-Shape. More details on the tracking system as well as the projectors and working station can be found in Table A.3 in the appendix.



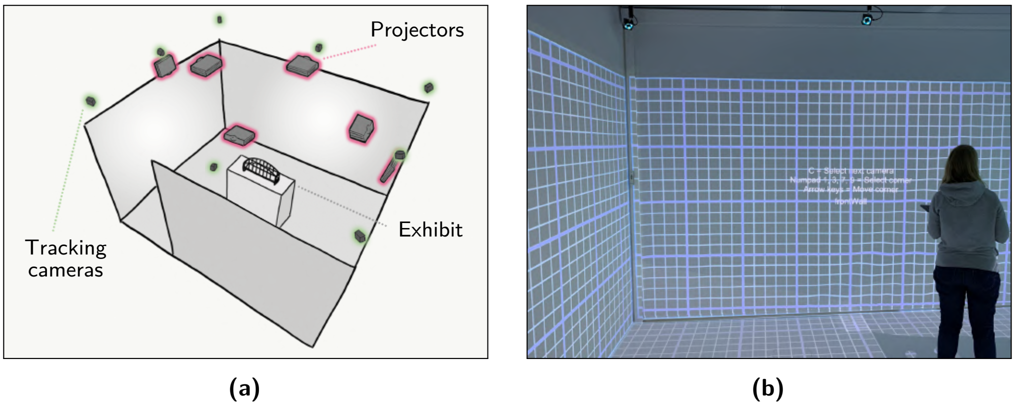
(a) (b)

Figure 5.2.: The 4-sided CAVE setup with (a) schematic illustration of different components, and (b) photo of an example museum application.

5.2. Case Study: 4-Sided CAVE

The starting point for our second implementation of a Blended Space is a cost-efficient CAVE, which includes three walls and the floor as projection surfaces (see Fig. 5.2). We used drywall panels to build a room that measures $3.15 \times 4.2 \times 2.36m$. The CAVE is equipped with four off-the-shelf Full HD Optoma projectors that are mounted on the ceiling (details on the used hardware can be found in Table A.3 in the appendix). Virtual content is displayed using front projection. In comparison to rear projection, the screens reflect most of the incoming light, resulting in a lower loss of brightness. Furthermore, less space is required to set up a front projection screen. To reduce the shadows thrown by users, the wall projectors are placed close to the screens, at a distance of around $0.8m$ (see Fig. 5.2a).

One or multiple exhibits can be placed inside the final CAVE. Using the floor projector to illuminate the exhibits from above increases the probability of casting shadows that interfere with the projection. To avoid such shadows and increase the pixel density, we use a fifth projector that is mounted on a swiveling arm. After an initial calibration of this projector within the CAVE, its pose can be tracked by the same system used for head tracking (see Sec. 6.1). Since the projector is not fixed, it can be individually adjusted for specific installations. For an optimal positioning of the mobile projector, some general constraints should be considered. In terms of the distance between the physical display surface and the projector, the projector frustum should barely cover the surface to ensure a high effective pixel resolution. Furthermore, the direction of the projector should match the user view for the key use cases. This reduces the probability of the projector to dazzle the user as well as the visibility of shadows, as they are cast behind the physical object. Finally, the height of the projector should be chosen above the human average height to prevent collisions between the user and the projector.



(a)

(b)

Figure 5.3.: The blended office setup with (a) schematic illustration of different components, and (b) photo of a calibration pattern.

Besides adding projection screens and projectors, the 4-sided CAVE also incorporates new technology for the stereo shutter glasses in comparison to the L-Shape. The DLP link technology, which was used in the L-Shape, displays a flash of white light between each projected frame to synchronize the shutter lenses with the projected content. This technique requires a direct line of sight between the user and the projection screen. Furthermore, synchronization may be lost in bright scenes. Both problems can be avoided by using stereo glasses that receive synchronization signals from a sender via Bluetooth.

5.3. Case Study: Blended Office

The third illustrated example can be considered as an extension to the previously explained 4-sided CAVE. We glance at this setup for two reasons. First, unlike the CAVE in the second case study, this Blended Space was integrated into a common office room. Second, due to a layout with four walls and the floor as projection displays, at least six projectors are required to cover all surfaces with reasonable quality. These constraints create new technological challenges regarding the projection as well as the synchronization between projectors. First of all, front projection is the only reasonable option if the existing walls should be used as projection surfaces. To connect and synchronize more than four displays, a cluster of two or more graphics cards is necessary. Furthermore, if the blended office is integrated into the daily workflow of users, the used projectors should be reliable and low-maintenance. A laser light source can make a significant contribution to this goal, as it loses operational brightness much more slowly (after more than 10,000 hours) than a projector lamp with lifetimes of 1,000 to 1,500 hours [JM09]. A possible hardware configuration that supports all of the mentioned features is listed in the appendix.

6 Chapter 6. Calibration

The overall goal of the calibration is to align physical surfaces with corresponding features of the projected computer-generated imagery to realize different levels of augmentation and diminishment, as illustrated in Figure 1.1. This alignment is usually denoted as the *geometric registration*. It can be approached in two different ways, depending on the spatial relationship between the virtual and real-world geometry.

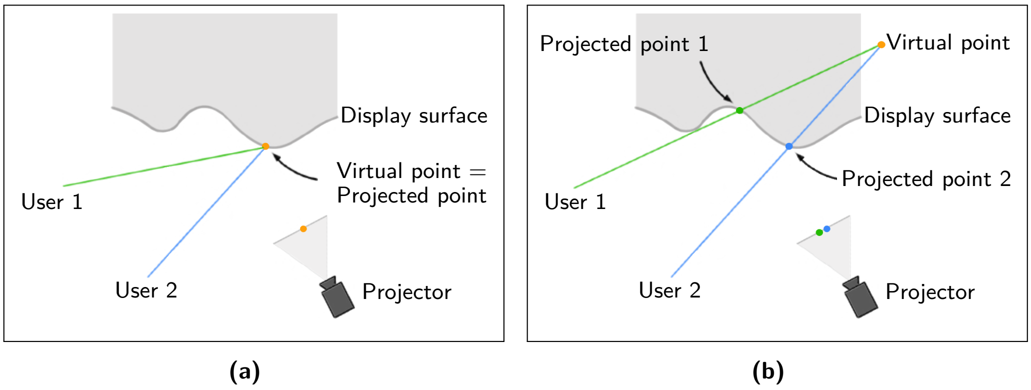


Figure 6.1.: Involved components when (a) flat textures and (b) virtual 3D objects are mapped onto a physical display surface.

If virtual objects are modeled as flat layers that appear like stickers on the physical geometry, they can be rendered independently from the current user (see Fig. 6.1a). This setup reduces the necessary efforts for calibration to a minimum, as spatial relationships between the user, projectors, and display surfaces do not have to be known. As no complex computations are needed, even non-professionals can perform the calibration steps, as we evaluated in a user study [SDBS15]. We developed a web application that is based on a feedback control system. The user manually transforms textured polygons in 2D space while every state is projected directly onto the 3D geometry on the fly (see Fig. 6.2 for examples). The projected images are adjusted continuously until the results meet the expectations of the user. This is a common approach in traditional projection mapping, for example, to turn building facades into large video projection surfaces. As mentioned before, this process is independent of the

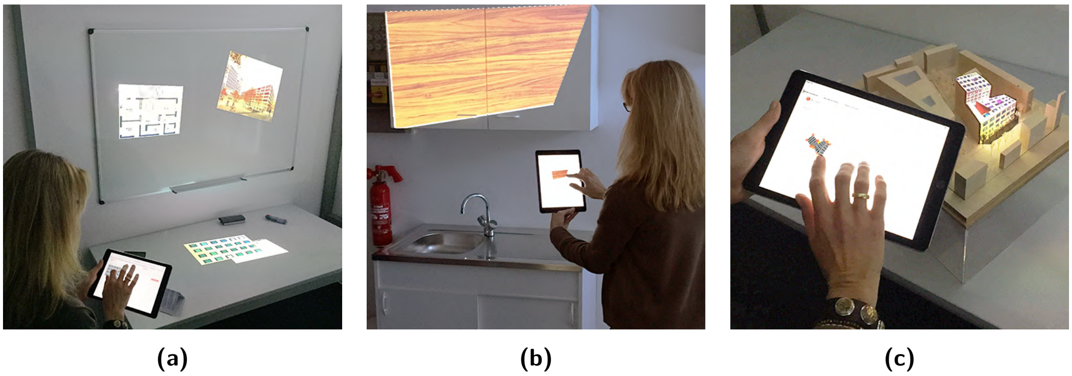


Figure 6.2.: Illustration of three texture mapping tasks that were tested as part of a user study: (a) desk organization, (b) furniture redesign, and (c) facade projection.

user perspective, since physical surfaces and virtual objects are layered directly on top of each other.

In contrast, all setups with virtual objects that are located in front of or behind the physical surfaces are view-dependent and therefore require spatial knowledge about several components, as illustrated in Figure 6.1b. Based on the depicted spatial relations, different parameters have to be estimated to perform a calibration:

1. Pose of the user's head
2. Display shape
3. Projector model

Depending on the applied calibration methods, these parameters may be approximated either within the same overall coordinate system or in relation to different systems. In the latter case, additional transformations are required to align all involved coordinate systems.

6.1. Pose of the User's Head

The positions of the user's eyes are crucial for all VR/AR systems that aim to display perspectively correct 3D content. As they have to be updated whenever the user's head pose is changing, a tracking system is required to locate the eyes over time. A common simplification of this tracking problem is to focus on the head instead of the eyes. By involving the head's position and orientation (= six degrees of freedom) as well as the inter-pupillary distance of the user, an approximation of the eyes' positions can be provided [SH16]. For projection-based Blended Spaces, a tracking system should fulfill several requirements:

(C1) No occlusion of the physical environment, as its surfaces simultaneously serve as display screens.

(C2) Low user instrumentation to preserve the low barrier to use spatial AR systems (see Sec. 2.2).

(C3) Proper functioning for mainly untextured environments such as CAVEs (without projection) as well as environments with dynamic texturing (with projection).

(C4) No interference with the projector light and resistance to changing light conditions in the scene (due to projections).

(C5) Low latency since temporal misalignments between virtual objects and their physical anchor points are even more perceivable than for VR systems.

(C6) High precision and sufficient accuracy, though the main focus is on the former.

(C7) Full coverage of compact working volumes.

Along with the emergence of VR/AR systems, several different tracking approaches were developed, which can be categorized according to various factors. Key components of each tracking system are system-specific sensors, which can detect signals that are either present in the scene by nature, or generated by corresponding emitters. The signals received by the sensors can take different physical forms, including mechanical, electromagnetic, optical, and acoustic forms. Each of these categories comes with a set of individual benefits and drawbacks that are surveyed in great detail in several reviews (see, for example, [MG96, SH16]). Due to some limitations, including violations of the stated conditions (C1) to (C7), mechanical, electromagnetic, and acoustic signals are rarely used in modern VR/AR hardware [SH16]. Instead, most consumer devices such as the *HTC Vive* series, *Oculus Rift* and *Oculus Quest*, both versions of *HoloLens*, and *Magic Leap One* use optical systems to implement positional tracking. Therefore, optical tracking is "one of the most important physical tracking principles used today for AR" and VR [SH16].

Optical tracking systems are based on light-sensitive sensors, such as IR or RGB cameras. Depending on the positioning of these sensors, we distinguish between *inside-out* and *outside-in* optical tracking systems. In inside-out systems, the cameras are mounted directly on the tracking target, for example, the user's head. Based on the camera images, technologies such as *Oculus Insight*, which is used in *Oculus Quest* and *Rift S*, search for distinctive key points within the scene that remain static over time. By establishing a connection between these key points, a digital map of the environment can be built, while the location of the user's head within this map can be tracked simultaneously. There are several algorithms to solve this *Simultaneous Localization And Mapping* (SLAM) problem [DWB06]. While we refer to the literature for details, some mathematical fundamentals are presented in Section 6.2.2. In general, such algorithms require different views of the same key points, which can be gathered by placing multiple cameras in a pre-defined rig, or by tracking 3D points over time and using the slight variations of the view that are caused by the user movement. In Blended Spaces, the identification of such keypoints can be difficult. Without projected content, CAVEs tend to provide monotonous surroundings, both in color and structure. However, when the Blended Space is in an AR or VR state, the surface color dynamically changes due to the projected objects (C3). While both problems could be avoided by adding static

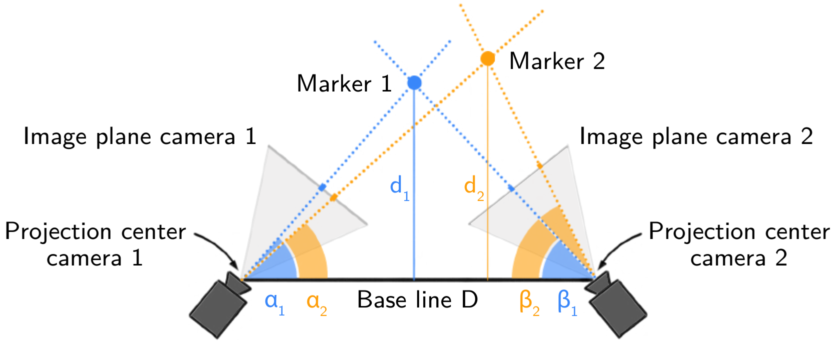


Figure 6.3.: Illustration of the triangulation principle to compute the distance d of a physical marker using two video cameras.

features to the scene, a general limitation of inside-out tracking systems remains. Cameras, as well as additional processing units, add a lot of weight and volume to the otherwise light-weight stereo glasses, leading to a violation of condition (C2).

The alternative approach, called outside-in systems, solves some of these challenges. In this kind of tracking system, the cameras are located around the tracking space. As the precise, accurate, and fast identification of the tracking target under varying light conditions and for different users is a challenging computer vision problem, its complexity is usually reduced by adding feature points to the target, which are very easy to detect. As these additional features should not interfere with the projections (C4) and be invisible to the user, corresponding devices are mainly relying on IR light. Lightweight IR markers can be easily mounted on the stereoscopic glasses of the user. To even remove the need for batteries to empower these markers, passive markers can be used. Such markers do not include an active light source but instead reflect IR light that is sent by the same devices that also include the sensors. A retroreflective surface ensures that the incoming light is reflected in the direction of the sender with minimum scattering. Therefore, most parts of the IR camera images are dark with the markers being exceptionally bright spots that can be identified easily using a computer vision algorithm such as the *Blob Detector* included in the *OpenCV* library [BK08]. Based on the 2D coordinates of a marker in at least two camera views, its position in 3D space can be computed using the triangulation principle. Figure 6.3 illustrates the involved trigonometric concepts. The depth d_i of a marker can be derived as follows:

$$D = \frac{d_i}{tan(\alpha_i)} + \frac{d_i}{tan(\beta_i)} \iff d_i = D \cdot \frac{tan(\alpha_i) \cdot tan(\beta_i)}{tan(\alpha_i) + tan(\beta_i)} \tag{6.1}$$

As can be seen from the equation, the spatial relationship between cameras has to be known to compute a value for d_i. The baseline distance D and the tilt angles of both cameras are measured in an initial calibration step, which is performed on the basis of multiple captures of a calibration tool with a known pattern of markers.

By applying the triangulation principle, we can compute the position of a marker but not its orientation. To be able to also track the direction a user is heading at, a rigid body of at least three markers has to be considered. For stability reasons, it is recommended to use 4 to 12 markers per rigid body, as this increases the number of samples to compute the pose of the rigid body and at the same time reduces the vulnerability to marker occlusions [Optb].

In summary, marker-based outside-in optical tracking systems can achieve sub-millimeter precision and accuracy (C6) with a latency of less than 5 ms (C5) and a tracking distance of up to 100 feet (C7) [Opta]. As only 4 to 12 small-sized markers have to be added on top of the stereoscopic glasses, they do not occlude the view of the user (as in contrast to, e.g., mechanical tracking systems) (C1). In conjunction with the afore-mentioned characteristics resulting from the minimal user instrumentation and the usage of IR light, they are well-suited for the proposed setup of Blended Spaces.

6.2. Display Shape

To achieve a pixel-correct alignment between projections and the display surface, a good approximation of the surface shape is essential. In general, we differentiate between two approaches to estimate this shape:

1. 3D production methods to build a physical representation of a virtual 3D model.
2. 3D capturing methods to create a virtual representation of the physical surface.

The choice depends on several factors, including the project schedule as well as available resources and complexity of the display shape. In the following two sections we will address both options, in particular with regard to the positive and negative aspects they introduce into the process of building a Blended Space.

6.2.1. Creating a Physical Representation of a Virtual Model

In many application fields such as rapid prototyping or building information modeling (BIM), projects start with the creation of a virtual model to envision a design before its actual realization. Based on this virtual model, a physical replica can be built to be placed within a Blended Space as described in Chapter 3. In domains such as architecture and arts, building hand-crafted physical 3D models is still common practice. These traditional techniques make use of low-cost and easy-to-process materials like cardboard, wood, foam, and cork, which are first cut into pieces and then assembled using material-dependent glue or tape. The entire modeling process requires a lot of manual effort, even though some subtasks such as cutting can be supported by machines, for example, cutting plotters and laser engravers.

A fully automated alternative to this approach emerged in the form of 3D printing technology, which gained much popularity within the last years [RSG17]. 3D printers allow for the manufacturing of complex geometries with sub-millimeter accuracy. The expenditure of time and money, as well as the achievable object size and surface smoothness, depend on a combination of the used materials and the applied 3D printing method. The majority of 3D printers are either using polymers in various forms such as solid filaments, powder, and resins, or metallic powders. The material is added layer by layer, which is why the term *additive manufacturing* is used synonymously with 3D printing. With current 3D printers, this additive process still takes some hours even for small objects [RSG17]. In Chapter 18, we envision how the future development of this technology may enable 3D printing of objects on the fly, to create advanced forms of Blended Spaces.

6.2.2. Capturing a Virtual Representation of a Physical Model

While virtual models can be a good starting point for building Blended Spaces, in some applications the physical object is what is available first. Examples include historic artifacts in museums or products in retail, that can serve as a display surface within a Blended Space. For this purpose, a virtual replica of the physical object has to be created. For objects of simple geometries and measurable sizes, such as boxes, tables, or the walls of a CAVE, a manual modeling process can be applied, using dedicated software such as *Blender* or *Autodesk 3ds Max*. To create virtual replicas with higher complexity, there are several (partly) automated processes that exploit (i) known geometric relationships within the scene, or (i) time measurements.

The first category of 3D capturing methods is based on the triangulation principle, which was already introduced in Section 6.1. In the same way as it was applied to measure the 3D position of single markers, it can be easily extended to cover the entire surface of physical objects. For this purpose, the 2D projections of *each* 3D point have to be identified in at least two (IR or RGB) camera images. This process of matching image points is called the *correspondence problem*. Approaches to solve the correspondence problem are either feature- or correlation-based [VKR05].

Feature-based algorithms were already mentioned in the context of tracking systems since they serve as the basis for many VR/AR glasses such as *Oculus Quest* or *HoloLens 2*. They identify distinctive image points, called features, that are stable under variations of the viewport, such as corners, edges, and line segments. For each feature, the characteristic information is encoded in a descriptor with a predefined format. Solving the correspondence problem based on features is equatable with finding a subset of features with matching descriptors as well as a similar layout in two or more camera images.

Correlation-based algorithms exploit image patches rather than single image points. For each pair of patches, a similarity function is computed, for example, by weighting the intensity values of all pixels within the patches [VKR05]. This approach relies on the assumption that corresponding image regions appear similar to each other if the baseline distance between the capturing cameras is shorter than the distance of both cameras to the captured object.

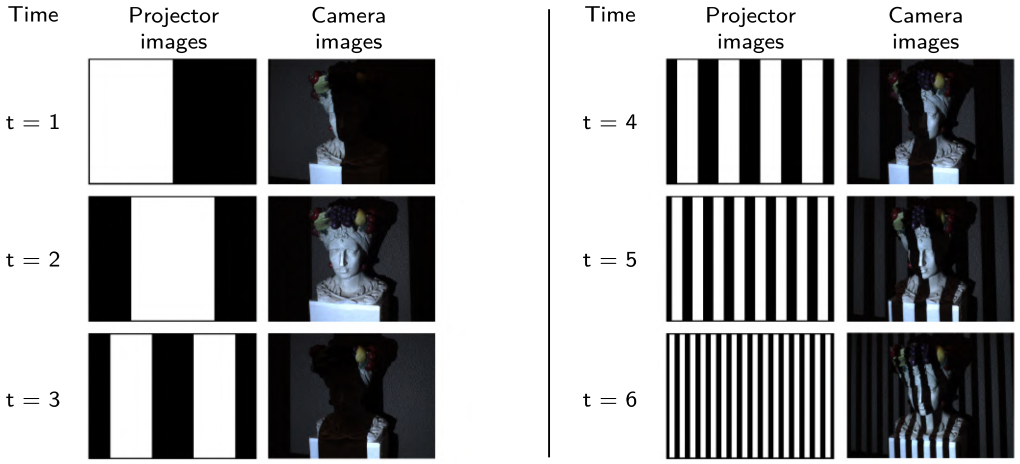


Figure 6.4.: Illustration of the Gray Code structured light approach for 3D capturing (the camera images are retrieved from [LT09]).

All techniques that are just observing the physical scene without adding features are called *photogrammetry* or *passive* 3D capture. In contrast, *active* techniques solve the correspondence problem by replacing one of the cameras by a projector. As the proposed setup for Blended Spaces already relies on projectors, they can be reused to not only serve for projecting the final virtual content but also for this prior calibration step. In contrast to the tracking process described in Section 6.1, the capturing of the display shape usually does not have to be hidden from the user as it is performed before the actual application starts. Therefore, projectors of visible light and standard color cameras can be used just the same as specific IR devices.

The main benefit of using active instead of passive 3D capturing techniques is the opportunity to decide on the features that are introduced into the scene by controlling the projected pattern. The pattern determines the duration of the 3D capturing process as well as its result, including the resolution of the depth image, depth accuracy, and noise sensitivity. A classical pattern, that is used in projects such as RoomAlive [JSM+14], is based on the Gray Code. A Gray Code pattern consists of multiple stripes of intensity 0 (i.e., black) and 1 (i.e., white) [Gen11]. During the 3D capturing process, a sequence of n patterns is projected onto the scene, as illustrated in Figure 6.4. For each pattern, the scene is captured with an RGB camera, and each image point has to be identified as *lit* or *unlit*. This results in a binary code for each image point, that can be mapped to one of the 2^n projected stripes to solve the correspondence problem. In the example illustrated in Figure 6.4, the resulting depth image features a horizontal resolution of $2^6 = 64$ pixels. By using a sequence of ten instead of six patterns, the resolution can be increased to $2^{10} = 1024$ pixels. The same procedure is repeated using patterns with horizontal stripes to compute a vertical correspondence for each 3D point in the scene. Besides the Gray Code sequence, there is a variety of structured light patterns that differ in terms of their color space (binary vs. gray scale vs. color) and their temporal dependency (static vs. dynamic). For an in-depth survey see, for example, Battle et al. [BMS98].

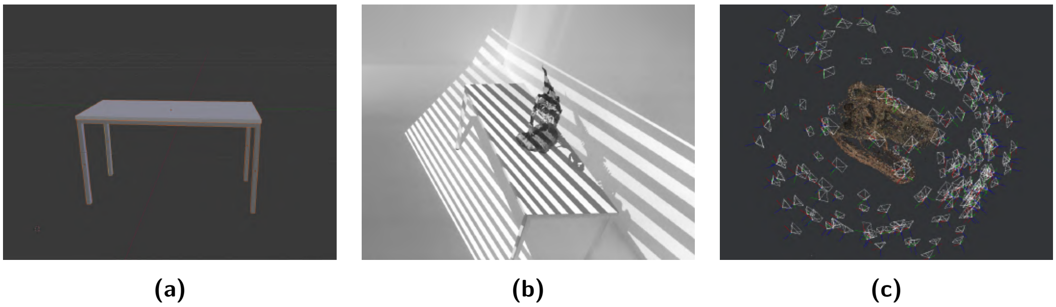


Figure 6.5.: A table and a dinosaur skeleton, either (a) modeled manually, or captured using (b) structured light, and (c) photogrammetry.

The second category of 3D capturing techniques is based on the *time of flight* (ToF) principle and usually requires a laser scanning device. The scanner is emitting a laser pulse into the scene, which is reflected at the physical object's surface and returned to the device. By measuring the round trip time t in s, the distance d of an object point in m can be computed using the following formula with c being the speed of light in m/s:

$$d = \frac{c \cdot t}{2} \tag{6.2}$$

A depth image with the mapping of all object points can be constructed either by steering the laser beam to pass through each image pixel sequentially or by sending a laser pulse that covers the entire scene at once. For the latter, a matrix of IR sensors is required to measure the depth of multiple object points within the same scanning cycle.

Each of the described 3D capturing methods was explored in the course of these PhD studies, as shown in Figure 6.5. Both triangulation and ToF methods result in a 2.5D point cloud if depth images are only captured from one specific viewport. 2.5D may be sufficient if this viewport is similar to one of the projector, and if the display surface and projector have a static relationship throughout the entire application. Otherwise, the capturing of multiple depth images from different viewpoints and subsequent registration of the resulting point clouds is necessary to construct a full 3D model. Common approaches to solve the so-called point set registration problem are the *Iterative Closest Point* algorithm or a feature-based algorithm, both of whom are implemented in the *Point Cloud Library* (PCL) [RC11]. After a final point cloud is composed of the collected depth data, another structural transformation is necessary to create a model representation that can be better processed in the up-following steps of the rendering pipeline. In computer graphics, the most common geometric description of virtual 3D models is the boundary representation, which relies on a polygonal mesh that describes the object's surface rather than its volume [Str06]. The polygonal mesh is composed of connected faces, usually triangles, that are bounded through edges, which in turn join two vertices each. Deciding on the number of used polygons is a matter of balancing, as more polygons usually come along with a better approximation of the surface but also at a higher computational cost. Reducing the executed operations per frame is especially important in the context of Blended Spaces, as most VR/AR applications require a continuous recomputation of the projected content, for example, to adapt to a moving user's perspective. Additional optimization strategies such as dynamic adjustment of the level of detail may be applied during runtime to further reduce the number of processed polygons without a perceivable loss of visual quality.

6.3. Projector Model

In a final calibration step, we have to describe the positioning of all projectors within the scene as well as their internal optics[1]. Due to the dualism between cameras and projectors, well-known principles of a pinhole camera can be applied to provide estimates for all projector parameters of interest (see [BR05, Ope, SH16]). The traditional camera model describes a mapping of homogeneous[2] 3D world coordinates $(x_{world}, y_{world}, z_{world}, 1)^T$ to homogeneous 2D image coordinates $(u, v, 1)^T$, using the camera matrix $\mathbf{P}$:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{P} \cdot \begin{bmatrix} x_{world} \\ y_{world} \\ z_{world} \\ 1 \end{bmatrix} \tag{6.3}$$

Inversely, $\mathbf{P}$ can also be used to describe which pixel of a projector's image has to be illuminated in order to project on a specific 3D point within the scene. $\mathbf{P}$ is a 3×4 matrix that does not contain a directly readable specification of the camera orientation or the internal geometry of the camera. This can be solved by decomposition of $\mathbf{P}$ into an intrinsic matrix $\mathbf{K}$ and an extrinsic matrix[3] $[\mathbf{R}|\mathbf{t}]$, with $\mathbf{P} = \mathbf{K} \cdot [\mathbf{R}|\mathbf{t}]$. For the mapping of 3D to 2D points we receive:

$$\underbrace{\begin{bmatrix} u \\ v \\ 1 \end{bmatrix}}_{\text{Image coordinates}} = \underbrace{\begin{bmatrix} f_u & 0 & c_u \\ 0 & f_v & c_v \\ 0 & 0 & 1 \end{bmatrix}}_{\text{Intrinsic matrix}} \cdot \underbrace{\begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix}}_{\text{Extrinsic matrix}} \cdot \underbrace{\begin{bmatrix} x_{world} \\ y_{world} \\ z_{world} \\ 1 \end{bmatrix}}_{\text{World coordinates}} \tag{6.4}$$

In the following sections, we will discuss the projector parameters which are included in each of the two matrices as well as their composition.

[1] Throughout this chapter, bold uppercase letters refer to matrices, bold lowercase letters to vectors, and italic letters to matrix/vector elements and scalars.

[2] Homogeneous coordinates add an extra dimension to matrices and vectors, thus allowing for a universal representation of affine transformations such as translation, rotation, and scaling. In this context, vectors can encode either directions or positions by setting the last (homogeneous) coordinate to 0 or a value ≥ 1. For details, see [MS15].

[3] Vertical and horizontal lines within a matrix indicate that it was constructed from multiple vectors and/or matrices (cf. [Ope, SH16]). For example, $[\mathbf{R}|\mathbf{t}]$ can also be denoted as $[\mathbf{R}_{3\times3} \;\; \mathbf{t}_{3\times1}]_{3\times4}$.

6.3.1. Extrinsic Projector Parameters

The extrinsic matrix of a projector describes a mapping of world space into camera space and therefore is equivalent to the view matrix used in the standard graphics pipeline.

It is composed of a translation vector $\mathbf{t}$, that describes the origin of world space in camera coordinates, and a 3×3 rotation matrix $\mathbf{R}$, whose columns are representing the world axes in camera coordinates. For subsequent processing steps, we extend the resulting 3×4 matrix by a fourth row[4]:

$$\left[\begin{array}{c|c} \mathbf{R} & \mathbf{t} \\ \hline \mathbf{0} & 1 \end{array}\right] \tag{6.5}$$

As stated before, this 4×4 matrix describes the transformation of the world in relation to the projector. To calibrate a Blended Space with multiple projectors, we usually need the inverse transformation as we aim to position all projectors within world space. In other words, we need to derive the projector pose $[\mathbf{R_p}|\mathbf{t_p}]$, where $\mathbf{R_p}$ denotes the orientation of the projector and $\mathbf{t_p}$ its position, both in world space. We compute the inverse transformation as follows:

$$\begin{aligned} \left[\begin{array}{c|c} \mathbf{R} & \mathbf{t} \\ \hline \mathbf{0} & 1 \end{array}\right] = \left[\begin{array}{c|c} \mathbf{R_p} & \mathbf{t_p} \\ \hline \mathbf{0} & 1 \end{array}\right]^{-1} &= \left[\left[\begin{array}{c|c} \mathbf{I} & \mathbf{t_p} \\ \hline \mathbf{0} & 1 \end{array}\right] \cdot \left[\begin{array}{c|c} \mathbf{R_p} & \mathbf{0} \\ \hline \mathbf{0} & 1 \end{array}\right]\right]^{-1} \\ &= \left[\begin{array}{c|c} \mathbf{R_p} & \mathbf{0} \\ \hline \mathbf{0} & 1 \end{array}\right]^{-1} \cdot \left[\begin{array}{c|c} \mathbf{I} & \mathbf{t_p} \\ \hline \mathbf{0} & 1 \end{array}\right]^{-1} = \left[\begin{array}{c|c} \mathbf{R_p}^T & \mathbf{0} \\ \hline \mathbf{0} & 1 \end{array}\right] \cdot \left[\begin{array}{c|c} \mathbf{I} & -\mathbf{t_p} \\ \hline \mathbf{0} & 1 \end{array}\right] \\ &= \left[\begin{array}{c|c} \mathbf{R_p}^T & -\mathbf{R_p}^T\mathbf{t_p} \\ \hline \mathbf{0} & 1 \end{array}\right] \end{aligned} \tag{6.6}$$

Based on the last equation we can derive the projector orientation $\mathbf{R_p}$ and position $\mathbf{t_p}$:

$$\begin{aligned} &\mathbf{R} = \mathbf{R_p}^T \implies \mathbf{R_p} = \mathbf{R}^T \\ &\mathbf{t} = -\mathbf{R_p}^T \cdot \mathbf{t_p} = -\mathbf{R} \cdot \mathbf{t_p} \implies \mathbf{t_p} = -\mathbf{R}^T \cdot \mathbf{t} = -\mathbf{R_p} \cdot \mathbf{t} \end{aligned} \tag{6.7}$$

6.3.2. Intrinsic Projector Parameters

The intrinsic matrix $\mathbf{K}$ is describing a transformation of 3D camera coordinates into homogeneous 2D image coordinates. In the standard graphics pipeline, this transformation is also known as the *perspective projection* [SH16]. The following schematic illustrates the projection of a 3D point $\mathbf{p_{cam}} = (x_{cam}, y_{cam}, z_{cam})^T$ in camera space onto the image plane. The resulting point $\mathbf{p'_{cam}} = (x'_{cam}, y'_{cam}, z'_{cam})^T$ is still linked to camera space and is therefore represented in meters (the same unit as for $\mathbf{p_{cam}}$) while $\mathbf{p_{image}} = (u, v)^T$ refers to the same point in pixel coordinates in image space.

[4] $\mathbf{0}$ refers to the zero vector, a column or row vector having all of its elements equal to zero. In the current context, the vector dimensions always match the dimensions of the adjacent matrix. In Equation 6.5, $\mathbf{R}$ refers to a 3×3 matrix, and therefore $\mathbf{0}$ has to be a 3-dimensional row vector to match the matrix.

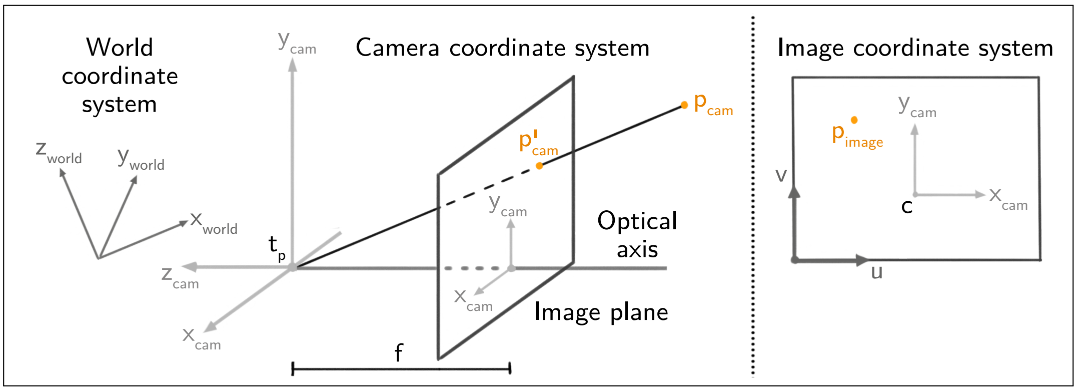


Figure 6.6.: Mapping of a 3D point in right-handed camera space to a 2D point in image space.

The illustration exemplifies multiple parameters that define a perspective projection: the *center of projection* $\mathbf{t_p}$, which is the intersection of all rays between pairs of corresponding points $\mathbf{p_{cam}}$ and $\mathbf{p_{image}}$ $(= \mathbf{p'_{cam}})$; the image plane, that denotes a plane parallel to the x- and y-axes of the camera coordinate system that includes all points $\mathbf{p_{image}}$; and the *principal point* $\mathbf{c}$, which is defined as the normal projection of $\mathbf{t_p}$ on the image plane. The line through the points $\mathbf{t_p}$ and $\mathbf{c}$ is called the *optical axis* of the projector, while their distance defines the *focal length* f.

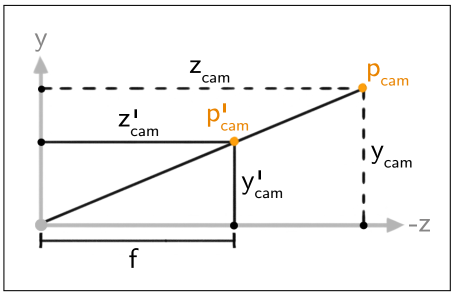


Figure 6.7.: Derivation of the perspective projection matrix based on the intercept theorem.

Figure 6.7 illustrates the correlation between camera and image coordinates by the example of the y value. For this purpose, we consider the points $\mathbf{p_{cam}}$ and $\mathbf{p'_{cam}}$ as seen from the positive x-axis. By applying the theorem of intersecting lines we can compute

$$\frac{y'_{cam}}{f} = \frac{y_{cam}}{z_{cam}} \quad \Longrightarrow \quad y'_{cam} = \frac{f \cdot y_{cam}}{z_{cam}} \tag{6.8}$$

In a last computational step, we have to transform the y coordinate of $\mathbf{p'_{cam}}$ from camera to image space. As the axes of camera and image space are aligned in parallel, we only have to apply a translation to the point $\mathbf{p'_{cam}}$. As illustrated in Figure 6.6, the translation vector corresponds to the principal point $\mathbf{c}$ of the projector. Before adding the principal point to $\mathbf{p'_{cam}}$ we have to perform a conversion between the used units of camera space (i.e., usually meters) and image space (i.e., pixels). This can be done by scaling the focal length f that is used in Equation 6.8 by the size of a pixel in the directions u and v, resulting in the values

f_u and f_v, respectively. Therefore, we can compute the coordinate v in image space as:

$$v = y'_{cam} + c_v = \frac{f_v \cdot y_{cam}}{z_{cam}} + c_v \tag{6.9}$$

The same rules apply to the u coordinate of $\mathbf{p_{image}}$. In summary, we can compute $\mathbf{p_{image}}$ as follows:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{f_u}{z_{cam}} \cdot x_{cam} + c_u \\ \frac{f_v}{z_{cam}} \cdot y_{cam} + c_v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{f_u}{z_{cam}} & 0 & c_u \\ 0 & \frac{f_v}{z_{cam}} & c_v \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_{cam} \\ y_{cam} \\ 1 \end{bmatrix} = \begin{bmatrix} f_u & 0 & c_u \\ 0 & f_v & c_v \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \frac{x_{cam}}{z_{cam}} \\ \frac{y_{cam}}{z_{cam}} \\ \frac{z_{cam}}{z_{cam}} \end{bmatrix} \tag{6.10}$$

with the final result:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_u & 0 & c_u \\ 0 & f_v & c_v \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_{cam} \\ y_{cam} \\ z_{cam} \end{bmatrix} \tag{6.11}$$

All of the previous calculations are based on the pinhole model, which usually cannot be assumed for real projectors. Through the use of lenses, aberrations in the form of radial and tangential distortions may be introduced to the projected image. In some of the calibration methods that will be discussed in the next section, those distortion coefficients can be identified along with the other extrinsic and intrinsic parameters during the calibration process.

6.3.3. Approaches for Projector Parameter Estimation

If known, the intrinsic and extrinsic matrices can be used to project any 3D point onto the image plane of the projector. Inversely, from a number of known correspondences between 3D world coordinates and 2D image coordinates, an estimate of both matrices can be derived. For this task, we can make use of a variety of approaches with individual technical requirements and different levels of automation. In the following, we want to discuss three of these approaches that were used for the calibration of Blended Spaces in the context of this thesis.

The RoomAlive Framework

The first presented method was released within the scope of Microsoft's *RoomAlive* project [JSM$^+$14] and utilizes the depth-sensing capabilities of the Microsoft *Kinect v2*. Although it is based on a combination of existing concepts, it may serve as an example for a fully automatic calibration process that covers all transformations between a projected pixel and its corresponding 3D point. A decomposition of this process results in the following steps:

(1) Mapping of 3D world coordinates to 2D color camera coordinates.

(2) Mapping of 2D color camera coordinates to 2D projector coordinates.

(3) Solving for the extrinsics and intrinsics of the projector.

While the first two steps aim to identify correspondences between points in the 3D world space and projector pixels, the last step makes use of these correspondences to estimate values for the intrinsic matrix $\mathbf{K}$ and the extrinsic matrix $[\mathbf{R}|\mathbf{t}]$. The first transformation can be subdivided into several substeps:

(1a) Kinect's ToF camera is used to capture a depth image of the scene.

(1b) Pixels of the depth image are transformed into precise 3D positions within the depth camera coordinate system.

(1c) Based on internal camera calibration information of the Kinect, a transformation between depth and color camera coordinates is performed.

The result of steps (1a) to (1c) is a set of 3D world points with their associated 2D color camera coordinates. Then, the color camera coordinates have to be mapped to pixel coordinates of the projector. For this purpose, a Gray Code mapping is used as described in Section 6.2.2. While the main procedure follows the same steps as for the capture of a 3D surface, there is a small but important difference in the setup. For the 3D capturing of a surface, the camera and projector need to be positioned in a known layout to compute the baseline as well as the tilt angles (see Fig. 6.3). In the context of projector calibration, this layout is not known but the subject of the calibration process. However, since the depth of 3D points is already measured in step (1), it is not of further interest for step (2). Instead, we make use of the first part of the Grey Code procedure to solve the correspondence problem, and therefore identify corresponding points in the camera and the projector image.

As a transformation between 3D world points and 2D camera pixels was already computed in step (1), and step (2) results in a mapping of camera pixels to projector pixels, an overall transformation between 3D world coordinates and 2D projector coordinates can be derived. From these 3D-2D point correspondences, both intrinsic and extrinsic parameters of the projector can be estimated using a Levenberg-Marquardt optimization [Row96].

The input of the algorithm is a self-defined function that computes the root-mean-square (RMS) error for a given set of 3D-2D correspondences and a current estimate of all intrinsic parameters (i.e., focal length and principal point) as well as extrinsic parameters (i.e., depth camera pose in the projector's coordinate space). When n_c denotes the number of correspondences, the RMS error is defined as follows:

$$rmsError = \sum_{k=0}^{n_c-1} \sqrt{(x_k - u)^2 + (y_k - v)^2} \tag{6.12}$$

The point (x_k, y_k) denotes the measured 2D image point, while (u, v) is computed by transforming the corresponding 3D world point with the estimated extrinsic and intrinsic matrix. The provided error function is minimized by variation of the given parameters. This is done iteratively, starting from an initial guess for all of the intrinsic and extrinsic parameters as well as a set of point correspondences.

As both the depth camera and the Gray Code procedure may be subject to inaccuracies, the full set of correspondences might include a considerable number of outliers. To exclude these outliers from the parameter estimation, and therefore increase the robustness of the method, the RoomAlive framework is embedding the optimization algorithm in a RANSAC procedure [JSM+14]. In each of the RANSAC iterations, a subset of 100 samples is randomly selected from the overall set of 3D-2D point correspondences. The selection process is repeated until the samples pass a test against co-planarity. This test is inevitable, as the estimation of a projector's intrinsics requires 3D points with different depths. For the resulting subset of non-planar points, the Levenberg-Marquardt optimization is performed, resulting in estimations for the matrices $\mathbf{K}$ and $[\mathbf{R}|\mathbf{t}]$. The estimated matrices are applied to the full set of 3D points and compared to the associated 2D image points. If the RMS error for a point correspondence is lower than a predefined threshold, this correspondence is supporting the estimated model of intrinsic and extrinsic parameters. All supporting correspondences, called inliers, are collected. If the number of 500 inliers is exceeded, the Levenberg-Marquard algorithm is performed for a second time, this time with all of the inliers instead of a subset of 100 samples only. The resulting model is assumed to be a good estimation for the intrinsics and extrinsics of the projector and is cached along with the according error. The entire procedure is repeated for up to 10 times and the model with the minimal error is used as the final result.

Although the RoomAlive method is both automated and flexible as it only poses a few requirements towards the calibration environment, it also has some limiting factors. The results of step (1) are limited by the quality of the ToF depth camera of the Kinect v2, which has a resolution of 512 × 424 pixels and a minimum detectable depth difference in the range of millimeters, that increases with distance [YZD+15]. These stats may be improved by using an alternative depth-sensing technology, such as a stereo color camera rig. Step (3) relies on the Levenberg-Marquardt minimization to solve for the projector parameters. Since the underlying problem is non-linear, the algorithm converges to the global minimum only if the initial guess is close to the final solution, or if there is only one minimum. Both limitations are addressed in the following semi-automated approach.

Zhang's Method

The following algorithm to estimate intrinsic and extrinsic parameters of a camera was suggested by Zhang [Zha00] and is included in several computer vision libraries such as *OpenCV* [BK08]. In contrast to the previous solution, the algorithm works with one color camera only. Zhang's algorithm is based on multiple views of a planar calibration pattern with known geometry. Traditionally, a checkerboard is used as its corners can be identified easily in the camera image. From the known 3D positions of the checkerboard corners as well as their corresponding projections onto the image plane of the camera, Zhang's algorithm calculates an estimation of an intrinsic matrix for the RGB camera as well as an estimated pose for each view of the calibration rig in camera space. In a second step, Gray code projections can be used to build a transformation matrix between camera and projector image space as described in Sections 6.2.2, and 6.3.3. By cascading the matrices gathered in both steps, we compute a final transformation from 3D world space to 2D image space of the projector. As the second part is executed in the same way as for the RoomAlive approach, we want to focus on the first part.

The calibration procedure according to Zhang includes the following steps [Bur16]:

1. Capture m views of the calibration pattern by moving the pattern to different poses in front of the camera.
2. For each of the m camera images, extract n feature points.
3. For each view $i = 1, ..., m$, estimate a homography $\mathbf{H}_i$ that maps 3D model points to 2D image points.
4. From the homographies $\mathbf{H}_1, ..., \mathbf{H}_m$ estimate the intrinsic parameters of the camera using a linear equation system.
5. Based on the estimated intrinsic parameters, compute the extrinsic parameters for each of the m views.
6. Refine the estimated intrinsic and extrinsic parameters by using them as an initial guess for a non-linear optimization algorithm.

The most important assumption for the algorithm results from the planarity of the calibration rig. By placing the x- and y-axis of the 3D world coordinate system at the plane of the checkerboard, the z values of all corners can be assumed to be 0. Due to this assumption, we get a simplified equation for the mapping between 3D object points and 2D image points:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{K} \cdot \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \cdot \begin{bmatrix} x_{world} \\ y_{world} \\ 0 \\ 1 \end{bmatrix} = \underbrace{\mathbf{K} \cdot \begin{bmatrix} r_{11} & r_{12} & t_1 \\ r_{21} & r_{22} & t_2 \\ r_{31} & r_{32} & t_3 \end{bmatrix}}_{Homography\ \mathbf{H}} \cdot \begin{bmatrix} x_{world} \\ y_{world} \\ 1 \end{bmatrix} \tag{6.13}$$

Im comparison to the mathematical approach used for RoomAlive, Zhang's algorithm reduces the number of unknown parameters for each homography matrix from 12 to 9:

$$\mathbf{H} = \begin{bmatrix} f_u & 0 & c_u \\ 0 & f_v & c_v \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} r_{11} & r_{12} & t_1 \\ r_{21} & r_{22} & t_2 \\ r_{31} & r_{32} & t_3 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \tag{6.14}$$

The homographies $\mathbf{H}_i$ vor views $i = 1, ..., m$ can be estimated with a Direct Linear Transformation. Following a sequence of steps (see [Zha00] or [Bur16]), the formula 6.13 is rearranged to a system of linear equations (for an example homography $\mathbf{H} = \mathbf{H}_i$)[5]:

$$\underbrace{\begin{bmatrix} -x_1 & -y_1 & -1 & 0 & 0 & 0 & u_1x_1 & u_1y_1 & u_1 \\ 0 & 0 & 0 & -x_1 & -y_1 & -1 & v_1x_1 & v_1y_1 & v_1 \\ -x_2 & -y_2 & -1 & 0 & 0 & 0 & u_2x_2 & u_2y_2 & u_2 \\ 0 & 0 & 0 & -x_2 & -y_2 & -1 & v_2x_2 & v_2y_2 & v_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -x_n & -y_n & -1 & 0 & 0 & 0 & u_nx_n & u_ny_n & u_n \\ 0 & 0 & 0 & -x_n & -y_n & -1 & v_nx_n & v_ny_n & v_n \end{bmatrix}}_{\mathbf{M}} \cdot \underbrace{\begin{bmatrix} h_{11} \\ h_{12} \\ h_{13} \\ h_{21} \\ h_{22} \\ h_{23} \\ h_{31} \\ h_{32} \\ h_{33} \end{bmatrix}}_{\mathbf{h}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \tag{6.15}$$

To create this system of equations, the values of the homography matrix $\mathbf{H}$ are rearranged into a 9-dimensional vector $\mathbf{h}$. Every two equations are built from one of the n 3D-2D point correspondences. Fur this purpose, the coordinates (x_j, y_j) with $j = 1, ..., n$ are computed from the physical size of the checkerboard as well as its number of squares, and therefore, are the same for all views. In contrast, the corners (u_j, v_j) are different for all views as they are extracted from the camera image j using a computer vision operator such as the Harris corner detector [BK08]. After assigning all of the parameters to the left matrix, the system of linear equations is solved using a singular-value decomposition [Bur16]. As we have to solve for nine parameters for each homography $\mathbf{H}$ and each point correspondence contributes to two equations, at least 5 points are needed for each view. Usually, many more corners are used to build an overdetermined equation system, as this compensates inaccuracies as well as occlusions during the computer vision process. A solution of the equation system is obtained by minimizing the residual $||\mathbf{M} \cdot \mathbf{h}||^2$. While this expression does not directly relate to the geometric projection error as for the RoomAlive method, the computed value for $\mathbf{h}$ can serve as a good initial guess for a subsequent non-linear Levenberg-Marquardt optimization [Row96]. The resulting homographies $\mathbf{H}_i$ for $i = 1, ..., m$ encode both the common intrinsic matrix $\mathbf{K}$ and the view-dependent extrinsic matrices $[\mathbf{R}_i|\mathbf{t}_i]$, which can be extracted as described in [Bur16].

Instead of a checkerboard, other calibration rigs can be used as long as their geometry is known and the 3D points are detectable in the camera image. Early methods usually relied on rigs that were made of two or three planes orthogonal to each other [Zha00]. However, due to their non-planarity, they require non-linear optimization strategies such as presented in Section 6.3.3.

[5] For the sake of readability, we use the short forms x and y to denote x_{world} and y_{world}.

Manual Mapping Without Cameras

The last presented method to calibrate the projector is relying on two requirements:

1. A virtual 3D model of the display surface exists.
2. A valid initial guess for the intrinsic parameters of the projector is available.

These requirements were naturally fulfilled for most of the experiments which were conducted in the context of this thesis.

Firstly, a virtual model of the display surface is one of the prerequisites to align the physical surface features with the projected content. In Section 6.2.2, different approaches to create such a virtual representation are introduced.

Secondly, intrinsic parameters of a projector are usually constant throughout its lifetime, provided that the optical zoom value is not adjusted. If calibrated once, the correspondence problem can be simplified by focusing on the extrinsic matrix, which maps the world space to projector space as described above. Even without knowledge of the exact values for the intrinsic matrix, the method can be applied if a rough guess for the viewing angle of the projector is known. From this angle (in degrees), initial values for the focal length f and the principal point $\mathbf{c} = (c_u, c_v)$ can be computed as follows:

$$\begin{aligned} f &= imageWidth \cdot deg2Rad(viewingAngle) \\ c_u &= 0.5 \cdot imageWidth \\ c_v &= 0.5 \cdot imageHeight \end{aligned} \tag{6.16}$$

During the calibration process, these initial guesses can be further optimized.

In comparison to the previously discussed calibration methods, the distinctive feature of this approach is the unnecessity of an additional camera. Instead, the user has to be involved to align physical and projected features. In the following, we explain the required sequence of user actions by taking the example of an application called *mapamok* [McD12]. It was developed by the YCAM InterLab and features a basic UI to manually select 3D-2D point correspondences that can be used to perform a projector calibration.

The calibration process starts with a wireframe view of the virtual surface model, that has to be constructed beforehand using one of the methods described in Section 6.2.2. This view is called *selection mode* and supports common 3D interactions such as rotating, panning, and zooming. In selection mode, the user has to pick a distinctive 3D point of the virtual model, for example, a corner, which is then highlighted and labeled with a unique ID. Afterward, the application toggles to the second view called *render view*. In this mode, the virtual model is hidden and a 2D representation of the selected point is displayed instead. In contrast to the selection mode, which displays the 3D world space, render mode represents the two-dimensional projector image space. Within this space, the user can drag the 2D point in x and y direction while observing its movement along the real display surface. This step intends to find the correct (u, v)-position of the projected 2D point so it is perfectly aligned with the corresponding 3D point of the physical object. Therefore, it is recommended to use feature points that are easy to match, in contrast to, for example, points within a planar surface area. This process of choosing a 3D point in selection mode and finding the corresponding 2D point in render mode has to be repeated at least 3 times to enable an initial estimate of intrinsic and extrinsic parameters. It is recommended to aim for 8 to 12 points that cover the entire scene and are neither close to each other nor co-planar [McD12]. From the identified 3D-2D point correspondences and the initial guess for the intrinsic parameters, estimates for the intrinsic and extrinsic matrix can be created and optimized iteratively using the Levenberg-Marquardt algorithm as described in Section 6.3.3.

The manual mapamok calibration finalizes the set of methods that were used in the course of these PhD studies to estimate projector intrinsics and extrinsics. All of them can be considered to have two subtasks, which address the following questions:

1. Which technique is used to identify 3D-2D point correspondences?
2. Which mathematical model is applied to solve for the intrinsics and extrinsics?

In general, the approaches to solve each of these two subtasks can be combined arbitrarily as long as the mentioned technical and mathematical requirements are fulfilled. One of these technical requirements refers to the availability of additional cameras. While the usage of one or more cameras promises some considerable advantages regarding the process automation, it also implies some negative effects. First of all, the complexity of the setup increases as some hardware components are used only during the calibration process but not necessarily for the final application. Furthermore, most calibration procedures that involve a camera impose additional requirements on the scene, such as proper lighting conditions and non-transparent surface characteristics. A series of tests concerning the methods of RoomAlive and Zhang showed that, for our setups and used devices, both accuracy and precision were in the range of centimeters, and therefore higher than demanded for most of the experiments. In contrast, the mapamok approach allows the user to precisely align features that are most important for the application, such as points around the exhibit in a museum-like setup. While this involves some manual effort, the time investment was consistent and within minutes. Compared to the manual approach, the automatic methods required several repetitions and optimizations of the scene. Therefore, the overall calibration time was in the range of hours, even if only little manual intervention was required. An advantage of the automatic processes is the possibility to calibrate multiple projectors simultaneously. If the camera and projector frustums are overlapping, all of them can be registered within the same global coordinate space. We did not take advantage of this in our experiments, since all of our setups use a 1:N mapping between projectors and physical objects. Therefore, a relative positioning in a global coordinate space is not necessary, and each physical object can be registered within the local space of the corresponding projector instead.

The minimization of projector overlap also reduces the need for additional blending procedures. Such techniques are used to compensate for intensity variations, which may appear between different parts of the Blended Space. The normalization of intensity values is the objective of the *photometric registration*, in opposition to the geometric registration, which was addressed in the previous sections. In the course of this thesis, photometric registration will not be considered in depth; for a formal overview see, for example, the surveys provided by Majumder et al. [MHTW00], or Bimber and Raskar [BR05]. In the context of all subsequently presented experiments, manual adjustments of the projectors' working color gamuts were performed to reduce the intensity mismatch between projectors. Besides, most of the used display surfaces featured a white, diffuse reflecting material. In conjunction with the reduction of overlapping projection regions, intensity variations could be reduced to a level that was not interfering with the presentation of the Blended Spaces.

7 Chapter 7. Rendering

If the Blended Space is calibrated once, it can display any virtual content to create all different states along the Mixed Reality continuum. For authoring the Blended Space, we used the game engine *Unity* [Uni], although other 3D graphics APIs such as the *Unreal Engine* [Unr] could be used as well. Unity supports the development of 2D and 3D real-time applications for different platforms such as PC, mobile devices, and VR glasses. Within a Unity scene, a precise virtual simulation of the Blended Space can be built, where virtual 3D models represent both the real-world display surfaces and the objects, which will be projected on top of these surfaces. Based on the geometric parameters that were estimated in the previous calibration process, virtual cameras with matching extrinsic and intrinsic values can be set up within the scene. Using these virtual cameras, a multi-step rendering process is performed to compute a 2D image for each of the projectors within the Blended Space. In this chapter, we first address the implementation of different physical and virtual components that are necessary to create a comprehensive simulated Blended Space. Afterward, we will present an overview of the rendering process that is performed to create the final projections.

7.1. Implementation of a Virtually Simulated Blended Space

Before the actual virtual content can be added to the Unity scene, a suitable representation of all physical components of the Blended Space has to be found. First of all, for each potential display surface in the physical scene, a polygonal mesh is imported to the Unity scene. For instance, this includes walls of the surrounding CAVE as well as virtual models that represent physical objects inside of the Blended Space. As we focus on the geometric registration of virtual and real content rather than a photometric one, the materials of all virtual models do not have to mirror the real surface structure of physical objects.

After a virtual replica of physical surfaces is set up, a representation of the user can be introduced to the scene. Which components are required to model a user, depends on the body parts that are intended to be tracked. For example, the positions of the user's hands could be mapped to a Unity collider to simulate interaction with virtual objects. If we only aim for a perspectively correct projection of the virtual content, a representation of the user's head is sufficient. For this purpose, the user's eyes are modeled as two virtual cameras, that are horizontally off-centered according to the user's interpupillary distance. Both virtual eyes are attached to a parent object that mimics real movements of the user's head. Most tracking system vendors such as OptiTrack already provide a Unity interface to transfer both the position and orientation of a tracked rigid body to a Unity object.

Like the user's eyes, the projectors of a Blended Space are also represented by virtual cameras. The details on this procedure are addressed in the next section as they can be deduced from the rendering process. A projector's intrinsic matrix, which was estimated in a previous calibration step, can be directly assigned to the projection matrix of the corresponding virtual camera. In contrast, the extrinsic matrix first has to be converted to be applicable to a Unity camera. If the extrinsic parameters describe a pose of the world origin in projector coordinates, an inverse transformation has to be computed as described in Equation 6.7. Besides, all of the presented calibration tools such as RoomAlive, OpenCV, and mapamok are using right-handed coordinates, while Unity is based on a left-handed coordinate system. A conversion between both systems can be performed by extracting the translation values $[t_x, t_y, t_z]$ and Euler angles $[\alpha, \beta, \gamma]$[1] from the extrinsic matrix and by applying the following transformations to them:

$$\begin{aligned}\mathbf{t}' &= [-t_x, t_y, t_z] \\ \mathbf{r}' &= [\alpha, -\beta, -\gamma]\end{aligned} \tag{7.1}$$

The results can be transferred to the camera's transform component that is mandatory to each Unity object to store its pose.

After building a replica of all physical components of the Blended Space, the virtual augmentations can be added to the scene. These may include virtual characters, props, and environments as well as additional lighting, sound sources, and visual effects that are not present in the real environment. There are no restrictions regarding whether to position the virtual objects within the borders of the physical room or beyond. The only limiting factor is the availability of a physical display surface, since, for example, no virtual contents can be displayed at the ceiling of a Blended Space if no such projection wall or projector was used in the CAVE configuration (see Fig. 5.1 to 5.3). Limitations of the displayable content should be taken into account during the authoring process to avoid undesired cropping of virtual objects. In particular in a CAVE without a ceiling, virtual objects that are taller than the user or floating above eye level usually should be placed behind the CAVE's walls. Otherwise, they increase their projection height drastically when the user moves towards the wall.

[1] Unity is using the Tait-Bryan convention since all three rotations are performed around distinct axes $(R_z(\gamma) - R_x(\alpha) - R_y(\beta))$ while Euler angles in the strict sense use the same axis for the first and the third rotation (e.g., $R_z(\gamma) - R_x(\beta) - R_z(\alpha)$) [Die06]. Rotations in Unity are extrinsic, i.e. they always use the fixed rotation axes of the parent object or the world coordinate system.

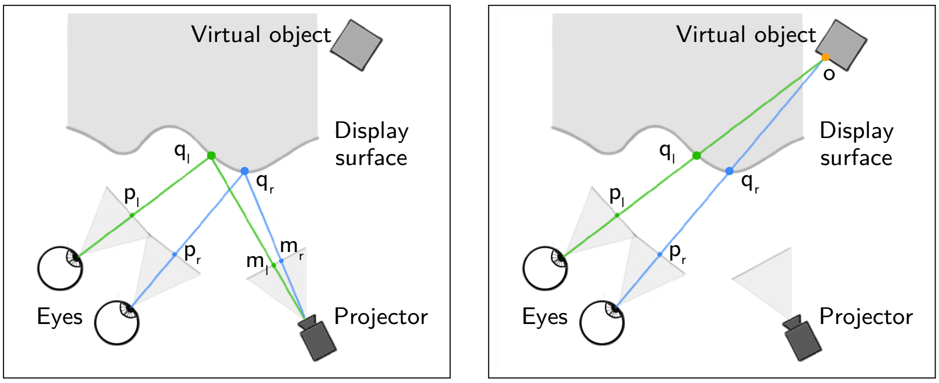


Figure 7.1.: Geometric relationships that are necessary to implement (a) the projection render stage, and (b) the user view render stage.

7.2. SAR Rendering Pipeline

For a derivation of the rendering steps to be performed, we first consider the geometric relationships that are illustrated in Figure 7.1. A projector pixel $\mathbf{m}_i$ is mapped to the point $\mathbf{q}_i$ on the physical display surface. For this purpose, each projector within a Blended Space projects a 2D image with virtual content onto the display surfaces of the physical scene. In case of stereoscopic display, two different images for the left and right eye have to be projected on top of each other. The physical display surfaces with projected textures are then viewed by the user. Stereo glasses can separate the overlayed images, so each eye is only receiving its corresponding image point $\mathbf{p}_i$. By interpreting the disparity information stored in the two received image points, the user's brain is constructing a virtual 3D point $\mathbf{o}$, that may be located at the same depth as the physical display surfaces, in front of them, or behind them. The actual and the perceived distance of $\mathbf{o}$ can deviate from each other due to conflicting depth cues such as convergence and accommodation. This effect increases with a higher distance between $\mathbf{o}$ and the display surface and is detailed in Section 8.4.2. In case of a monoscopic projection, the same process is applied with the difference, that there are no binocular depth cues to be used to position the virtual object within the Blended Space.

To compute the initially projected 2D images, we can reverse this step sequence. Due to the dualism of cameras and projectors, all capturing entities can be replaced by projecting entities and vice versa. The result is a rendering process with three semantic steps:

1. Off-screen rendering of virtual scene objects from the user's viewpoint.
2. Projection of the user view back onto physical scene objects.
3. On-screen rendering of newly textured physical scene objects from the projector's viewpoint.

The second and third steps can be implemented within one rendering pass as we will show below. We therefore end up with two rendering stages (see Fig. 7.2), that will be discussed in the following two sections.

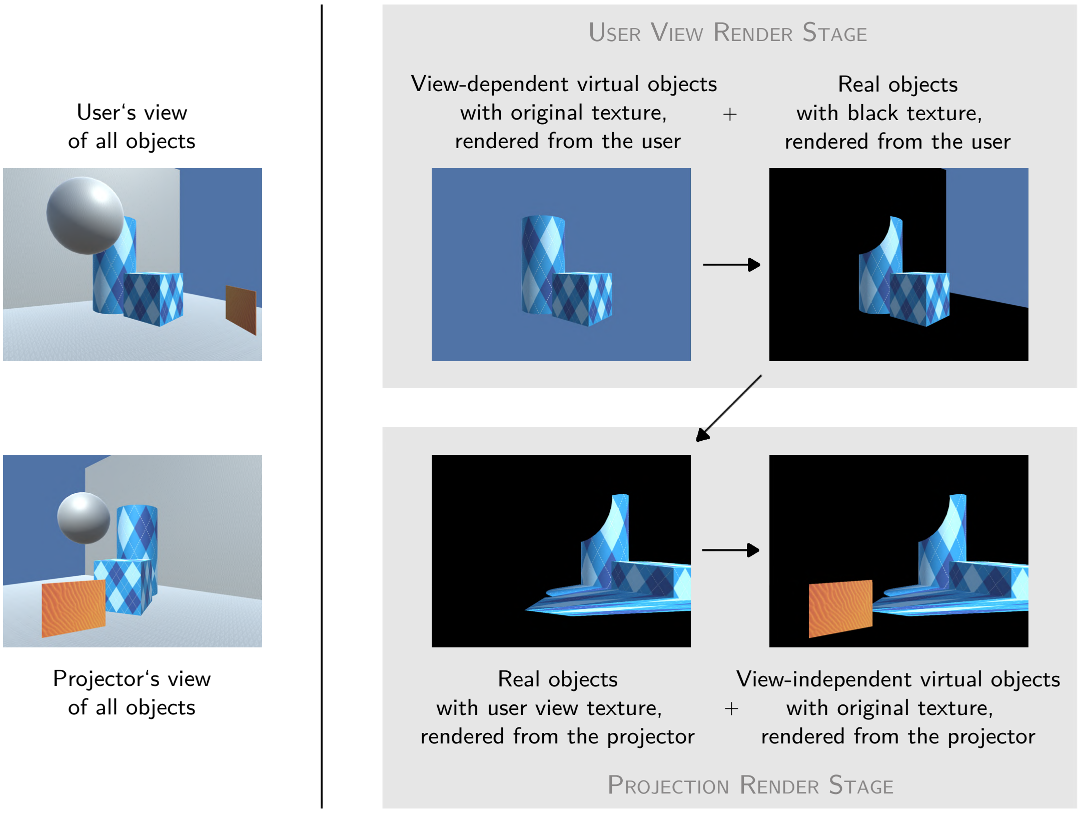


Figure 7.2.: Two stages of the SAR rendering pipeline. Real objects are depicted gray, view-dependent virtual objects blue, and the view-independent virtual object orange-colored. The two images on the left are only for illustration purposes but are not used in the final pipeline.

7.2.1. User View Render Stage

The first render stage captures the user view of the Blended Space. For rendering purposes, this should only involve additional virtual objects that are not already present in the Blended Space. Consequently, representations of physical objects have to be excluded. This is done by assigning all virtual objects to a specific layer and by restricting the culling mask of the user cameras to this layer. In a first draw call, this virtual layer is rendered to an off-screen texture target. A second, optional draw call handles occlusions of the virtual objects. This is only necessary for an augmentation scenario, where real-world geometry occludes virtual objects that are located at a greater distance. To achieve the desired effect, occluded parts of virtual objects have to be culled in the user view. This can be implemented by rendering all occluding real-world objects with a black unlit shader. The real texture of physical objects is not captured as it would be re-projected onto the real objects later on. Since the render buffer is not cleared between the previous and this draw call, the results of both rendering processes are combined. If real-world objects should be diminished as illustrated in Figure 1.1, the second draw call can be skipped. The resulting off-screen target texture, that is either the result of the first draw call or a combination of both calls, provides the basis for the subsequent stage.

7.2.2. Projection Render Stage

The second rendering stage targets the assembly of one 2D image per projector within the Blended Space. A projector image consists of multiple rendering layers, that focus on different objects of the scene. In the first step, a black background layer is rendered. For projection systems, the color black corresponds to the absence of light, and therefore no virtual content is projected onto the physical surfaces. If the displayed state of the Blended Space does not involve any virtual objects, which is the case for the left extreme of the RV continuum, the rendering process ends here.

For all other states that involve virtual objects, a second rendering layer is created. In this layer, all physical display surfaces are rendered from the projector's viewpoint. The original color of a surface point $\mathbf{q}_i$ is ignored, as we are only interested in the color of $\mathbf{o}$, which is the virtual point mapped onto $\mathbf{q}_i$. The point $\mathbf{o}$ corresponds to a pixel $\mathbf{p}_i$ in the target texture of the previous user view render stage. To replace the original color of $\mathbf{q}_i$ with the color stored in $\mathbf{p}_i$, a specific shader is used for all physical objects in the scene. This shader is different from the one used in the first rendering stage, which colored the same objects black to model occlusions. The texture coordinates of $\mathbf{p}_i$ can be computed by multiplying the point $\mathbf{q}_i$ with the view-projection matrix of the user camera. This matrix is known due to the continuous tracking of the user. With the resulting texture coordinates, a lookup in the user view texture can be performed and the corresponding color is mapped to the object point $\mathbf{q}_i$. Afterward, each projector applies a standard perspective projection to map all retextured physical display surfaces to a projector-specific 2D image.

In a third, optional step, view-independent virtual objects can be handled. As described at the beginning of Chapter 6, those objects appear like stickers in the scene. Therefore, they can be rendered directly from the projector's view without the need for a lookup in the user view texture.

The combination of all three draw calls results in a set of 2D images that can be projected by the corresponding calibrated projectors to create a Blended Space with both view-dependent and -independent content.

7.2.3. Special Case: Planar Quadrangular Surfaces

The described render stages can be applied to any combination of real and virtual geometry. However, if the display surface is both planar and quadrangular, such as a CAVE wall, the rendering process can be further simplified. First of all, the edges of the projector image and the projection surface have to be aligned. This can be done with a four-corner correction that is built in the projector or added to the Unity scene. The huge benefit of this approach is that the exact pose of the projector does not have to be known. In a second step, the user camera frustums have to be matched to the four edges of the projection surface. The frustums are usually asymmetric and therefore implemented with an off-center projection matrix. The resulting user view texture directly corresponds to the on-screen texture of the projector, so the projection render stage is redundant.

Part III.

Visual Perception in Blended Spaces

When real objects and 3D virtual projections are blended, spatial consistency is a crucial factor to present a convincing seamless environment to the user. This involves a correct spatial layout of objects, for example, through natural occlusions, as well as smooth transitions at the edge between surfaces. While there is a large body of literature in the field of depth perception in VR as well as traditional AR, so far only little research has been conducted on the characteristic features of projection-based spatial AR setups and their implications on the perception of such environments.

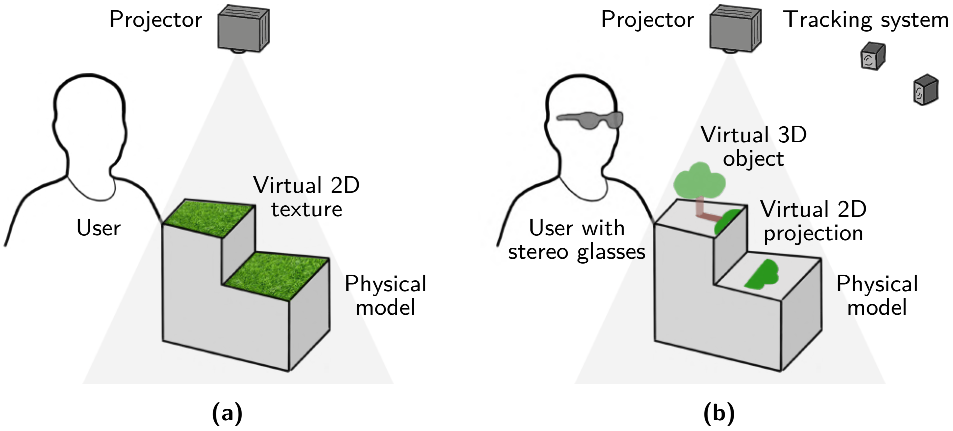


Figure III.1.: A virtual (a) 2D texture or (b) 3D object is projected onto two different surfaces.

As introduced in Section 6, Blended Spaces can feature two different kinds of projection (see Fig. III.1). The first category involves projections that are not affected by the current position of the observer. By this means, the surface of 3D objects can be augmented with virtual 2D textures, while the overall layout of scene objects remains unaffected. However, visual arts prove that even small changes of an object's surface material, such as modifications of its color temperature or brightness, can cause changes in its perceived depth. It is an interesting question whether similar strategies can be applied to Blended Spaces to manipulate perceived relationships between augmented objects. In the second category, 3D virtual objects can be detached from real-world geometry, displaying them at any position within the Blended Space. In this case, even a single object could be projected onto different surfaces with varying depths, orientations, or forms. As the projected parts might differ with regard to their luminance contrast, blur, and stereoscopic parallax, a consistent spatial impression may not be guaranteed. In particular, a mismatch between the distances of a virtual 3D object and corresponding projection surfaces causes the so-called *vergence-accommodation conflict*, which can lead to focusing difficulties (see Sec. 8.4.2). As this conflict does not exist in REs, it is one of our major research interests to what extent users can adapt to the conflicting depth cues in projection-based Blended Spaces. Therefore, in this part of the thesis, we present a series of studies, which investigate the perceived spatial relationships between users and their projection-based environment.

Based on the two introduced use cases, we address the following research questions:

1. Can projections modify the way users perceive spatial relationships in Blended Spaces?
2. Are projections of 3D stereoscopic content on different display surfaces perceived consistently in Blended Spaces?

We discuss practical implications and individual differences in the perception of depth between observers, and we outline future directions to influence and improve human depth perception in the real world.

8 Chapter 8. Psychophysical Fundamentals of Human Depth Perception

In this chapter, we introduce the physiological and optical fundamentals that are most important to understand human depth perception in Blended Spaces. We will review several cues that are used by the human visual system to derive depth information from 3D scenes. Based on these depth cues, we will discuss implications on the perception of virtual objects, including a number of perceptual illusions as well as cue conflicts that might occur in Blended Spaces.

8.1. The Visual System

In order to study the human depth perception in Blended Spaces, we first have to develop a basic understanding of how our visual system processes stimuli from the surrounding environment. This process of light reception, retinal image formation and extraction of depth information involves several psychophysical and optical mechanisms, which are specified in the following two sections based on the fundamental work of Palmer [Pal99], Bruce, Green, and Georgeson [BGG03] as well as Goldstein [Gol13]. First, we will take a closer look at different physiological components of our visual system, which play a crucial role in spatial perception. Afterward, some optical principles to map a 3D environment on 2D images are considered.

8.1.1. Anatomy of the Human Eye

When light enters the eye, it first passes a transparent front layer, the cornea, which accounts for about two-thirds of the overall refractive power of the eye. The iris is placed underneath the cover of the cornea. It has a circular opening, which is called “pupil”. Since the diameter of the pupil regulates the amount of light entering the eye, the pupil acts as an aperture of the eye [BGG03].

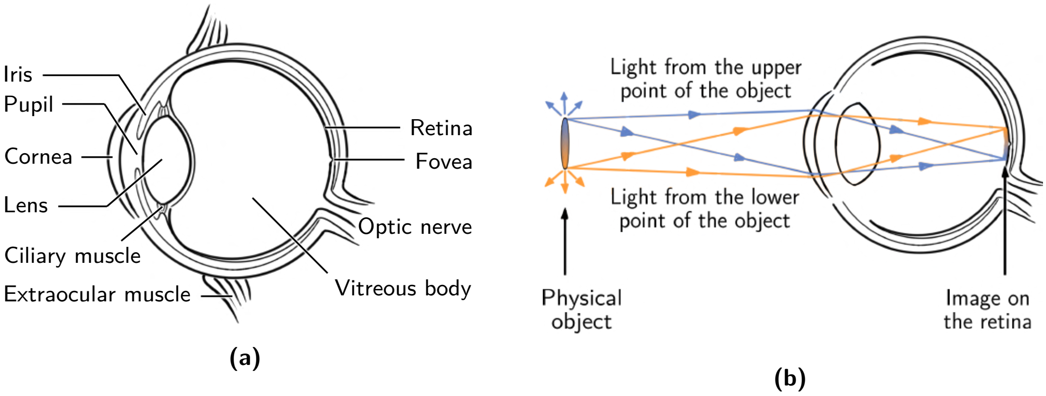


Figure 8.1.: (a) Anatomic model of the human eye, and (b) image formation process (based on [BGG03]).

Behind the iris, the eye's lens is situated. Along with the cornea, this capsule-shaped, transparent structure forms the refractive part of the eye. Its shape can be altered by contraction or relaxation of the connected ciliary muscles. These changes in lens curvature are necessary to allow the eye to focus on objects at different distances. For this reason, the entire process known as accommodation plays an essential role in human depth perception. It is explained in detail in Section 8.2.1.

After incoming light passed the first layers of the eye including cornea, pupil, and lens, it has to traverse the transparent, gelatinous vitreous body before reaching the inner layer of the eye, the retina. The retina contains millions of light-sensitive cells, called cones and rods, which are distributed across the surface with varying density [Gol13]. Rods are most sensitive to low light intensities, whereas cones respond to bright light, and therefore are responsible for color perception as well as the discrimination of fine details. The term fovea is used to refer to the point on the retina with the densest concentration of cones whereas no rods are present. Furthermore, the blind spot is a point on the retina without photoreceptor cells. It corresponds to the position of the optic nerve, which connects the retina with the brain. Therefore, the highest visual acuity can be reached at the fovea, while the visual field is obscured at the blind spot. By integrating the sensed light at all photoreceptor cells, a 2D image of the visual world is formed at the retina. This process involves multiple chemical and electrical mechanisms, however, we will focus on higher-level optical principles to explain the mapping between the 3D world and the 2D retinal image.

In this context, we introduce two additional axes that serve as important references in the following considerations. The *visual axis* is a line connecting the fixation point and the fovea. The *optical axis* is defined as the line that is normal to the lens surface, along which incoming light is not deviated by the lens. As the fovea is not located on the optical axis but is displaced around 5 degrees on the temporal side of the retina, the visual and optical axes are misaligned.

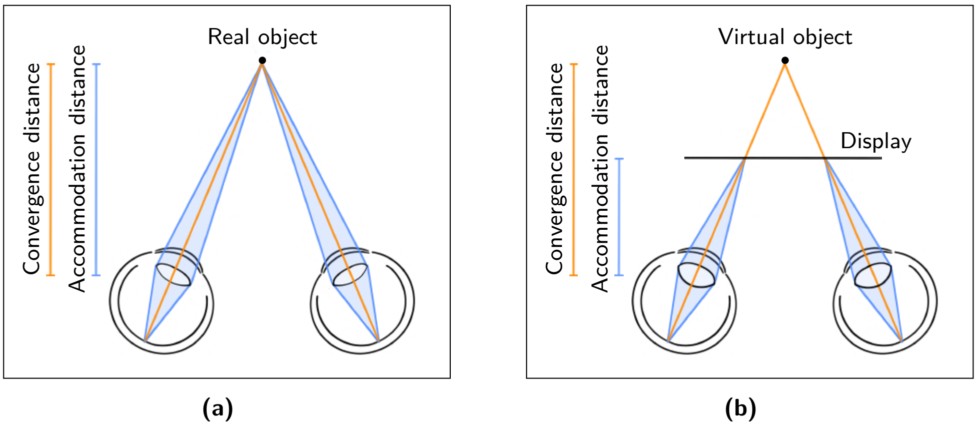


Figure 8.2.: Illustration of the vergence-accommodation conflict, which (a) does not occur in REs (accommodation distance = convergence distance), but (b) can occur in VEs (accommodation distance ≠ convergence distance).

8.1.2. Image Formation

When light reaches the surface of an object, it is scattered in different directions. A small portion of the light is reflected into the eyes as illustrated in Figure 8.1b. These light rays are refracted at the cornea and the lens and bend towards the retina. Without any eye disorders, all rays leaving a single object point should be focused on one image point at the retina. As depicted in Figure 8.1b, rays that are traced from the top and the bottom of an object produce an inverted image on the retina. The brain reverses this effect for the retinal images of both eyes, resulting in an upright representation of the visual world.

8.2. Depth Cues

In this section, we will summarize several factors, which contribute to the spatial impression of a viewed scene. We consider a set of nine so-called depth cues as suggested by Cutting and Vishton [CV95]. As the authors argue in their article, other commonly accepted cues such as linear perspective, texture gradient, and shading can be created by a systematic combination of these nine cues, and are therefore not considered separately in the following list.

8.2.1. Oculomotor Cues

The first category of depth cues is strongly related to the internal structure of the human eye, as discussed in Section 8.1.1. Before light rays reach the retina and therefore produce image points that can be analyzed by the visual cortex, several elements in the frontal part of the eye have to adjust themselves to focus the object point of interest. This is done by ciliary muscles, which are attached to the lens, and results in two different but related effects, that are described in the following.

Accommodation

Accommodation denotes the mechanism that allows the lens of the human eye to change shape to keep the retinal image sharp, regardless of whether the focused object is close or far. If a close object should be brought into focus, the ciliary muscles contract, the lens takes on a more rounded shape and, by this means, the refractive power of the eye increases. The reverse effect occurs when the muscles relax.

Convergence

Another measure of the distance to the focused object is the current angle between the optical axes of both eyes. To focus on a close object, extraocular muscles turn the eyes inward. Accommodation and convergence reflexes are neurally coupled, and therefore, both systems are responding if one of them is stimulated. In natural viewing conditions, these responses indicate a consistent depth of the focused object. In VEs, the latter is not true in the majority of cases, as we will discuss in the context of cue conflicts (see Sec. 8.4.2). Figure 8.2 illustrates the conflicting depth information in VEs in comparison to the natural coupling of accommodation and convergence in REs.

8.2.2. Monocular Cues

Monocular cues allow observers to derive a spatial impression of the viewed scene with only one eye. This naturally involves the aforementioned oculomotor accommodation cue as well as the two additional categories of pictorial and motion-induced cues.

Pictorial cues receive their name due to the fact that even flat pictures can provide these depth cues to the viewer. Indeed, visual artists make use of a variety of techniques to induce a sensation of depth in their 2D paintings (for a review see [HR12]). Some of them, such as height in the visual field, relative size, and occlusion, are related to the object's size or its relative positioning. Besides, there are tone-related cues, which are inferred from luminance distributions in a scene, such as shading and aerial perspective [TI12].

Occlusion

Occlusion occurs when an object is fully or partially hidden by another object that is placed in front. The (partially) occluded object appears further away than the occluding object, resulting in a relative order of scene objects in space. However, besides this ordinal information, no absolute depth values can be inferred. Nevertheless, occlusion is a powerful depth cue that typically dominates the following cues.

Relative Size

When at least two objects are mapped to the retina, the ratio of their retinal images' height is called relative size. To suffice as a depth cue, it requires the observer to have a rough understanding of the objects' sizes when viewed at the same distance. This may be due to similar sizes of the objects, as for a tree-lined road, or because the viewed objects are priorly known to the observer. The latter is usually denoted as familiar size and allows humans to extract absolute depth information from the scene. Additionally, it works if one object is present in the scene only. Without prior knowledge of the objects, relative size contains ordinal information as well as scaling factors, however, without an absolute point of reference in the scene. For example, a relative size of 2 only implies that for equal object sizes the second object's egocentric distance is twice as big as the first object's distance, but there is no indication of the exact depths of both objects (cf. Emmert's law [Gol13]).

Relative Density

Relative density as a depth cue is strongly tied to relative size since both cues are based on the concept of linear perspective. When a group of objects moves further away, it appears denser to the viewer. A similar effect can be observed for the surface textures of objects. When textures are viewed from a larger distance, their details become finer. As for relative size, relative density does not contain any absolute depth information unless the horizontal distance of objects or texture details is known beforehand.

Height in the Visual Field

When objects move away from the observer, they get smaller and elevate in the visual field. However, this is only true for objects that are not floating and therefore are based on a non-transparent floor plane. If the former conditions are met and the height of the observer's eyes above the ground level is known, this cue cannot only serve as a source of relative depth information but also allows humans to derive absolute egocentric distances of objects.

Aerial Perspective

In a real environment, light is scattered by particles in the atmosphere, resulting in a reduction of contrast when the viewing distance increases (e. g., [BF86]). This effect is called *aerial perspective* and can be observed for an object's texture and shading as well as the contrast between an object and its background. Its strength depends on the condition of the atmosphere and therefore can vary from area to area.

All of the monocular cues that were presented so far do not require any relative motion between scene objects and the observer and therefore also work for static 2D images. However, there is another category of monocular cues that integrates this additional source of information and uses it to improve the spatial impression of the scene.

Motion Parallax

Motion parallax occurs when the observer is moving by an otherwise stationary scene. Due to the observer's movement, the images of different objects on the retina are moving as well, with higher speeds for closer objects than for distant objects. Several research projects demonstrated that even microscopic head movements greatly enhance the spatial perception of the viewer, emphasizing the importance of motion parallax as a depth cue (e.g., [RG79]).

8.2.3. Binocular Cues

The last considered category of depth cues arises from the fact that our two eyes have a slightly different view of the world due to the interpupillary distance. The human brain uses the differences between the retinal images of the left and the right eye to infer the egocentric distance of objects.

Disparity

When an object point is viewed with both eyes, it is mapped to a specific location in the retinal image of each eye. The distance of these two corresponding image points, which is also called binocular disparity, can then be used to triangulate the egocentric distance of the object point. Binocular disparity is a quantitative representation of depth and therefore is often claimed to be one of the strongest depth cues.

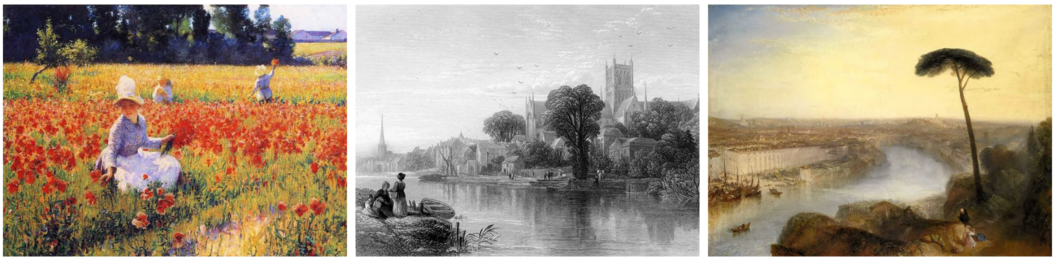
8.2.4. Range of Cue Effectiveness

According to Cutting and Vishton [CV95], the importance of the various depth cues depends on the distance of the viewer to the object of interest. To quantify the effectiveness of different depth cues, the viewer's environment is divided into three circular, egocentric regions: personal space $(0 - 2m)$, action space $(2 - 30m)$, and vista space $(> 30m)$.

As a general rule, occlusion typically dominates the other cues in all of the three zones. Beyond that, if none of the objects are overlapping, binocular disparity often provides the most accurate depth cue in personal space. Convergence and accommodation also indicate a close distance to the object since these depth cues are effective within arm's reach and slightly beyond.

Conversely, aerial perspective only gets effective in vista space, meaning the object of interest is located at a distance of more than 30 meters to the viewer. In general, oculomotor and binocular cues are greatly diminished in vista space, and therefore the human visual system has to rely mostly on pictorial as well as motion-induced cues for such large distances.

For the intermediate zone, the action space, disparity still contributes to the overall depth impression, however, pictorial cues such as height in the visual field and occlusion outperform this binocular cue. At about 30 meters, the border between action and vista space, the utility of binocular disparity declines to an effective threshold value of 10%, which Cutting and Vishton assume to be a margin for considering a depth cue effective for determining the layout of objects.



(a) *Coquelicots* by Robert Vonnoh.

(b) *Worcester* by William Miller.

(c) *Rome, From Mount Aventine* by J. M. W. Turner.

Figure 8.3.: Artwork featuring illusions of depth caused by variations of (a) color temperature, (b) luminance contrast, and (c) blur.

8.3. Depth Illusions

In addition to classical depth cues, which are used by the visual system to infer depth when viewing a natural 3D environment, visual artists make use of several depth illusions that mimic the effects of depth cues or take advantage of the optical structure of the human eye to create apparent depth in 2D paintings. In the following, we will present three such factors that can create illusions of depth, and we will discuss their correlation to the previously introduced depth cues.

8.3.1. Color Temperature

When viewing an image that shows different colored objects within dark surroundings, most people observe what is called positive chromostereopsis: warm colored objects tend to appear closer to the viewer while cool colored objects appear further away (see Fig. 8.3a). The opposite effect can be perceived when the background is white instead [DN93, Gol13]. Most researchers, who considered the phenomenon in the past, indicate a physiological cause for this visual illusion. When light enters the human eye, it is refracted depending on its wavelength. Shorter wavelengths, for example, blue light, are refracted more than longer wavelengths such as red light. This effect is also referred to as *chromatic aberration* and is a reason why long-wave light sources occur nearer than short-wave light sources when placed at the same distance to the viewer. Instead of the positive effect, some observers also experience the opposite negative chromostereopsis. Previous research projects attributed such individual differences in the perception of colored objects to physiological characteristics of the observer, for example, the location of the pupillary center [TMS93, Vos08]. In the past, several studies addressed the impact of an object's color to its perceived depth, although most of them focused on stimuli presented on a 2D display [BGDA07, Fau95]. In addition, a few perceptual experiments were conducted to investigate chromostereopsis in a real environment with different colored test objects [Atl10] or light sources [Hua07]. In all referenced setups a measurable effect of color on depth perception could be found.

8.3.2. Luminance Contrast

A second depth illusion that is related to an object's surface characteristics is based on the luminance contrast between the object and its background. By manipulating the luminance values of adjacent regions, the characteristics of aerial perspective can be simulated and therefore an illusion of depth can be added to an image (see Fig. 8.3b). The systematic correlation between perceived depth and luminance contrast was first revealed by Egusa [Egu83] and has been confirmed in several perceptual studies since then (e. g., [GD04, IKA07, TI12, OBO94]). In his studies with two achromatic hemifields, Egusa also found that perceived depth increased with increasing brightness difference, however, some participants tended to judge the brighter side nearer, and others the darker side.

8.3.3. Blur

A further important depth factor is the amount of blur perceived from an object, which has different underlying interpretations that the visual system might use to infer the distance of an object from the observer.

On one side, the aforementioned aerial perspective also accounts for the fact that the relative sharpness of an object's outline decreases with an increasing distance to the observer. Blur from aerial perspective usually means that the object in focus is far away; however, illusionary depth estimates are known to be caused even in closer distances in situations with unusual aerial effects, such as fog. Studies suggest that such blur can create a sensation of depth even in the absence of any other cues and under monocular conditions [Mat97]. Blur is a commonly used technique in arts, photography and video editing (see Fig. 8.3c).

Another cause of blur in the near field is related to the accommodation of the human eye. The primary stimuli that drive accommodative blur are the blur gradient in depth (changing focus sharpens the image) and changing vergence (convergence—accommodation). The extent of blur perceived in a real-world environment depends on the distance between the object in focus that is driving accommodation and the visual periphery, known as depth of field effects, which can thus act as a relative measure of depth. Accommodative blur is only effective in the near field, i. e., the contributions diminish after about six meters in common viewing conditions.

8.4. Depth Perception in Virtually Enhanced Environments

Depth perception in VR/AR was addressed by many studies in the past and most of them concluded that egocentric distances tend to be underestimated in virtual and augmented environments [CRWGT05, LK03, SJK$^+$07]. In the following sections, we will probe the causes of this discrepancy between real and virtual worlds, with a strong focus on projection-based environments. For a comprehensive literature review of distance perception in traditional AR environments see, for example, Swan et al. [SJK$^+$07] or Kruijff et al. [KSF10].

8.4.1. Parallax

One key aspect of most applications along the RV continuum is the capability to display virtual 3D content that is spatially integrated into the real environment of the user. To convey a stronger sense of a spatial presence, virtual objects are commonly displayed stereoscopically in such applications. In this context, objects may be rendered with *negative*, *zero*, or *positive* parallax, depending on their distance to the display. In the case of *zero parallax* objects appear on the projection surface and can be naturally viewed, i. e, the eyes focus and converge to the same points on the surface. In contrast, objects that appear in front of or behind the projection surface are rendered with *negative* and *positive parallax*, respectively, usually resulting in cue conflicts that are described in the next section.

8.4.2. Cue Conflicts

As the previous sections showed, many individual depth cues are used by the visual system to provide an estimate of relative and absolute distances of objects within a scene. The integration and interpretation of these cues usually create a consistent spatial impression of real environments. However, in VR and AR not all of the cues can be provided correctly, resulting in inconsistent or even conflicting depth information that can cause spatial misperceptions.

One of the most researched conflicts originates from different depth information provided by the accommodation and convergence cues. In a natural viewing situation, the vergence stimulus and focal stimulus are at the same distance and therefore the *convergence distance*, i. e., distance to the object to which the eyes converge, and the *accommodation* or *focal distance*, i. e., distance to the object at which the eyes focus to sharpen the retinal image, are consistent with each other (see Fig. 8.2a).

In a virtual scenario, convergence distance depends on the position in space where the object is simulated and therefore is the same as in natural viewing. However, the focal distance is fixed at the distance from the eyes to the display at which the two images for left and right eye are presented, resulting in discrepant results and the well-known *vergence-accommodation conflict* [DM96]. In order to see an object sharply without double vision, the human viewer must counteract the neural coupling between convergence and accommodation by accommodating to a different distance than the distance to which the eyes converge (see Fig. 8.2b). Unfortunately, this vergence-accommodation conflict may result in visual fatigue, visual discomfort, and spatial misperceptions as previous work has shown [HGAB08]. In particular, several studies reported a tendency towards depth underestimation in VEs where virtual objects are displayed with positive parallax as in head-worn AR and HMDs as well as projection-based VR (for a review see [RVH13]). Furthermore, Bruder et al. [BSOL15] reproduced this effect in a large projection system and also revealed a distance overestimation for close objects at negative parallaxes. Nevertheless, the influence of different technical and human factors on depth perception in VEs is still an object of investigation.

While most of the previous studies on depth perception in AR focused on see-through displays or handheld devices, only a small number of experiments have been conducted in projection-based environments. Considering the characteristics of Blended Spaces, there are several indications that perceived spatial relationships in such an environment differ significantly from those in traditional AR environments. In monoscopic projection-based systems, virtual content is displayed and viewed on the same depth plane, and hence a vergence-accommodation conflict does not occur [KSF10]. For stereoscopic projections, the difference between accommodation and convergence distance depends on the distance between the virtual objects and the physical surface. In many applications, such as the projection of surface details, this distance can be assumed to be comparatively small. A depth perception study that was performed by Benko et al. [BJW12] found that participants were reasonably accurate in their depth estimates, with a corrected average estimation error of ca. $1.3cm$. Participants of the study were able to perceive a virtual object's 3D shape and position even when projected on geometrically distorted backgrounds of varying color.

Another cue conflict that could affect depth perception in Blended Spaces was investigated by Broecker et al. [BST14]. They addressed the conflicting depth information provided by the physical surface of an object and virtual content, which is projected onto this surface depending on the current view of the user. In their studies, the authors compared perceived distances for physical and virtual objects that were providing different depth cues. Although no significant effects of the tested depth cues could be found, the results of the project indicate that monoscopic projections can produce a strong impression of depth, even if they have to compete with other cues such as accommodation.

Overall, the results of these studies suggest that both binocular disparity and accommodation provide important depth information in projection-based AR, and therefore can be used to resolve ambiguities created by other perceptual cues. So far, the interaction with other depth cues such as brightness differences or blur, which are inevitable when virtual content is projected onto different depth planes via one single projector, is mostly unknown. To fully understand how depth is perceived in projection-based environments both with monoscopic and stereoscopic projections, further investigations are needed. Hence, the next two chapters focus on spatial relationships as well as spatial consistency in Blended Spaces.

9

Chapter 9.

Effects of Projection-Based Illusions on Depth Perception

In this chapter, we present and evaluate multiple approaches to manipulate perceived spatial relationships between the user and real-world objects in a Blended Space. In particular, we focus on the following research questions:

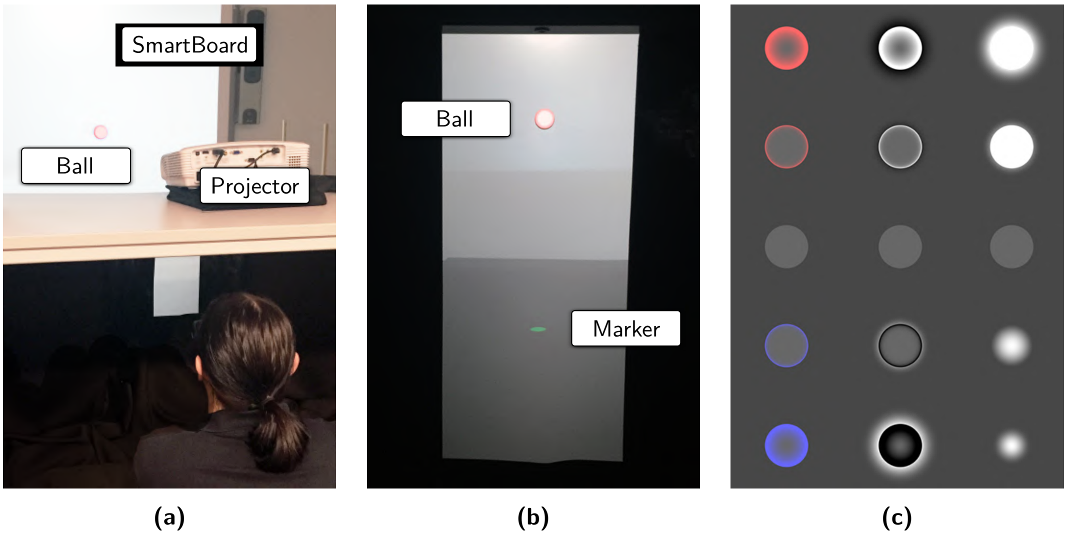
1. Can projection-based techniques significantly change the perceived depth of real-world objects?
2. Which of four common techniques used in visual arts (color temperature, luminance contrast, and blur) and filmmaking (binocular disparity) can cause significant changes in depth perception?

We present two user studies in which we evaluated the research questions. The results indicate that projected illusions can significantly change the perceived depth of real-world objects even in a complex environment with diverse distance cues.

9.1. Depth Manipulation in AR

A number of artificial depth cues have been proposed to aid the estimation of depth relationships between physical and virtual objects. For example, several distance indicators such as a virtual grid on the ground [TYK06], layers with different levels of opacity [LSIG$^+$03], and the tunnel cut-out [AST09] were used to improve the x-ray visualization of occluded objects. Furthermore, Wither and Höllerer [WH05] evaluated a set of monoscopic depth cues including vertical and horizontal shadow planes, top-down maps, and depth labels for an accurate positioning of virtual annotations at physical target points. All of these techniques have in common that they provide additional depth information by adding artificial elements to the scene that are clearly distinguishable from real objects.

In contrast, visual artists make use of various techniques to create apparent depth in paintings without adding artificial elements as presented in Section 8.3. The two introduced



(a) (b) (c)

Figure 9.1.: Illustration of (a) the experimental setup in the pre-study, (b) the user's view, and (c) the three projected monoscopic illusions with four different levels; the middle row shows the baseline.

subcategories of size-related and tone-related pictorial cues differ in terms of their suitability for projector-induced modifications. Manipulation of size-related cues in AR is difficult since this requires deforming the apparent shape of the affected objects. In contrast, luminance variations only affect the surface characteristics of scene objects while their shape remains unchanged, and therefore tone-related cues could be modified via projections.

Hence, for our preliminary study, we made a selection of three pictorial cues, which were rated as most practical in a Blended Space: (i) color temperature, (ii) luminance contrast, and (iii) blur. From the category of binocular cues, we chose a fourth depth cue, the (iv) retinal disparity, to compare it with the monocular cues.

9.2. Preliminary Study

We performed a pre-study to get a first impression of whether projected illusions can influence perceived spatial relationships between users and their environment [SBS16]. Our study involved the projection of different stimuli onto the surface of a white ball and its immediate vicinity (see Fig. 9.1a). The participants' task was to adjust a virtual marker in depth until they estimated the depths of both the marker and the ball to be identical.

9.2.1. Participants

We invited 17 participants, 14 male and 3 female (aged from 24 to 64, $M = 36.5$). The participants were members of the department of informatics at our university. All of them had normal or corrected-to-normal vision. We confirmed each participant's ability to perceive binocular depth with stereograms before the experiment.

9.2.2. Materials

Since there is only a small number of reported perceptual studies in spatial projection-based environments, we had to build our own prototype setup that meets the requirements of reproducible and ecologically valid studies (see Fig. 9.1a). We started from a slightly dimmed room, i. e., no direct sunlight interfered with the projection. For the purpose of testing the different illusions, any solid object with a light-colored, non-textured surface is suitable, provided that its material reflects light diffusely in order to avoid specular highlights. For our prototype, we chose a styrofoam ball that was illuminated by a projector (Optoma HD20). To avoid reference points that might influence the viewer's depth perception, we used a transparent fishing line to place the ball between two poles and therefore create the illusion that it is levitating. Overall, there were three pairs of poles, which made it possible to place the ball at three different distances from the viewer as well as the background. In order to realize changes in the background's luminance and color, we used a second projector in the form of a smartboard short-throw projection setup, which was placed behind the ball. We decided against using the same projector for both the foreground and background object, since the offset between the user's eyes and the center of projection produced shadows that potentially bias the results. By using two separate projectors, we were able to reduce this effect to a minimum. The third projector (Acer H5360) that can be seen in Figure 9.1 projected onto a table, which was situated beyond the levitating ball. It had a single purpose, namely the rendering of a small marker, which could be used to communicate the perceived depth of the ball later on. Via a connected mouse, the marker could be shifted along the z-axis, which corresponds to a movement towards the user or away from him. The user looked at the entire setup from the front side while all construction details and projectors were hidden by a mask. In a final step, a chin rest was mounted in front of the setup to relieve users and fix their head's position at the same time. The final setup for our perceptual pre-study is depicted in Figure 9.1.

9.2.3. Methods

At the beginning of the study, each participant was guided with a blindfold to the position shown in Figure 9.1a. Afterward, the experimenter provided detailed instructions on the depth estimation task the participant was required to perform.

For the main part of our study, we followed a repeated-measures within-subjects design. The independent variables were the number of used eyes (E), the real distance between the user and the ball (D), the applied illusion (I) and the gain of this illusion (G). To test the illusions both under binocular and monocular conditions, the experiment was divided into two blocks. In the first block, the participants wore shutter glasses while in the second block they had to cover their non-dominant eye with an eyepatch. Within each block, three distances at steps of 0.5 meters in two meters were considered. We did not alternate between these distances, since the ball had to be moved manually in our prototype setup, resulting in a time delay for every distance change. Instead, all conditions for a specific configuration $(E_i, D_j)_{i \in \{1,2\}, j \in \{1,2,3\}}$

were processed in a row, while the order of the distances was randomized between participants.

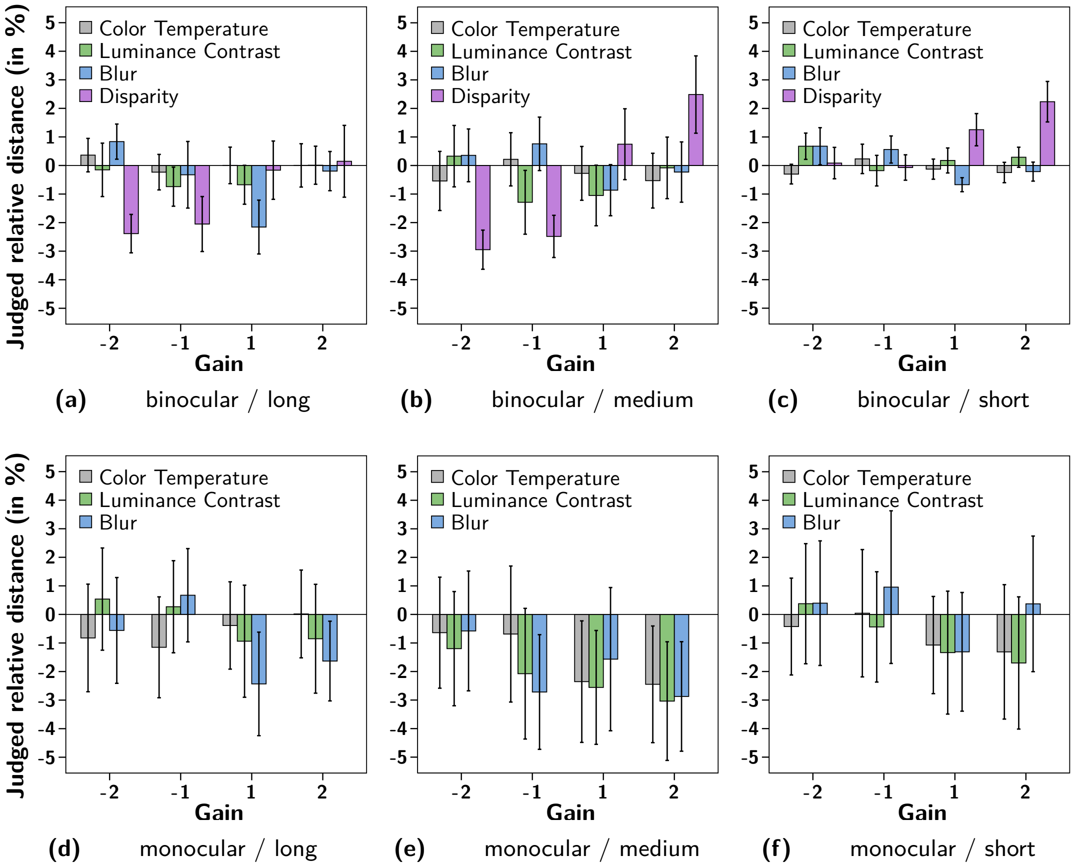
For the implementation of the illusions that were introduced in Section 8.3, we followed the approach of Bailey and Grimm [BG06]. In a perceptual study, the authors showed that modifying only the boundary of an object as well as the background can be sufficient to change the observer's depth perception. Since we intended to manipulate the real scene as little as possible, we decided to adopt this technique for the color temperature, luminance contrast, and blur illusion. For the remaining depth-from-disparity cue a golf ball texture was applied to the ball. This effect was skipped in the second part of the experiment, which relied on monocular vision. For every illusion, four gains were chosen to represent different effect levels, as illustrated in Figure 9.1c. It should be noted that the gains can not be considered to be equidistant. As discussed in Section 8.3, it can be expected that some of the pictorial cues are interpreted differently from observer to observer, for example, whether the brighter or the darker of two test objects appear closer. In addition to the illusion-specific gains, one baseline condition was inserted for every distance. For this condition no effect was applied to the ball, so systematic underestimations or overestimations of depth can be revealed and subtracted from the results later on.

Overall, every participant completed 90 different conditions (E, D, I, G) and, since every condition was repeated, 180 trials in total. At the beginning of each trial, the marker was located at a random distance, and the ball, as well as the background, were illuminated according to a randomly selected condition. Afterward, the participants had to position the marker exactly underneath the ball by scrolling the mouse wheel. When they confirmed the marker position, a new condition was selected. For every trial, we logged the estimated distance, which acts as the dependent variable in our study. After completing all trials, every participant filled in a questionnaire that collected both demographic data and qualitative feedback. Overall, the study took around 40 minutes per participant.

9.2.4. Results

We analyzed the results with repeated-measures ANOVAs and multiple comparisons with Bonferroni correction at the 5% significance level. We confirmed the assumptions of the ANOVA for the experiment data. Degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity when Mauchly's test indicated that the assumption of sphericity had been violated. Figure 9.2 shows the pooled responses plotted as judged distances relative to the baseline distance in percent. The baseline is the distance that participants indicated in the real-world condition without illusory stimulation.

We found no significant difference between the relative judged distances in the monocular and binocular conditions. In the binocular condition, we found a significant interaction effect between illusion and distance on relative judged distances, $F(6,96) = 4.59$, $p < .001$, $\eta_p^2 = .223$, and between illusion and gain on relative judged distances, $F(3.7,59.4) = 10.39$, $p < .001$, $\eta_p^2 = .394$. Furthermore, we found a significant main effect of gain on relative judged distances, $F(1.8,28.1) = 4.88$, $p = .018$, $\eta_p^2 = .234$. In the monocular condition we found a significant main



(a) binocular / long (b) binocular / medium (c) binocular / short

(d) monocular / long (e) monocular / medium (f) monocular / short

Figure 9.2.: Results of the estimated distances for binocular and monocular vision with three different distances each. The vertical bars show the standard error.

effect of gain on relative judged distances, $F(3,48) = 5.18$, $p = .003$, $\eta_p^2 = .245$. No other main effect or interaction effect was significant.

9.2.5. Discussion

Overall, the results of our preliminary study provide positive indications that distance judgments in a spatial projection-based environment can be changed with illusion techniques. In particular, under the binocular conditions, retinal disparity showed the strongest effect for all tested distances, which aligns well with results from previous studies in AR and VR environments. Moreover, for the longest distance, blur seems to be the most effective pictorial technique. Based on this observation, it can be hypothesized that the application of blur will allow the modification of perceived depth for even longer distances. However, further perceptual studies under SAR conditions are necessary to fully understand the approaches and their effects on spatial estimation.

Under the monocular conditions, we observed a comparably high standard error. This confirms previous studies, which claim that binocular disparity is the most important depth cue for the human visual system in close range [Jul64]. The correct estimation of the distance of an object proved difficult in the conditions in which the visual system could not make use of this cue in the study. In general, for most conditions, a depth underestimation can be noticed when one eye of the participant was covered.

A few limitations of our prototype setup can be inferred from the informal qualitative feedback, which was given in the concluding questionnaire and which provides practical insights for the development of future SAR experimental setups for perceptual studies.

According to the qualitative feedback, five of our participants stated that in some trials they estimated the ball's position even closer or farther than the marker could be moved on the horizontal table below the floating ball. We did not anticipate this for the tested distances before running the experiment since this would correspond to an underestimation or overestimation of more than half a meter in depth.

Moreover, one participant reported in the condition with the retinal disparity technique that he perceived a superimposed golf ball in front of the physical one and therefore estimated the depth of this virtual ball. This indicates that the manipulation of perceived depth in this condition might additionally be limited by whether or not participants perceive one or two targets; a large discrepancy in depth might favor the perception of different objects.

Furthermore, the participants were asked if they used or developed any particular cognitive strategy to complete the depth estimation task. Five participants answered that they compared the current illusion to that seen in the previous trial and tried to judge the relative difference in depth, which implies that future studies should include higher interstimulus intervals, and therefore make use of change blindness [ROC97] to reduce such effects.

9.3. Main Study

In the pre-study, we intentionally chose a setup with a high degree of abstraction to control depth cues that might affect the distance judgment. Since the results indicate that projected illusions can influence the perceived distance of objects, we decided to conduct a follow-up study within a more realistic and plausible environment. The main intention of this study is to investigate whether and to what extent these results can be replicated in such an environment or if the applied illusions are dominated by additional depth cues that are provided by a real scene such as the familiar size of objects, texture gradients, or shading. Retinal disparity dominated the other cues for all tested distances in the pre-study. However, we decided to neglect the binocular disparity technique to avoid the requirement for users to wear 3D stereoscopic glasses. This allows us to focus on monoscopic illusions and to provide a more natural viewing condition without an additional technical instrumentation of the user. Three different illusions that influenced (i) color temperature, (ii) luminance contrast, and (iii) blur were projected onto two real persons, whose relative distance had to be judged in a two-alternative forced-choice task (2-AFCT).

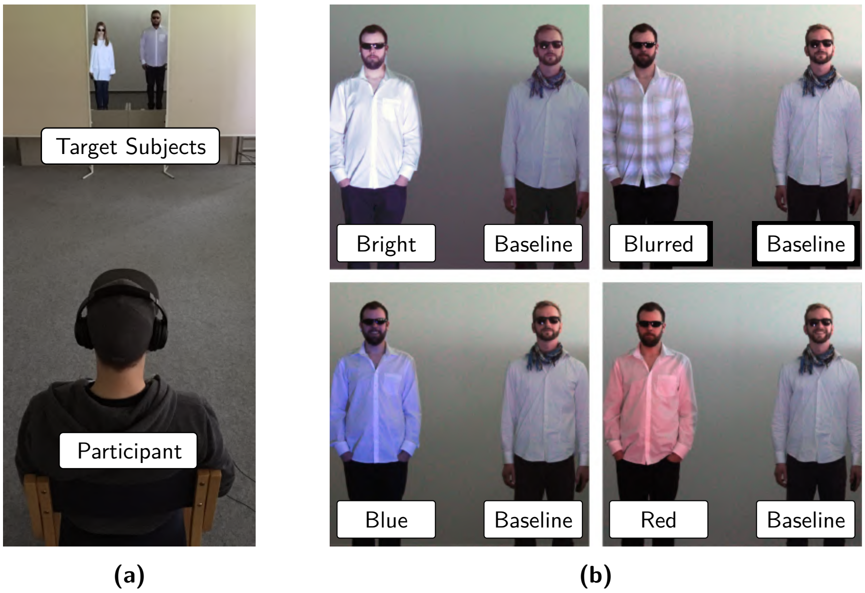


Figure 9.3.: Illustrations of (a) the experimental setup, and (b) some of the conditions from the user's perspective, i. e., left-to-right, top-to-bottom: bright, blurred, blue and red illusions always displayed against the baseline condition.

9.3.1. Participants

For the main study, we invited 12 male and 8 female participants, aged 20 to 58 ($M = 30.8$). All of them were students or members of our local department of informatics. Every participant had normal or corrected-to-normal vision. One participant reported a night blindness and five of the participants stated to have a limited or no stereo vision. This was confirmed by a graded circle test to evaluate each participant's stereoscopic acuity [FS97] and was incorporated in the analysis. No other known vision disorders, such as color blindness, dyschromatopsia, or an impaired eyesight were reported.

9.3.2. Materials

As illustrated in Figure 9.3, the study was conducted in a $17.4 \times 7.0m$ sized room, which was furnished with translucent window blinds to regulate daylight. During the study, the participants were seated on a chair, which was placed in the middle of the room with at least 2 meters clearance in each direction. Participants were facing two target subjects, which were actual persons standing at a distance of 10 meters. Both target persons wore white, untextured shirts and were illuminated with an Optoma GT1080 projector. The projector provided 2800 ANSI lumens and a contrast ratio of 25000:1 and was able to reproduce the standardized Rec. 709 color gamut. Partition walls that hid both the floor and the side walls were placed between the observer and the target subjects to restrain the participants from applying a strategy such as using reference points or comparing the tiptoes of both target subjects. In order to avoid acoustic feedback about the spatial relations in the environment, participants had to wear noise cancellation headphones during the study.

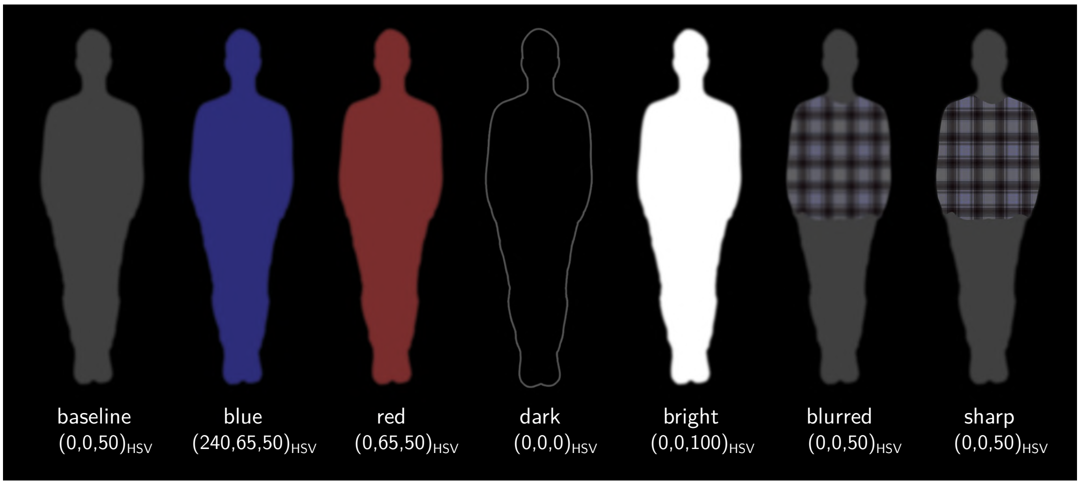


Figure 9.4.: Illustration of the six projected illusions and the baseline condition.

The main reason for choosing real persons as targets was the familiarity with the task of judging distances to them. All participants of the study should have a more or less good understanding of interpersonal distances as they experience situations that involve other people on a daily basis. This is not true for most artificial objects such as the ball, which was used in the pre-study. Also, a real person usually provides a variety of different depth cues that may conflict with the projected illusions. While we used white shirts for reasons of comparability and repeatability, we intentionally did not choose conditions that are best for projections, since those would not apply to real environments and prohibit any form of generalization. Therefore, it can be assumed that parts of the projections that cover the face or the legs of a target are less effective than those on the upper body. The research question is whether such realistic conditions still enable the manipulation of perceived distances or if the applied illusions are dominated by other natural depth cues.

9.3.3. Methods

Prior to the study, all participants filled out an informed consent form. Afterward, the experimenter guided the participant to the seat and provided detailed instructions on how to perform the given task. To familiarize them with the task, every participant passed an initial training phase. These training trials were excluded from the analysis.

In the main part of the study, the participants were required to perform a 2-AFCT by judging which of two target persons appeared closer to them [Fer08]. The target persons were augmented with different illusions. In the pre-study, an arbitrary percentage of the object's surface was affected by the projected illusions. For future applications, it might be interesting to figure out the best trade-off between the effect of projected illusions on the estimated depth and the conspicuousness of these projections. However, since the intention of the study is to investigate the general applicability of the illusions, we focused on the two extreme values for every illusion. This resulted in a total of six different effects as illustrated

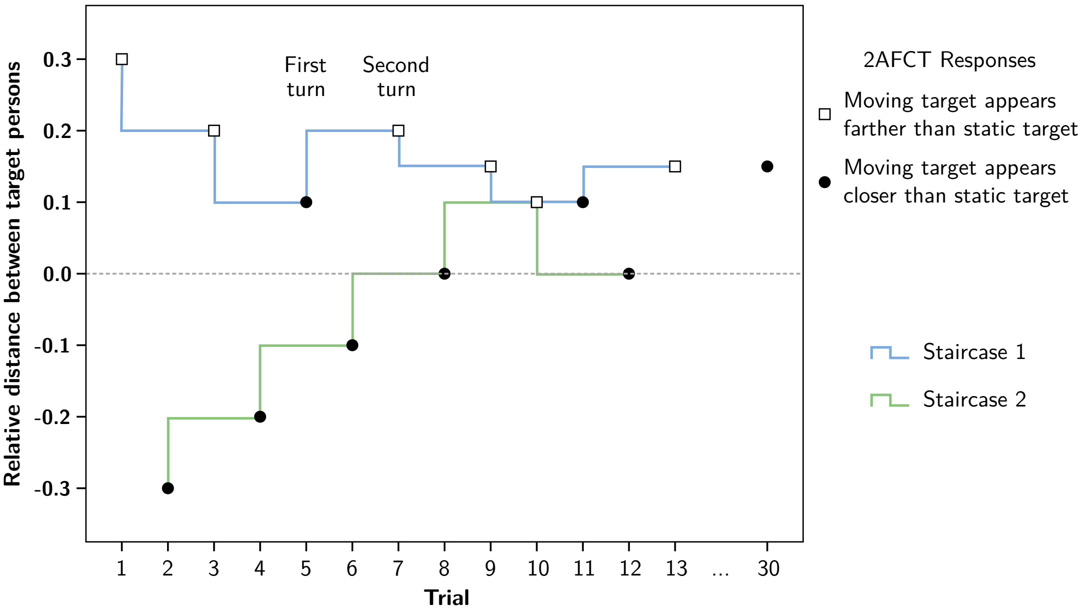


Figure 9.5.: Illustration of the applied adaptive staircase design.

in Figure 9.4. All projections were two-dimensional layers, which were adapted to fit the different silhouettes of the target persons.

The six effects were presented to the participants pairwise, with a direct comparison of the two color temperature values (blue vs. red), as well as the luminance contrast values (dark vs. bright), and the blur effects (blurred vs. sharp). In addition, each effect was compared to a baseline condition as shown in Figure 9.3b. For the baseline condition, a gray color was projected onto one of the target persons, simulating a low illumination of the scene. The gray value was chosen to be the mean of the values used in the bright and dark condition and also served as the basic color in the sharp and blurred illusions. In the latter illusions, a sharp or blurred checked pattern was projected onto the shirt of the target subject. Overall, nine conditions (three for effect vs. effect and six for effect vs. baseline) were presented fully randomized to the observer.

To analyze the effects of relative size on judged distances, two different pairs of target persons were used in the experiment. While two target persons were similar in terms of their body heights, the other pair had a height difference of approximately 20 centimeters. The presented pair of target persons as well as the positioning of these two persons and the decision, which of them is moving and which is static, were counterbalanced between all participants.

To obtain an absolute measure for the misjudgment of depth caused by a particular illusion, we altered the relative distance between both target persons following an adaptive staircase design [Cor62, Lee01]. For this purpose, one target person remained at a fixed position d_0 while the second person moved forward and backward depending on the participant's decision in the 2-AFCT. The moving target person started at an initial position $d_1 = d_0 + \Delta d_1$ with $d_0 = 10m$ and a randomly chosen $\Delta d_1 = +0.3m$ or $\Delta d_1 = -0.3m$. Based on the results

	M	SE	t	df	Sig.
$G_{red<blue}$	.136	.036	3.771	9	.004
$G_{blue<red}$	-.089	.023	-3.835	9	.004
$G_{bright<dark}$	.126	.024	5.219	17	.000
$G_{dark<bright}$	-.065	.020	-3.250	1	.190
$G_{sharp<blurred}$	.078	.034	2.336	9	.044
$G_{blurred<sharp}$	-.076	.028	-2.727	9	.023

(a)

	M	SE	t	df	Sig.
$G'_{red<blue}$	-.098	.028	-3.473	10	.006
$G'_{blue<red}$	.129	.040	3.224	8	.012
$G'_{bright<dark}$	-.132	.034	-3.917	13	.002
$G'_{dark<bright}$	.129	.034	3.816	5	.012
$G'_{sharp<blurred}$	-.086	.032	-2.667	9	.026
$G'_{blurred<sharp}$	.125	.033	3.772	9	.004

(b)

Table 9.1.: Mean and standard error as well as the results of (a) the paired t-tests and (b) the one-sample t-tests ordered by subgroups. The mean and the standard error for the paired t-tests refer to the paired differences.

of the pre-study we expected this relative distance of 0.3 meters to be easy to detect for all applied illusions. After the initial condition, the moving target person walked to the opposite position $d_2 = -d_1$, starting a second interleaved staircase. For all following trials i, $3 \leq i \leq 30$, Δd_i was computed as $\Delta d_i = \Delta d_{i-2} + \mathit{step}$, where the step width was positive in case the participant judged the moving target person to be closer than the static target person in trial $i - 2$ and negative, otherwise (see Fig. 9.5). The step width was initialized with 0.1 meters and halved after two turns in response behavior (cf. [BS14, Lee01]). By this iterative refinement, the interleaved staircase design converges to a distance difference Δd for which the participant perceives both target persons at the same egocentric distance. The mean difference was determined by the average of the last ten estimates [Cor62, Lee01]. Afterward, the initial state was restored and a new effect was projected onto both target persons. Between two trials, observers were required to close their eyes, which was signaled by a sound on their headphones and ensured by the experimenter.

After performing the 2-AFCT 30 times for every tested pair of illusions, participants were asked to fill out a demographic questionnaire and to give some qualitative feedback on their experiences. Overall, the study took around 45 minutes per participant.

9.3.4. Results

Figure 9.6 shows the pooled responses plotted as judged distance differences in meters. A judged distance difference near 0 is veridical, while a judged distance difference > 0 indicates an overestimation and a judged distance difference < 0 indicates an underestimation.

We analyzed the results with a repeated-measures ANOVA and multiple pairwise comparisons with Bonferroni correction at the 5% significance level. We confirmed the assumptions of the ANOVA for the experiment data. Degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity when Mauchly's test indicated that the assumption of sphericity had been violated. We found a significant main effect of the projected illusion on estimated distance differences, $F(2.94, 55.86) = 3.716$, $p = .017$, $\eta_p^2 = .164$. Post-hoc tests revealed significant differences between the illusions *dark* and *bright* ($p = .007$) as well as *blurred* and *bright* ($p = .048$) (see Fig. 9.6a).

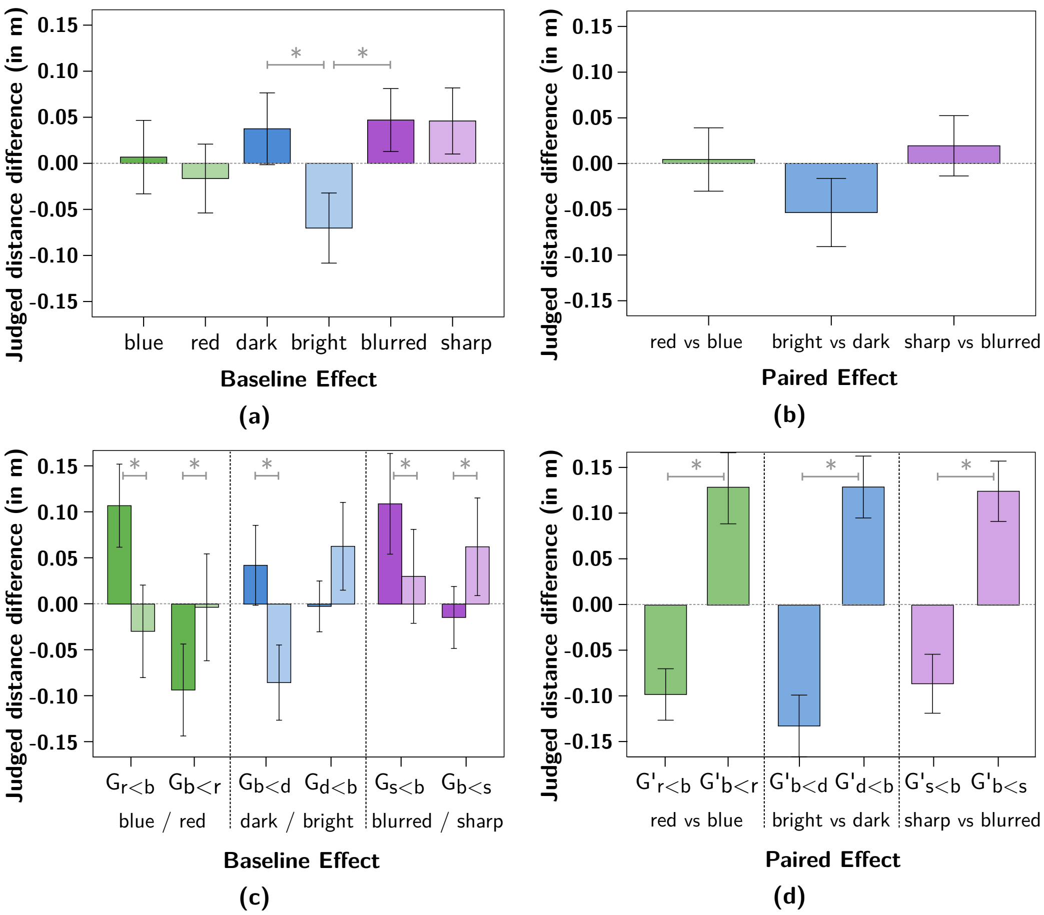


Figure 9.6.: Pooled results of the estimated distances measured in the 2-AFCT with (a) effects compared with the baseline and (b) pairs of two effects compared with each other. Subfigures (c) and (d) illustrate the results of the subgroup analysis. The vertical bars show the standard error.

We could not find significant differences between the illusions *blue* and *red* or *blurred* and *sharp*, however, we observed two different behaviors when comparing the estimated distance differences for the two color temperature values and the two blur effects. In particular, only ten out of twenty participants estimated the target that was illuminated red to be closer than the target that was illuminated blue (always in relation to the baseline illusion). For further analyses, we divided these participants into subgroups $G_{red<blue}$ and $G_{blue<red}$. Furthermore, half of all participants estimated the target with a sharp projection to be closer than the target with a blurred projection and vice versa. Hence, we divided them accordingly into subgroups $G_{sharp<blurred}$ and $G_{blurred<sharp}$. By comparison, only two participants judged the dark target to be closer than the bright target. For consistency, we formed two more subgroups, i. e., $G_{bright<dark}$ and $G_{dark<bright}$.

Based on these groups, we performed a subgroup analysis with one paired t-test for each of the groups. We found significant effects for the groups $G_{red<blue}$, $G_{blue<red}$, $G_{bright<dark}$, $G_{sharp<blurred}$ as well as $G_{blurred<sharp}$, as shown in Figure 9.6c. We could not find a significant effect for $G_{dark<bright}$. However, this result was expected due to the small sample size

of $G_{dark<bright}$. The results of the paired t-tests are summarized in Table 9.1a.

The analysis of the pairwise comparison of color temperature, luminance contrast, and blur values did not reveal any significant effects (see Fig. 9.6b). For further insights, we grouped participants in the same manner as for the effect-baseline comparisons. For every subgroup, we ran a one-sample t-test to determine whether the projection of illusion pairs results in a perceived distance offset different to 0.0. As illustrated in Figure 9.6d, we found statistically significant differences within all groups ($p < .05$). The exact individual values are listed in Table 9.1b.

In the post-questionnaire participants were asked to rate as how subtle they perceived the different projected effects on a five-point Likert scale, where 1 means the effect was very obvious to them whereas 5 means they did not notice the manipulation at all. According to the results, the most obvious effect was the manipulation of the color temperature with a mean value of $M = 1.50$ followed by the manipulation of luminance contrast ($M = 1.85$) and blur ($M = 3.85$). Six of the participants did not notice the blurring effect at all.

9.3.5. Discussion

The results of the follow-up study indicate that visual changes induced in a projection-based environment can manipulate perceived spatial relationships between users and their surroundings. In particular, the modification of the luminance contrast between a target and its background had a significant effect on its judged distance. On average, participants underestimated the egocentric distance of a target person when the person was brightly lit. Although the majority of participants perceived the brighter target to be closer than the darker one, we also observed the opposite effect in some cases, which underlines the inter-individual differences. A similar finding was described by Egusa [Egu83] as discussed in Section 8.3.2.

For the illusions based on color temperature, we also observed two reverse effects. Half of the participants judged the target person to be closer in the red condition than in the blue condition while the other half judged the situation in the opposite way. Again, this finding is in line with previous research that attributed such individual differences in the perception of colored objects to physiological characteristics of the observer, as well as to the background color (see Sec. 8.3.1). With a model that also includes such calibration parameters, a classification of participants into two groups and the according adaptation of the projected illusion could be feasible. However, further research studies are required to understand how stable these biases are, for example, related to perceptual learning [RPAG95], and whether individual or environmental effects might cause observers to change between the $red < blue$ group and the $blue < red$ group.

Assuming that such a classifier also exists for the depth perception of blurred objects, the significant effects achieved by grouping the participants would be practically relevant. However, to our knowledge, there is no evidence that humans systematically underestimate or overestimate the distance of blurred objects in contrast to the theory of color perception as described above. As discussed in Section 8.3.3, visual blur can have different causes.

The visual system adapts to some of them over time, for example, with non-corrected vision disorders. Others are related to relative and absolute depth cues, which are dependent on the distance from the observer. For instance, blur from aerial perspective usually involves longer distances than tested in our study. Conversely, blur from accommodation is known to have an effect within a distance of about six meters, while contributions to depth perception diminish for longer distances. It also serves as a relative depth cue, i. e., it indicates that a considered object is in front of or behind the focal point and may be interpreted either way as being closer or further away, even by the same observer. This might explain the differences observed in our study, but warrants more future work to correlate these effects in comparative studies. In general, we found that on average the absolute distance difference achieved by applying a blurring effect to the target subject was small in comparison to the effect observed in the pre-study. Considering the results of the post-questionnaire, this could be accounted for by the fact that only a small part of the target subject, the shirt, was blurred in the experiment, whereas the remaining parts were not manipulated. Facial features and the silhouette of the target person were not affected by the projected illusion. As a consequence, it might be that only observers who focused the shirt perceived an effect, and the overall effect might be reduced. Therefore, applying a blur illusion might be less useful in a real environment. We suggest that future work should perform these studies with eye tracking devices to gain information about where participants are exactly looking at when judging the distances.

When replicating the experimental task in future studies, the setup can be improved in terms of the general projection of the visual illusions. Instead of using predefined projection patterns, which fit the silhouettes of the target persons at specific positions, real-time tracking of the targets could increase the accuracy of the projection. This can be implemented using skeletal tracking in combination with a segmentation approach based on depth images of the environment. For real-world applications that are meant to incorporate AR illusions, such a method for aligning projections and targets would be inevitable.

9.4. Conclusion

In this chapter, we presented first approaches to manipulate perceived spatial relationships between the user and real-world objects by introducing perceptual illusions to a projection-based environment. For this purpose, we analyzed the effect of three monoscopic illusions, which are well-known from visual arts, i. e., color temperature, luminance contrast, and blur. The results suggest that the perceived depth of objects can be affected by projected illusions, even in a complex environment with diverse distance cues. In particular, we found that an increase of the luminance contrast between an object and its surroundings made the object appear closer to the observer. Manipulations of color temperature also caused significant effects in the perceived depth, however, they strongly depended on the observer. In order to account for individual differences between observers, even a short calibration routine with test stimuli would be sufficient, without the need for measuring specific values as the stereo acuity or the location of the pupillary center.

In the future, such individualizations might be of particular interest for AR devices that allow a private augmented view of the real environment, such as smart glasses or contact lenses. While we built on projection-based technology for our pilot study, the proposed perceptual illusions could also be transferred to other AR environments. However, this requires to factor in general depth misperceptions, which usually appear in such environments [SJK$^+$07, KSF10]. Improvements in AR hardware might also pave the way for other perceptual illusions. For instance, most AR solutions, except for video see-through AR devices, are not designed to remove light from the scene and are therefore incapable of darkening objects or placing virtual shadows on real surfaces. In particular, the effect of luminance contrast manipulations could be amplified by overcoming these limitations.

Another direction of future research is to extend investigations of projected illusions to dynamic scenes. Movements between scene objects and the observer result in motion parallax as a cue of relative depth. In the context of projection-based environments, motion parallax may interact with other depth cues and therefore could affect the perception of projected illusions.

Finally, though the observed effects were in the range of centimeters, we still have to consider possible consequences on the perception of Blended Spaces. In such environments, constant luminance distributions across different projection surfaces usually cannot be assumed. On the one hand, the physical objects, as well as the surrounding CAVE, may differ in their surface color and texture. On the other hand, affordable off-the-shelf projectors usually have a fixed focal plane, causing varying amounts of blur when surfaces of different depths have to be covered by the same projector. Another effect of using a single projector for multiple surfaces is a light fall-off for increasing distance to the projector. In the next chapter, we investigate whether all of these factors interfere with a consistent depth perception of Blended Spaces and if perceptual compensation techniques are necessary to improve the spatial impression in such environments.

10

Chapter 10.

Consistent Depth Perception in Projection-Based Blended Spaces

In the last chapter of this part, we intentionally induced perceptual illusions to a projection-based AR environment to modify the spatial relationships between physical objects. However, in Blended Spaces, such perceptual differences between display surfaces may occur even without intervention from the outside. In such spaces, virtual objects are projected onto 3D real-world geometry, which usually consists of several surfaces with various depths, orientations, and forms. Depending on the user's viewpoint, even a single virtual object could be displayed at multiple of these surfaces.

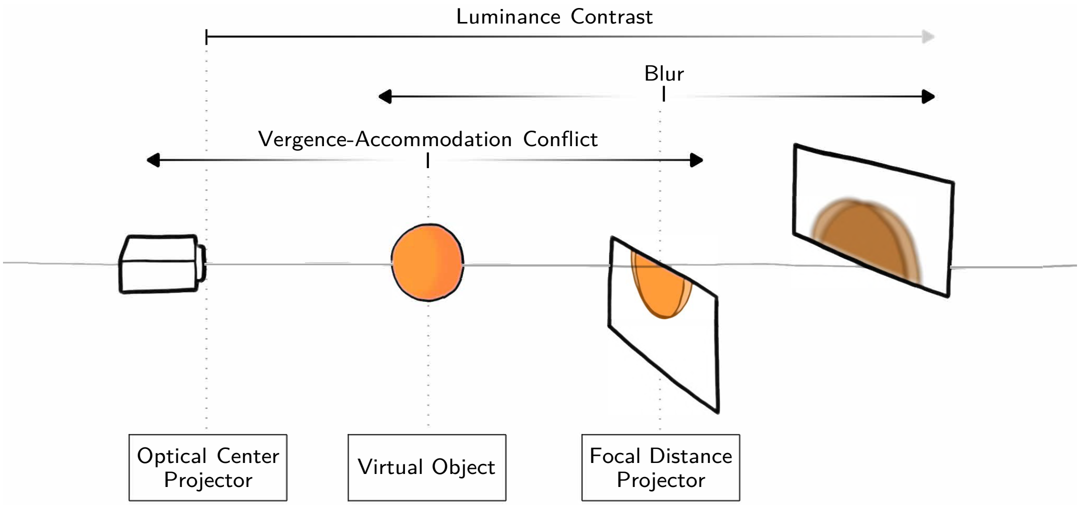


Figure 10.1.: Effects of the position of a display surface on the luminance contrast, blur, and vergence-accommodation conflict of the projection.

Figure 10.1 illustrates three potential cue conflicts, which might arise due to different distances between the projector and two involved display surfaces. To some extent, any projector shows a light fall-off for increasing distances, which results in diverging luminance contrasts for projections at different depths. Since most projectors can fix their focus only

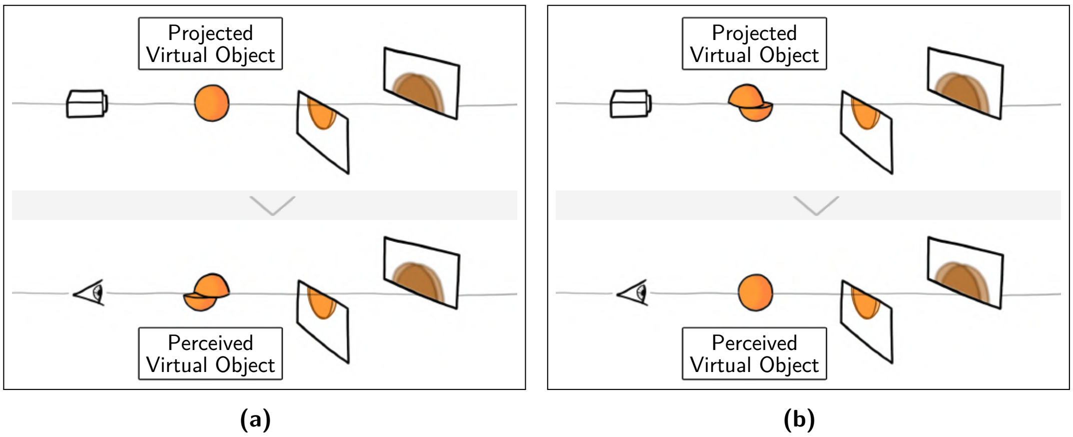


Figure 10.2.: Projection of a virtual ball onto two display surfaces with different depths. (a) shows a geometrically correct projection, that may result in an inconsistent spatial impression of the ball, and (b) shows a perceptually adapted projection with an expected consistent perception of the ball.

on a single distance at a time, projections on display surfaces with a different distance will be blurred. In addition, varying stereoscopic parallaxes at the involved display surfaces may cause different degrees of the vergence-accommodation conflict, as introduced in Section 8.4.2. Without regulation, these perceptual differences may cause undesired effects, such as an inconsistent spatial impression of virtual objects at the edges of real-world surfaces [RVH13, SBS20].

It is a challenging question how such conflicts affect the spatial perception of stereoscopically presented 3D objects. Furthermore, so far it is unknown if those visual conflicts could be reduced by perceptually-adapted projections, which compensate how objects are projected onto the surface; even if the perceptually-adapted manipulations lead to geometrically incorrect projections (see Fig. 10.2). To investigate these research questions, we performed two studies, in which we analyze the effects of stereoscopic parallax on human perception of consistent depth when stimuli are projected onto different surfaces. The contribution of this work is twofold:

1. A psychophysical study to validate the effects of stereoscopic parallax, varying luminance contrasts, and blur differences on depth perception when stimuli are projected onto multiple surfaces.

2. A confirmatory study to verify the findings with more ecologically valid stimuli.

The results provide important insights in how depth is perceived in Blended Spaces between different user groups.

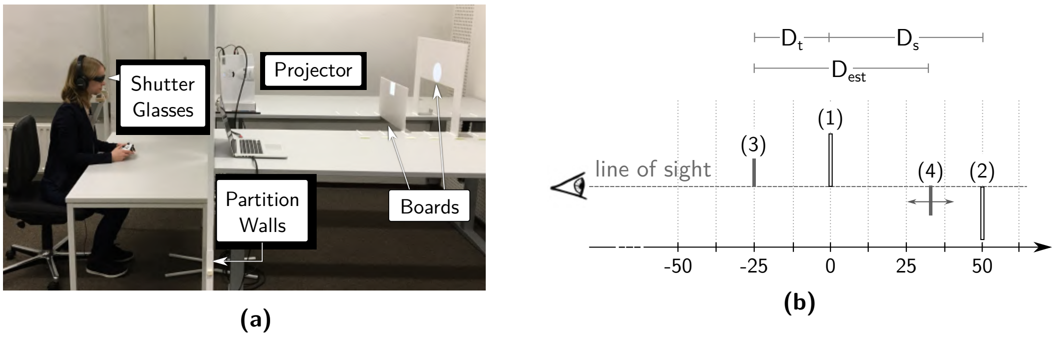


Figure 10.3.: Illustration of (a) the experimental setup and (b) the analyzed distances. In the shown configuration the (1) upper projection surface is stationary. The (2) lower projection surface and the (3) target object are placed at one of the illustrated distance marks at the beginning of each trial, while the (4) object was controlled by the user.

10.1. Main Study

In this section, we describe the study that was conducted to analyze the effects of stereoscopic parallax, luminance contrast, and blur on depth perception in projection-based AR environments. The study involved a perceptual matching task, in which participants were shown two connected visual stimuli top-bottom at different depths. Their task was to adjust the depth of one of the stimuli until they estimated that depths of both stimuli matched.

10.1.1. Participants

A total of 20 participants, 18 male and 2 female (aged from 20 to 38, M =28.2) were recruited through advertisements. All participants were students or members of the local department of informatics. All participants had normal or corrected-to-normal vision; six wore glasses during the study. One participant suffers from a mild kind of astigmatism. No other vision disorders, such as color or night blindness, dyschromatopsia, or a strongly impaired eyesight, were reported.

10.1.2. Materials

The experimental setup is illustrated in Figure 10.3a. During the study, participants were facing a real scene, which was augmented with two virtual objects via 3D projection mapping. To display the stereoscopic imagery we used an active shutter 3D system, including a 3D-capable Optoma HD20 projector as well as compatible RF shutter glasses. The projector was placed out of view behind two partition walls that also hid the mounting of the projection surfaces (see Fig. 10.3a). Through the restriction of their view, participants were restrained from applying a strategy such as using reference points for their depth estimation. To keep the projection center and the participant's eyes vertically aligned throughout all trials, a chin rest was used to fix the participant's head position. The virtual scene was projected onto two planar foam boards with a smooth, white-colored, diffusely reflecting surface. The boards were placed one above the other, both facing the participant. The initial alignment of the projector and the boards was performed in a one-time calibration step that utilized the RoomAlive framework (see Sec. 6.3.3) as well as a custom marker tracking implementation. During the main experiment, the boards were shifted manually using attachment points at pre-defined positions.

Since the boards were shifted in depth during the experiment, we aligned the focal plane of the projector with the board at medium distance. Hence, all objects projected onto this board appeared sharp, whereas objects projected onto the other board were slightly out of focus depending on the current surface distance. Although this blur effect was barely noticeable with the naked eye, a possible correlation between defocus and the estimated distance is considered in the analysis. The difference of illuminance between an object projected at a maximum distance and one projected at a minimum distance was $400lx$.

For the visual stimuli, we used a flat textured square and a circle with a randomized size between $8cm$ and $12cm$. We chose these reduced-cue stimuli in order to focus on binocular disparity and accommodation as the mainly available depth cues. The different shapes were used because of a pre-test, which revealed that participants heavily focused on the edges instead of the objects' depths when two squares had to be matched. These pre-tests also showed that a texture helped the participants to focus on the virtual objects, which is also in line with existing literature on surface textures rendered with stereoscopic displays [TGCH02]. Furthermore, in our first experimental setup, we dynamically adapted the size of the virtual objects in such a way that the retinal size was kept constant regardless of their current distance. However, this was rated as unnatural by multiple testers and therefore was discarded in the current setup.

During the experiment, the participant was required to control one of the projected virtual objects along its z-axis via a connected gamepad. The movement was not limited to a minimum or maximum value. The smallest achievable change of position of the virtual object with the gamepad was $1mm$.

10.1.3. Methods

Prior to the study, the interpupillary distance of every participant was measured to provide a correct stereoscopic rendering of the virtual scene in the trials. We verified each participant's ability to perceive binocular depth with the Titmus test, followed by a graded circle test to evaluate their stereoscopic acuity [FS97]. After passing all pre-tests, participants were instructed to sit in an upright position as explained in Section 10.1.2. They received detailed instructions on how to perform the required task. To familiarize with the setup and the stereoscopic stimuli, every participant passed an initial training phase before the actual experiment started. These trials were excluded from the analysis.

For the main part of the study, we followed a mixed factorial design with the three independent variables *surface offset*, *target offset*, and *moving board*. We define the *surface offset* as the relative distance from the movable board to the stationary board with a positive offset indicating that the movable board is further away than the stationary one as seen from the participant. Similarly, we define the *target offset* as the relative distance from the controllable object to the target with a positive offset indicating that the controllable object is further away than the target as seen from the participant. For every participant, the moving board was chosen randomly at the beginning of the experiment and afterward kept constant during the entire session. Overall, the decision if the upper or the lower board was moving was counterbalanced between all participants.

To investigate possible correlations between depth estimation and the offset between projection surfaces as well as the relative position of the virtual stimuli compared to the surfaces, we used two different configurations. In both of them, the stationary board was placed at a fixed egocentric distance of $200cm$ in front of the participant. In the first configuration, the target was always projected onto the stationary board with zero parallax. The dynamic board was moved between nine pre-defined positions with the following relative offsets to the stationary board: $D_s \in \{-50, -37.5, -25, -12.5,\ 0, 12.5, 25, 37.5, 50\}cm$ (see Fig. 10.3b). In the second configuration, the stationary board was still placed at its initial position while the second board was positioned at a relative offset of 50cm behind the first one. Furthermore, the target was moved between five different locations with offsets $D_t \in \{-25, 0, 25, 50, 75\}cm$. This corresponds to a relative positioning of the target in front of both boards, at the same depth as the first board, between the boards, on the same depth as the rear board and behind both boards, respectively.

According to the current condition, both the movable board and the target were moved to one of the pre-defined distances D_s and D_t before each trial. The shape of the target (circle or square), as well as its size, were chosen randomly. The second virtual object, which was controlled by the participant, was initialized at a random offset in the range of $-60cm$ to $60cm$. For each trial, participants had to move the controllable object along its z-axis to the perceived depth of the corresponding stationary target with the gamepad. Since the target and the controllable object differed in their size, participants had to rely on their depth perception instead of matching the objects' edges. The relative distance between the estimated depth of the controllable object and the target position was recorded as the dependent variable of the study.

To restrain participants from comparing the target position between trials, all conditions of both configurations were presented fully randomized. In addition, we introduced two trials at the beginning of each experiment session (after the training phase) to verify the participant's ability to perform the task correctly. In these two verification trials, both boards were placed at the same depth and the target was projected $25cm$ in front of or behind the boards. Thus, the task was reduced to adjust the same parallax for both visual stimuli and therefore should be solvable for every person with normal stereoscopic vision and correct understanding of the task. Including the verification trials, we tested 15 different conditions. This is because one condition was equivalent in the first and second configuration ($D_s = 50cm$ and $D_t = 0cm$) and therefore was only included once. After presenting every condition, they were repeated a second time, again in randomized order. This results in an overall number of 30 trials per participant. Between two trials, the participants had to close their eyes, which was signaled to them via headphones. The headphones also used active noise cancellation to minimize the bias through noise caused by the repositioning of the boards.

After the study, the participants completed a questionnaire to provide qualitative feedback as well as some demographic information. The total time per participant including questionnaires, instructions, training, experiment, and debriefing was around half an hour.

Considering previous results in the literature and the depth cues described in Section 8.2 our hypotheses are:

H1 Increasing the surface offset leads to increased absolute errors in matching depths estimates.

H2 Increasing the target offset leads to increased absolute errors in matching depths estimates.

H3 Participants experienced in the usage of stereoscopic 3D glasses provide more accurate estimates for matching depths.

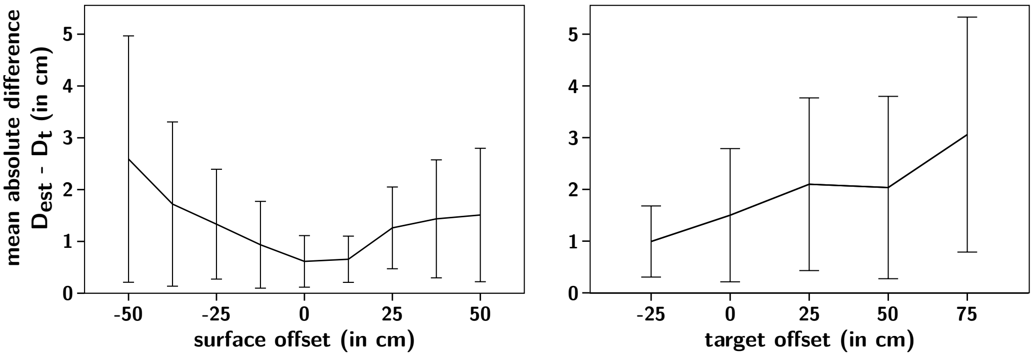
Although other, sometimes contrary cues also affect the depth perception as mentioned in Section 8.2, we still expect convergence and accommodation to be the most dominant cues in near-field AR, resulting in an underestimation of the distance to objects exhibiting positive parallax and overestimation of the distance to objects exhibiting negative parallax. The third hypothesis is mainly based on observations made in previous experiments. Participants who experienced stereoscopic display only very occasionally often reported difficulties in judging distances or focusing on virtual 3D objects, especially when these objects exhibited a strong parallax.

10.1.4. Results

For analyzing the results of the psychophysical experiment we discarded two participants from the data since their estimated depth extremely deviated from the target depth in the verification trials. Besides, five data points with values more than three times the interquartile range were considered as extreme outliers and were therefore also excluded from the analysis. On the resulting data set, we conducted multiple JZS Bayes factor ANOVAs [RMSP12]. Over the last years it has become increasingly apparent that the Bayesian approach to data analysis comes with considerable advantages over classical statistics, both theoretical and practical (e.g. [Die11]; for a systematic overview of more than 1500 articles reporting Bayesian analyses in psychology see [vdSWR$^+$17]).

Surface Offset

Figure 10.4(a) shows the mean absolute differences between estimated distances D_{est} and target distance D_t for the experiment conditions in which the target object was always presented at an egocentric distance of $200cm$, i. e., the target distance always matched the distance of the stationary board. The surface offset on the x-axis indicates the relative distance from



(a) **(b)**

Figure 10.4.: Pooled differences between the estimated distance and target distance ($D_{est} - D_t$) (a) for surface offsets D_s, and (b) for target offsets D_t measured on a 5-point Likert scale.

the movable board to the stationary board with a positive offset indicating that the movable board is further away than the stationary one as seen from the participant. Since the target object was always displayed at a distance of $200cm$, it was presented with zero parallax on the stationary surface. The vertical bars show the standard deviation.

A JZS Bayes factor ANOVA with default prior scales revealed that $H1$ was preferred to the null model by a Bayes factor of $B_{10} = 3353.142$. Therefore, the data provides very strong evidence for the hypothesis $H1$ that larger surface offsets lead to increased absolute errors in matching depths estimates. However, considering the results of every participant separately, we observed individual trends. Participants, who adjusted the object in front of the target for negative surface offsets, moved the object behind the target for positive surface offsets and vice versa. While the tendency towards underestimation of the distance to target objects that are displayed with positive parallax can be explained by the vergence-accommodation conflict, the opposite trend has to be induced by depth cues other than accommodation. We evaluated a correlation of the moving board and the reported strategies with the individual trends of participants, but could not find a reportable effect.

Target Offset

Figure 10.4(b) shows the mean absolute difference $D_{est} - D_t$ pooled over target offsets D_t. The vertical bars show the standard deviation. According to our hypothesis $H2$ we expected increased absolute errors in matching depths estimates with increasing target offsets D_t. To evaluate this, we performed another Bayes factor ANOVA with default prior scales, resulting in a Bayes factor B_{20} of 1501.625. According to Raftery [Raf95] this corresponds to a very strong evidence against the null model in favor of $H2$. Furthermore, we expected an underestimation of depth at all target offsets due to the vergence-accommodation conflict. Although this trend can be observed for a subgroup of participants, we also registered an opposite trend as in the previous configuration. In general, we observed a higher standard deviation for target positions further away from the user, which could indicate a dependency of the estimates from the egocentric distance rather than the target offset.

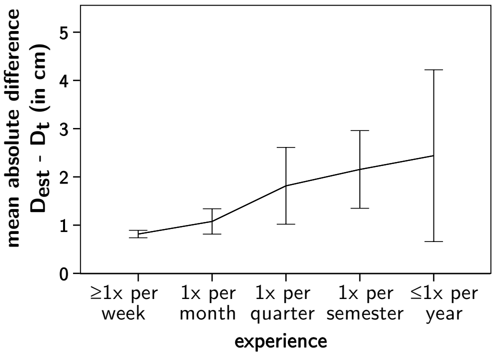


Figure 10.5.: Mean absolute differences between estimated distance and target distance $(D_{est} - D_t)$. The x-axis shows the experience of participants with stereo 3D glasses measured on a 5-point Likert scale.

Experience

For measuring the experience of participants with stereo 3D glasses, we used a 5-point Likert scale with values 'once a week or more', 'once a month', 'once a quarter', 'once a semester' and 'once a year or less'. Each option was chosen by 2 to 5 participants. For every participant, we averaged the means of the absolute error in the estimated distance of the 13 different conditions. The results are plotted in Figure 10.5. For the 18 participants of the experiment, a trend can be observed, suggesting a higher accuracy of distance estimation with increasing experience with stereo 3D glasses. To validate this assumption, we formed two subgroups of regular users, who reported to wear stereo 3D glasses at least once a month, and occasional users. A two-sample JZS Bayes factor t-test with default prior scales [RSS^{+}09] resulted in a Bayes factor of 1.799, suggesting a weak evidence in favor of $H3$ against the null model. To clarify this result, we analyzed the experiment data with an additional t-test, which revealed a significant difference between regular users (M=0.98, SD=0.24) and occasional users ($M = 2.16, SD = 1.20$); $t(16) = 2.15, p = .047$. This further supports our hypothesis $H3$; however, a larger sample could be considered in future experiments to allow a more differentiated analysis for several levels of experience.

10.1.5. Discussion

Overall, we observed large variance in the responses, which increased with larger offsets between the projection surfaces, whereas depth estimation was more accurate for small surface offsets. However, the study revealed different trends that are not correlating with the strategies, which were reported in the post questionnaire. To verify this observation, three participants repeated a shortened version of the experiment, in which they had to wear an additional Pupil Labs headset for binocular eye tracking. An analysis of the sample eye tracking data did not reveal a dependency between the focused board and the observed trends. Participants moved the controllable object back and forth until it leveled off at a depth they perceived as correct. In particular during the fine tuning at the end of each trial, their gaze switched between the visual stimuli several times.

While different strategies to solve the task do not seem to have an impact on the estimation error, the results of the experiment indicate a correlation between the experience of participants with stereo 3D glasses and the absolute difference between estimated depth and target depth. In particular, for an increasing surface offset the measured absolute error turned out to be higher for inexperienced participants. Including the finding that distance estimation was nearly veridical for experienced participants, although they were confronted with the same vergence-accommodation conflict, another interpretation is admitted that considers additional depth cues to binocular disparity. Participants with less experience in wearing stereo 3D glasses could have difficulty focusing the two visual stimuli with an increasing difference of the parallaxes and could therefore make use of other cues, consciously or unconsciously. This would be in line with the qualitative feedback that was provided in the questionnaires. The integration of luminance contrast between both visual stimuli could cause an underestimation of the distance to objects exhibiting positive parallax and an overestimation of the distance to objects exhibiting negative parallax as observed for a subgroup of participants. The opposite results may be caused by a subliminal depth compression. A common technique in 3D filmmaking is to reduce depth differences of the scene in order to minimize visual discomfort of the viewers [LHW^{+}10]. Regarding the conducted study, the parallax difference between the controllable object and the target could be reduced by moving the object closer to the projection surface instead of further away as expected due to the vergence-accommodation conflict. A smaller parallax difference results in a more comfortable viewing experience and could therefore influence the participant's depth estimation. Considering the overall results of the first study, the question arises, whether a perceptually motivated correction of object depths improves the spatial perception of a projection-based AR environment or if no perceptual inconsistencies occur for geometrically correct projections. For further investigation of this question, we decided to perform a follow-up study described in the next section.

10.2. Confirmatory Study

The results of the first psychophysical study suggest a correlation of the distance between both projection surfaces and the depth estimation error, which strongly depends on the participant's experience with stereoscopic 3D. We conducted a confirmatory study to test whether a compensation of this error results in a perceivable improvement of the spatial impression in near-field projection-based AR. Using the setup described in Section 10.1.2 the participants saw two virtual 3D objects; one was displayed without any modifications whereas the other's halves were shifted against each other according to the depth estimate error of the participant in the first experiment. The participants then performed a 2AFCT, deciding for which of the two presented objects they had a more consistent spatial impression.

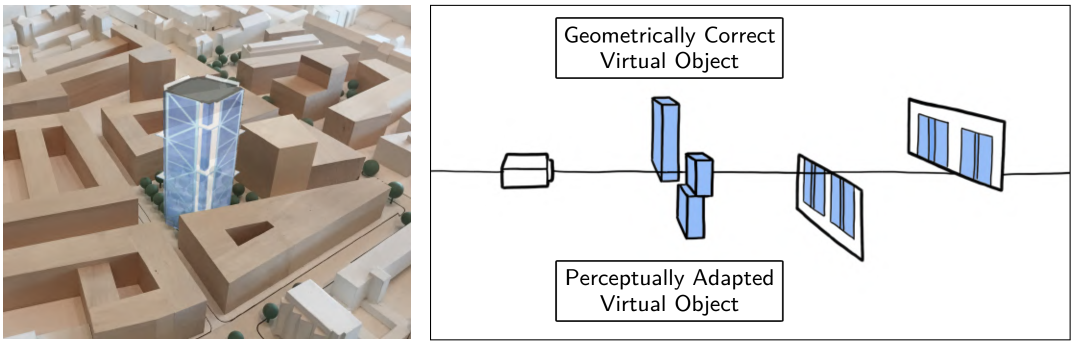


Figure 10.6.: (a) Illustration of an application of stereoscopic 3D projection mapping in an urban planning process, with a virtual 3D skyscraper that was used as the visual stimulus in the confirmatory study as depicted in (b).

10.2.1. Participants

From the set of participants of the first study, we recruited 12 participants, 11 male and 1 female (aged from 20 to 38, M=29.25). The sample equally represented the three different behaviors, which were identified in the first psychophysical study.

10.2.2. Methods

In the confirmatory study we followed a repeated-measures within-subjects design, which involved the surface offset ($D_s \in \{-50, -37.5, -25, -12.5, 0, 12.5, 25, 37.5, 50\}$cm), target offset ($D_t \in \{-25, 0, 25, 50, 75\}$cm) and moving board (upper/lower) as independent variables. Possible combinations of surface distance and target distance were the same as in the first experiment. However, this time each participant performed trials both with the upper and the lower board moving. To reduce the time for changing the boards' positions between the trials, all conditions were grouped according to the moving board. Therefore, the moving board only switched once after the participant finished all conditions of the first, randomly chosen group.

To simulate a realistic projection scenario, we used a virtual textured 3D model of a skyscraper as visual stimulus for the confirmatory study. For a better comparability, the size of both skyscrapers was kept constant through the experiment. As described in Section 10.1.3, both projection surfaces, as well as the virtual objects, were positioned according to one of 13 possible configurations before each trial. Unlike the first experiment, the upper and lower parts of both objects were not positioned independently of one another. Instead, they were placed either exactly one above the other or with a slight depth shift as described before. Each configuration was repeated twice, once with the geometric correct skyscraper projected on the left and once on the right.

In summary, participants performed 13 conditions with 2 × 2 repetitions each, resulting in 52 presented trials, which were randomly presented. Overall, one session took around 20 minutes to complete.

Table 10.1.: Bayes factors for comparisons of the models $M1$ and $M0$ as well as $M3$ and $M2$. The first row represents different levels of (top) surface offset D_s and (bottom) target offset D_t.

$\boldsymbol{D_s}$	-50	-37.5	-25	-12.5	0	12.5	25	37.5	50
$\boldsymbol{B_{10}}$	0.520	0.416	0.572	0.298	10.903	0.581	0.581	0.572	0.312
$\boldsymbol{B_{32}}$	5.956	3.980	6.913	1.525	235.048	7.097	13.533	0.145	1.876

$\boldsymbol{D_t}$	-25	0	25	50	75
$\boldsymbol{B_{10}}$	0.287	0.312	0.295	0.298	0.378
$\boldsymbol{B_{32}}$	1.000	1.878	1.435	1.525	3.258

10.2.3. Results

For analyzing the results of the 2AFCT we ran one-sample JZS Bayes factor t-tests with default prior scales [RSS+09] and a null value of the mean of 50 for each level of D_t and D_s.

To investigate whether the participants perceived a qualitative difference between the perceptually adapted and the geometrically correct projection, we compared the following two models:

M0 Random decision.

M1 Non-random decision.

The resulting Bayes factors are listed in Table 10.1. For different levels of the target offset D_t the t-tests resulted in Bayes factors B_{10} ranging from 0.287 to 0.378. According to Raftery [Raf95] this corresponds to a positive evidence of the hypothesis that participants were guessing randomly in case the target was positioned at $D_t \in \{-25, 0, 25, 50\}cm$ and only a weak evidence when $D_t = 75cm$.

For surface offsets D_s the t-tests revealed Bayes factors B_{10} between 0.298 and 0.915, suggesting only a weak evidence against $M1$. One exception is the Bayes factor for a surface offset of 0, which is in favor of the alternative model $M1$ against the null model $M0$ by a factor of about 10.902. However, this result was predictable, since no perceptual adaption of the virtual content should be necessary when both boards are positioned at the same depth.

Since there is no strong evidence in favor of either model $M0$ or $M1$, we additionally considered the following models:

M2 Preference of perceptually adapted projection.

M3 Preference of geometrically correct projection.

$M2$ and $M3$ were tested against one another for varying surface offsets D_s. This allows investigating the participants' preferences of either a perceptually-adapted or the geometrically correct projection, assuming a non-random decision. We conducted one-tailed t-tests with a

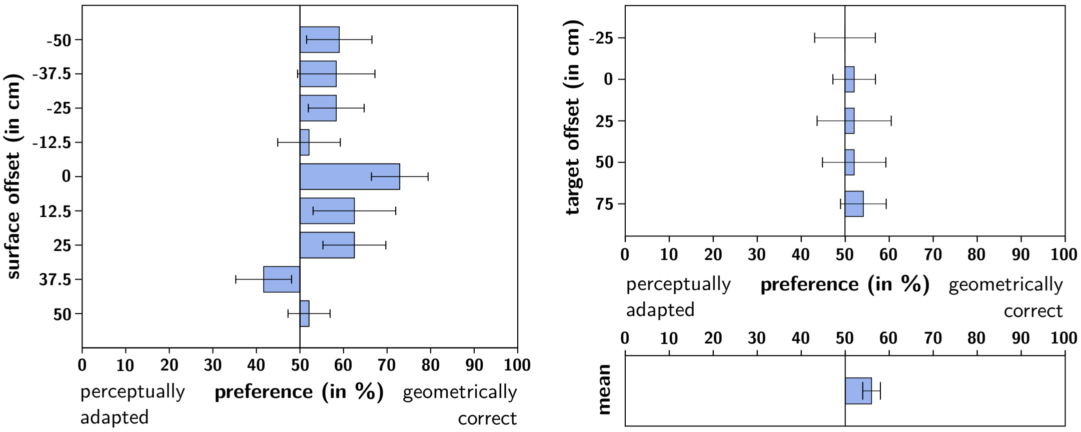


Figure 10.7.: Pooled results of the 2AFCT: (left) for the surface offsets D_s, (right top) the target offsets D_t and (right bottom) a mean value across all surface and target offsets. The x-axis shows a percentage value of how often the geometrically correct projection was judged as more consistent.

null interval of $(0\ ,\ \infty)$ or $(-\infty\ ,\ 0)$. By dividing the resulting values B_{30} by B_{20}, we got Bayes factors B_{32} as shown in Table 10.1. They suggest that the data favors $M3$ over $M2$ for six out of seven surface offsets.

For the sake of completeness Bayes factors B_{32} for different target offsets D_t are also listed in Table 10.1, although no strong evidence against one model or the other could be found. All results are illustrated in Figure 10.7.

10.2.4. Discussion

The Bayesian analysis provides indications that the perceptually adapted projection is not preferred to a geometrically correct projection of the visual stimuli, independent from the individual behavior in the first experiment. Decision rates close to 50% suggest that the 2AFCT was approaching our participants' sensitivity to differences in depth when using stereoscopic display. However, we also observed a tendency towards the geometrically correct projection, indicating that variances in adjusted distances from the first experiment may be caused by uncertainties and do not reflect a real perceptual difference between the depths of both visual stimuli. It still has to be investigated if the same results can be reproduced in a full-cue environment when the user's perception is influenced by other depth cues such as motion parallax.

10.3. Conclusion

In this chapter, we presented a psychophysical experiment and a confirmatory study to investigate the effects of stereoscopic parallax on the human depth perception in projection-based Blended Spaces. Such environments typically contain several surfaces with various depths, orientations or forms and, therefore, perceptual differences might occur when virtual objects are stereoscopically projected over multiple surfaces at different depths. To evaluate differences in depth perception and consistency of stereoscopically presented depth of virtual objects, we projected visual stimuli at two different surface planes with varying distances to the user. A perceptual matching task was performed, which gives indications on the depth perception in a spatial projection-based environment.

First, the results support the hypotheses that increasing offsets between multiple projection surfaces as well as the projection surfaces and projected targets lead to increased absolute errors in estimated depths. However, the relative errors differ between participants and therefore cannot be explained by the vergence-accommodation conflict in each individual case. The observed trends could be caused by individually perceived and weighted characteristics of a projection-based environment such as luminance differences of visual stimuli projected onto different surfaces. Considering the variance in the responses, it can be assumed that for most participants estimation of target distances was more difficult for larger surface offsets.

Furthermore, the results indicate that the effect of parallax on the estimation of matching depths strongly depends on the participant's experience with stereoscopic 3D. Participants, who wear stereo 3D glasses at least once a week, were able to match the depths of both stimuli with a mean error of less than one centimeter. This is an interesting result since it suggests that more experienced users perceive projected VEs in a different way than less experienced users. However, the confirmatory study revealed a tendency towards the preference of a geometrically correct projection of the visual stimuli, independently from the individual behavior of the participants in the first experiment. Considering practical applications of Blended Spaces, this could indicate that there is no need for a complex perceptual adaptation of the visual stimuli to create a spatially consistent environment. However, it also implies that offsets between physical projection surfaces and stereoscopically projected objects should be reduced to a minimum to facilitate perceptual integration of stimuli, in particular for users who are less experienced in the usage of stereo 3D glasses.

Future work should focus on the analysis of the learning curve for reliable depth estimations in stereoscopic environments. Furthermore, we would like to explore full-cue environments, in which the user's perception is influenced by other depth cues such as motion parallax.

Part IV.

Interaction Techniques for Shared Blended Spaces

In the previous part, we conducted a series of user studies to gain a deeper understanding of the perception of projection-based Blended Spaces, in particular with regard to stereoscopic projections. As the results indicate that depth misperceptions are in the range of centimeters, they should not crucially interfere with the interaction of users and their 3D mixed environment. Such interactions in Blended Spaces occur on two different levels: (i) to transition between scenes, and (ii) to operate within scenes.

Scene transitions are a basic functionality of Blended Spaces, as one of their three constitutive features is the seamless traversability of the RV continuum, according to the definition in Section 1. As many VR/AR applications also feature a scene graph, established transitions can serve as an inspiration for Blended Spaces. Examples include *smooth transitions*, that gradually add or remove elements of the next scene [SBH$^+$09, VF17]. In order to exploit the full potential of the environment characteristics, we developed a set of transitions in consideration of the structural conditions of CAVE-based Blended Spaces [NCSS18]. For one of these transitions, we create a virtual container with walls that are aligned with the CAVE. During the transition, the virtual walls tilt backward and thereby reveal the next scene. In a second example, we are using an elevator metaphor, with each floor representing a different scene. The elevator's movement is controlled by the user, and by opening the doors, the next scene is revealed. We compared these natural transitions with more classic image effects such as fade, glitch, and vortex. The natural transitions were yielding significantly higher presence scores and were also subjectively preferred by the users over the classic ones. On the other hand, they are more time-consuming, and may not be applicable to each scenario. Therefore, the context defines which of the available transitions are best for a specific application.

With regard to the particular scenes, standardized 3D interaction techniques can be applied, for example, based on controllers, natural gestures, or speech input [BKLJP04]. The presence of physical objects within Blended Spaces also enables tangible user interaction, which couples manipulations of the real world with changes of connected digital information [Ish08]. Furthermore, the input mode may be adjusted between scenes to match the current state of the Blended Space.

As single-user interaction techniques for different stages of the RV continuum are already well-researched, our investigations focus on the interaction with other real as well as virtual cooperation partners within Blended Spaces (see Fig. IV.1). As discussed in Section 2.2, such environments are characterized by a low level of user instrumentation and the capability of providing a shared interaction space for multiple users. However, projector-based systems using stereoscopic display are usually single-user setups, since they can provide the correct perspective for only one tracked person. Exceptions are projector systems as presented by Kulik et al. [KKB$^+$11], which use a high frequency to render different views for up to six users. However, such systems are customized, highly complex, and usually not affordable for applications that have to cover larger areas and therefore are based on more than one

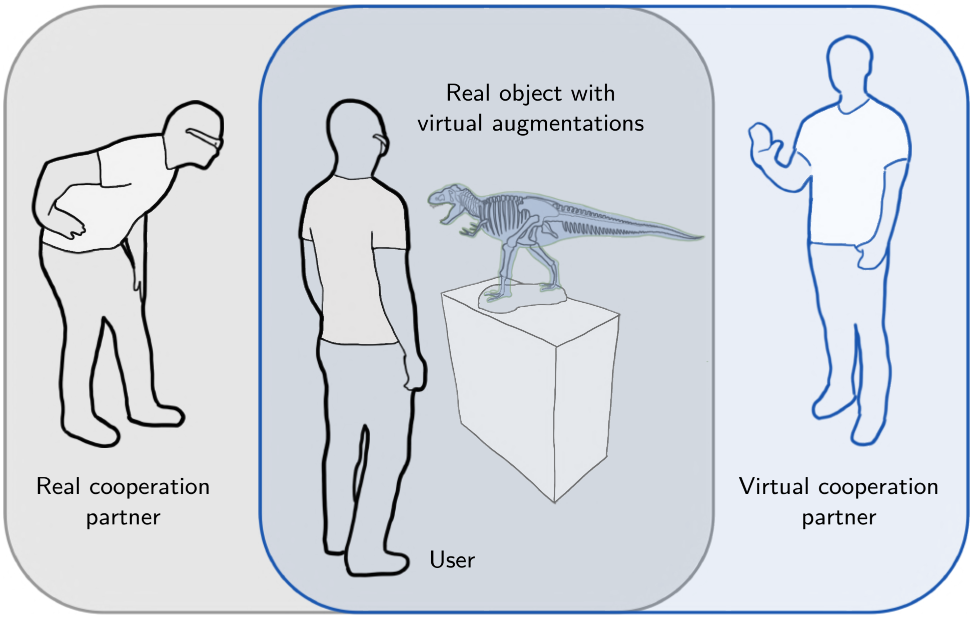


Figure IV.1.: Overview of participating entities in a shared Blended Space. The left section refers to the interaction between multiple users and the objects of a Blended Space, while the right section covers the interaction of a user with a virtual cooperation partner.

projection unit. To avoid the necessity of such specific projection hardware, multi-user support can also be approached with customized UIs, or by adjusting the projected content. An introduction to previous research projects that considered one of these options is given in Chapter 11. In the subsequent Chapters 12 and 13, we will present two own approaches to address user collaboration in projection-based environments with view-dependent 3D content. Both approaches take advantage of the structural conditions of the proposed Blended Space setup as described in Section 5 by utilizing multiple projection screens for different views.

While the first half of Part IV (i.e., Chapters 11 to 13) is focusing on the collaboration with other real users, the second half (i.e., Chapters 14 to 16) is considering the blended interaction with intelligent virtual agents (IVAs). Through the separation of display and user in projection-based Blended Spaces, the user instrumentation can be kept to a minimum, and there are no technical limitations of the field of view (see Sec. 2.2). Both factors could improve the perceived realism of IVAs; first, since the agents seem to exist in the environment instead of being displayed only on a user-worn device, and second because they are not cut off due to a small field of view. In Chapter 14, we provide an overview of related work in the context of IVAs, in particular with a focus on the factors of an agent's human likeness. The following Chapter 15 examines one of these factors, the agent embodiment, more closely and brings it into the context of Blended Spaces. Finally, Chapter 16 pursues the research question of whether IVAs are perceived as more realistic when they are able to manipulate not only their VE but also real-world objects within the Blended Space.

11 Chapter 11.

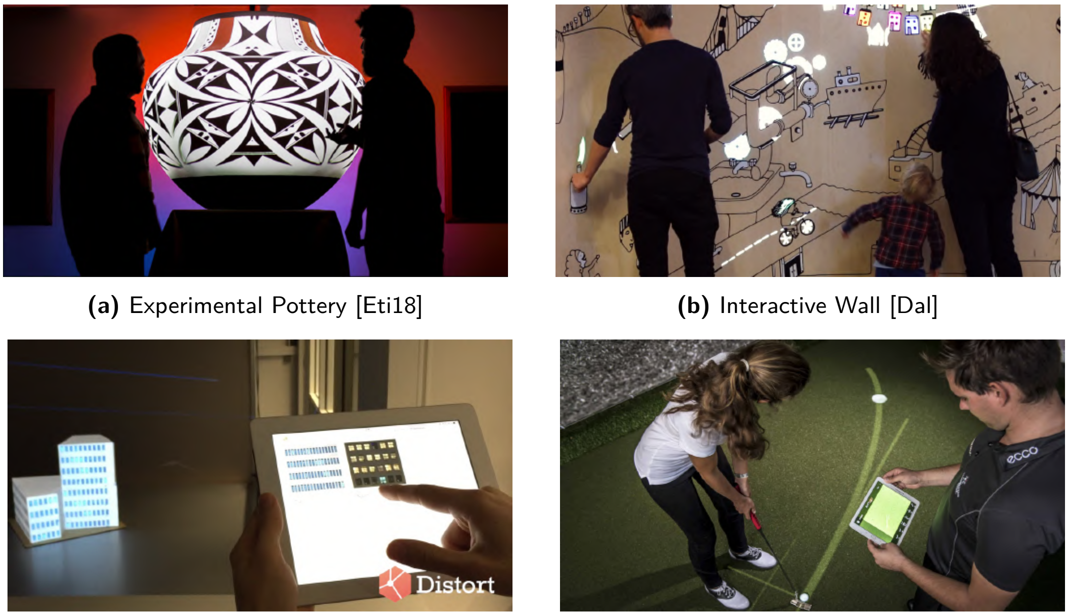
Multi-User Collaboration in Projection-Based Environments

There are several examples for collaborative projection-based environments that support multiple co-located users. Early work in this scope focused mainly on the augmentation of flat object surfaces with projected textures. *Augmented Surfaces* [RS99], the *Digital Desk* [Wel93], and the *Office of the Future* [RWC+98] are just three examples of workspaces that are virtually augmented via a system of projectors. By aligning virtual documents with the surfaces of tables and walls, they can be viewed independently from the current position of the user, and therefore information sharing among multiple participants is facilitated. The conceptualized *Office of the Future* additionally employed the floor as an extension of the projection space. In the following years, this concept was incorporated in several collaborative projects, both as an input and an output medium. The *iFloor* developed by Krough et al. [KLLO04] visualizes Q&A that were sent in by users via SMS and email. The selection of a particular question is implemented through a shared cursor, whose position is a weighted aggregation of all participating users. Inspired by the *iFloor*, Grønbæk et al. [GIK+07] developed a floor-based environment for collaborative gaming. More than ten users can interact with the rear-projected platform through the tracking of their limbs, such as feet, hands, or knees. As for the *iFloor*, the communication between co-located users is an integral part of the interaction design, as each user influences the shared state of the system.

One of the first applications that extended projections beyond planar surfaces was *Shader Lamps* by Raskar et al. [RWLB01]. While the projected content was still monoscopic with zero parallax, the augmented physical object could take non-trivial shapes. By this means, the material properties of real-world objects could be varied without the need to exchange the underlying object. The authors propose that this technique could be used to communicate ideas, for example in architectural teams or for city planning tasks. The idea to support multi-user experiences by projecting view-independent monoscopic content onto non-planar geometries was adopted many times. Prominent examples include the Disney theme parks, which use projection-based AR to virtually enhance buildings as well as exhibits and to implement interactive applications for groups of visitors [MvBG+12]. Besides public installations, systems such as the *IllumiRoom* [JBOW13] demonstrate the potential of this technology to enhance conventional media experiences such as gaming or home cinema. The prototype developed by Jones et al. is projecting virtual illusions onto the physical environment surrounding a television. By this means, the user's peripheral vision is stimulated, and the virtual field of view can be extended. An extension of this technology called *RoomAlive* [JSM+14] was already introduced in the context of projector calibration in Section 6.3.3. The mentioned calibration concept was developed in the scope of a proof of concept to transform an entire room into an immersive entertainment environment using multiple units of cameras and projectors. To support multi-user rendering, the authors suggest averaging the head positions of all viewers. The resulting viewpoint offers a satisfying approximation for all users when the virtual content is close enough to physical projection surfaces. If this requirement is not fulfilled, the system is designed to provide a single-viewer experience only.

Despite the restriction of distances between virtual and real geometry, *RoomAlive* presents one possible solution to support multiple viewers, even when the virtual objects are not projected with zero parallax. *ScreenX* [LLKN16] follows a similar approach, as the same virtual projections are presented to multiple viewers, or more specifically to visitors of a public cinema. The conventional screen on the front is extended by two screens on the left and right sides. To ensure a minimum average distortion of the images that are displayed at the side screens, they are deformed based on a mathematical model that samples viewing directions from each seat in the audience. User studies indicate that the proposed optimization of the projected content results in a more uniform movie viewing experience regardless of the seating locations in the cinema. Another solution for perspective display is demonstrated in the form of *Mano-a-Mano* [BWZ14], a projection-based face-to-face AR system that supports the collaboration of two users. A virtual object, that is floating between the users, is projected once for each of the two viewpoints. This concept is based on the assumption that the projections do not overlap if users are standing opposite to each other.

As can be seen from the various examples, different categories of projections impose varying requirements on user collaboration. In the most trivial case, 2D textures are mapped directly to the surfaces of physical objects. As projections can be viewed from any direction without distortions, multiple users can observe the virtual content simultaneously and without any dependencies from each other. Therefore, even single-user interaction techniques could be applied, as long as the collaborating users coordinate their actions to create the desired overall system state (example interactions are illustrated in Figure 11.1). We developed an authoring application for texture mapping that allows users to precisely align projected images and physical object surfaces via a mobile or desktop device [SDBS15]. By connecting multiple input devices to the projection system, the application could be easily extended to suit more than one active user. While this is a simple approach to manipulate 2D surface characteristics of real-world objects, it is severely limited in terms of stereoscopic 3D display. Depending on the current viewpoint of the observing user, different parts of the virtual object have to be visible to create a plausible illusion of spatiality. Moreover, the viewpoint also



(a) Experimental Pottery [Eti18]

(b) Interactive Wall [Dal]

(c) Distort [SDBS15]

(d) PuttView [Put]

Figure 11.1.: Examples of collaboration in 2D texture mapping applications, using (a) a shared tablet, (b) direct touch input, (c) multiple networked mobile devices, and (d) a combination of 3D input via a golf club and 2D input via a tablet.

influences on which physical surfaces the virtual content has to be projected. The latter may be diverging for users with different positions, particularly for high distances between the projected object and the display surface. As the independent arrangement of virtual and real 3D objects is a vital part of Blended Spaces, we developed two UIs to address multi-user support without limiting the displayed content. In the scope of the projects presented hereafter, we investigated the following approaches:

(Chapter 12) An operator-follower system with separate 3D and 2D views.

(Chapter 13) An operator-follower system with a shared 3D view.

In the context of these projects, we also showcase different forms of user interaction, including both 2D input (via mouse and touch), and 3D input (via controllers, customized tracked devices, gestures, and user movement).

12 Chapter 12. Layer-Based 3D Virtual Environment for Architectural Collaboration

In this chapter, we introduce a method for the collaborative exploration of projected stereoscopic 3D content, which is based on two different views of the same model. The proposed interface allows one user to experience a perspectively correct visualization of the 3D model. By specifying a region of interest within the 3D view, other participants can follow the first user's perspective in a second 2D representation that is not view-dependent.

The collaboration method is evaluated against the backdrop of a cooperative architectural design process. As discussed in Section 3.2, such processes involve a variety of users with different levels of expertise such as architects, engineers, investors, or end customers. To obtain a common understanding of the architectural models is an ambitious task as architects, as well as other involved parties, often need to work with 2D floor plans. While these plans are meaningful and easy to interpret for professionals, non-expert users often face problems when deducing 3D properties of a building. We explore a layer-based visualization method, which stacks 2D floor plans in 3D space providing a simple 3D impression without actually using a 3D model.

12.1. Layer-Based 3D Virtual Environment

In this section, we describe our layer-based 3D VE and user interface including the hardware and software components as well as the visualization and exploration techniques.

12.1.1. Hardware Setup

For the visualization of the architectural model as well as the 2D view, we use an *L-Shape* projection setup as described in Section 5.1. While this gets along with only two projection screens, other configurations for Blended Spaces are also supported.

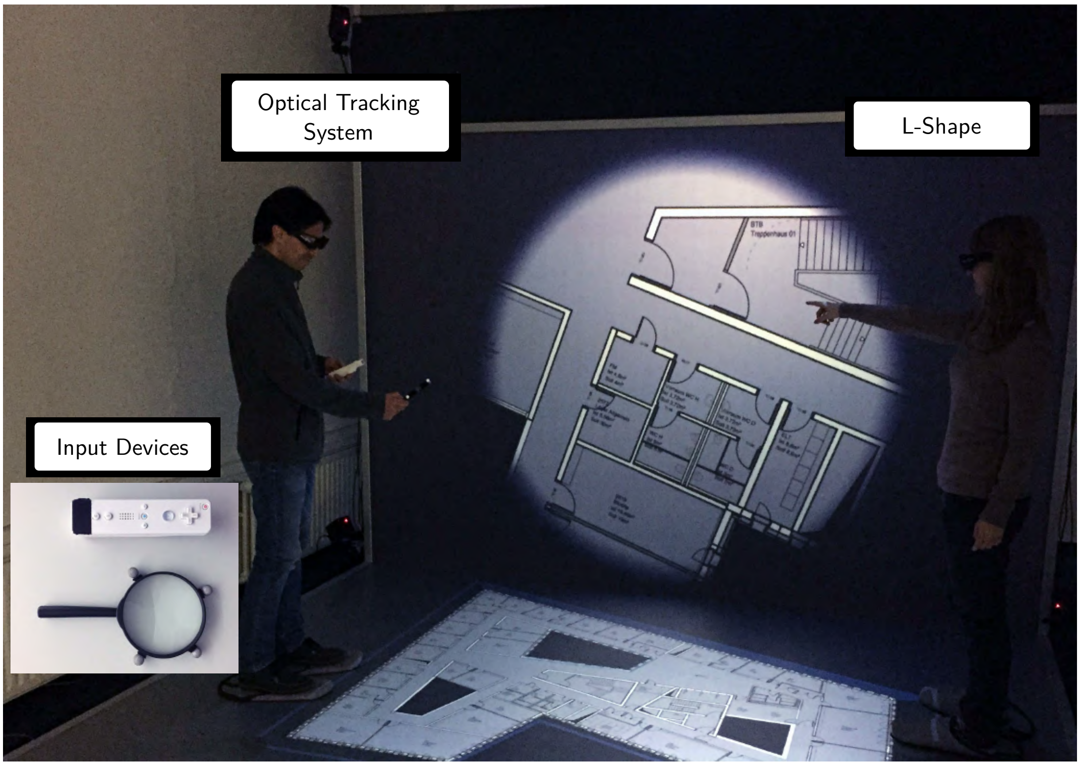


Figure 12.1.: Illustration of the experimental setup with an L-Shape consisting of a front and a floor screen. The inset shows a Wiimote and a magnifier lens with attached tracking markers.

We use two input devices for interacting with the architectural 3D models. First, a Wiimote controller is connected to the workstation via Bluetooth. In the current setup, only the buttons of the Wiimote are used for input. The second input device is an off-the-shelf magnifier lens that is equipped with 4 additional passive markers as illustrated in Figure 12.1. Its position and orientation are tracked by the ARTTRACK2 system and can be assigned to an object in the Unity scene.

12.1.2. Layer Construction

In architectural design processes, 3D data for a planned building is often not available in the early development phases. To allow users to get a spatial impression of the building even without existing 3D data, we stacked the existing 2D floor plans as illustrated in Figure 12.2a. Sectional views can serve as an additional source for proper ceiling heights as those are essential for obtaining an intuitive understanding of proportions and dimensions. While the stacking process is done manually in our prototype, it can be easily automated as long as the 2D input data complies with some basic formatting rules. Besides the building itself, the Unity scene contains a virtual representation of both the front and the floor screen. Hence, after calibration of the projectors and tracking system, the scene is an exact virtual replica of the L-Shape and therefore facilitates proper scaling and positioning of the building.

To support the collaborative process, we render the 3D building model on the floor screen creating the illusion of displaying a 3D block model standing on the floor (or alternatively on a pedestal in mid-air). In this setup, the front screen can display additional information like a more detailed or labeled 2D plan of the currently selected floor.

12.1.3. Collaborative Layer-based Interaction

The goal of this project was to create a user interface for exploring 3D models with a strong focus on collaboration. The described L-Shape setup is a suitable environment for this purpose as it allows multiple users to share one single interaction space. The challenge is to provide interaction concepts that ensure that all involved persons concentrate their attention on the same part of the model. To avoid the emergence of disorientation, all scene navigation tasks are executed by a single person whom we refer to as the *operator*. This position is typically held by the architect or the person currently guiding the conversation. The operator's head is tracked and the head's pose is coupled to the virtual camera. Combined with shutter glasses the operator gets a realistic sense of perspective when walking around the 3D model. Other users can wear shutter glasses as well, allowing them to perceive the model stereoscopically. However, in order to get a less distorted perception of the model, they have to stay close to the operator.

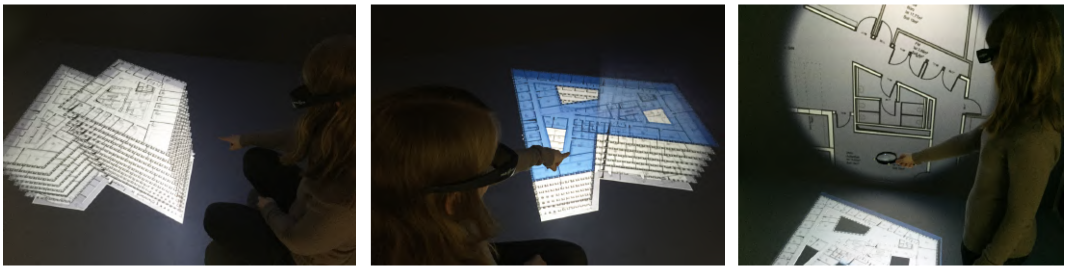
In a focus group meeting with architects, three different exploration modes were identified as particularly helpful. Using the buttons of the Wiimote, the operator can switch between these modes:

1. In the **overview mode**, the building is treated as a single uniform 3D object, which can be rotated around the y-axis (see Fig. 12.2a). In this mode, the overall appearance of the 3D building, as well as its context, can be examined without showing too much detail concerning the interior. Additional virtual contents like 2D facade textures applied to the outer walls of the building or a road map projected onto the floor screen of the L-Shape can support this stage of exploration.

2. In the **highlight mode**, one particular floor can be highlighted while all overlying floors appear transparent. Additionally, all floors move upward or downward to guarantee that the active floor is always displayed on an easily accessible level. This mode is well suited for focusing on a specific floor without losing the general view of the building.

3. In the **focus mode**, the building can be folded to the currently selected floor in order to focus the user's attention on this specific floor. As single rooms and their labels are clearly visible in this mode, it enables a more detailed exploration of the building's interior.

The last mode provides a second 2D view of the active floor on the front screen. Thus, users can take part in the discussion of a specific floor without the need of following the operator's movements.



(a) Overview mode **(b)** Highlight mode **(c)** Focus mode

Figure 12.2.: Photos of the three different visualization modes.

Furthermore, they do not have to enter the L-Shape or wear shutter glasses since the projection of the active floor plan on the front screen is displayed at zero parallax.

To guide the users' attention to what the operator is currently focusing on, we use a second input device – the tracked magnifier lens. The purpose of the lens is to allow the operator to specify a region of interest in the 2D floor plan. By moving the lens above the 3D view of the building projected onto the floor screen, a circular area is highlighted in the 2D floor plan on the front screen; all parts of the floor that are not within this region are culled out. The interaction model of the magnifier adopts the typical real-world interaction. Moving the lens in the horizontal plane results in a motion of the spot on the vertical front screen. While the 2D floor plan is magnified by an initial factor the operator can increase the zoom level by moving the lens downward.

12.2. Pilot Study

In this section, we describe the evaluation of the current design of the user interface in the iterative human-centered design process.

12.2.1. Participants

To gather information about the usability of the interface, we conducted a pilot study with 5 future inhabitants (1 female and 4 male, aged from 28 to 45, $M = 35$) of a recent building planning process of the University of Hamburg. For the project, we received the actual architectural documents and, in particular, the 2D floor plans as annotated PDF documents. The building consisted of a basement as well as 11 stories as shown in Figure 12.2.

12.2.2. Methods

We performed the study with a two-stage procedure:

1. In the first stage, we displayed the 2D floor plans using Adobe Acrobat on a 55-inch multi-touch tabletop, around which the participants were gathered.

2. The second stage consisted of the participants moving over to the L-Shape projection setup, in which the floors were displayed using the layer-based VE described in Section 12.1.

We purposely chose a simple PDF viewer to receive an impression of our prototype's features and interaction techniques rather than comparing our system to a professional CAD software with a specific toolset.

The tasks we gave the participants consisted of finding their future office rooms in the building and following the paths they had to take to reach the rooms from the main entrance of the building. We asked the participants to discuss the observations they made about the spatial properties of the building during the collaborative process using the *think-aloud* protocol [Lew82]. While the task mainly involved collaborative touch interaction in the tabletop setup, one participant assumed the position of the operator in the L-Shape environment. After 20 minutes of active discussions in each phase, we asked them to come to a conclusion. After the collaborative phases, we asked them to fill out NASA Task-Load-Index (TLX) questionnaires [Har06] and performed a debriefing with the participants, encouraging them to comment on the advantages and disadvantages of the user interfaces. The study took approximately one hour in total.

12.2.3. Results and Discussion

In the following, we present the questionnaire results and subjective comments during the pilot study.

Task Load We analyzed the questionnaire data with Wilcoxon signed ranks tests [Ott15]. The results of the NASA TLX questionnaire show a significant difference for mean task load of 67.59 ($SD = 21.29$) for the table condition and 26.97 ($SD = 21.69$) for the L-Shape condition, $Z = 2.02$, $p = .04$. In particular, we found a significant difference for mental demand between the table ($M = 68.60$, $SD = 23.01$) and L-Shape ($M = 29.00$, $SD = 27.63$) conditions, $Z = 2.02$, $p = .04$. We found no significant difference for physical demand between the table ($M = 35.20$, $SD = 28.35$) and L-Shape ($M = 24.40$, $SD = 25.16$) conditions, $Z = .944$, $p = .35$. We found a significant difference for temporal demand between the table ($M = 59.20$, $SD = 24.77$) and L-Shape ($M = 23.20$, $SD = 18.57$) conditions, $Z = 2.02$, $p = .04$. Also, we found a significant difference for performance between the table ($M = 70.80$, $SD = 30.34$) and L-Shape ($M = 24.40$, $SD = 19.26$) conditions, $Z = 2.02$, $p = .04$. Moreover, we found a significant difference for effort between the table ($M = 69.00$, $SD = 15.54$) and L-Shape ($M = 28.40$, $SD = 23.27$) conditions, $Z = 2.02$, $p = .04$. Additionally, we found a significant difference for frustration between the table ($M = 72.20$, $SD = 26.25$) and L-Shape ($M = 21.60$, $SD = 21.30$) conditions, $Z = 2.02$, $p = .04$. The results indicate that completing the task in the L-Shape was significantly less demanding than interpreting the floor plans when they were displayed on the table.

Subjective Comments We grouped the comments during the think-aloud and debriefing sessions, and identified four main topics:

1. *Misinterpretations*: Throughout the exploration process, a number of uncertainties regarding the architectural annotations occurred. In the layered setup, most of these uncertainties, for example, concerning the role of a room, could be resolved by the participants by magnifying the region of interest. However, the zooming tool in the PDF was rarely used. The participants also approved the cut-outs in the layered visualization as they helped them to identify wall penetrations. A recurrent point of discussion in the PDF visualization was the currently selected floor since it did not match the page number. Besides these differences, there were also some annotations in the 2D floor plans that could not be interpreted without help by an architect, for example, the markings that indicated the direction of stairs. In the layered view, these questions could partially be resolved due to contextual information or switching back and forth between different layers.

2. *Navigation*: The process of navigating through the building and finding the entrance took much longer in the PDF than in the layered visualization. Besides the learning effect that appears in the second phase of the experiment, this can be reasoned by the just mentioned misinterpretations regarding the current floor level. The navigation was further aggravated by the fact that all scrolling and zooming operations in the PDF caused a sequential reload of the page content. Additionally, multiple participants stated that switching and selecting floors in the L-Shape is much easier and faster than in the PDF and therefore improves the navigation through the building.

3. *Sense of space*: After every phase, we asked the participants to show the actual location of a specific room in a physical 3D model of the building. In both versions, the participants were able to point to this location correctly. However, the ceiling height of a floor could not be inferred from the PDF plans, which was criticized by one participant in the debriefing. In general, the test group stated that the L-Shape setup provided a better spatial impression of the building than the PDF.

4. *Collaboration*: Regarding the collaborative aspect of the two compared user interfaces, the opinions in our test group were divided. While some participants felt that it is easier to directly point on a specific location in the PDF, others preferred the magnifier interface for showing something to their group partners. The designation of an operator in the L-Shape setup was judged favorably as it prevents conflicting user inputs.

In conclusion, one participant remarked, that he would prefer to have an actual physical 3D model compared to both visualizations. However, in the absence of such a physical model, the layered virtual view was preferred over the PDF.

12.3. Conclusion

In this chapter, we presented a combination of 3D and 2D representations, which allows the collaborative exploration of 3D models in the context of architectural review and design processes. We proposed a stacked layout based on 2D floor plans to facilitate the discussion and evaluation of building designs in early development stages when 3D models are not available yet. The stereoscopic 3D view of the building was supplemented by a monoscopic representation to support interactions with multiple users. To attain a more natural interaction we also introduced different input devices such as a magnifier lens.

In the future, it is important to evaluate the comfort and effectiveness of the proposed interaction concepts. For that purpose, a more extensive study involving other existing CAD tools can be conducted. A further comparison with a fully-fledged 3D model would also be useful to investigate which spatial characteristics can be explored in the layer-based visualization and which cannot. An improvement of the multi-user capabilities of the system could be achieved by lifting the projection surface to a higher level, for instance by placing a purpose-built table on top of the L-Shape's floor panel. Aside from reducing the vergence-accommodation conflict, this setup allows projecting the active floor onto the table with zero parallax, which means it is displayed perspectively correct for all users. Finally, we will expand our work concerning shared Blended Spaces with an alternative collaboration technique in the next chapter.

13 Chapter 13. Floor-Projected Guidance Cues for Collaborative Exploration of Blended Spaces

In this chapter, we present a floor-based UI, which was developed to emphasize the positive effects of Blended Spaces and to reduce possible complications, which may emerge in everyday use of the technology. In detail, the interface is designed to address the following three issues:

1. **User Interaction**
   Allow users to seamlessly transition between different states of the system, without needing to use additional input devices or to learn application-specific interaction methods.

2. **User Guidance**
   Support the storytelling by guiding users to regions of interest and ideal viewpoints.

3. **User Collaboration**
   Extend the system to multiple users by introducing a master-follower concept and corresponding visualizations.

The UI is adapted for, but not limited to, domains as exhibitions or architectural meetings, in which Blended Spaces can be used to present different aspects of a physical object. We developed a set of guidance cues, which are projected onto the floor to assist multiple users in the above-mentioned tasks. In a user study with 40 participants all cues were evaluated and a set of feedback elements, which are essential to guarantee an intuitive self-explaining interaction, was identified. The results of the study also indicate that the developed UI guides users to more favorable viewpoints and therefore is able to improve the experience in a multi-user Blended Space.

13.1. Design of a Floor Based UI

The floor interface was designed to address three design goals: supporting (i) user interaction, (ii) user guidance, and (iii) user collaboration. In the following, detailed information on the design goals and derived interface elements are provided.

13.1.1. User Interaction

As a Blended Space covers different stages of the RV continuum, the question arises how to switch between these stages while preserving the simplicity of the system. In terms of the basic interaction paradigm, related projects cover a wide range of input methods for floor UIs, including gestures (e.g., [AKM+10, BHH+13, GIK+07]), specific slippers (e.g., [CR97, LJFKZ01]), or additional tools (e.g., [SRP+14]). In contrast to these examples, our system resembles the idea of proxemic interaction, based on the *position*, *identity*, *movement* and *orientation* of entities in the scene [BMG10].

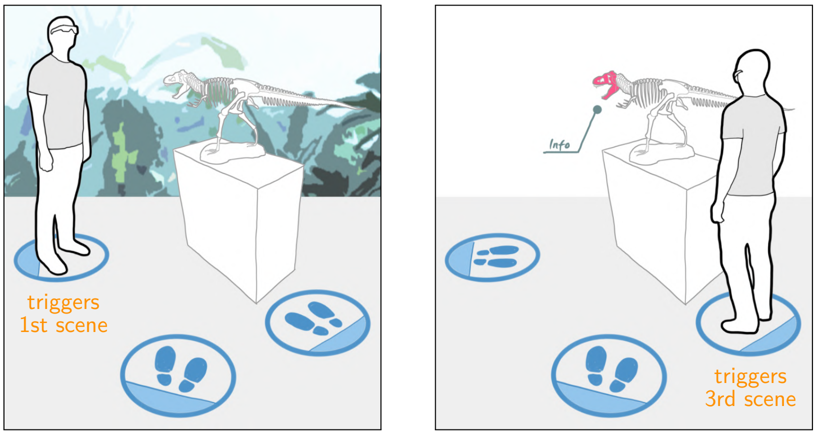


Figure 13.1.: Interface elements to support user interaction. Each button triggers a different scene with specific virtual content.

The basis for our UI is formed by a scene selection menu, which consists of several buttons that are projected onto the floor (see Fig. 13.1). Each button represents one content-related scene and can be labeled with the according topic. When a user steps into a specific button, a circular progress bar frames the button. When the loading is complete, the application transitions into the selected scene and the button transforms into a larger floor area. This area indicates the walking zone and is discussed in greater detail in the following section. While the user moves within the boundaries of the walking area, he can explore the scene autonomously. To exit the current scene and return to the scene selection menu, the user just has to step out of the walking area. To prevent an accidental leaving, for example, when the user is moving backward, the entire floor UI is vibrating as soon as the user is approaching the boundaries.

13.1.2. User Guidance

As introduced in Section 2.2, one limitation of projection-based AR systems is the existence of shadows that might interfere with the projection. Particularly for objects with complex shapes as a dinosaur skeleton, self-shadowing is usually inevitable, unless an excessive number of projectors is used. One option to reduce the visible shadows from the user's point of view is to correlate his position with the projectors' frustums. If the user's head is close to the optical center of the regarding projector, shadows are occluded by the physical object itself and therefore do not disrupt the projection.

Restricting the user's movement within the scene can also be of practical value from a narrative point of view. As in all immersive setups, Blended Spaces also face the challenge to allow users an autonomous exploration of the scene and to present pre-configured story elements at the same time. Indeed, users might miss important elements of the story because they are looking in a different direction.

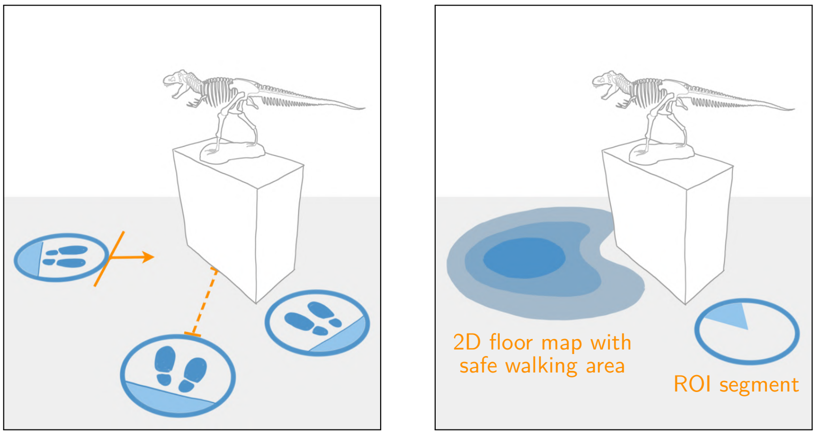


Figure 13.2.: Interface elements to support user guidance. The buttons' orientation and distance to the exhibit ensure an optimal view at scene entry (left) while a walking area and a circular segment guide the users to regions of interest within the current scene (right).

To ensure that users can make the best of an application within the Blended Space, our floor UI suggests particularly favorable viewpoints via customized UI elements. Every scene is connected to a pre-defined area that is safe to walk in, both with regard to the storyline and the technical limitations such as self-shadowing. At scene entry, a visual representation of this area gradually fades in. This leads to a color-coded 2D floor map showing the quality of different viewpoints, where only areas with a minimum quality level are included. To make sure the user is entering a scene with the ideal viewpoint, the buttons' layout can be constructed accordingly. For that purpose, the individual position of a button can factor in different aspects of the linked scene. First of all, the scene's level of virtuality usually has a strong impact on which real and virtual elements are of utmost interest to the user. If no virtual images are overlayed, the objects themselves are brought into focus. Therefore, a close distance to the object could help the user to discover details in terms of shape, material qualities, and surface texture. In contrast, if the object's context is displayed, a comprehensive view of the entire surroundings might give a better impression of the scene in its entirety. Overall, the distance of the buttons to the physical exhibit can reflect the relevance of specific scene elements and therefore allows the designer of the system to draw the users' attention to them. This proximity-based approach can be expanded by the direction the user is facing at the beginning and throughout a scene. The goal is to provide unobtrusive cues that suggest regions of interest (ROIs) to the user, rather than forcing him to look in a specific direction. This is achieved by two different interface elements, which are used according to the current state of the system. When the scene selection menu is displayed, footprints inside each button indicate the ideal direction of view for every scene. After a scene was loaded, the footsteps disappear and are replaced by a circular segment which represents the current ROIs within the scene. All UI elements for guidance are illustrated in Figure 13.2.

13.1.3. User Collaboration

At this point, a single user can take full advantage of the system, explore the different scenes and transition between them using the floor interface. It is also easy to extend the system to multiple users if the current scene only contains virtual objects that are independent of the viewpoint (as in the case of the states in Figures 3.1a, 3.1b and, depending on the scene, also 3.1d). However, for scenes that contain 3D content as in Figure 3.1c, the virtual cameras have to be coupled to the pose of an observer's head to convey a realistic sense of perspective when the observer walks through the scene. Therefore, even for multiple users, a correct perspective can only be provided for the position of one observer.

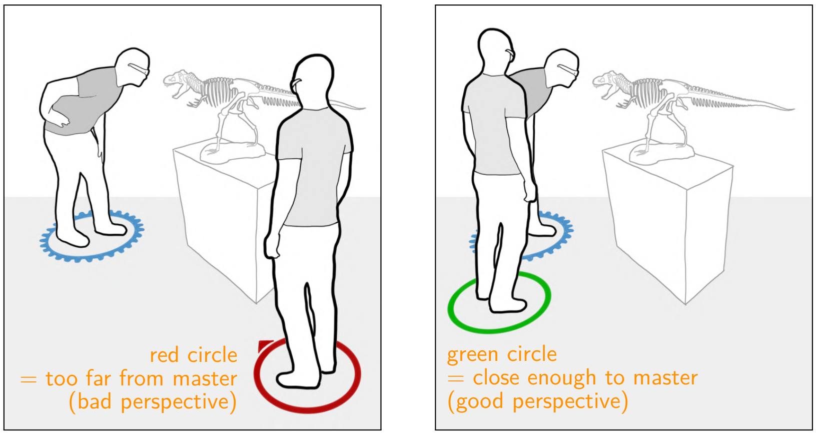
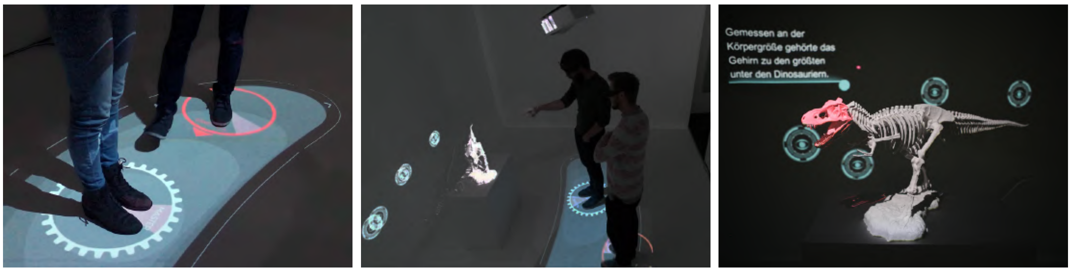


Figure 13.3.: Interface elements to support user collaboration. The optimal distance to a selected master (the user who controls the perspective) is indicated via color-coded circles with an additional arrow.

To prevent conflicting user inputs during the exploration of such a scene, one user is dedicated to the master task and therefore controls the perspective for all observers of the scene. This concept to support user collaboration was already introduced in the previous chapter and was judged favorably in the according user study that focused on subjective measures.



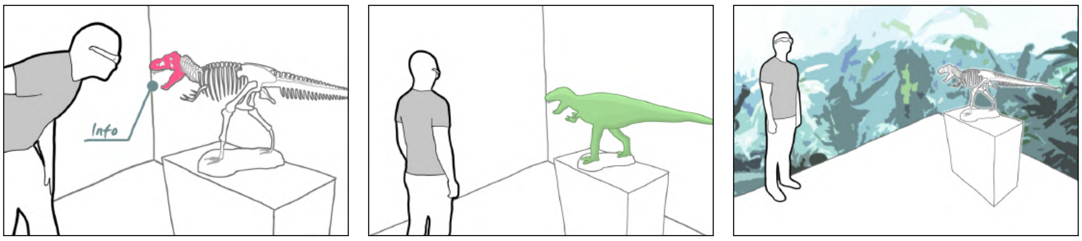
(a) (b) (c)

Figure 13.4.: Experimental setup with (a) the extended floor UI, (b) participants discussing a virtual scene, and (c) the virtual scene from the master's point of view.

Which user is chosen to be the master is decided in the moment of selecting the next state in the main menu of the system. Other users can wear shutter glasses as well, allowing them to perceive the scene stereoscopically. However, to get a less distorted perception of virtual 3D models they have to stay close to the master. To inform users about this, we introduce different UI elements as shown in Figure 13.3. First of all, the master is identified with a gearwheel around the feet along with the lettering 'master'. Every other user is surrounded by a colored circle, which represents a rating of the user's current position. In scenes without 3D content, the circle is always green. If the current scene contains 3D content and the user is too far from the master to have a good viewpoint, the circle turns reddish. In addition, an arrow appears to show which direction the user has to go in order to improve the perspective. By this means, the master can move freely within the scene while other users are encouraged to follow the master's movement. The master also decides when to leave the current scene and return to the scene selection menu. In the menu, the master task can be passed on to another user as described before. Whether this approach is feasible in a realistic scenario or if an automatic timer to leave the scene as well as a balancing strategy to assign the master task should be implemented, are two of several questions we wanted to investigate in a user study.

13.2. User Study

For the evaluation of our proposed floor interface, we simulated an exhibition scenario in a CAVE, with a physical dinosaur skeleton serving as the central exhibit. To emphasize the social aspects of the interface, participants completed the study pairwise, as illustrated in Figure 13.4. Following a between-subjects design, the interface was compared to a control condition, which was reduced to basic UI elements. The control condition involved plain floor-projected buttons that pulsated to gain the users' attention and stopped pulsation after one of the users stepped in. In contrast to the developed extended UI, the buttons' locations were not adapted to the scene content in terms of distance and direction. Also, none of the previously described guidance cues were used in the basic UI.



(a) (b) (c)

Figure 13.5.: Illustration of three different views of the same exhibit, with virtual overlays showing (a) details of the exhibit, (b) the original appearance of the exhibit, and (c) the original context of the exhibit.

13.2.1. Participants

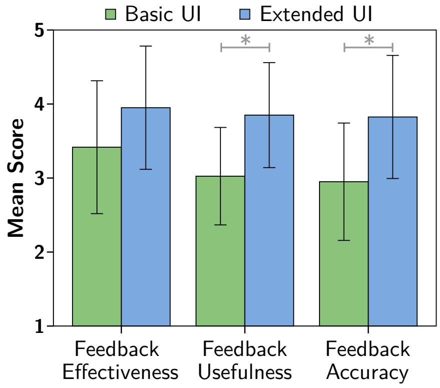
We invited 40 participants to our study, 26 male and 14 female (aged from 19 to 65, $M = 29.5$) and assigned them to 20 experiment sessions. 32 of the participants were students or staff members of the local department of informatics, while 8 participants stated to pursue a non-technical profession. To model a natural situation in a museum, half of the participants already knew their partner while the other half of the participant pairs were strangers before the beginning of the study. This differentiation will be taken into account during the analysis, to identify possible issues of the interface when two strangers have to interact with each other in a shared space. In order to qualify for participation in the experiment, each user had to confirm to be unfamiliar with the CAVE. This prerequisite was used to ensure that participants had no preknowledge over the functionality of the CAVE and its limitations regarding the multi-user capacity.

13.2.2. Materials

For conducting the user study we used the setup described in Section 5.2. Inside the CAVE, close to the front wall, a replica of a dinosaur skeleton was positioned on a white box. To augment the physical object as well as its environment, the CAVE was equipped with five 3D projectors. Both participants of an experiment session had to wear 3D shutter glasses in order to experience the stereoscopic content.

13.2.3. Methods

Prior to the study, both participants had to fill in a consent form, including a declaration of the planned video recording. After a small introductory story to stage the experience, participants were instructed to put on the shutter glasses, enter the CAVE and explore the presented exhibition as during a normal museum visit. Apart from this, no specific tasks were given and the used technology was not introduced.



UI Element	M_{basic}	SD_{basic}	$M_{extended}$	$SD_{extended}$
Footsteps	3.85	1.089	4.80	0.410
Progress bar	4.45	0.759	4.26	0.991
Buttons' layout	3.70	0.865	3.17	1.150
2D floor map	3.60	0.995	3.50	1.318
ROI segment	2.90	1.021	2.19	1.276
Follower arrow	3.80	1.240	2.82	1.590

Figure 13.6.: Mean scores (left) for the projected feedback in general, and (right) for specific UI elements.

In total, three scenes could be selected as shown in Figure 13.5. This included a presentation of the dinosaur's anatomy with 2D highlighting and 3D textual annotations, a 3D projection of the skin around the skeleton, and a stereoscopic 360-degree video that showed the habitat of dinosaurs. All scenes were accompanied by an audio commentary. At any time, participants were free to talk to each other and to move through the CAVE, however, as in usual museums the skeleton must not be touched. The behavior of the participants, as well as their conversations, were recorded using a video camera. Furthermore, additional data was stored for later analysis, including the distance between users, the distribution of the master and follower roles, and targeted objects. After 8 minutes of free exploration, participants were asked to move into two separate rooms and to fill in some post questionnaires. This included scales regarding usability and subjective communication. In total, one experiment session took around 30 minutes.

13.2.4. Results

A variety of subjective and objective measures was used to evaluate different aspects of the developed UI. In the following, we refer to participants who used the basic UI as the control group and to participants with the extended UI as the test group.

User Interaction

The usability of the presented UIs was investigated both with the System Usability Scale (SUS) [Bro96] and the AttrakDiff questionnaire [HBK03]. We analyzed the results with five unpaired t-tests at the .05 significance level. Since no significant differences between the test and the control group could be found, we pooled the results of both groups. The average SUS score adds up to $M = 73.250$ ($SD = 14.212$), which can be interpreted as a grade of a B [Sau11]. On a scale of -3 to +3, the pragmatic quality was rated with $M = 0.996$ ($SD = 0.806$), the hedonic quality (identity) with $M = 1.004$ ($SD = 0.0.907$), the hedonic quality (stimulation) with $M = 1.061$ ($SD = 0.748$) and the attractiveness with $M = 1.600$ ($SD = 0.819$).

In addition to the usability scales, we adopted a questionnaire from [MBSG09] to measure the usefulness, accuracy, and effectiveness of the floor-projected feedback on a 5-point Likert scale. The results were analyzed using three Mann-Whitney-U tests, since the assumption for normality could not be assumed. We found significant effects at the .05 significance level for the feedback usefulness ($U = 83.000$, $p = 0.001$, $r = 0.510$) and the feedback accuracy ($U = 86.000$, $p = 0.002$, $r = 0.496$). The effect of the UI type on feedback effectiveness was not significant ($U = 129.500$, $p = 0.055$, $r = 0.304$), but showed a trend towards the extended UI. The results are illustrated in Figure 13.6.

Besides this general evaluation of the provided feedback, participants who tested the extended UI were asked to rate the usefulness of specific UI elements on a 5-point Likert scale. Since participants of the control group did not experience these UI elements, we asked them to rate the desirability of such additional cues instead. The resulting scores are listed in Figure 13.6.

User Guidance

To investigate to what extent the system allowed the users to follow the story, we set up a questionnaire that asked for four different aspects of storytelling, including the clarity of the storyline, the obviousness of where to look at, the feeling of disorientation and the fear of missing important story elements. Each item was measured on a 5-point Likert scale that ranged from *Strongly disagree* to *Strongly agree*. We ran four unpaired t-tests to analyze the questionnaire's results, however, no significant effect was found.

In addition to the subjective rating of the storytelling, we used an objective measure to evaluate the view direction of the participants during story-driven scenes. For every scene, regions of interest (ROIs) were defined that changed with regard to their size and position over the course of the story. While a floor-projected circular segment pointed to these regions in the test condition, they were invisible in the control condition. For each participant, a ratio was calculated that describes to which percentage the participant's view matched the intended view direction. The difference between the values of participants testing the basic UI ($M = 0.588$, $SD = 0.084$) and the extended UI ($M = 0.563$, $SD = 0.129$) was compared with an unpaired t-test, however, no significant effect was found.

Besides the ROI segment, which was designed to indicate a good view direction, two additional floor-projected cues directed users to favorable viewpoints, the buttons themselves and scene-dependent 2D maps that emerged around the buttons. Both elements got neutral to positive reviews by users of the extended UI (see Fig. 13.6).

Table 13.1.: Categories that were used to analyze the speech data.

Speech Data	Description
Interaction-related	Discussion or interpretation of interaction methods and UI elements.
Social or Emotional	Social or emotional utterances, such as laughing or expressions of excitement.
Technical	Hardware- or software-related discussions.
View-related	Discussions related to the perspective and visual perception.
Action-related	Planning of the own or the partner's next actions.
Content-related	Discussion of elements, which are presented visually or auditory in the scenes.

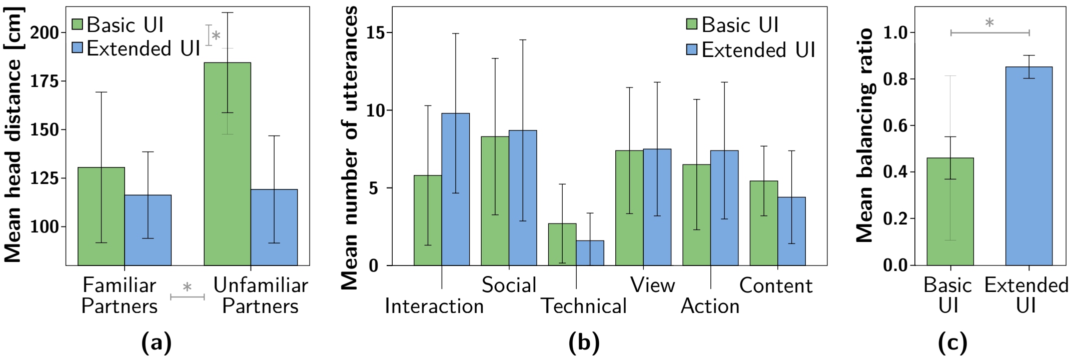


Figure 13.7.: Pooled results of (a) the distance between the partners' heads, (b) the categorized speech data, and (c) the balancing of the master and follower roles (min/max). The vertical bars show the standard deviation.

User Collaboration

To analyze the master-follower concept, which was introduced to support user collaboration, we used both subjective and objective measures.

A questionnaire to measure the group accord was used as suggested by Slater et al. [SSUS00]. For each participant, we constructed an overall score from six questionnaire responses: the degree of enjoyment, the desire to meet the study partner again, the extent of perceived isolation, the degree of comfort with the partner, the degree of embarrassment induced by the partner, and the extent of perceived cooperation. Analysis of the results of the questionnaire with a two-way ANOVA did not reveal any significant effects. Group accord scores were similar for both the basic UI ($M = 80.278$, $SD = 14.247$), and the extended UI ($M = 77.778$, $SD = 12.998$). There was also no significant difference between unfamiliar partners ($M = 81.25$, $SD = 12.254$) and familiar partners ($M = 76.806$, $SD = 14.651$), although unfamiliar partners even reached a slightly higher score.

To gain further insight into the group behavior of users, we measured the distance between the users' heads during the scenes and calculated a mean value for each pair of participants. Requirements for normally distributed data were fulfilled and the assumption of equal variances was not rejected by Levene's test, so we ran a two-way ANOVA. We found a main effect of the type of UI on the mean head distance ($F(1, 16) = 9.224$, $p = 0.008$, $\eta_p^2 = 0.366$), indicating a significant difference between users of the basic UI ($M = 1.575$, $SD = 0.421$) and the extended UI ($M = 1.177$, $SD = 0.237$). The familiarity also showed a main effect ($F(1, 16) = 4.718$, $p = 0.045$, $\eta_p^2 = 0.228$), indicating a significant difference between partners, who knew each other ($M = 1.234$, $SD = 0.307$) and strangers ($M = 1.518$, $SD = 0.426$). No significant interaction effect was found between type of UI and familiarity ($F(1, 16) = 3.801$, $p = 0.069$, $\eta_p^2 = 0.192$).

Besides logging the participants' positions, we also captured their behavior on video. This allowed us to analyze the communication between partners. Based on an approach used by Smith and Neff [SN18], we considered utterances of the participants during the study. In comparison to the stated paper, we slightly extended the definition of an utterance to an individual word, sentence, or even a small unit of a conversation between the participants, as long as their statements directly correlate. All utterances were assigned to the categories that are defined in Table 13.1. 10 of overall 763 utterances had to be discarded because they were too low-voiced or slurred. The remaining utterances were analyzed using a two-way ANOVA with the type of UI and the familiarity of partners as independent variables. Although no significant effect of the overall communication could be found, the distribution of utterance types differed between the two UIs, as illustrated in Figure 13.7b.

To address the questions that are stated in Section 13.1.3, we asked participants how they subjectively perceived the balancing of the master and follower roles as well as the master-driven leaving of a scene. Concerning the latter issue, 32 of the participants opted for the current solution, while only 7 participants would have preferred an automatic timer to leave the scene. One participant suggested introducing a consensus mechanism. The opinions regarding the role assignment were divided. 23 of the participants decided for the current mechanism on a first-come-first-serve basis while 17 participants preferred an automatic assignment of the master role that is balanced between the users. However, only 8 of the 23 proponents of the first-come-first-serve technique were members of the control group.

In addition to this subjective evaluation, we also analyzed the actual balancing of roles between the two partners of a study session. For this purpose, we determined how often each partner held the master role. Afterward, the ratio of both values was calculated by dividing the minimum by the maximum. Therefore, a balancing ratio of 1.00 corresponds to a perfectly balanced role assignment, while the balancing gets worse with lower ratios. The results, as shown in Figure 13.7c, were analyzed with a Mann-Whitney-U test, since a normal distribution of the data cannot be assumed. We found a significant effect at the .05 significance level ($U = 14.000, p = 0.005, r = 0.620$).

As for user interaction and guidance, we also let the participants rate the UI elements, which were designed to support user collaboration. Participants of the test group rated the colored follower circle that pointed to the master with a mean score of $M = 3.80$, which is the third-best value of the six tested UI elements.

13.2.5. Discussion

For the interpretation of the results, we will again consider the three groups of guidance cues separately.

User Interaction

Both tested UIs achieved over-average usability scores, which is a point in favor of floor-based interaction techniques in Blended Spaces. However, in the concluding questionnaire, six participants of the control group reported their confusion about who was able to control the scene elements, which is in line with the oral feedback after the study sessions. A reason why this difference between the UIs is not reflected in the usability scores might lay in the between-subjects design, since participants did not have any reference value. Regarding the feedback quality of the two tested UIs, we found an increased usefulness and accuracy of the extended UI. Particularly, both cues that were designed to support the interaction with the system, namely the footsteps and the progress bar, achieved the best ratings of all UI elements. Pre-tests revealed that footsteps are inevitable to be able to interact with the system on one's own, especially for users who are not experienced in video games. We therefore decided to tell participants of the control group that they can interact with the scene via the pulsing buttons on the floor. Nevertheless, participants often left the buttons before the loading was finished, although the pulsing animation stopped when users entered a button. A progress bar is an easy way to provide more detailed information on the current state of the system and to reduce unintended abortions of the loading process. From a technical point of view, the tracking of the users' head poses could be replaced by a more fine-grained capturing of floor-based touch input. Other floor UIs already demonstrated the recording of floor interaction through contact sensing (e.g., [CHK$^+$10, VSL$^+$10, VSC13]), and by under-floor camera tracking (e.g., [GIK$^+$07, AKM$^+$10, BHH$^+$13, SRP$^+$14]). Both methods add complexity to the overall setup but, on the other hand, allow for a precise interaction with UI elements.

User Guidance

While both the buttons' layout and the 2D floor maps were well received, the ROI segment attained the lowest score of all feedback cues. Moreover, 4 of 20 participants of the test group did not notice the element at all. This was also confirmed by users of the basic UI. With an average score of $M = 2.90$, the support of storytelling through a floor cue was the least wished feature in a future UI. One reason for the low ratings may be the manageable size of the 4-sided CAVE, since all projection surfaces could be observed at once, either directly or peripheral. Besides, the story was rather simple with most story elements being presented in the center of the CAVE. Regions of interest inherently attracted the attention due to the movement within the scene. This assumption is also supported by the storytelling questionnaire, which resulted in scores around 4 out of 5 for both UIs. Moreover, participants noted that they had to decide whether to focus on the floor-projected segment or the story since it was difficult to bring both parts into view simultaneously. Therefore, future UIs might prefer storytelling cues that are directly integrated into the scene instead of floor-projected cues.

User Collaboration

During the study, we observed a highly collaborative behavior of the participants, both for familiar and unfamiliar partners as well as for both UI conditions. This impression matches the results of group accord scores and the measured amount of communication. However, the results also reveal differences in some measures, including head distances, communication subjects, and role balancing.

The mean head distance for participants using the extended UI was significantly smaller than in the basic UI, which indicates a positive effect of the floor-projected cues. Since users with the follower role were standing closer to the master, it can be assumed that they had more favorable viewpoints. Although no significant interaction effect between the UI type and the familiarity of partners on head distance was found, Figure 13.7a shows an interesting trend. While the head distance is almost the same for familiar and unfamiliar partners in the extended UI, unfamiliar partners kept a bigger distance from each other in the basic UI. This could indicate a positive effect of the extended UI elements on the collaboration between users who meet each other for the first time.

Categorizing the speech data during the study revealed that users of the extended UI talked more often about the interaction with the system. This could be interpreted in two ways. On the one hand, users were more focused on the UI, since more elements that gave room for interpretation were present. Therefore, the participants might be more distracted from the actual content that was presented, and therefore, the learning outcome could be reduced. However, responses of the users of the extended UI also suggest a lower level of frustration while interacting with the system, which, on the other hand, could improve the learning experience. Future investigations should focus on the effects of different UIs on learning before such systems can be used in an educational context.

Another significant difference was found regarding the user-driven assignment of the master and follower roles. While the distribution of roles was almost balanced between both partners in the extended UI, a high disparity could be observed in the basic UI condition. However, video inspection indicates that this is not caused by an unfair behavior of users, but by difficulties in understanding the interaction mechanism. Several groups who tested the basic UI had bad guesses on how to enter a scene, including the simultaneous standing of both partners on one or even two different buttons, the malfunction of one of the trackers, and the idea that only one user is qualified to be the master. Consequently, the partner who seemed to be tracked more reliable in the early stage of the study was chosen to be the master in a group consensus. This confusion could also be the primary reason why only 8 members of the control group voted for a user-driven balancing approach and the majority preferred an automatic, fair assignment of the roles instead.

13.3. Conclusion

In this chapter, we presented a multi-user floor-based interface for Blended Spaces. We developed multiple floor-projected cues that aim to support users in several aspects when interacting with such systems. To evaluate the effectiveness of the cues, we performed a user study with 20 pairs of participants. The results indicate that the interface is self-explanatory and easy to use, and therefore could be used in public environments such as museums without the need for support of an additional instructor. It also fostered communication between partners, both between friends and strangers, and encouraged users to move closer together. By this means, better viewpoints could be ensured for multiple users. From a narrative perspective we could not observe significant improvements, however, high scores were achieved even in an application without additional feedback cues. When faced with a more complex story, users might need more support to keep track of where to look at within the scene. Furthermore, future studies should focus on the learning success achieved by using the system. After the present study, it is still an open question whether the floor interface distracts users from the presented content or if it sparks the users' interest in a new topic. Further investigations could pave the way for floor-projected UIs to be used in real scenarios.

14 Chapter 14. Virtual Agents in VR/AR

Inspired by science fiction media, such as the movies *Her* (2013) and *Blade Runner* (2017) – stories that show the potential of virtual agents (VAs) integrated into our daily social life – we have seen a large public interest in related technologies. Different forms of VAs were proposed and evaluated throughout Milgram's RV continuum, as surveyed, for example, by Holz et al. [HCO+11, HDO09] and Norouzi et al. [NKH+18]. These projects show the potential of VR/AR agents but also challenges related to creating a high sense of social interaction and connection between users and VAs. For instance, Obaid et al. [ONP11, ODK+12] showed that the physiological arousal of users in VR/AR depends on an agent's behavior associated with cultural differences, for example, related to gaze behavior and interpersonal distances. Furthermore, studies of Kim et al. [KMB+17, KBW17] indicate that visual conflicts in AR such as occlusion and dual occupancy between VAs and physical objects can significantly impair their social connection with users. However, despite the challenges related to realistic and/or effective social interaction, a large number of applications could benefit from VAs [NKH+18, KBB+18]. In the next sections, we will present some of the most promising application fields for VAs before discussing by which factors the human likeness of VAs is influenced, and how it can be measured in human-subject studies.

14.1. Applications of Virtual Agents

Over the last years, voice-controlled agents were embedded in consumer devices such as Amazon's Echo or Apple's HomePod and connected to home appliances to provide an intuitive and natural form of interaction with their smart home environments and as a means to access information from the internet [KBB+18]. Beyond home uses, smart services provided by VAs are popular as they can be accessed through ubiquitous smartphone technologies and can be implemented for professional applications such as in the form of educational audio guides in museums or audio-visual presentations for mixed media installations or exhibits. In particular, in situations where the demands for individual support or care exceed the supply of specialized trained personnel, such as museum guides, caregivers, or private assistants, these VAs are a promising solution that can complement human professionals [NKH+18]. With the current convergence of different research fields such as Machine Learning, Internet of Things, and VR, it seems reasonable to assume that people will be confronted with an increasing amount of such services, which poses new challenges to the interface designers, particularly in terms of social interaction and integration. For further information, we refer to Magnenat-Thalmann et al. [MTPC08], who provide a literature review of promising application fields for VAs including interactive virtual guides in cultural heritage sites, museums, art installations, and related fields.

14.2. Indicators of Human Likeness

The realism of a VA was initially considered as a synonym for its human-like visual appearance. However, advanced models from research fields such as speech synthesis, motion capture, and non-verbal communication led to an extended definition that involves VAs with natural language, realistic behavior, and attention towards their environment as well as their human communication partners, resulting in the notion of intelligent virtual agents (IVAs). This conceptual change raises interesting research questions regarding the correlation between the manifold dimensions of the human likeness of VAs and social factors such as perceived copresence in a shared MR environment. In the following, we summarize a selection of previous studies that investigated this correlation, grouped by the dimension of human likeness they considered.

14.2.1. Agent Embodiment

Though personal digital assistants have become widespread in the context of smart homes as well as professional environments, most of the current implementations rely on audio output or displayed text only. By means of VR/AR technology, such voice-based VAs can be supplemented with a humanoid virtual body. Several research projects addressed the question of whether and how agent embodiment affects the social interaction between VAs and real humans.

A literature meta-review by Yee et al. [YBR07] suggests that the inclusion of any visual representation of a VA's face leads to higher task performance measures. The presence of a face also seems to be much more important than its visual quality. Therefore, even a representation with low realism can provide important social cues for human-agent interactions.

Positive effects of agent embodiment on the users' sense of trust, social richness, and social presence with the VA could be found in a human-subject study by Kim et al. [KBH^{+}18]. In addition, participants of the study reported an increased confidence in the agent's ability to influence the real world and to react to real-world events, when the VA was embodied and showed natural social behaviors.

A literature review on early embodied VAs was presented by Dehn and van Mulken [Dv00]. Their meta-analysis revealed some inconsistent findings regarding the effects of embodied

VAs on user experience. While some of the analyzed empirical studies report benefits of embodied VAs, others conclude that agent embodiment only showed little or even negative effects on the users' responses. Dehn and van Mulken hypothesize that these different outcomes may be attributed to varying degrees of the agent's appearance, both in terms of visual fidelity and natural voice.

Our own research regarding the embodiment of VAs in the context of Blended Spaces is presented in Chapter 15.

14.2.2. Agent Appearance

By means of continuous advancements in technology, real-time AR applications can make use of increasingly realistic VAs, both with regard to the visual appearance and quality of synthesized speech. While a positive correlation of a VAs fidelity and the elicited sense of anthropomorphism seems to be reasonable at first glance, several studies demonstrate that the agent's appearance cannot be considered in isolation, since it strongly correlates with other indicators of agent realism.

For instance, Bailenson et al. [BSH+05] investigated the effects of visual and behavioral realism of VAs on perceived copresence. They conclude that both types of realism should be considered in conjunction as large disparities between them resulted in lower levels of copresence. These results are consistent with a previous study conducted by Garau et al. [GSV+03], which also revealed a significant interaction effect between agent appearance and behavior. A mismatch between visual fidelity and behavioral realism might also explain an observation made by Nowak and Biocca [NB03]. To their surprise, VAs with a higher level of anthropomorphism caused a decrease in reported copresence and social presence. The authors argue that anthropomorphic VAs may raise expectations about their behavioral realism and should only be used if the system is able to meet these expectations.

14.2.3. Behavioral Realism

As already discussed in the previous section, realistic behavior, including natural gestures, body, and eye movements, as well as lip syncing, is a crucial factor for VAs to be perceived and treated as if they were human.

A study of Gratch et al. [GWG+07] revealed that even simple non-verbal reactions to the user such as gaze shifts or head nods can cause feelings of rapport. Moreover, Demeure et al. [DNP11] showed that appropriate emotional verbal and non-verbal behaviors of VAs can evoke a higher sense of perceived believability, competence, and warmth. Further studies with a focus on objective measures found that users maintain a greater distance from VAs who engage them in mutual eye contact [BBBL03], and that culturally inconsistent gaze behavior of VAs results in higher heart rates [ODK+12].

14.2.4. Environmental Awareness

A special type of behavioral realism relates to the degree to which VAs are aware of their physical environment (for a review see [NBB$^+$19] or [HCO$^+$11]). Different approaches are possible to endow a VA with knowledge about the physical world, for example, extracting information from a pre-populated database, or analyzing dynamic sensor data.

The embodied agent *MACK* [CSB$^+$02], for example, is using a static knowledge base including the VA's fixed location and orientation, as well as the layout of the physical building in which it is located. Based on this information, *MACK* is able to provide context-sensitive and spatially referenced information, such as directions to a specific room in the MIT Media Lab. Barakonyi et al. [BPS04] developed a framework for autonomous AR agents, which can monitor and thus react to changes of real-world attributes. The presented *AR Puppet* cannot only avoid collisions with physical obstacles but is also able to support a user in a physical construction task by tracking the user's progress. In a recent study, Kim et al. [KBW18] demonstrated that by responding to subtle environmental events such as the airflow of a real fan, VAs can create a higher sense of copresence. Despite the positive results of their study, the authors also state that the VA's awareness behavior was less effective than techniques that involve the active participation of the user, such as the *wobbly table experience* [LKD$^+$16]. In the latter project, Lee et al. introduced the concept of *virtual-physical interactivity*, which will be considered in detail in the following section.

14.2.5. Virtual-Physical Interactivity

In AR environments, the creation of a human-like VA turns out to be even more challenging than in VR environments. Even with a maximum degree of visual fidelity, natural voice, realistic behavior, and environmental awareness, the created illusion can suddenly break if the agent is not following the same laws as its physical surroundings. Potential physicality conflicts include unnatural occlusions between the VA and physical objects as well as implausible physical-virtual collisions.

Negative implications of such conflicts on human-agent interactions were demonstrated by Kim et al. [KMB$^+$17, KBW17]. In their studies, the participants observed a VA encountering a physical obstacle such as a door or a chair. In different conditions, the VA either avoided collisions with the obstacle, asked the participant to move it out of the way, or passed through the obstacle. Subjective responses indicate that physical-virtual conflicts reduced the sense of copresence while proactive behavior asking help from the users to avoid implausible conflicts increased the ratings of copresence.

Instead of avoiding collisions between VAs and physical objects, another approach is to allow VAs to actually interact with these objects, for example, to move them to a different location. In Chapter 16, we introduce the concept of *blended agents* – VAs that are not only capable of influencing their virtual surroundings but also of performing *virtual-physical interactions.*

The latter concept was first investigated by Lee et al. [LNB$^+$18]. They implemented a tabletop game that can be played by a real human and a VA. Via an actuator system underneath the surface of the table, the VA is able to move not only virtual tokens but also a physical token. In a within-subjects study, the authors were able to show benefits of the VA moving a physical token both with regard to subjective and some behavioral measures. In this condition, participants reported a higher sense of copresence and physicality, as well as higher expectations regarding the virtual human's abilities.

In a recent paper, Lee et al. [LBW17] demonstrated that even subtle tactile footstep vibrations induced via the floor can increase subjective estimates of copresence in an AR environment.

Another example of VAs that are capable of influencing their physical environment was implemented by Lee et al. [LKD$^+$16]. Their custom-made *Wobbly Table* crosses the boundary between the physical and virtual world. While the physical half is standing in front of a projection screen, a virtual counterpart is visually extended into the VE with the VA. If the VA is leaning on the table, a virtual-physical interaction occurs as both the virtual and the physical parts are slightly tilted.

14.3. Measures of Realism and Effectiveness of Social Interaction

Various metrics can be applied to evaluate social interaction with VAs. Some of them consider the subjective qualities of the agent, while others aim to measure their effects on the user's behavior. In the following, we will discuss both subjective and objective metrics that were used in the scope of the subsequently described studies.

14.3.1. Social and Co-Presence

A generalizable metric for the effectiveness of VAs in VR/AR is their ability to convey an illusion of being perceived as a real social entity sharing the same space with a real person, called *social presence* and *co-presence*. Co-presence denotes the sensation of "being together", while social presence is the sense of "being socially connected" [HB04]. Blascovich et al. define social presence as "the degree to which one believes that he or she is in the presence of, and dynamically interacting with, other veritable human beings" [Bla02, BLB$^+$02].

Various studies on VAs in VR/AR environments aimed at identifying effects of agent characteristics on the sense of social and co-presence during an interaction, using measures such as questionnaires, physiological responses, and behavioral differences [BAB$^+$04, RDI03]. For example, Lee et al. [LBHW18] found that the proxemics during interaction with VAs in AR differs significantly from those between real humans, with users giving agents more space than they would give a real person. Therefore, the trajectory of users can give some indication of how the VA is perceived. Bailenson et al. [BBBL03] investigated the effects of behavioral realism on both the distance maintained between users and VAs, and self-reported social presence. In this context, they designed a questionnaire with five items, each with a Likert-type scale from -3 to $+3$. For an overall social presence score, responses to the five questions were added, meaning a positive score indicates the perception of a conscious and aware VA, and vice versa.

Another tool to assess social presence exists in the form of the *Temple Presence Inventory* (TPI) [LDW09]. It covers multiple aspects of social presence, including *presence as social actor*, and *active social presence*. The former, sometimes also referred to as parasocial interaction, addresses whether the border between the actual physical environment and the mediated environment is crossed in order to interact with the VA in real time [LDC+00]. The corresponding seven items directly evaluate the interaction with the VA, for example with regard to the establishment of eye contact or the VA's awareness of the user. In contrast, the three items in the category of active social presence are related to the extent of the user's emotional responses to the VA's actions, for example, in the form of laughing, smiling, or even speaking to the VA. Responses to all items are measured on 7-point Likert scales and are averaged to calculate an overall score per dimension.

14.3.2. Spatial Presence

Social presence is just one aspect of the more general concept of *presence*, which is mostly used to describe the sensation of being in a VE in spite of the knowledge that it is not real. To avoid ambiguities, Slater termed this feeling of "being there" *place illusion* [Sla09], while other literature uses *spatial presence* to describe the same concept [HWV+15].

As for social presence, we use a subscale of the TPI as a subjective measure for spatial presence [LDW09]. The resulting score relates to the realism of a VA as it incorporates the extent to which a user perceives the agent to be a physical entity. This includes the self-reported avoidance behavior when the VA is approaching, as well as the impression that the VA can be touched by reaching out. The scale consists of seven items, which show the same structure as for the dimensions of social presence, and therefore can be aggregated similarly.

14.3.3. Ecological Validity

Strongly related to the *place illusion* is the notion of *plausibility illusion*, which was also introduced by Slater [Sla09]. Plausibility illusion indicates that "the scenario being depicted is actually occurring" with a "credible scenario and plausible interactions between the participant and objects and virtual characters in the environment."

In the scope of this thesis, the concept is particularly important to assess the interaction between VAs and their physical environment; a research topic that we will focus on in Chapter 16. To evaluate whether a VA's actions are perceived as believable and natural, we used the *ecological validity* subscale of the *ITC Sense of Presence Inventory* (ITC-SOPI) [LFKD01]. An overall score can be calculated by averaging the responses to five statements with five levels each (1 = strongly disagree; 5 = strongly agree).

14.3.4. Agent Anthropomorphism

As discussed in Section 14.2, the anthropomorphism or human likeness of a VA has a determining influence on the perceived quality of social interactions between a real user and an agent. Anthropomorphism may involve the attribution of various human traits to a non-human entity, for example, with regard to its appearance, behavior, or even emotional responses. Therefore, a variety of questionnaires can be used to assess different aspects of an agent's anthropomorphism.

For our studies, we adopted a metric from the field of human-robot interaction, called the Godspeed questionnaire [BKCZ09]. It contains five semantic differential scales: "fake to natural", "machinelike to humanlike", "unconscious to conscious", "artificial to lifelike", and "moving rigidly to moving elegantly". The questions are relevant to VAs as well, as the development of robots and VAs presents similar challenges, for example, in terms of rigid movements, an artificial appearance, or unnatural speech. Each rating is indicated on a 5-point Likert scale, and therefore, the overall mean score is a value in the range of 1 to 5 (1 = not anthropomorphic; 5 = highly anthropomorphic).

14.3.5. Agent Credibility

In both introduced application fields for Blended Spaces (see Chap. 3), VAs may be deployed to impart knowledge or to inform the user, be it in form of a museum guide or consultant. To create added value by entrusting a VA with such a task, all of these roles require the user to accord a certain amount of credibility to the agent.

To evaluate the credibility of VAs, we use a scale developed by McGloin et al. [MNW14], that was initially introduced in the scope of assessing online peer reviews. For each participant, an overall score can be constructed from five questionnaire responses to the following bipolar adjective items: "unintelligent to intelligent", "uninformed to informed", "unreliable to reliable", "incompetent to competent", and "untrustworthy to trustworthy". Each item is measured on a 7-point Likert scale, and therefore, the overall mean score is a value between 1 and 7, with 7 representing the highest possible agent credibility.

14.3.6. User Experience

Besides the previous metrics which are specifically designed for the assessment of VAs, several existing questionnaires address the general user experience while working with an interactive application.

The *User Experience Questionnaire* (UEQ) [LHS08] asks participants of a study to provide ratings on 26 items using a 7-point Likert scale. The UEQ is in the form of a semantic differential and allows for an evaluation of both the hedonic and the pragmatic quality of a system. For this purpose, all items are assigned to six dimensions of user experience, including attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty. A transformation of the collected data is required to bring all items to a common polarity (negative term left; positive term right) and shift the values to a range of -3 to +3 (-3

= most negative; +3 = most positive). Finally, a mean score can be computed for each dimension by averaging the corresponding item values. Other questionnaires introduce the factor of *engagement* to refer to the user's emotional involvement and interest in the VE. In our user studies, we used the engagement subscales of both the ITC-SOPI [LFKD01], and the TPI [LDW09]. From the ITC-SOPI, we selected a subset of the three top-loading items, each measured on a 5-point Likert scale (1 = strongly disagree; 5 = strongly agree). The engagement subscale of the TPI involves six questions, that are answered on a 7-point Likert scale (1 = not at all; 7 = very much). In both cases, a mean score is created by averaging the ratings for all items.

In the subsequently described user studies, we utilize a combination of the presented objective and subjective measures to assess the quality of user-agent interactions.

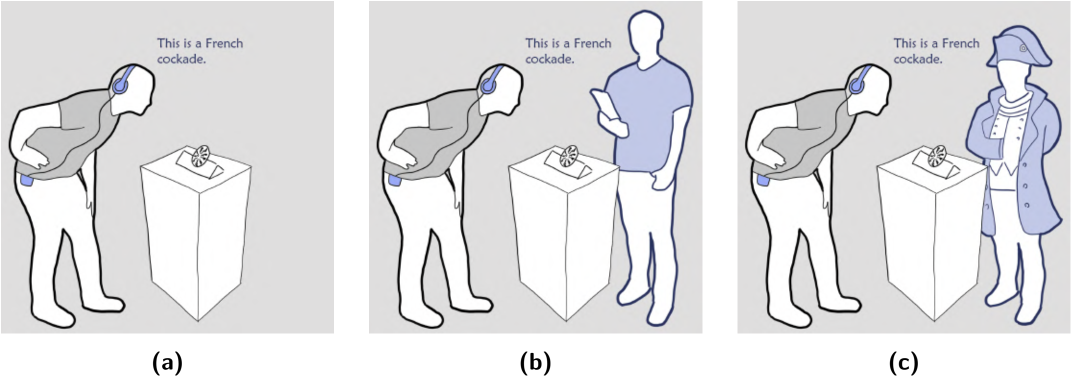
15

Chapter 15.

Effects of Virtual Agent and Object Representation on Experiencing Blended Spaces

With the emergence of speech-controlled IVAs in consumer devices, we have seen a large public interest in related technologies. While most of the currently used services are limited to audio or flat 2D visual representations, VR and AR technology can add a new dimension by providing a 3D virtual body to complement the voice. Human-like VR/AR representations can enrich the communicative channels that convey the agent's status and intentions to interlocutors with gestures and other forms of social behaviors. Moreover, they can be registered spatially with their environment, which enables a more direct form of spatial interaction compared to voice-only interaction. This is particularly interesting in situations that have a strong spatial component such as art installations and museum exhibitions since spatial relations are usually harder to communicate via speech than with gestures [Ali05]. Therefore, it may be beneficial to provide an IVA with a virtual body, which could also increase the user's feeling of co-presence, i. e., raising the visitor's sense of being together with the content on display. For museum exhibits, this could be strengthened, for instance, by choosing a historical person as the agent's representation, as exemplified in Figure 15.1. Through the encounter with a contemporary witness, visitors get to know the subject matter from a personal perspective, which may increase interest in the historical events as well as empathy with the people involved.

In this chapter, we present two human-subject studies that were performed in a historical exhibition context to understand the importance of different representations of IVAs. In the first study, we analyze the effectiveness of virtual museum guides with varying embodiment (embodied vs. disembodied) and thematic closeness (astronaut vs. museum guide) in the scope of a simulated exhibition related to the Apollo 11 mission. In particular, we are interested in the effects on the elicited sense of social presence, knowledge transfer, and the ability to communicate a sense of social competence and trust. In a follow-up study, we extend


This is a French cockade.
This is a French cockade.
This is a French cockade.

(a) (b) (c)

Figure 15.1.: Example museum application with (a) a traditional audio guide, (b) a generic embodied virtual guide, and (c) a content-related embodied virtual guide.

this work by further analyzing the effects of the representation (i.e., virtual vs. physical) of the exhibit in focus. By including this additional factor, we aim to increase the ecological validity of the results, since most traditional museums place real exhibits on display rather than relying on purely virtual visualizations.

Throughout the chapter we evaluate the following three research questions:

1. Do embodied virtual guides perform significantly better than voice-only guides in terms of co-presence, social presence, credibility, and the ability to impart knowledge?

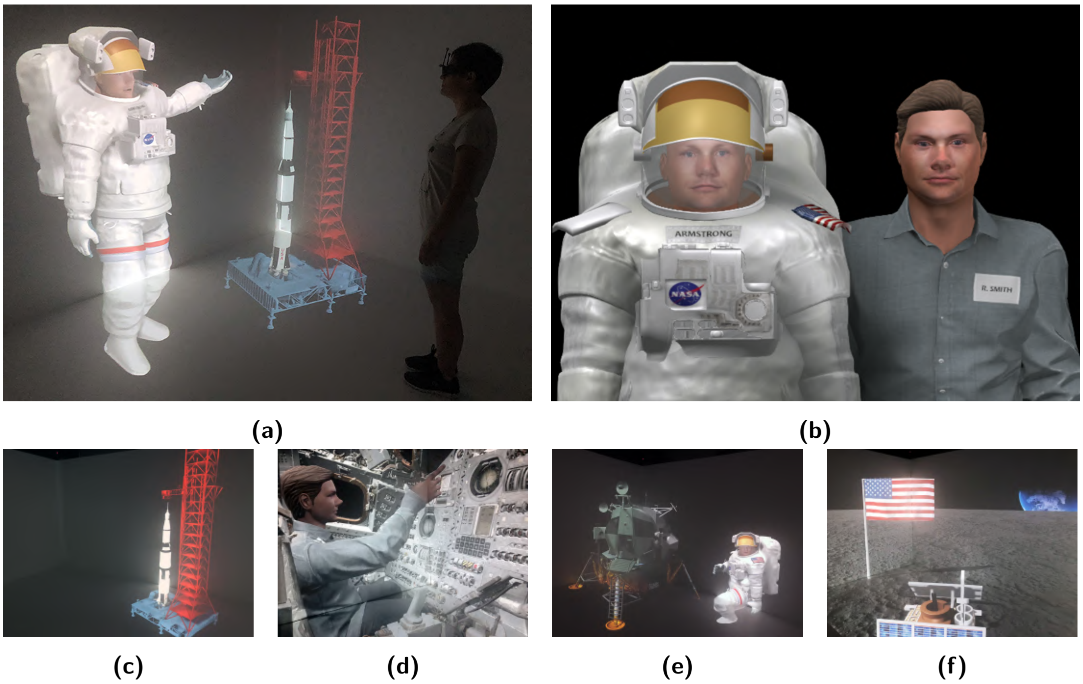
2. Do thematically close content-related guides perform better than generic guides in terms of the above-mentioned metrics?

3. Is the performance of virtual guides affected by the physicality of surrounding objects?

The results of the user studies indicate benefits of embodied as well as thematically close audio-visual representations of virtual guides, both in the presence of virtual and physical exhibits. Higher scores in terms of user engagement and knowledge transfer also suggest advantages of including a virtual component in educational applications, either in the form of an embodied agent or as a virtual exhibit. We discuss implications and suggestions for user interface and content developers to design believable IVAs in the context of both virtual and physical installations.

15.1. User Study with Virtual Exhibits

In this section, we describe a user study that we conducted to investigate the effectiveness of virtual museum guides with varying embodiment relative to the thematic context. In our simulated case study, we explore a virtual exhibition that addresses four episodes of the first manned moon landing. Each episode was presented by a different virtual guide in randomized order: (i) a generic virtual character or (ii) a thematically close content-related astronaut, each presented either as (iii) a disembodied voice (as known from voice-controlled agents such as Amazon's Echo) or (iv) a stereoscopic 3D embodied representation.



(a) (b)

(c) (d) (e) (f)

Figure 15.2.: (a) Photo showing the experimental setup, (b) the two guides in their embodied version, and (c) - (f) photos of the four episodes with exemplary guides.

15.1.1. Participants

In total, 24 participants (17 male and 7 female, aged from 19 to 39, $M = 25.1$) participated in our experiment. All of them were students or staff members of the local department of engineering and informatics. None of the participants reported any visual or motor impairments that could affect the results of our experiment.

15.1.2. Materials

The study was conducted in a four-sided CAVE-like environment with four projectors as described in Section 5.2. Participants wore tracked shutter glasses to experience the stereoscopic content that was displayed with a correct perspective. The voice of the virtual guides was presented to participants via noise-cancelling headphones of type *Bose QuietComfort 25*, with a compatible Bluetooth receiver to make them wireless. Hence, participants were not restricted in their movement and were able to walk around virtual objects in the CAVE freely. Figure 15.2a shows the experimental setup.

For our case study, we presented four episodes of the Apollo 11 mission, for which we used different models of a scaled-down Saturn V rocket with a launch pad, the interior of the Columbia command module, a scale model of the lunar module, and the moon surface with the American flag as well as scientific experiments (see Fig. 15.2c to 15.2f). All of the shown virtual models are of historical relevance, and both originals and physical replicas are currently on display at museums across the U.S., including the National Air and Space

Museum and the Kennedy Space Center. If available, original footage such as a 3D scan of the command module and NASA photographs of the lunar surface was used to build detailed models. We created four versions of the IVA used in the experiment (see Fig. 15.2b):

1. The **embodied thematically close** character was modeled as an astronaut with a space suit. The astronaut's face was generated using original footage of Neil Armstrong.

2. The **embodied more generic** virtual character was designed to match a museum guide wearing a shirt and dress pants. To prevent any preference towards one of the guides due to sympathy, we used similar basic facial characteristics for both embodied guides. However, variations of the textures, facial hair, and general hairstyle were made to ensure that the civilian and the astronaut were not perceived as the same person.

3. The **disembodied voice of the thematically close** astronaut character was identical to that condition except for the visual feedback of the agent.

4. The **disembodied voice of the more generic** character matched the embodied condition except for the visual feedback.

To increase the level of realism, we added idle behaviors to the embodied virtual guides. They also made eye contact with the user as a real guide would do in a one-on-one conversation. Furthermore, we anticipated that, in order to obtain meaningful results, the types of IVAs have to be perceived as distinct characters and not as the same character dressed differently. Hence, we performed a small survey with 20 respondents before the study to fine-tune and validate this aspect of the study.

In the embodied conditions, the agent's lip movements were matched with the spoken text via the Oculus Lip Sync plug-in. The audio track of the guides was created with the Oddcast Vocalware text-to-speech engine. For the astronaut, additional post-processing in Audacity was applied to simulate the sound of radio transmissions at that time.

The four episodes of the Apollo 11 mission provided educational information to the participants, narrated by the virtual guides. The assignment of a virtual guide to the four episodes was randomized. The educational content differed between the four episodes, but it was the same for all guides, except for the narrative point of view: The thematically close astronaut told the story from a first-person perspective and called "his" companions by their given names, while the more generic museum guide told the story from a third-person perspective. The four episodes included the following content (see Fig. 15.2c to 15.2f):

- **Episode 1:** The preparation of the Apollo 11 mission and its launch at the Kennedy Space Center as well as technical details on the Saturn V rocket.

- **Episode 2:** The three-day journey of the crew to the lunar orbit inside the Command Module with a focus on the roles of Armstrong and Collins.

- **Episode 3:** The descent of Aldrin and Armstrong to the lunar surface using the lunar module as well as the first steps of a man on the moon.

- **Episode 4:** The duties of the astronauts at the landing site, including the flag planting and scientific experiments.

15.1.3. Methods

For the first study, we used a within-subjects design based on two factors with two levels each: *agent embodiment* (embodied vs. disembodied) and *thematic closeness* (astronaut vs. museum guide). Each participant experienced all four episodes and all four agents described above in randomized order.

Prior to the study, each participant completed a consent form and a demographics questionnaire. Afterward, participants were guided into the CAVE-like environment by following a virtual 3D floating globe. Participants were introduced to the display technology, had time to familiarize themselves with the system and the stereoscopic display, and were informed about the context of the study and the Apollo 11 mission scenario.

After this introductory phase, the main study started with the first of the four episodes of the Apollo 11 mission. Each episode took around three minutes to complete. Participants were allowed to move about the space in the experimental room freely. During the episodes, one of the four guides was present and gave a presentation on the virtual space models on exhibition in the CAVE.

After each episode, the participants were asked to rate their experience using subscales of the Temple Presence Inventory [LDW09] as well as questionnaires that address the agent's credibility and the subjective knowledge gain.

We further ran participants through an "exam" on the presented educational content of the episode they just experienced, assessing how much of the information they actively perceived and could remember. The exam was chosen as a meaningful measure of the guides' quality since museums usually have an educational mandate. While the visitor is not expected to learn all facts that are presented within an exhibition, the ability to provide interesting information that sticks in the visitors' minds is of great value to any public educational institution. In this sense, the exam should give an idea of how successful a guide was to tell a memorable story rather than providing a generalizable percentage of learned facts. Initially, we planned for the exam to be completed without prior notice of the participants at the end of the study. However, a pre-study with ten participants revealed that only a minority of the users paid attention to any of the spoken text and the majority understood it more as an educational entertainment experience. We therefore decided to announce the exam before the study. For each episode, a set of 12 questions was prepared, which were similar in terms of their memorizability. They were grouped into four categories: numerical, spatial, social, and visual facts. Numerical questions included sizes, weights, quantities, and periods of time. In spatial tasks, participants had to point at a specific location within a picture of the according scene. This location was described during the episode and was usually supported by a gesture in the embodied conditions. Social facts referred to stories that were experienced by the crew and members of the mission. Visual features were not mentioned by the guide but could be

	Presence				Learning					User Experience					
	spatial	active social	social actor	engagement	numerical	spatial	social	visual	credibility	attractiveness	perspicuity	efficiency	dependability	stimulation	novelty
Embodiment	***	**	***	*	-	-	-	**	*	**	**	-	-	***	***
Closeness	-	-	*	-	-	-	-	-	-	*	-	-	-	**	**
Emb. * clos.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 15.1.: Main and interaction effects of the two factors *agent embodiment* and *thematic closeness* on each dependent variable. Asterisks indicate a statistically significant effect (* significant at .05 level, ** significant at .01 level or lower, *** significant at .001 level or lower).

observed in the presented scene. The exam was conducted orally to ensure that responses, which were guessed or already known before the study, could be identified.

After the exam was finished, participants were guided to the next episode and all steps were repeated. The second episode differed from the other scenes since participants were seated in the center of the CAVE. At the end of all episodes, participants were confronted with all four guides for a second time and had to compare them in an additional questionnaire. The entire study took around 45 to 60 minutes per participant.

15.1.4. Results

We evaluated the effect of the two factors *agent embodiment* and *thematic closeness* on several subjective and objective measures using multiple two-way repeated-measures ANOVAs. The normality assumption was not met in a few cases, however, the ANOVA tolerates moderate deviations from normality, as was shown in several studies [GPS72, HRHO92, LKK96]. A summary of all main and interaction effects can be found in Table 15.1. The calculation of each score is addressed in Section 14.3.

Presence

Different aspects of presence were assessed using the TPI: spatial presence, active social presence, presence as social actor, and presence as engagement. Each dimension involved three to seven items that were measured on a 7-point Likert scale. We ran a two-way repeated-measures ANOVA that revealed a significant main effect of agent embodiment on spatial presence ($F(1,23) = 25.822, p < 0.001, \eta_p^2 = 0.529$), active social presence ($F(1,23) = 12.181, p = 0.002, \eta_p^2 = 0.346$), presence as social actor ($F(1,23) = 299.404, p < 0.001, \eta_p^2 = 0.929$), and presence as engagement ($F(1,23) = 7.516, p = 0.012, \eta_p^2 = 0.246$). Thematic closeness only showed one significant main effect on presence as social actor ($F(1,23) = 4.420, p = 0.047, \eta_p^2 = 0.161$). No other main effect or interaction effect was significant. The results of the TPI are illustrated in Figure 15.3a.

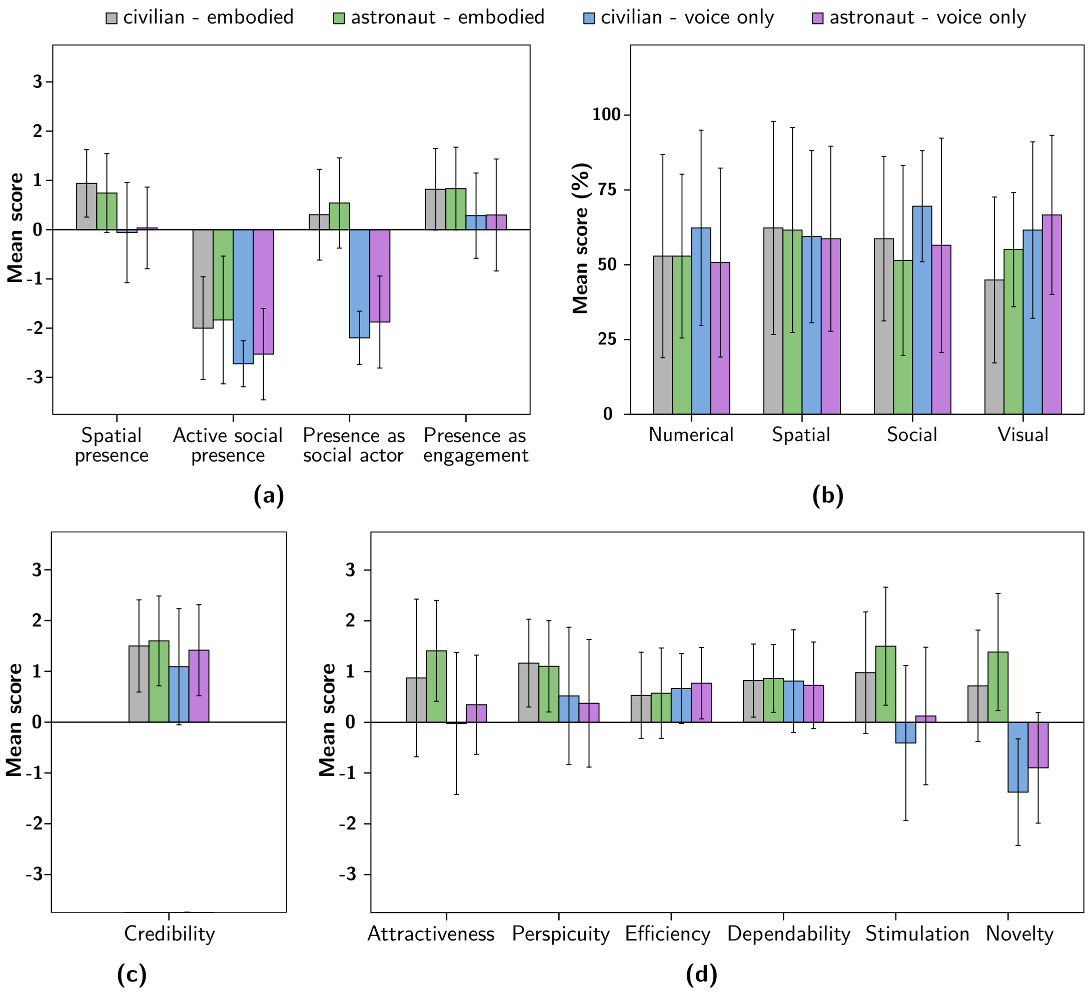


Figure 15.3.: Pooled results of (a) different presence measures, (b) learning results in four categories, (c) agent credibility, and (d) six dimensions of user experience. The vertical bars show the standard deviation.

Learning

Scores of the oral exam were added up per participant and category, with a score of 3 corresponding to the maximum value of 100%. The results of one participant had to be removed from the data because he admitted knowing several of the tested facts even without the guides due to prior knowledge on the moon landing. The remaining scores were pooled according to the four categories as illustrated in Figure 15.3b. An ANOVA revealed a significant main effect of agent embodiment on the test scores in the category of visual facts $(F(1, 22) = 8.933, p = 0.007, \eta_p^2 = 0.289)$. Apart from this, no other effects on the learning results could be found.

In addition to the objective exam, we also wanted to learn more about the subjective impression of the participants regarding their knowledge gain through the guided presentations. After each episode, before the oral exam, we asked them to make a rough estimate of how many facts they are still able to recall now and in one week. We ran another ANOVA and found a significant main effect of embodiment on the perceived number of long-term memorized facts ($F(1,23) = 16.403, p < 0.001, \eta_p^2 = 0.416$), but not on the number of short-term memorized facts ($F(1,23) = 3.185, p = 0.088, \eta_p^2 = 0.122$).

Credibility

For evaluation of the credibility of guides, we used a scale introduced by McGloin et al. as described in Section 14.3.5. We analyzed the results with a two-way repeated-measures ANOVA. The analysis revealed a main effect of agent embodiment on credibility ($F(1,23) = 5.842, p = 0.024, \eta_p^2 = 0.203$), indicating a significant difference between embodied guides ($M = 5.550, SD = 0.888$), and guides with voice only ($M = 5.254, SD = 1.030$).

User Experience

Besides the aforementioned influence of agent embodiment and thematic closeness on perceived presence, agent credibility, and learning, we were also interested in the general experience of users while interacting with the guides. For this purpose, we measured six dimensions of user experience with the UEQ. We stressed the point that all responses should be based on the impression of the guide only, without including the virtual scene. This is because the virtual objects were only used in the context of the first study and are no inherent part of applications with IVAs in general. For example, a museum could also incorporate a virtual guide to present real physical exhibits instead of virtual ones; a scenario that was investigated in the follow-up study. We ran two-way repeated-measures ANOVAs for the six dimensions of the UEQ. We found a significant main effect of agent embodiment on attractiveness ($F(1,23) = 8.837, p = 0.007, \eta_p^2 = 0.278$), perspicuity ($F(1,23) = 8.307, p = 0.008, \eta_p^2 = 0.265$), stimulation ($F(1,23) = 26.527, p < 0.001, \eta_p^2 = 0.536$), and novelty ($F(1,23) = 82.786, p < 0.001, \eta_p^2 = 0.783$). Thematic closeness also showed a main effect on attractiveness ($F(1,23) = 7.212, p = 0.013, \eta_p^2 = 0.239$), stimulation ($F(1,23) = 10.291, p = 0.004, \eta_p^2 = 0.309$), and novelty ($F(1,23) = 10.505, p = 0.004, \eta_p^2 = 0.314$). No significant interaction effects between agent embodiment and thematic closeness were found. The results are illustrated in Figure 15.3d.

After participants experienced all conditions, they were asked for a subjective ranking of the four different guides. Embodied guides were preferred by most of the participants, with 6 votes for the generic museum guide and 14 votes for the astronaut. In comparison, the unembodied generic guide took the last place for 12 and the unembodied astronaut for 9 of the participants.

In a pre-study, a participant pointed out an unfair inequality between guides, because he perceived the condition with an embodied astronaut to be the only one dubbed by a real person, while the others were assumed to be generated by a text-to-speech engine. Since even the astronaut guides with and without body were rated differently, although the same artificially generated voice was used for both of them, we decided to pursue investigations on this aspect in the main study. For each guide, participants had to decide whether the spoken text seemed to be produced by a text-to-speech engine or by a real speaker. For the unembodied astronaut, 45.8% of the participants assumed that the agent was synchronized by a real person. For the embodied astronaut, this was the case for even 62.5% of all participants. In contrast, the option of a real speaker was chosen by 37.5% of the participants for the embodied generic guide, and only by 33.3% for the unembodied generic guide.

15.1.5. Discussion

Even though our exemplary museum application did not include any forms of active interaction between the participants and the guide, the agent's embodiment had a positive effect on all measured presence dimensions. Through the presence of a second individual within the CAVE, participants felt significantly more spatially involved in the VE. Participants also reported that the embodied guides caused more emotional responses such as laughing or smiling. In general, there was only little active interpersonal communication between users and guides in all conditions, however, this could be regulated by the introduction of additional interaction mechanisms such as voice commands. Whether this is desirable strongly depends on the application itself. In public settings such as a museum, speaking with a VA may make users feel uncomfortable. In contrast, speaking with a personal assistant at home is already common practice and generally accepted. The most remarkable difference between embodied and unembodied guides can be observed in the scores of presence as social actor, sometimes also referred to as parasocial interaction. This measure of presence contains items that are related to crossing the border between the actual physical environment and the mediated environment in order to interact with the agent in real time [LDC^{+}00]. Higher scores for embodied guides indicate that participants felt that their presence was noted by the agent and that he was establishing a connection to them. Although no complex reactions of the agent to the user's behavior were implemented, a feature as simple as making eye contact seems to be an effective method to create a sense of responsiveness and intimacy. Not only the agent's embodiment but also his thematic closeness had a main effect on presence as social actor. This positive effect could be caused by the first-person perspective of the astronaut since the guide was not only imparting knowledge but was inviting the user to take part in his personal story.

Regarding credibility, all guides got mean scores in the upper range of the 7-point scale. Besides the realism of guides, this could also be attributed to the fact that users do not expect museum guides to lie to them about the chronological order of historical events. Nevertheless, we found a significant effect of agent embodiment on the perceived credibility, indicating that embodied guides seemed to be even more competent and trustworthy.

Despite the exam was announced beforehand to the participants of the study, we expected different learning results for the four types of agents, in particular with regard to the different categories of information. However, this hypothesis could be confirmed only to some extent. Visual details such as the color of specific objects could be remembered better in conditions with a voice-only guide than in scenes with an embodied guide. We expected this outcome, since users tend to follow the agent's lip movements in the embodied condition and therefore could be more distracted from the actual scene. On the other hand, we hypothesized a positive effect of embodiment on the memorization of spatial information, however, such an effect could not be found in the data. In contrast to the results of the objective oral exam, participants subjectively perceived their gain of knowledge to be higher in the conditions with embodied guides, in particular in the long term. Indeed, the involvement of multiple modalities in the learning process, as well as an increased presence in virtual environments, were related to better learning results in previous studies [Mik06]. A follow-up study that focuses on long-term effects of learning could resolve the question of whether the subjective impression of participants can be supported by an objective test.

We also expected the scores for social questions to be higher for guides with a personal connection to the stories. While no significant effect was found between the generic museum guide and the astronaut, Figure 15.3b even indicates a trend in favor of the generic guide. The comment section of the questionnaires could give some indication of possible reasons for the observed behavior. Some participants stated that the astronaut was harder to understand due to the applied radio transmission effect.

Besides the problems in understanding the astronaut due to the added distortions, it was also mentioned that this effect made the astronaut sound more realistic than the generic guide. This impression was also confirmed by the responses to the question of whether the audio was generated with a text-to-speech engine or spoken by a real person. Besides the thematic closeness of the agent, his embodiment also affected the perceived realism of his voice positively. Despite identical audio tracks, the presence of an embodied agent seems to distract the user from artifacts of speech synthesis and made the voice sound more natural. Therefore, the embodied astronaut was perceived to have a real voice by the majority of participants.

The perceived realism of the astronaut could also contribute to his positive reception by the participants of the study. In the usability questionnaires, the astronaut guides were rated as significantly more attractive, exciting, and motivating, as well as innovative and creative. This is also true for embodied guides in comparison to voice-only guides. These results indicate that the extra effort that has to be made to implement a customized agent could be worthwhile to increase user satisfaction and improve the overall user experience.

15.2. Follow-Up Study with Real Exhibits

The user study described in Section 15.1 provides insights into the effects of an IVA's embodiment in museum exhibitions with virtual exhibits. However, it remains open if a physical exhibit in combination with a virtual guide could further enrich the user experience. Hence, we conducted a follow-up study to replicate the scenario from the first experiment in an environment in which the IVA and real objects are blended into the same space.

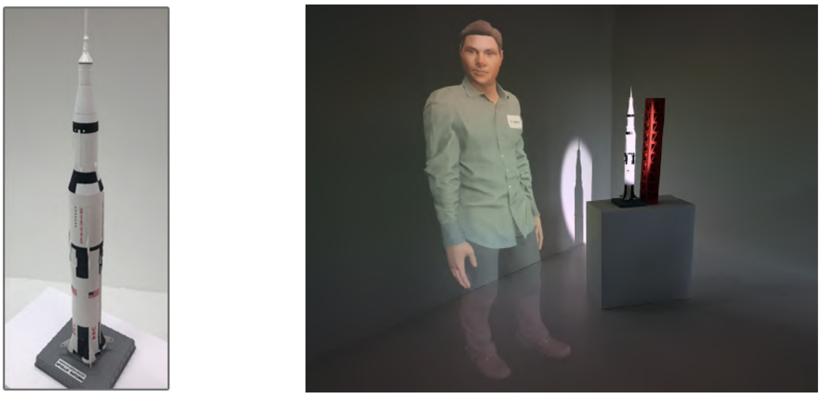


Figure 15.4.: (a) Physical scale model of the Saturn V, and (b) experimental setup.

15.2.1. Participants

For the follow-up study, we recruited 24 participants (15 male and 9 female, aged from 20 to 46, $M = 26.5$), who did not participate in the first experiment. All participants were students or staff members of the local department of informatics. Most of them already had some experience with VR/AR since only two of them participated in a study involving VR or AR for the first time. As for the first experiment, we verified that participants do not suffer from any visual disorders that could interfere with the study procedure.

15.2.2. Materials

To ensure comparability between both the first experiment and the follow-up study we used the identical technology setup at the same location as described in Section 15.1.2. Due to the different situation in this study, it was required to slightly modify the scene. This particularly involved the presented exhibit, while the guides remained unchanged. For the follow-up study, we decided to recreate the first episode using a plastic scale model of the Saturn V and its launcher as illustrated in Figure 15.4a. The physical rocket featured the same visual details as its virtual equivalent with half the overall size. Due to the smaller height of $77.5cm$, it was placed on a white box and therefore could be examined by the participants of the study on eye level similar to a real exhibit in museums. As in the first experiment, the scale model was positioned in a corner of the CAVE, as illustrated in Figure 15.4b.

15.2.3. Methods

Due to the hardware constraints, we focused on the first episode only, and hence, the follow-up study followed a between-subjects design with two independent variables: *agent embodiment* (embodied vs. disembodied) and *thematic closeness* (astronaut vs. museum guide). Participants were randomly assigned to one of the four resulting conditions. The introduction was carried out as in the first experiment, including a consent form, a demographic questionnaire, and the staging of the exhibition scenario. For participants who experienced a voice-only

	Presence			Learning					User Experience					
	active social	social actor	engagement	numerical	spatial	social	visual	credibility	attractiveness	perspicuity	efficiency	dependability	stimulation	novelty
Agent embodiment	***	***	***	-	-	-	-	-	***	**	-	-	***	***
Thematic closeness	-	-	-	-	-	-	-	-	*	-	-	-	*	*
Exhibit virtuality	-	-	*	-	-	-	**	-	-	-	-	-	-	-
Emb. * clos.	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Clos. * virt.	-	-	-	-	-	-	-	*	-	*	*	-	-	-
Emb. * virt.	-	*	*	-	-	-	-	-	-	-	-	-	-	-
Emb. * clos. * virt.	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 15.2.: Pooled results of the second study and the first episode of the initial study. Asterisks indicate a statistically significant main or interaction effect of the three factors *agent embodiment*, *thematic closeness*, and *exhibit virtuality* on the corresponding dependent variable (* significant at .05 level, ** significant at .01 level or lower, *** significant at .001 level or lower).

condition, we omitted a demonstration of the stereoscopic display since none of the presented objects were virtual. All of the participants were instructed to imagine visiting a real space museum and to behave as naturally as possible. As in a real museum, participants were allowed to move freely within the exhibition space but were prohibited from touching the exhibit. After the introduction, the selected guide appeared and presented the first episode of Apollo 11 as described before. As in the first study, the episode took around 3 minutes and was followed by several questionnaires that addressed different presence scales, the guide's credibility, and the subjective knowledge gain. Afterward, we performed the oral test using the same questions as in the first iteration. The experiment was concluded by some final questions regarding the user experience. As described above, the procedure was slightly different from the first experiment since participants experienced only one of the guides before rating their experience.

15.2.4. Results

To evaluate the effect of presenting a real exhibit instead of a virtual one, we compared the observations of the first experiment's first episode with the follow-up study, which used the identical material and methods. Therefore, the data gathered in the first episode can be treated as obtained in a between-subjects design like in the follow-up study. Hence, we also considered the factor called *virtuality* with two levels *virtual exhibit* and *real exhibit*, and ended up with a $2 \times 2 \times 2$ design. The data was analyzed using a three-way ANOVA with the three factors *agent embodiment*, *thematic closeness*, and *virtuality*. An overview of all main and interaction effects is presented in Table 15.2.

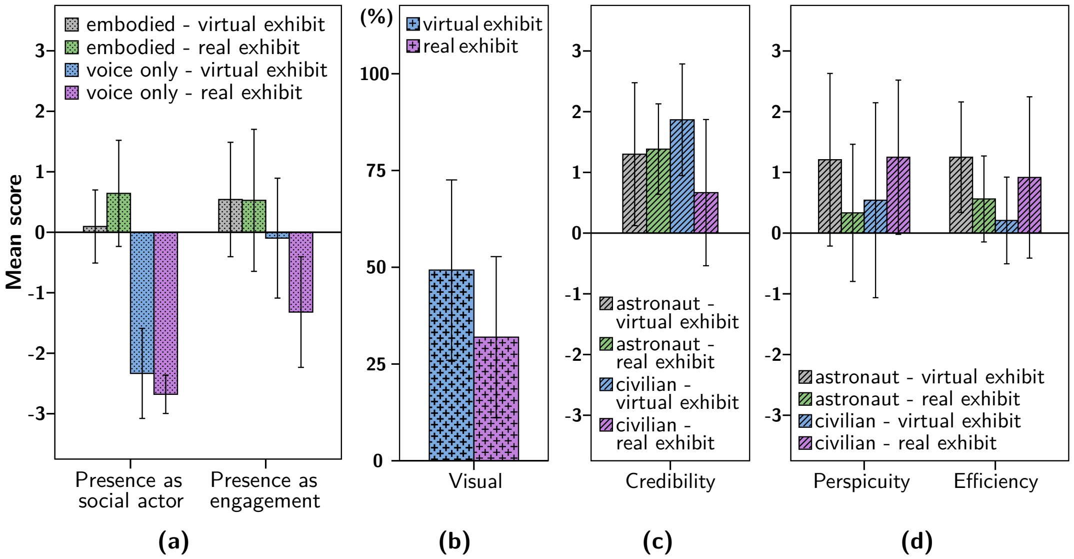


Figure 15.5.: Pooled results of the merged studies including (a) different presence measures, (b) learning results in the visual category, (c) agent credibility, and (d) two dimensions of user experience. The vertical bars show the standard deviation.

Presence

For evaluating the presence we excluded the subscale of spatial presence since half of the participants in the second experiment did not experience virtual content at all and were therefore not able to make valid statements on this dimension of presence. We found significant main effects of embodiment on all remaining subscales, namely active social presence ($F(1, 40) = 12.813, p = 0.001, \eta_p^2 = 0.243$), presence as social actor ($F(1, 40) = 212.119, p < 0.001, \eta_p^2 = 0.841$), and presence as engagement ($F(1, 40) = 17.982, p < 0.001, \eta_p^2 = 0.310$). In contrast to the first experiment, thematic closeness did not show any significant main effects.

Virtuality also showed a significant main effect on engagement ($F(1, 40) = 4.446, p = 0.041, \eta_p^2 = 0.100$). Furthermore, the ANOVA revealed significant interaction effects between virtuality and embodiment both on presence as social actor ($F(1, 40) = 5.115, p = 0.029, \eta_p^2 = 0.113$), and presence as engagement ($F(1, 40) = 4.248, p = 0.046, \eta_p^2 = 0.096$). The results involving virtuality are illustrated in Figure 15.5a.

Learning

The results of the oral exam were prepared for the analysis as described in Section 15.1.4. A three-way ANOVA resulted in a significant main effect of virtuality on test scores in the visual category ($F(1, 39) = 7.574, p = 0.009, \eta_p^2 = 0.163$). No other significant effects on objective and subjective learning results could be found.

Credibility

Credibility scores, which were again computed using the approach suggested by McGloin et al., were also analyzed using an ANOVA. While no significant main effect could be found for any of the three factors, the ANOVA revealed a significant two-way interaction effect between virtuality and thematic closeness ($F(1,40) = 4.398, p = 0.042, \eta_p^2 = 0.099$). The interaction between both factors is visualized in Figure 15.5c.

User Experience

As in the first experiment, we found significant main effects of agent embodiment on attractiveness ($F(1,40) = 20.169, p < 0.001, \eta_p^2 = 0.335$), perspicuity ($F(1,40) = 9.158, p = 0.004, \eta_p^2 = 0.186$), stimulation ($F(1,40) = 36.507, p < 0.001, \eta_p^2 = 0.477$), and novelty ($F(1,40) = 106.997, p < 0.001, \eta_p^2 = 0.728$). Thematic closeness also showed a main effect on attractiveness ($F(1,40) = 5.322, p = 0.026, \eta_p^2 = 0.117$), stimulation ($F(1,40) = 4.711, p = 0.036, \eta_p^2 = 0.105$), and novelty ($F(1,40) = 4.644, p = 0.037, \eta_p^2 = 0.104$). In addition, two significant interaction effects between virtuality and thematic closeness on perspicuity ($F(1,40) = 4.535, p = 0.039, \eta_p^2 = 0.102$) and efficiency ($F(1,40) = 6.058, p = 0.018, \eta_p^2 = 0.132$) could be found.

15.2.5. Discussion

We found significant differences for the virtuality of the exhibit as well as interactions between the virtuality and both embodiment and thematic closeness of the agent, as summarized in Figure 15.5. In contrast to the first experiment, no significant main effects of thematic closeness on presence as social actor as well as embodiment on learning of visual facts and credibility were found in the aggregated data, which might be due to the reduced sample size.

One of the most interesting results from the study is the observed interaction effect between the agent's embodiment and the exhibit's virtuality on the subjective measure of presence as social actor, or parasocial interaction. As described above, this dimension of presence relates to a cross-over between the actual physical environment of the user and the mediated environment. Our analysis of the collected data of both experiments revealed that embodied guides achieved higher scores when displayed alongside a physical exhibit. Therefore, the physical exhibit may have supported the transfer of the virtual guide to the real environment of the participant. On the other hand, the parasocial presence was rated higher for the virtual exhibit than for the real one for conditions featuring a voice-only guide. This is interesting as it may indicate a reverse effect compared to the previously reported effect. As the audio guide was not embodied in the actual physical environment, a virtual exhibit may have helped the user feel more present in the virtual environment of the guide, therefore again bridging the gap between the user and the guide.

Another interaction effect between the embodiment of the agent and the virtuality of the exhibit was found for engagement. While the engagement ratings were similar for the embodied guides, they were significantly lower for audio guides in conjunction with real exhibits. Participants assigned to this condition did not experience any virtual content, therefore being the closest to a traditional exhibition scenario. Though participants of the first experiment were explicitly asked to focus on the guide during their evaluation, the overall context with the virtual exhibit seemed to have an influence on their engagement as well.

A lack of engagement in the group of participants experiencing an audio guide with a real exhibit may also have contributed to the lower performance in the oral test with regard to visual facts. Overall, participants in the conditions with a virtual exhibit could remember more visual facts than participants in the conditions with a real exhibit. Though this effect could also be attributed to the differences in size and visual details of the real exhibit, the informal comments during the oral exam support another conclusion. In the second study, several participants who experienced an embodied guide reported that they were more interested in the virtual guide than the physical rocket, therefore not paying attention to visual features of the latter. Furthermore, in the first experiment, some participants who were assigned to a condition with an audio guide stated that they paid less attention to what was said since they preferred to explore the virtual rocket. It can be assumed that both reported behaviors eventually caused the effect which is shown in Figure 15.5b.

Another significant difference between the first and the second experiment was found for two of the six user experience scales. In the second study using a real exhibit, the generic museum guide was rated higher while participants of the first experiment with virtual exhibits provided higher scores in favor of the astronaut. This interaction between thematic closeness and virtuality may provide an indication that a generic museum guide, which actually could be found in a real museum, fits in better with a real exhibition room than a content-related guide, whose presence is unusual for visitors of a museum. On the other hand, an environment with a virtual exhibit already is an exception to the norm and therefore, the presence of Neil Armstrong as a tour guide might be more relatable. However, this interpretation is limited in view of the fact that the results only apply to the scales of perspicuity and efficiency, while participants of both experiments preferred the astronaut in terms of stimulation and novelty.

15.3. Conclusion

In this chapter, we summarized two user studies, which investigated the effectiveness of different representations of IVAs in an exhibition scenario. We analyzed the effects of three factors, agent embodiment (embodied vs. disembodied), thematic closeness (astronaut vs. museum guide), and exhibit virtuality (virtual vs. physical) on a number of variables that are relevant to the museum domain, including social presence, guide credibility, knowledge transfer, and visitor experience. In this context, we aimed to examine whether the costly and time-consuming implementation of embodied agents and their customization to a specific application give a competitive edge over common IVAs with audio only. The first study was conducted in a virtually simulated exhibition room addressing the Apollo 11 mission. To ensure ecological validity, we replicated the scene in a real exhibition space and analyzed the effects in a second study.

In the pooled data of both experiments, we found significant differences between audio guides and embodied guides with regard to all presence measures as well as a subset of user experience scales, including perceived perspicuity, attractiveness, stimulation, and novelty.

All effects were in favor of the embodied guide and therefore could justify the extra effort that is necessary to model and animate such an agent. This option should be taken into consideration for all applications, in which user experience is of top priority and the usage of additional technology such as a projector is reasonable, for example, in public installations.

The content-specific guide in the form of an astronaut achieved higher scores in the dimensions of attractiveness, stimulation, and novelty. As a representative of historically relevant guides, we also expected an increase in both the credibility and the knowledge transfer, since visitors may emphasize with the guide's feelings and emotionally engage with him because of the personal connection to the story that was told. This hypothesis could not be confirmed based on the results of both user studies, however, credibility was rated slightly higher for the astronaut guide than for a generic museum guide. While no positive effects on learning could be found for the astronaut, visitors of a museum might be attracted by the more innovative guide representation and therefore pay more attention to the related exhibit. Furthermore, different results may be achieved in a field study within a real museum since, according to the qualitative feedback, many participants were not eagerly interested in the Apollo 11 mission and instead participated because of their interest in VR technology.

The virtuality of exhibits showed significant main effects on presence as engagement as well as the rate of remembered visual facts. The first result emphasizes general advantages of using virtual content in the museum context. Participants assigned to the condition with a real exhibit and a voice-only guide were significantly less engaged than participants of any other condition. This lack of engagement may also have contributed to the latter result since participants who experienced a real exhibit apparently paid less attention to the visual details and therefore could remember only a few. Besides these differences between a virtual and real exhibition space, most of the results of the first study could be reproduced within the second study, suggesting that the described positive effects of both embodied and content-related guides also apply to traditional museums with real exhibits.

Though IVAs are emerging in various domains, we chose the environment of an exhibition to gain initial insight into the effectiveness of different agent representations. Some of the results may apply to other domains, too, as the considered aspects are relevant not only in the context of exhibitions. For example, high spatial and social presence values increase the IVA's ability to be perceived as a real social entity and therefore contribute to any social experience that involves IVAs. The positive effects on variables such as attractiveness and stimulation are also of high value for other applications, since user experience is a key aspect for most human-computer interfaces. On the other hand, knowledge transfer is one of the more specific aspects in the presented studies, which may be less relevant to other domains. Instead, there might be additional application-specific factors to be included. For example, in health care for children, a virtual expert such as a doctor could be compared to a less intimidating agent such as a mascot. In this domain, agent credibility and social presence are still of great importance, but other variables such as the release of fears should also be brought into focus. Additional studies are necessary to fully answer the questions regarding which agents perform best in different scenarios.

16 Chapter 16. Blended Agents: Manipulation of Physical Objects Within Blended Spaces and Beyond

In modern MR environments, the links between real users and IVAs are bidirectional in many regards: By means of advanced display technology, IVAs can be rendered as 3D spatial entities within the same environment as the user, while head tracking allows agents to also detect and react to the user's position within this environment. Natural language processing enables IVAs to understand their human communication partners while speech synthesis and natural dialogue systems generate human-like responses. By contrast, modern sensing technology such as head or hand tracking systems allows virtual objects to show physically correct reactions to actions of the real user, while the manipulation of real objects through IVAs is only possible in a very limited scope, for example, in the form of coherent global illumination. Other virtual-physical interactions such as simulated collisions between IVAs and physical objects need more complex actuators, in particular if the technology behind this interaction should be hidden from the user to create an advanced illusion of plausible human-like agents. This additional complexity is leading to an asymmetry between real and virtual interaction partners as illustrated in Figure 16.1. As a consequence, only a few examples of virtual-physical interactions can be found in the literature (examples are discussed in Section 14.2.5). Instead, most of the current MR applications accept implausible virtual-physical collisions to a certain extent or try to avoid them when possible [KBW17, KMB^{+}17].

In this chapter, we introduce the concept of blended agents, which are able to manipulate real-world objects in an interactive way. Throughout the following sections, we focus on two different forms of virtual-physical interactions:

1. Manipulations of physical properties related to the object's location.
2. Manipulations of physical properties related to the object's surface material.

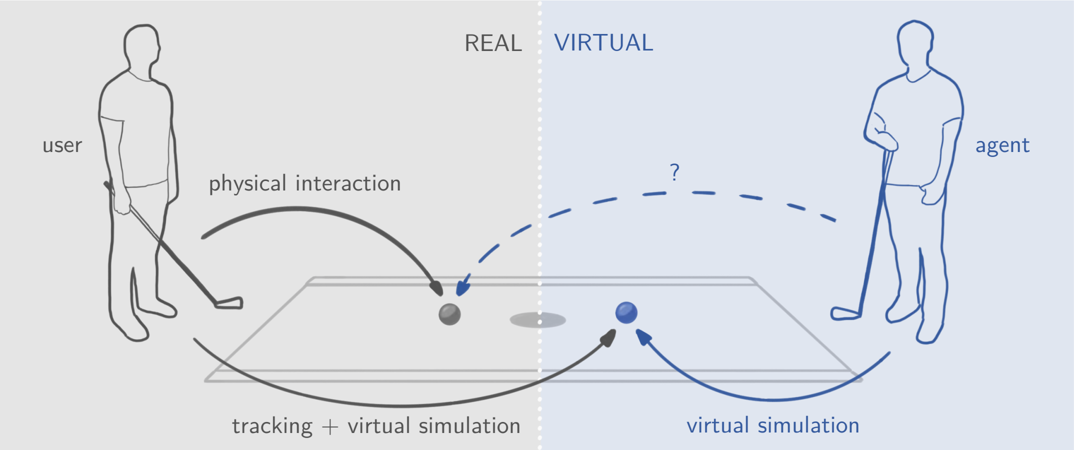


Figure 16.1.: Asymmetric interaction between real users and IVAs. While users can influence both physical and virtual objects, actions of agents usually only affect their virtual surroundings.

To address the first form of virtual-physical interactions, we utilized an off-the-shelf robotic golf ball that moves along a scripted path to simulate interaction with a virtual golf player. For the second form, we designed a novel device that uses temperature variation to activate thermochromic ink on a sheet of paper. In the presented prototype setup, a pre-defined score appears virtually before it is replaced by a physically persistent version. Synchronized with the animations of a blended agent, the illusion of a virtual human writing on a physical piece of paper can be created.

Both forms of virtual-physical interactions not only differ from each other with regard to the underlying physical change, but may also cause varying user responses due to differences in what we call *explicability*, *observability*, and *persistence*. In terms of explicability, we assumed that manipulations of an object's position achieved via motors or magnetic actuators are both more common and, depending on the implementation, more conspicuous as they might also influence other properties of the prepared object such as its weight. Changing an object's surface material, however, usually requires chemical reactions, for example, to temperature variations, pressure, or UV light. Such chemical changes are uncommon in other application fields and usually do not interfere with other object properties. If users are not capable of finding an obvious explanation for an effect, this might support the illusion of an interaction between the blended agent and the physical object. Furthermore, during the blended experience, manipulations of an object's position can be observed directly, while changes of the surface material might not be easily detectable by users as similar effects can be achieved by overlaying virtual projections. Therefore, we hypothesize that during the experience changes of the material might be not recognized by the user at all. On the other hand, since such changes can be persistent they could be observed by the user outside of the Blended Space. Whether such a long-term manipulation of real-world objects could change the perception of the IVA's realism retrospectively is one of the questions we intended to answer in a user study.

Taking these considerations into account, we formulate the following hypotheses:

(H1) Virtual-physical interactions improve the user experience in terms of social and spatial presence, ecological validity, perceived anthropomorphism of the blended agent, and engagement.

(H2) Virtual-physical interactions related to the surface material of an object have a stronger positive impact on the aforementioned metrics than those related to the object's position.

(H3) Chemical changes of an object's surface material can be hidden from the user, while mechanical manipulations of the object's location are more explicable.

(H4) Manipulations of the object's position are observed by the users directly, while manipulations of the object's surface material are not observed before the end of the MR experience.

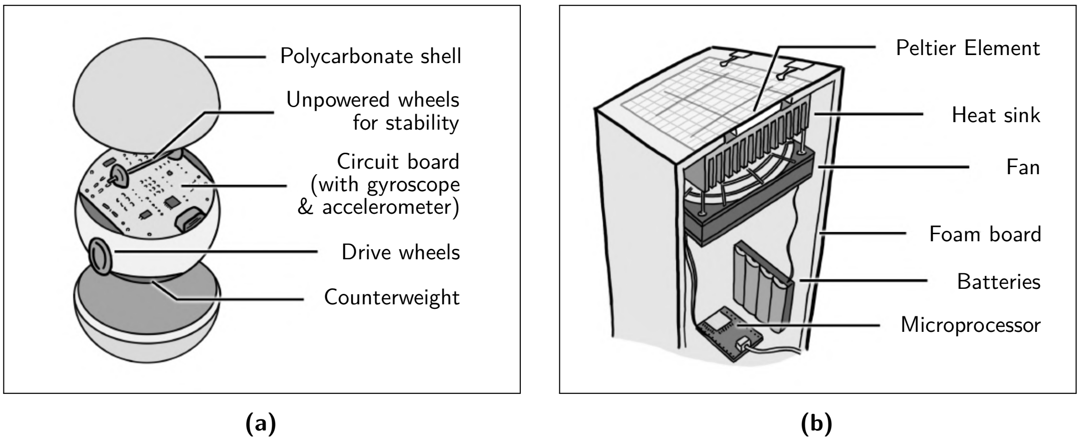
To our knowledge, no prior work has investigated similar manipulations and their effects on agent-human interaction. While the related research projects presented in Section 14.2.5 already demonstrate the potential of blended agents for enhancing MR experiences, they show considerable differences to our project. Firstly, each of the previously implemented virtual-physical interactions is based on manipulations of the physical object's pose. As the range of physical changes covers many other effects, for example, related to the object's shape and surface material, it is an interesting question whether the previously observed effects are generalizable to these forms of physical manipulations, and if there are individual differences between them. Secondly, all of the previously presented studies rely on a within-subjects design, therefore allowing the participants a direct comparison of agents with and without virtual-physical capabilities. As users in real-world applications usually do not have this comparison and virtual-physical manipulations are not yet common enough to be assumed to be the norm, the question arises whether users expect IVAs to be capable of physical manipulations to appear human-like. Therefore, the contributions of this project are:

- Implementation of an API for an off-the-shelf robotic ball to simulate collisions with a virtual golf club.[1]

- Development of a proof of concept for a thermal table that allows blended agents to persistently write on physical sheets of paper.

- Collection of subjective quantitative and qualitative user responses to compare both forms of virtual-physical interactions in a between-subjects design.

[1] https://github.com/augmentedrealist/spheromini.js



(a)

(b)

Figure 16.2.: Schematics of (a) a robotic ball and (b) a thermal table, which implement two different virtual-physical interactions.

16.1. Apparatus

Our primary goal in this project was to gain insight into MR experiences that involve different forms of interactions between blended agents and physical objects. For this purpose, two different setups were implemented to exemplify virtual-physical interactions in the form of (i) movements of a physical ball, and (ii) writing on a physical sheet of paper.

16.1.1. Robotic Ball

To create the illusion that a blended agent is moving a physical ball, we utilized an off-the-shelf *Sphero mini*. This motorized ball has a diameter of 42 mm and therefore matches the size of a customary minigolf ball. The locomotion system, as well as additional sensors, are hidden inside an outer shell. Two motor-driven wheels move along the inner surface of the shell to actuate a weight at the opposite side (see Fig. 16.2a). Due to the resulting weight shift, the ball can move with three degrees of freedom (i.e., rotation around the y-axis and x/z translation) with a maximum speed of $1mps$. During navigation, the ball is stabilized using an inertial measurement unit (IMU) that contains both an accelerometer and a gyroscope.

While we could take advantage of the mechanics of the Sphero mini, the control from within a game engine required some implementation effort. Since at the time of writing this chapter no SDK for this Sphero model was available, we implemented a custom JavaScript library based on existing SDKs of the preceding models. The library contains commands to establish a connection to the robotic ball and to control its motors. The code is executed on a dedicated NodeJS server, which can receive motion commands from any Unity application via HTTP requests. The commands are translated to Sphero mini machine code and sent to the device via Bluetooth LE. The full source code has been uploaded to GitHub.

16.1.2. Thermal Table

To address virtual-physical manipulations on a material level, we did a lot of research and testing to find a dye that is invisible to the eye under normal conditions and can be activated by external stimuli such as light or temperature. Photochromic dyes turned out to be unsuitable for Blended Spaces with real users as they usually have to be exposed to high-intensity UV light in order to change their color within seconds. We therefore decided in favor of the thermochromic ink and evaluated four different activation temperatures ($-10°C$, $31°C$, $47°C$, and $90°C$) in advance of the user study to find the best tradeoff between practicability and user experience. The inks with the highest and lowest activation temperature featured an irreversible color change but had to be discarded, nonetheless, as touching the activating thermal element creates a strong sensation of cold/heat, which may reveal the hidden technology and even pose a danger to the uninformed user. For this reason, we narrowed down the choice to the two inks with medium activation temperatures and reversible behavior. To still convey the user the impression of permanent writing, we finally chose the $47°C$ dye[2], as body temperature is not high enough to unintentionally trigger another transition from black to transparent after the sheet of paper was taken by the user.

For the activation of the thermochromic ink, we built a thermal device composed of a Peltier element that can display temperature variation (see Fig. 16.2b). The device features the TEC1-12706 thermoelectric plate, it measures $40 \times 40mm$, operates from $0 \sim 15V$ and $0 \sim 6A$, and the temperature range goes from $-30°C$ to $70°C$. We used the plate in conjunction with a heat sink and a fan, as a mechanism to dissipate the heat from the bottom side. In order to variate the temperature on the top side, to be a cold or a hot end, we use an H-Bridge circuit that switches the polarity of the voltage applied. As a result, we can transfer heat with a cooling rate of $4°C/s$ and a heating rate of $8°C/s$.

The plate is installed on the top of a pedestal ($1.0m$), offering access to the temperature-switching side of the Peltier device with a tilt angle of 22.2°. An Espressif ESP32 controls the thermoelectric plate, a dual-core microprocessor clocked at $240Mhz$ with $520Kb$ RAM and powered by rechargeable 18650HG2 Li-Ion batteries ($3.7V$). All the electronic components are mounted inside the pedestal and connected to the control PC using a USB cable. The cable handles the serial communication with the device at $57Kbps$, enabling the PC to send three commands in order to enable/disable the thermoelectric plate, switch the temperature from high ($65°C$) to low ($10°C$) or vice versa, and activate the thermo-active ink accordingly.

In the current implementation of the thermal table as a proof of concept, all written text has to be prepared before the MR experience. Only when placed on the powered table, the prepared text turns invisible until the polarity of the thermoelectric plate is changed and the sheet of paper is cooled down. In Section 16.2.6, we will discuss some thoughts on possible enhancements of the setup to allow flexible writing instead of pre-defined text only.

[2] SFXC thermochromic color changing screen ink for paper and board, Black 47C.

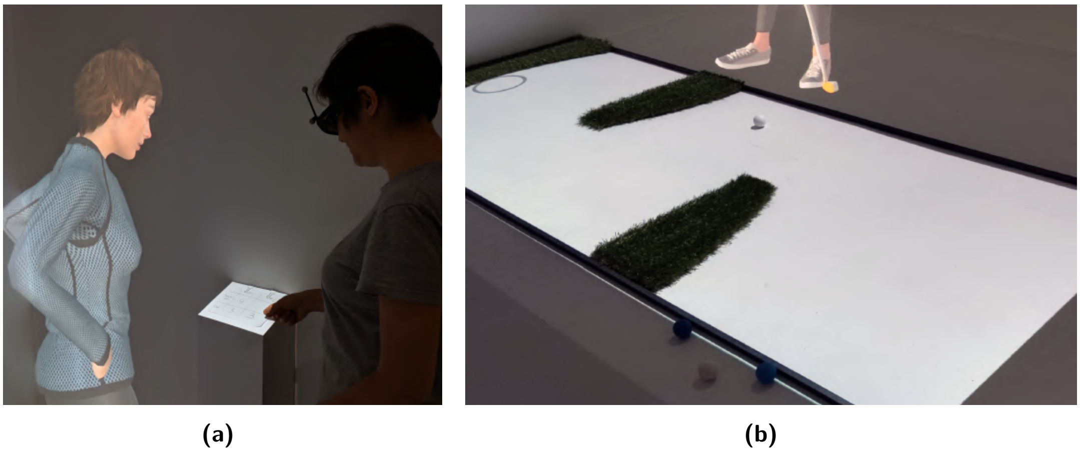


Figure 16.3.: Photos showing the experimental setup, with (a) the thermal table, and (b) the minigolf course with a robotic ball.

16.2. User Study

As the proposed systems implement two different forms of virtual-physical interactions – one similar to previously tested techniques and one that features a novel approach – we conducted a comparative study to collect subjective responses of users. For this purpose, we designed an experimental environment that naturally embeds both implementations, as described in detail in the following section.

16.2.1. Materials

The Blended Space for the user study was inspired by a minigolf course, with an IVA acting as the opposing player. The course with a length of 2.9 meters and a width of 1.05 meter was set up within a four-sided CAVE, which is described in detail in Section 5.2. Two heavy ropes marked the edges of the course and pieces of artificial turf served as obstacles. The hole was marked with a slightly raised ring. To experience the view-dependent stereoscopic content, users had to wear shutter glasses with passive markers that were tracked by a five-camera OptiTrack system. Furthermore, the voice of the IVA was presented to participants via wireless noise-cancelling headphones. Another purpose of the headphones was to block ambient noise that was created by the thermal table's internal fan. In contrast, sounds caused by the friction between the golf ball and the floor were still audible. All of the actions of the participants were monitored by the experimenter using a camera at the ceiling of the CAVE. By this means, the experimenter was able to trigger particular reactions of the blended agent from a neighboring room, without being visible to the participants.

As we learned from related research projects, consistency between different dimensions of realism seems to be crucial for the perceived human likeness of IVAs. For this reason, the IVA used in the experiment has to meet several expectations regarding her appearance, speech, and behavior. For a high degree of visual fidelity, we used a 3D scanned female head model that features a highly detailed mesh, 4K PBR textures, and multiple facial expressions.[3] To add vividness to the IVA's face, her eyes were not moving randomly but were focusing occasionally on special points of interest such as the golf ball or the user. Furthermore, micro movements, i.e., saccades, as well as blinking reflexes were performed. The agent's body was created and rigged using Adobe Fuse as well as Mixamo. As retargeted keyframe animations were described as stiff and artificial in a pre-study, we decided to replace them with motion-captured material. All animation sequences were performed by a female actor and recorded with an 8-camera Qualisys Miqus M3 system. A path of the robotic golf ball was programmed accordingly to match the animations of the blended agent. In addition, a native speaker provided the voice of the IVA, and matching lip movements were created via the Oculus Lip Sync plug-in. To improve the realism of sound propagation, we used the Unity MS HRTF spatializer plug-in, which incorporates the binaural head-related transfer function. Besides the voice, no additional sound effects were included in the current experiment.

16.2.2. Methods

For the user study, we followed a between-subjects design with two independent variables and two levels each. The resulting four conditions, all in relation to the IVA's interactions, are:

- (B_vH_v) Virtual golf ball and virtual handwriting.
- (B_vH_r) Virtual golf ball and real handwriting.
- (B_rH_v) Real golf ball and virtual handwriting.
- (B_rH_r) Real golf ball and real handwriting.

The random assignment of conditions was counterbalanced among participants. Before a new participant arrived, the golf course was prepared according to the selected condition. In the conditions (B_vH_r) and (B_rH_r) the thermal table was turned on and a sheet of paper prepared with invisible ink was placed on its top. In preparation for the conditions (B_rH_v) and (B_rH_r), the NodeJS server was started and a connection of Unity to the robotic ball was established. Since the Sphero mini is not able to provide a global orientation value, the initial rotation of the golf ball had to be determined by hand. A manual correction was performed until the ball moved perfectly along a test track. Afterward, the ball was positioned at its starting slot along with three other physical golf balls.

Before they entered the previously described minigolf course, participants had to fill in a consent form as well as a pre-questionnaire to provide demographic information. Afterward, each participant was guided to the CAVE and the procedure, as well as general minigolf rules, were explained. In this introductory phase, participants were able to examine the golf course as well as the scorecard with their naked eye. Also, the preparation of the scorecard with a table was executed in sight of the participants to make sure that they realize it was empty when they entered the room. After all questions were resolved, the participant had to wear shutter glasses and the experimenter started with the first round.

[3] Animatable Digital Double of Louise by Eisko©(www.eisko.com).

In total, four rounds of play were performed: (1) The experimenter playing with a physical ball, (2) the participant playing with a physical ball, (3) the (blended) agent playing with a virtual/robotic ball, and (4) the participant playing with a physical ball. After rounds (1) and (2), the experimenter and participant filled in one blank of the scorecard each. Afterward, participants were introduced to the IVA by the experimenter. They were given noise-cancelling headphones and the experimenter left the room. The IVA then started a conversation with the participant and putted either a virtual or a robotic ball into the hole, according to the selected condition (see Fig 16.3b). After finishing round (3), the IVA walked to the scorecard and asked the participant where to fill in her score, as illustrated in Figure 16.3a. The animation was only resumed by the experimenter if the participant was close to the table. This artificial pause should ensure that all participants witnessed the handwriting of the IVA and therefore do not make false assumptions about how and when the VA's score was added to the scorecard. If the participant was in sight of the thermal table, the IVA virtually wrote a predefined score of 4 in the dedicated blank space. The IVA then challenged the participant to bet her score in round (4). During this second round of the participant, all hits were counted by the experimenter using the live camera view. In the conditions (B_vH_r) and (B_rH_r), when the participant was close to the hole, the temperature of the thermal table was switched from high to low and the thermochromic ink below the projected score became visible. At the same moment, the projected score was faded out with the result that the participant could only see the physically written score when he returned to the table.

After round (4) was finished and the participant filled in the last blank of the table, the IVA started a final evaluation of the match. Based on the number of strokes that were digitally logged by the experimenter, the IVA announced the winner of the game. Finally, the IVA said goodbye and suggested to the participant to take the scorecard as a souvenir. The user left the CAVE and was asked to fill in a series of questionnaires. Overall, the study took around 20 to 25 minutes to complete.

16.2.3. Participants

We invited 40 participants to our study, 27 male and 13 female (aged from 18 to 41, $M = 25.35$). 36 of them were students or staff members of the local department of informatics, while 4 stated to pursue a non-technical profession. According to the pre-questionnaire, 7 participants took part in a study involving VR or AR for the first time. None of the 40 participants reported any visual impairments that could affect the results of our experiment.

16.2.4. Results

During the user study, we collected both quantitative and qualitative subjective data that can give some indication of how different virtual-physical interactions affect a MR experience. The results are presented in the following section.

Table 16.1.: Mean scores and standard deviations for each of the 4 conditions and 5 dependent variables.

		Social Presence (1 - 7)		Spatial Presence (1 - 7)		Ecological Validity (1 - 5)		Anthropo-morphism (1 - 5)		Engagement (1 - 5)	
Golf ball	Writing	M	SD	M	SD	M	SD	M	SD	M	SD
virtual	virtual	4.74	1.139	4.600	0.824	3.840	0.506	3.620	0.745	4.567	0.522
virtual	real	4.20	1.178	4.643	0.678	3,520	0,738	3.320	0.694	4.233	0.545
real	virtual	4.80	0.660	4.671	1.048	3,720	0,634	3.320	0.655	4.500	0.503
real	real	4.16	0.970	4.586	0.995	3,700	0,620	3.260	0.795	4.467	0.477

Quantitative Analysis

Each experiment session was concluded with five questionnaires that addressed different aspects of the experience:

- Social presence (= Social Presence Questionnaire by Bailenson et al. [BBBL03])
- Spatial presence (= subscale of the Temple Presence Inventory [LDW09])
- Ecological validity (= subscale of the ITC Sense of Presence Inventory [LFKD01])
- Perceived anthropomorphism (= subscale of the Godspeed questionnaire [BKCZ09])
- Engagement (= top-loading items of the engagement subscale of the ITC Sense of Presence Inventory [LFKD01])

Results were measured on 7/5-point Likert scales, as noted in the head of Table 16.1. For each participant, average scores were formed according to the computation models that are suggested in the original papers (see Sec. 14.3). We analyzed the data using multiple two-way ANOVAs, but could not find any significant main or interaction effects. The mean values, as well as standard deviations for all conditions and each dependent variable, are also summarized in Table 16.1.

Qualitative Analysis

In addition to the mentioned Likert scales, we were also asking participants some open-ended questions about experienced or expected effects of blended agents, depending on the tested condition:

(B_r) *"How did the agent's interaction with a real golf ball affect your experience?"*

(B_v) *"Imagine the agent interacting with a real golf ball instead of a virtual one. How would this interaction have affected your experience?"*

Table 16.2.: Categorization of the users' utterances in open-ended questions related to the user experience.

			Count		Count	
	Sub-category	Example	B_r	B_v	H_r	H_v
Realism	More realistic	'It made me feel like I am playing against a real human.'	6	10	7	5
	Less realistic	'It looked a little unrealistic how the golf club was hitting the ball.'	2	1	0	0
	No difference	'I don't think that it would have changed much.'	2	4	3	4
Emotions	Disconcertment	'Would've probably felt more weird.'	3	5	2	3
	Confusion	'Initially there was a bit of confusion whether it is the real ball.'	4	2	3	2
	Surprise	'It was a fun surprise to see the actual ball be moved.'	7	0	6	3
	Enjoyment	'It made the experience more fascinating and memorable.'	11	3	11	7

(H_r) *"How did the agent's persistent handwriting affect your experience?"*

(H_v) *"How would your experience have been changed, when the handwritten score of the agent would be still visible on the paper?"*

We directed similar questions at participants assigned to the virtual and the real conditions as we were interested in the users' perception of individual potentialities of both virtual and blended agents. By this means, we also aimed to investigate whether expectations for virtual-physical interactions and the reality diverge to some extent. To extract comparable data, we assigned utterances to seven different categories using an open coding strategy. The first three categories denote opinions regarding the perceived realism of blended agents in comparison to IVAs without virtual-physical capabilities. Another four categories cover different dimensions of emotional responses. If a single response of a participant included multiple utterances within the same category (e.g., "fascinating and memorable"), they were still counted only once. As each scenario (B_r, B_v, H_r, and H_v) was experienced by 20 participants, 20 is the maximum value in each category. The resulting frequency distribution is illustrated in Table 16.2.

In addition to the categorized utterances, four participants acknowledged the increased fairness when both the user and the blended agent have to play with a physical golf ball. Regarding the scorecard, four of the participants with a (H_v) condition reported their initial surprise when the projected score disappeared. One participant even admitted feeling disappointed as he could not take the completed scorecard as a souvenir. In general, three participants mentioned that a scorecard with physically written text feels more like a trophy.

We were also interested in whether reactions to the persistent handwriting were different for users assigned to a (B_r) or (B_v) condition. Experiencing a blended agent that is interacting with a real golf ball might have raised the expectations regarding the agent's capabilities. However, no such interaction effect could be found as both user groups showed similar, mainly positive, responses to the persistent handwriting.

Finally, we asked participants of the (B_r) or (H_r) conditions about the used mechanism to test our hypothesis (H3). For the physical golf ball, 8 of the participants stated that they figured out the mechanism or at least got an idea of how it worked. Surprisingly, only one of the ideas was correct while most participants suspected a magnetic track behind the ball movement. In contrast, none of the (H_r) participants perceived the mechanism behind persistent handwriting as obvious. Due to a lack of explanations, two participants were convinced that another person entered the room to replace the virtual score, while another two felt uncertain about the fact whether the scorecard was empty at the beginning of the study.

Observational Data

In addition to feedback obtained through the questionnaires, we also made some observations regarding the participants' behavior, both during the study and directly afterward.

When the agent was approaching the scorecard to fill in the blank, six of the participants (five of H_r and one of H_v) used their own pen to write down the agent's score in her stead.

After the MR experience but before completing the post-questionnaires, all participants of the conditions (B_rH_v), (B_vH_r), and (B_rH_r) were asked whether they noticed the physicality of the golf ball and handwriting, respectively. 16 participants who experienced a B_r condition realized that the ball was real during the experiment, while 4 reported having doubts whether the ball was real or not. In contrast, only 7 participants of the H_r conditions noticed that the handwritten score was still persistent after they left the MR environment. These results support our initial hypothesis (H4) regarding the observability of both forms of virtual-physical interactions. It was surprising, though, as we expected that users will realize the persistent handwriting at the latest when they leave the MR environment.

After completion of the entire experiment, 20% of participants with one of the H_v conditions took their scorecard home. In the H_r conditions featuring persistent handwriting, 40% of participants kept their scorecard.

16.2.5. Discussion

Against the in (H1) and (H2) hypothesized positive effects of blended agents on several social and spatial factors, no significant differences could be found in the collected data. These statistical results can be interpreted in different ways. Two possible implications might be that (i) there actually are no differences between IVAs without virtual-physical capabilities and blended agents, and (ii) some users react positively to blended agents while others show negative reactions, compensating one another. In contradiction to both approaches to an explanation are the qualitative comments that were collected at the end of the study. The majority of participants reported a positive influence of both virtual-physical interactions in terms of perceived realism and/or user experience. The question arises why the quantitative ratings do not reflect these subjective impressions. In the following, we discuss several potential influencing factors and rate their respective impact.

Limited Expectations on IVAs Starting point of the following discussion is the basic question “Do users expect IVAs to have physical capabilities?”. The fact that the majority of participants did not notice the persistent handwriting at all and showed emotional responses of surprise and confusion when they were made aware of it is indicative of rather low expectations on the IVA. Therefore, users assigned to a virtual condition most likely compared the displayed IVA to agents they experienced in the past, without a negative impact of missing physical capabilities. This impression is supported by the qualitative feedback as users of the virtual condition felt positive about the realistic body movements, the natural voice, and individual reactions of the IVA. Therefore, the reported effects found in previous within-subjects studies might be only due to the direct comparison between agents.

Low Granularity of Used Scales Three of the questionnaires used 5-point Likert scales as we complied with the standards. As current VAs are still far from being indistinguishable from real humans, only a few users will rate items that are related to the human likeness with a maximum value of 5. Therefore, only one of the remaining options refers to a positive response. A higher scale granularity that allows participants to rate their experience more precisely might reveal significant differences between the conditions. That these differences are expected to be rather small was already indicated by the results of a similar study with a within-subjects design [LNB$^+$18]. For the ratings of engagement, an additional ceiling effect can be observed as mean scores are already close to the maximum for conditions without blended agents.

Limited Importance of the Physical Reactions Although both the golf ball and the scorecard were designed to be an integral part of the interaction between the participants and the MR environment, they might have been less meaningful than other interactions with physical objects. For example, if a blended agent moves a real chair towards the user to take a seat, this physical manipulation has an impact on the subsequent actions while the physical golf ball could only be observed without any direct contact. In another example, a blended agent could mark a location on a physical map to direct the user to a place. In contrast, the scorecard was only given as a souvenir without any future purpose. More meaningful interactions might have increased the perceived value of the physical (persistent) manipulations.

Distrust of the Experimental Environment An observation that might have influenced the results without being the sole reason was that some participants conjectured that somebody entered the room and replaced the virtual score by a physical one when the participant was distracted by the golf match. Even the fact that the experimenter was neither in the CAVE nor the directly neighboring room could convince them of the contrary. Two other participants mentioned that they were sure that the sheet of paper was empty at first but were skeptical about this in retrospect as they would not know how this could have been done.

16.2.6. Limitations

Both the interactions with the robotic ball and the thermal table are proof of concepts with some limitations.

As the robotic ball has no global tracking capabilities, its position and orientation have to be determined manually. A computer vision algorithm could be used to compute the current position of the ball and to match the subsequent actions of the blended agent. Such a tracking algorithm could also compensate for the limited precision of the Sphero mini. In the current implementation, the robotic ball ends up in slightly varying places. While the blended agent always putted the ball into the hole, the golf club and the ball movements were not always perfectly in sync; an observation that was also shared by some of the participants. Another limitation is related to the ball physics. As any motorized objects need an acceleration phase to be set in motion, the initial impulse imparted to the golf ball by hitting it with a golf club cannot be simulated completely.

The current implementation of the thermal table also requires some preparations to create a convincing illusion of a blended agent. First of all, the ink can only be made visible at once. This is why the handwritten text has to be prepared before the MR experience. For the same reason, the writing path has to be simulated virtually before it is replaced by the physical writing. To solve both problems, coated paper could be used. In this case, the thermal mechanism has to be changed from a heating plate to a heated metal tip as used for soldering irons. Furthermore, as the used thermochromic ink is visible at room temperature, it always has to be placed on top of the thermal table at the beginning of a MR experience. By using a dye with different characteristics users could bring a sheet of paper to the MR environment, which might further increase the believability of the blended agent.

16.3. Conclusion

All actions performed by IVAs are usually restricted to computer-generated objects, resulting in an asymmetry between real and virtual interaction partners. In this chapter, we investigate the concept of *blended agents*; IVAs that can cross the boundary between a virtual and the physical world. We implemented two exemplary manipulations that are affecting real-world objects either within a MR environment (i.e., moving a physical golf ball) or even outside of a MR environment (i.e., writing on a physical sheet of paper).

To compare both forms of virtual-physical interactions we conducted a user study with 40 participants. Although a statistical analysis of subjective data obtained through several questionnaires did not yield any significant differences between blended agents and IVAs without physical capabilities, user responses still provide insight into the potential of virtual-physical manipulations. Users described their interaction with blended agents as an “amazing, very surprising and immersive experience”, a “fascinating magic trick”, or the sensation of “being inside the Holodeck”. The agent’s physical manipulations “made the agent appear more present”, and created a “more enjoyable” and “more memorable” MR experience. In spite of little divergences between natural and simulated behavior of the physical golf ball, the majority of participants appreciated the virtual-physical interaction. Some participants also mentioned effects on their behavior inside the MR environment, as they “felt the urge to respond” to the blended agent or avoided any collisions.

To take up these points, we want to collect objective data such as the users’ avoidance behaviors or points of interest in a future study. In terms of subjective responses, we plan to use scales with higher granularity to obtain more differentiated feedback. Furthermore, a future study design should focus on virtual-physical interactions that are more meaningful to the users, for example, a persistent writing that serves as a reminder note. By this means, blended agents could make a real contribution to the tourism center, museum, or office of the future.

Part V.

Conclusions and Future Work

17 Chapter 17. Summary and Guidelines

In this thesis, we introduced the concept of Blended Spaces, which allow users to fully exploit the RV continuum seamlessly within the boundaries of a single environment. We considered different aspects that are relevant to designers and developers of such a space and collected empirical data to gain an understanding of how it distinguishes from conventional VR or AR environments. From the main findings of each conducted study, a set of guidelines can be derived with the objective of enhancing the user experience within Blended Spaces.

Part II laid the foundation for all subsequent user studies we conducted. To adapt the calibration and rendering pipelines to fit the special needs of our setup, we adjusted algorithms known from related research fields such as computer vision, and mobile AR. In addition, we presented different hardware configurations to implement a Blended Space on the basis of projection-based SAR technology. Our extended CAVE setups demonstrated the usage of low-cost projectors for building room-sized immersive environments, that involve up to four walls as well as the floor as projection screens. The corresponding hardware list in the appendix sums up to less than $75,000 for a blended office including six synchronized laser projectors, an 8-camera tracking system, and a powerful workstation. We also found that it is not required to provide special screens for front projection, as matte colored walls can also display projections with sufficient brightness. Nevertheless, the best visual quality was achieved for scenes with high contrast and saturation, as well as a low overall brightness. Developers of Blended Spaces should have in mind, that this effect becomes more apparent with a higher number of projectors as the ambient light intensity within the room also increases.

Part III addressed depth perception within Blended Spaces, in particular with a focus on the spatial consistency of projected 3D objects. In this context, we first presented early approaches to manipulate perceived spatial relationships between the user and real-world objects by introducing perceptual illusions to a projection-based Blended Space. We analyzed the effects of three monoscopic illusions, which are inspired by visual arts, i.e., color temperature, luminance contrast, and blur. The results provided positive indications that perceived depth of objects can be affected by computer-generated projected illusions, even in a complex environment with diverse distance cues. Manipulations of both color temperature and luminance contrast caused significant changes of perceived depth, although the latter turned out to be the more effective illusion in the tested environment.

Although the observed effects were in the range of centimeters, we considered possible consequences on a consistent depth perception of Blended Spaces in a second series of user studies. In the collected data, we observed a trend suggesting that the effect of parallax on depth estimation strongly depends on the user's experience with stereoscopic 3D. Participants, who wear stereo 3D glasses at least once a week, were able to perform the perceptual matching task with a mean absolute error of less than one centimeter, whereas less experienced users made errors in the range of more than 2cm on average. In spite of these individual differences, a confirmatory study revealed a general tendency towards the preference of a geometrically correct projection of the visual stimuli rather than a perceptually adapted alternative. This could obviate the need for a complex perceptual adaptation of the visual stimuli to compensate for spatial inconsistencies. However, it also implies that offsets between physical projection surfaces and stereoscopically projected objects should be reduced to a minimum to facilitate perceptual integration of stimuli, in particular for users who are less experienced in the usage of stereo 3D glasses.

Part IV of the thesis emphasized the potential of Blended Spaces as shared environments for collaboration between multiple users, as well as the user and virtual cooperation partners. The conducted research projects aimed to emphasize the social characteristics of stereoscopic projection-based Blended Spaces and to reduce multi-user limitations that result from the perspective display of view-dependent content. For this purpose, we introduced two operator-follower systems, one with a shared 3D view, and a second with separate 3D and 2D views. Both proposed systems were well received by participants of the user studies, and therefore demonstrated the potential of customized UIs to compensate for perceptual difficulties when view-dependent content is displayed for multiple users. The suggested interaction paradigms fostered communication between interaction partners, even when they did not know each other before the experiment. We also suggested the floor as an efficient extension of the 3D interaction space within Blended Spaces. Guidance cues that are projected to the ground plane can encourage users to move closer together, and therefore ensure better viewpoints for multiple users. Developers of Blended Spaces should also have in mind that virtual UI elements can distract users from the actual scene, and therefore should be reduced to a minimum. As Blended Spaces are not restricted to a single stage within the RV continuum, developers could provide contextual UI elements and disable them when not needed.

Regarding the collaboration with a virtual partner, we focused on the effectiveness of different representations of IVAs, as well as their capability to manipulate not only their VE but also real-world objects within the Blended Space. Based on the analysis of collected empirical data, we derived several suggestions for the design of IVAs in the special context of educational Blended Spaces. Applications with a high priority on user experience such as museums or other educational facilities should consider customization of IVAs in preference to generic agents. A higher effort to model content-specific agents can be justified by better ratings of attractiveness, stimulation, and novelty. A similar result was found regarding embodied IVAs, which outperformed audio guides in terms of user experience as well as the perceived presence of the agent. In general, our user studies indicated that incorporating virtual content into the scene, either in the form of a virtual embodied IVA or virtual objects, has positive effects on the users' engagement, which is an important aspect in educational applications.

With regard to the usage of *blended agents* that can affect real-world objects either within a MR environment (e.g., moving a physical ball), or even outside of a MR environment (e.g., writing on a physical sheet of paper), we did not find any significant differences to IVAs without physical capabilities. Nevertheless, we encourage developers to create new ways of virtual-physical interactions, since the qualitative feedback of study participants both was exceptionally positive and indicates advantages of persistent manipulations with regard to the perceived realism of blended agents and the overall user experience. Finally, according to the collected user responses, little divergences between the natural and simulated behavior of physical objects do not inherently break the illusion of interactions between IVAs and their real-world surroundings. In spite of these limitations, blended agents have the potential to create an "amazing, very surprising and immersive experience" that is "more enjoyable" and "more memorable".

18 Chapter 18. The Future of Blended Spaces

While this thesis draws on the technical capabilities of current AR displays, future technological developments will have a direct impact on the implementation of Blended Spaces. Head-worn devices already evolved considerably in terms of display quality, size, and wearing comfort, and will do so further on. Smart glasses and contact lenses could establish Blended Spaces as ubiquitously embedded systems instead of stationary communal facilities. If personal VR/AR displays gained acceptance as mass media just as smartphones did, a number of the previously discussed aspects, both perception- and interaction-related, would have to be reconsidered. For each user, a private view of the world that is matching the user's interests and current schedule could be generated. Distortions that are caused by different viewers' perspectives would vanish as private view-dependent imagery could be displayed for each user. Furthermore, the varying perception of depth cues such as color differences or luminance contrasts could be compensated by individual modifications of the computer-generated content. On the other hand, even under the assumption of technological improvements of HMDs, projection-based SAR setups can still turn out to be beneficial regarding conflicts between the accommodation and convergence depth cues. Virtual 2D textures can be projected directly onto the surfaces of physical objects, and therefore do not cause a vergence-accommodation conflict at all. This is contrary to HMDs, for which 2D textures have to be displayed with positive parallax to be perceived as spatially connected to the physical objects' surfaces. For stereoscopically displayed objects, the mismatch between accommodation and convergence is smaller for projection-based setups than for HMDs as well, if the virtual objects' distance to physical projection surfaces is kept smaller than to a display worn directly in front of the user's eyes. As sensory conflicts are assumed to cause visually induced motion sickness (VIMS) and associated symptoms such as nausea, disorientation, headaches, and blurred vision [Auk16], a reduction of those conflicts is an argument in favor of projection-based Blended Spaces.

Moreover, the shift from public to private displays could also diminish the social potential of Blended Spaces. In contrast to the social behavior we observed in our studies, users may be less encouraged to collaborate. The reception of *Google Glass* demonstrates that HMDs with the eyeglasses form factor can even raise privacy issues when worn in public spaces, and therefore have to be treated with caution [KV19]. Furthermore, in the course of numerous demonstrations of our VR/AR setups, particularly users without a profound technical understanding appreciated the low barrier to enter the projection-based Blended Space. However, this argument in favor of SAR technology has to be reassessed in a couple of years, when the number of digital immigrants will have been declined and digital natives have been incremented [Pre01, Ste16].

Technological advances not only influence the display hardware for Blended Spaces but may also lead to an altered understanding of the general concept of state transitions along the RV continuum. Current implementations of Blended Spaces are based on pre-existing physical and virtual objects, that are superimposed upon each other to create the impression of an overall MR or VR state. Therefore, an object's transition from real to virtual does not destroy its physical representation but overlays it with matching virtual content. With future developments of computer vision and 3D printing technologies, real-time changes of an object's state could be within the realms of possibility. 3D depth cameras, like those integrated into current smartphone models, already allow users to create 2.5D reconstructions of real-world objects within seconds. A full 3D virtual replica of the physical object can also be generated by registering multiple point clouds from different views. In contrast, the instant creation and dissolution of objects from and to physical matter are still subject to active research. Commercially available 3D printing devices already allow materializing a 3D model that was only existing in a virtual form. While this process usually takes minutes to hours depending on object size, researchers of the Lawrence Livermore National Laboratory developed a printing technique with a special resin that solidifies within seconds when exposed to intense light [KBH$^+$19]. By this means, instead of adding material gradually layer by layer, a 3D structure can be printed at once, promoting the idea of instant transitions from virtual to real objects.

Finally, the advancements in computer science will add to the quality of displayed content, and allow for the creation of photorealistic, vivid, and more life-like experiences. The natural interplay of virtual objects and their real environment was already subject to several research projects, which thereby reveal future prospects of Blended Spaces (e.g., the simulation of realistic illumination effects [RPAC17] or shadows [PML$^+$10] in MR environments). Considering the virtual content separately, the film industry demonstrates algorithms to create computer-generated content that is already close to being indistinguishable from a real video recording. While the rendering of these graphics still takes a lot of resources in terms of time and computational power, it is reasonable to assume that real-time renderings with an equal quality will be achievable with improving computing systems [BJTK12, Ste16]. In particular, IVAs could benefit from advancements in the field of computer graphics, as in many applications their appearance is still referred to be uncanny.

VAs will not only advance in terms of their appearance but also regarding their behavior. Constant improvements in fields such as computer vision and machine learning enable developers of VR/AR applications to factor in the current state of users and their real-world surroundings. Through an interpretation of this state, IVAs and other virtual objects can react more naturally to changes in their environment and are therefore able to create more realistic blended interactions. In future applications, these interactions will most likely be characterized by a multi-sensory integration of stimuli. In the context of this thesis, we solely relied on the visual and auditory senses as well as passive haptic feedback provided by tangible objects. Advances regarding kinesthetic feedback, olfactory, and even gustatory displays could complement these stimuli to create even more convincing VR/AR experiences [SH16].

Appendix

Table A.1.: Comparison of three AR display types with regard to technological factors [BCL15, SH16].

	OST-HMD	**VST-HMD**	**SAR**
Technological Factors			
Ocularity	mono and stereo display	mono and stereo display	mono and (with 3D glasses) stereo display
Vergence-Accommodation Conflict (see Sec. 8.4.2)	occurs for both 2D textures and 3D virtual objects	occurs for both 2D textures and 3D virtual objects	occurs for stereoscopic 3D virtual objects only
Occlusion	virtual objects appear translucent, dark virtual objects cannot occlude brighter real objects	virtual objects can fully occlude real objects and vice versa	virtual objects can interfere with color and texture of physical surface; real shadows occlude virtual objects
Field of View (FOV)	narrow FOV but full view of real environment	wide FOV but limited view of real environment (depends on camera)	natural FOV possible by using multiple projectors
Viewpoint Offset	none	discrepancy between the camera's position and the user's eyes	none
Resolution	limited for virtual objects only	limited for all scene objects	limited for virtual objects only; scalable with multiple projectors
Brightness / Contrast	limited visibility in bright environments; real objects cannot be darkened, and virtual shadows cannot be displayed	no dependency on environmental light; full control of virtual shadows	limited visibility in bright environments; starting from a dark environment virtual shadows can be displayed by illuminating their surroundings
Latency	a lagged generation of virtual elements (e.g., due to tracking system latency) causes a temporal misalignment between real and virtual content	the video can be delayed to decrease temporal misalignments between virtual and real content (at the cost of an increased overall lag)	a lagged generation of virtual elements (e.g., due to tracking system latency) causes a temporal misalignment between real and virtual content

Table A.2.: Comparison of three AR display types with regard to human and economic factors [BCL15, SH16].

	OST-HMD	**VST-HMD**	**SAR**
Human Factors			
Ergonomics	big size & high weight	big size & high weight	no instrumentation for mono, light-weight glasses for stereo
Accessibility	instruction of technical person for individual adjustments and calibration	instruction of technical person for individual adjustments and calibration	no support needed, even for non-technical users
Social Acceptance	still low, especially when worn in the public (due to privacy concerns)	more common for private use; uncommon for use in the public, since eyes and large parts of the face are covered	high, because display is detached from user
Multi-User Support	single-user system by design, but expandable using multiple devices	single-user system by design, but expandable using multiple devices	shared system by design, but view-dependency for perspectively projected 3D content
Economic Factors			
Set-up Fees	high price per unit	medium price per unit, can be reduced using private phones and cardboard viewers	high due to projector costs but scales well with additional users
Maintenance Costs	potentially high, since glasses are prone to damage	potentially high, since glasses are prone to damage	low, since users do not get into contact with expensive components
Hygiene	high number of different users come into contact with the HMD's frame, that has to be cleaned manually	high number of different users come into contact with the HMD's foam pads, that are difficult to sanitize (improvable by using disposable masks)	high number of different users come into contact with the 3D glasses, which, however, can be cleaned automatically such as in cinemas and theme parks
Space Requirement	low, especially for integrated systems	low, especially for integrated systems	high, due to required distance between projectors and projection surfaces

Table A.3.: Hardware and software configurations for three CAVE models.

	L-Shape	**4-sided CAVE**	**Blended Office**
Projector			
Model	2× ProjectionDesign F10 AS3D	3× Optoma EH320UST, 2× Optoma GT1080(e)	4× Optoma ZH500UST, 2× Optoma ZH406ST
Light Source	Lamp	Lamp	Laser
Resolution	1400×1050px (SXGA+)	1920×1080px (Full HD)	1920×1080px (Full HD)
3D			
Stereo Capability	Yes	Yes	Yes
Stereo Glasses	RealD CE5 (DLP Link)	Optoma ZF2300 (RF)	Hi-SHOCK Oxid Diamond (RF)
Update Rate	60 Hz per eye	60 Hz per eye	60 Hz per eye
Tracking			
Hardware	A.R.T ARTTRACK2	A.R.T ARTTRACK2 / OptiTrack Prime 13W	OptiTrack Prime 13W
Software	DTrack2	DTrack2 / OptiTrack Motive	OptiTrack Motive
Number of Cameras	7	7 / 5	8
Workstation			
CPU	Intel Core i7-4930K	Intel Core i7-6900K	Intel Core i9-9900K
Graphics Card	Nvidia Quadro K5000	2× Nvidia GeForce GTX 1080	2× Nvidia Quadro RTX 6000 with Quadro Sync II

Bibliography

[AEI03] L. Anderson, J. Esser, and V. Interrante. A Virtual Environment for Conceptual Design in Architecture. In *Proceedings of the Eurographics Workshop on Virtual Environments (EGVE)*, pages 57–63, 2003.

[AKM+10] T. Augsten, K. Kaefer, R. Meusel, C. Fetzer, D. Kanitz, T. Stoff, T. Becker, C. Holz, and P. Baudisch. Multitoe: High-Precision Interaction with Back-Projected Floors Based on High-Resolution Multi-Touch Input. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*, pages 209–218, 2010.

[Ali05] M.W. Alibali. Gesture in Spatial Cognition: Expressing, Communicating, and Thinking About Spatial Information. *Spatial Cognition and Computation*, 5(4):307–331, 2005.

[AMR06] D. Aliakseyeu, J.-B. Martens, and M. Rauterberg. A Computer Support Tool for the Early Stages of Architectural Design. *Interacting with Computers*, 18(4):528–555, 2006.

[Ara14] The Colours of the Ara Pacis. `http://www.arapacis.it/en/mostre_ed_eventi/eventi/i_colori_dell_ara`, August 2014. Accessed: 2020-02-04.

[AST09] B. Avery, C. Sandor, and B.H. Thomas. Improving Spatial Perception for Augmented Reality X-Ray Vision. In *Proceedings of the IEEE Conference on Virtual Reality (VR)*, pages 79–82, 2009.

[Atl10] D. Atlı. *Effects of Color and Colored Light on Depth Perception.* PhD thesis, Bilkent University, 2010.

[Auk16] S. Aukstakalnis. *Practical Augmented Reality: A Guide to the Technologies, Applications, and Human Factors for AR and VR.* Addison-Wesley Professional, 2016.

[BAB+04] J.N. Bailenson, E. Aharoni, A.C. Beall, R.E. Guadagno, A. Dimov, and J. Blascovich. Comparing Behavioral and Self-Report Measures of Embodied Agents' Social Presence in Immersive Virtual Environments. In *Proceedings of the International Workshop on Presence (PRESENCE)*, pages 1864–1105, 2004.

[BBBL03] J.N. Bailenson, J. Blascovich, A.C. Beall, and J.M. Loomis. Interpersonal

Distance in Immersive Virtual Environments. *Personality and Social Psychology Bulletin*, 29(7):819–833, 2003.

[BCL15] M. Billinghurst, A. Clark, and G. Lee. A Survey of Augmented Reality. *Foundations and Trends® in Human–Computer Interaction*, 8(2-3):73–272, 2015.

[BF86] C.F. Bohren and A.B. Fraser. At What Altitude Does the Horizon Cease to Be Visible? *American Journal of Physics*, 54(3):222–227, 1986.

[BG06] R.J. Bailey and C. Grimm. Perceptually Meaningful Image Editing: Depth. *All Computer Science and Engineering Research*, (Report Number: 2006-11), 2006.

[BGDA07] R. Bailey, C. Grimm, C. Davoli, and R. Abrams. The Effect of Object Color on Depth Ordering. *All Computer Science and Engineering Research*, (Report Number: 2007-18), 2007.

[BGG03] V. Bruce, P.R. Green, and M.A. Georgeson. *Visual Perception: Physiology, Psychology, & Ecology.* Psychology Press, 2003.

[BHH$^+$13] A. Bränzel, C. Holz, D. Hoffmann, D. Schmidt, M. Knaust, P. Lühne, R. Meusel, S. Richter, and P. Baudisch. GravitySpace: Tracking Users and their Poses in Smart Room using a Pressure-sensing Floor. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 725–734, 2013.

[BJTK12] M. Borg, S.S. Johansen, D.L. Thomsen, and M. Kraus. Practical Implementation of a Graphics Turing Test. In *Proceedings of the International Symposium on Visual Computing (ISVC)*, pages 305–313. Springer, 2012.

[BJW12] H. Benko, R. Jota, and A.D. Wilson. MirageTable: Freehand Interaction on a Projected Augmented Reality Tabletop. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 199–208, 2012.

[BK03] A. Butz and A. Krüger. A Generalized Peephole Metaphor for Augmented Reality and Instrumented Environments. In *Proceedings of the International Workshop on Software Technology for Augmented Reality Systems (STARS)*, pages 1–4, 2003.

[BK08] G. Bradski and A. Kaehler. *Learning OpenCV: Computer Vision with the OpenCV Library.* O'Reilly Media, Inc., 2008.

[BKCZ09] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, 1(1):71–81, 2009.

[BKLJP04] D. Bowman, E. Kruijff, J.J. LaViola Jr, and I.P. Poupyrev. *3D User Interfaces: Theory and Practice.* Addison-Wesley, 2004.

[BKP01] M. Billinghurst, H. Kato, and I. Poupyrev. The MagicBook - Moving Seamlessly between Reality and Virtuality. *IEEE Computer Graphics and Applications*, 21(3):6–8, 2001.

[Bla02] J. Blascovich. Social Influence within Immersive Virtual Environments. In *The Social Life of Avatars*, Computer Supported Cooperative Work, pages 127–145. Springer London, 2002.

[BLB^{+}02] J. Blascovich, J. Loomis, A.C. Beall, K.R. Swinth, C.L. Hoyt, and J.N. Bailenson. Immersive Virtual Environment Technology as a Methodological Tool for Social Psychology. *Psychological Inquiry*, 13(2):103–124, April 2002.

[BMA12] D. Benyon, O. Mival, and S. Ayan. Designing Blended Spaces. In *Proceedings of the BCS Conference on Human Computer Interaction (HCI)*, pages 398–403, 2012.

[BMG10] T. Ballendat, N. Marquardt, and S. Greenberg. Proxemic Interaction: Designing for a Proximity and Orientation-Aware Environment. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces (ITS)*, pages 121–130, 2010.

[BMS98] J. Batlle, E. Mouaddib, and J. Salvi. Recent Progress in Coded Structured Light as a Technique to Solve the Correspondence Problem: A Survey. *Pattern Recognition*, 31(7):963–982, 1998.

[BPS04] I. Barakonyi, T. Psik, and D. Schmalstieg. Agents that Talk and Hit Back: Animated Agents in Augmented Reality. In *Proceedings of the IEEE/ACM International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 141–150, 2004.

[BR05] O. Bimber and R. Raskar. *Spatial Augmented Reality: Merging Real and Virtual Worlds.* CRC press, 2005.

[Bro96] J. Brooke. SUS - A Quick and Dirty Usability Scale. *Usability Evaluation in Industry*, 189(194):4–7, 1996.

[BS14] G. Bruder and F. Steinicke. Threefolded Motion Perception During Immersive Walkthroughs. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology (VRST)*, pages 177–185, 2014.

[BSH^{+}05] J.N. Bailenson, K. Swinth, C. Hoyt, S. Persky, A. Dimov, and J. Blascovich. The Independent and Interactive Effects of Embodied-Agent Appearance and Behavior on Self-Report, Cognitive, and Behavioral Markers of Copresence in

Immersive Virtual Environments. *Presence: Teleoperators & Virtual Environments*, 14(4):379–393, 2005.

[BSH09] G. Bruder, F. Steinicke, and K.H. Hinrichs. Arch-Explore: A Natural User Interface for Immersive Architectural Walkthroughs. In *Proceedings of IEEE Symposium on 3D User Interfaces (3DUI)*, pages 75–82, 2009.

[BSOL15] G. Bruder, F.A. Sanz, A.-H. Olivier, and A. Lécuyer. Distance Estimation in Large Immersive Projection Systems, Revisited. In *Proceedings of the IEEE Conference on Virtual Reality (VR)*, pages 27–32, 2015.

[BST14] M. Broecker, R.T. Smith, and B.H. Thomas. Depth Perception in View-Dependent Near-Field Spatial AR. In *Proceedings of the Australasian User Interface Conference (AUIC)*, pages 87–88. Australian Computer Society, Inc., 2014.

[BSVH10] G. Bruder, F. Steinicke, D. Valkov, and K.H. Hinrichs. Immersive Virtual Studio for Architectural Exploration. In *Proceedings of IEEE Symposium on 3D User Interfaces (3DUI)*, pages 125–126, 2010.

[Bur16] W. Burger. Zhang's Camera Calibration Algorithm: In-depth Tutorial and Implementation. Technical Report HGB16-05, 2016.

[BWZ14] H. Benko, A.D. Wilson, and F. Zannier. Dyadic Projected Spatial Augmented Reality. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*, pages 645–655, 2014.

[Cas] CastAR: The Most Versatile AR & VR System. `http://www.kickstarter.com/projects/technicalillusions/castar-the-most-versatile-ar-and-vr-system`. Accessed: 2020-02-04.

[Cas15] M. Casperson. Interactive Terracotta Warrior. `http://www.projection-mapping.org/interactive-terracotta-warrior/`, October 2015. Accessed: 2020-02-04.

[CGRR18] P. Cipresso, I.A.C. Giglioli, M.A. Raya, and G. Riva. The Past, Present, and Future of Virtual and Augmented Reality Research: A Network and Cluster Analysis of the Literature. *Frontiers in Psychology*, 9:2086, 2018.

[CHK$^+$10] S. Chang, S. Ham, S. Kim, D. Suh, and H. Kim. Ubi-Floor: Design and Pilot Implementation of an Interactive Floor System. In *Proceedings of the International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, volume 2, pages 290–293. IEEE, 2010.

[CNSD$^+$92] C. Cruz-Neira, D.J. Sandin, T.A. DeFanti, R.V. Kenyon, and J.C. Hart. The CAVE, Audio Visual Experience Automatic Virtual Environment. *Communications of the ACM*, 35(6):64–72, 1992.

[CNSD93] C. Cruz-Neira, D.J. Sandin, and T.A. DeFanti. Surround-Screen Projection-based Virtual Reality: the Design and Implementation of the CAVE. In *Proceedings of the ACM Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 135–142, 1993.

[Cor62] T.N. Cornsweet. The Staircase-Method in Psychophysics. *The American Journal of Psychology*, 75(3):485–491, 1962.

[CR97] I. Choi and C. Ricci. Foot-mounted Gesture Detection and its Application in Virtual Environments. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, volume 5, pages 4248–4253, 1997.

[Cra02] B. Craddock. In the Museum: The Rock. *Air & Space Magazine*, 16(6), 2002.

[CRWGT05] S.H. Creem-Regehr, P. Willemsen, A.A. Gooch, and W.B. Thompson. The Influences of Restricted Viewing Conditions on Egocentric Perception: Implications for Real and Virtual Environments. *Perception*, 34(2):191–204, 2005.

[CSB$^+$02] J. Cassell, T. Stocky, T. Bickmore, Y. Gao, Y. Nakano, K. Ryokai, D. Tversky, C. Vaucelle, and H. Vilhjálmsson. MACK: Media Lab Autonomous Conversational Kiosk. In *Proceedings of IMAGINA*, volume 2, pages 12–15, 2002.

[CV95] J.E. Cutting and P.M. Vishton. Perceiving Layout and Knowing Distances: The Integration, Relative Potency, and Contextual Use of Different Information about Depth. *W. Epstein & S. Rogers (Eds.)*, Vol 5; Perception of Space and Motion:69–117, 1995.

[Dal] Our Engaging Space at RDE. `http://www.dalziel-pow.com/news/interactive-animations-retail-design-expo`. Accessed: 2020-02-04.

[Die06] J. Diebel. Representing Attitude: Euler Angles, Unit Quaternions, and Rotation Vectors. *Matrix*, 58(15-16):1–35, 2006.

[Die11] Z. Dienes. Bayesian Versus Orthodox Statistics: Which Side Are You On? *Perspectives on Psychological Science*, 6(3):274–290, 2011.

[DM96] D. Drascic and P. Milgram. Perceptual Issues in Augmented Reality. In *Stereoscopic Displays and Virtual Reality Systems III*, volume 2653, pages 123–134. International Society for Optics and Photonics, 1996.

[DN93] M. Dengler and W. Nitschke. Color Stereopsis: A Model for Depth Reversals Based on Border Contrast. *Perception & Psychophysics*, 53(2):150–156, 1993.

[DNP11] V. Demeure, R. Niewiadomski, and C. Pelachaud. How Is Believability of a Virtual Agent Related to Warmth, Competence, Personification, and Embodiment? *Presence: Teleoperators & Virtual Environments*, 20(5):431–448, 2011.

[DRHL+03] L. Davis, J. Rolland, F. Hamza-Lup, Y. Ha, J. Norfleet, and C. Imielinska. Enabling a Continuum of Virtual Environment Experiences. *IEEE Computer Graphics and Applications*, 23(2):10–12, 2003.

[DSD+02] J.M.S. Dias, P. Santos, N. Diniz, L. Monteiro, R. Silvestre, and R. Bastos. Tangible Interaction for Conceptual Architectural Design. In *Proceedings of IEEE International Workshop on Augmented Reality Toolkit (ART)*, pages 1–9, 2002.

[Dud18] S. Dudley. *Materiality Matters: Experiencing the Displayed Object*, pages 418–428. 1 2018.

[Dv00] D. Dehn and S. van Mulken. The Impact of Animated Interface Agents: A Review of Empirical Research. *International Journal of Human-Computer Studies*, 52:1–22, 2000.

[DWB06] H. Durrant-Whyte and T. Bailey. Simultaneous Localization and Mapping: Part I. *IEEE Robotics & Automation Magazine*, 13(2):99–110, 2006.

[Egu83] H. Egusa. Effects of Brightness, Hue, and Saturation on Perceived Depth Between Adjacent Regions in the Visual Field. *Perception*, 12(2):167–175, 1983.

[Eti18] A. Etienne. Ideum's Experimental Pottery Brings Native American Culture to Life. `http://www.projection-mapping.org/ideums-experimental-pottery-brings-native-american-culture-life/`, January 2018. Accessed: 2020-02-04.

[Fau95] J. Faubert. Colour Induced Stereopsis in Images with Achromatic Information and Only One Other Colour. *Vision Research*, 35(22):3161–3167, 1995.

[Fer08] J.A. Ferwerda. Psychophysics 101: How to Run Perception Experiments in Computer Graphics. In *Proceedings of the ACM Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, number 87, 2008.

[FS97] T.R. Fricke and J. Siderov. Stereopsis, Stereotests, and their Relation to Vision Screening and Clinical Practice. *Clinical and Experimental Optometry*, 80(5):165–172, 1997.

[GD04] C.R.C. Guibal and B. Dresp. Interaction of Color and Geometric Cues in Depth Perception: When Does “Red” Mean “Near”? *Psychological Research*, 69(1-2):30–40, 2004.

[Gen11] J. Geng. Structured-Light 3D Surface Imaging: A Tutorial. *Advances in Optics and Photonics*, 3(2):128–160, 2011.

[GIK+07] K. Grønbæk, O.S. Iversen, K.J. Kortbek, K.R. Nielsen, and L. Aagaard. IGameFloor: A Platform for Co-located Collaborative Games. In *Proceedings of the*

International Conference on Advances in Computer Entertainment Technology (ACE), pages 64–71. ACM, 2007.

[Gol13] E.B. Goldstein. *Sensation and Perception.* Cengage Learning, 2013.

[GPS72] G.V. Glass, P.D. Peckham, and J.R. Sanders. Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance. *Review of Educational Research*, 42(3):237–288, 1972.

[GSV$^+$03] M. Garau, M. Slater, V. Vinayagamoorthy, A. Brogni, A. Steed, and M.A. Sasse. The Impact of Avatar Realism and Eye Gaze Control on Perceived Quality of Communication in a Shared Immersive Virtual Environment. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 529–536, 2003.

[GWG$^+$07] J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy. Creating Rapport with Virtual Agents. In *Proceedings of the International Workshop on Intelligent Virtual Agents (IVA)*, pages 125–138. Springer, 2007.

[Har06] S.G. Hart. NASA-Task Load Index (NASA-TLX) 20 Years Later. In *Proceedings of Human Factors and Ergonomics Society Annual Meeting (HFES)*, pages 904–908, 2006.

[HB04] C. Harms and F. Biocca. Internal Consistency and Reliability of the Networked Minds Measure of Social Presence. In *Proceedings of the International Workshop on Presence (PRESENCE)*, pages 246–251, 2004.

[HB10] J. Herling and W. Broll. Advanced Self-Contained Object Removal for Realizing Real-Time Diminished Reality in Unconstrained Environments. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 207–212, 2010.

[HBK03] M. Hassenzahl, M. Burmester, and F. Koller. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Proceedings of the Mensch & Computer (MuC)*, pages 187–196. Springer, 2003.

[HCO$^+$11] T. Holz, A.G. Campbell, G.M.P. O'Hare, J.W. Stafford, A. Martin, and M. Dragone. MiRA - Mixed Reality Agents. *International Journal of Human-Computer Studies*, 69(4):251–268, 2011.

[HDO09] T. Holz, M. Dragone, and G.M.P. O'Hare. Where Robots and Virtual Agents Meet: A Survey of Social Interaction Research Across Milgram's Reality-Virtuality Continuum. *International Journal of Social Robotics*, 1(1):83–93, 2009.

[Hei62] M.L. Heilig. Sensorama Simulator, August 28 1962. US Patent 3,050,870.

[HGAB08] D.M. Hoffman, A.R. Girshick, K. Akeley, and M.S. Banks. Vergence–Accommodation Conflicts Hinder Visual Performance and Cause Visual Fatigue. *Journal of Vision*, 8(3):33–33, 2008.

[HR12] I.P. Howard and B.J. Rogers. *Perceiving in Depth: Other Mechanisms of Depth Perception*. Oxford Psychology Series. Oxford University Press, 2012.

[HRHO92] M.R. Harwell, E.N. Rubinstein, W.S. Hayes, and C.C. Olds. Summarizing Monte Carlo Results in Methodological Research: The One- and Two-Factor Fixed Effects ANOVA Cases. *Journal of Educational Statistics*, 17(4):315–339, 1992.

[HSS17] F. Heinecke, S. Schmidt, and F. Steinicke. Präattentive Wahrnehmung von Farbe bei der Gestaltrichtung Flat Design. In *Proceedings of the Mensch und Computer (MuC)*, pages 391–394, 2017.

[Hua07] K.-C. Huang. Effects of Colored Light, Color of Comparison Stimulus, and Illumination on Error in Perceived Depth with Binocular and Monocular Viewing. *Perceptual and Motor Skills*, 104(3c):1205–1216, 2007.

[HWV$^+$15] T. Hartmann, W. Wirth, P. Vorderer, C. Klimmt, H. Schramm, and S. Böcking. Spatial Presence Theory: State of the Art and Challenges Ahead. In *Immersed in Media*, pages 115–135. Springer, 2015.

[IKA07] S. Ichihara, N. Kitagawa, and H. Akutsu. Contrast and Depth Perception: Effects of Texture Contrast and Area Contrast. *Perception*, 36(5):686–695, 2007.

[Ish08] H. Ishii. The Tangible User Interface and its Evolution. *Communications of the ACM*, 51(6):32–36, 2008.

[Jac13] P.J. Jacobowitz. Hands On: Vuforia ”Smart Terrain”. `http://www.qualcomm.com/news/onq/2013/09/06/hands-vuforia-smart-terrain/`, September 2013. Accessed: 2020-02-04.

[JBOW13] B.R. Jones, H. Benko, E. Ofek, and A.D. Wilson. IllumiRoom: Peripheral Projected Illusions for Interactive Experiences. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 869–878, 2013.

[Jer15] J. Jerald. *The VR Book: Human-Centered Design for Virtual Reality*. Morgan & Claypool, 2015.

[JM09] P. Janssens and K. Malfait. Future Prospects of High-End Laser Projectors. In *Emerging Liquid Crystal Technologies IV*, volume 7232. International Society for Optics and Photonics, 2009.

[Jon] B. Jones. Projection Mapping Central. http://www.projection-mapping.org/. Accessed: 2020-02-04.

[JSM+14] B.R. Jones, R. Sodhi, M. Murdock, R. Mehra, H. Benko, A.D. Wilson, E. Ofek, B. MacIntyre, N. Raghuvanshi, and L. Shapira. RoomAlive: Magical Experiences Enabled by Scalable, Adaptive Projector-Camera Units. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*, pages 637–644, 2014.

[Jul64] B. Julesz. Binocular Depth Perception Without Familiarity Cues. *Science*, 145(3630):356–362, 1964.

[Kau03] H. Kaufmann. Collaborative Augmented Reality in Education. Technical Report 188-2-2003-1, Technische Universität Wien, 2003.

[KBB+18] K. Kim, M. Billinghurst, G. Bruder, H. Been-Lirn Duh, and G.F. Welch. Revisiting Trends in Augmented Reality Research: A Review of the 2nd Decade of ISMAR (2008-2017). *IEEE TVCG*, pages 1–16, 2018.

[KBH+18] K. Kim, L. Boelling, S. Haesler, J. Bailenson, G. Bruder, and G.F. Welch. Does a Digital Assistant Need a Body? The Influence of Visual Embodiment and Social Behavior on the Perception of Intelligent Virtual Agents in AR. In *Proceedings of the IEEE Symposium on Mixed and Augmented Reality (ISMAR)*, pages 105–114, 2018.

[KBH+19] B.E. Kelly, I. Bhattacharya, H. Heidari, M. Shusteff, C.M. Spadaccini, and H.K. Taylor. Volumetric Additive Manufacturing via Tomographic Reconstruction. *Science*, 363(6431):1075–1079, 2019.

[KBW17] K. Kim, G. Bruder, and G.F. Welch. Exploring the Effects of Observed Physicality Conflicts on Real–Virtual Human Interaction in Augmented Reality. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology (VRST)*, pages 1–7, 2017.

[KBW18] K. Kim, G. Bruder, and G.F. Welch. Blowing in the Wind: Increasing Copresence with a Virtual Human via Airflow Influence in Augmented Reality. In *Proceedings of the International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments (ICAT-EGVE)*, pages 183–190, 2018.

[KKB+11] A. Kulik, A. Kunert, S. Beck, R. Reichel, R. Blach, A. Zink, and B. Froehlich. C1x6: A Stereoscopic Six-User Display for Co-located Collaboration in Shared Virtual Environments. 30(6):188, 2011.

[KLLO04] P. Krogh, M. Ludvigsen, and A. Lykke-Olesen. “Help Me Pull That Cursor” A Collaborative Interactive Floor Enhancing Community Interaction. *Australasian Journal of Information Systems*, 11(2), 2004.

[KMB+17] K. Kim, D. Maloney, G. Bruder, J.N. Bailenson, and G.F. Welch. The Effects of Virtual Human's Spatial and Behavioral Coherence with Physical Objects on Social Presence in AR. *Computer Animation and Virtual Worlds*, 28(3–4):e1771, 2017.

[Kre] The Kremer Museum. `http://www.thekremercollection.com/the-kremer-museum/`. Accessed: 2020-02-04.

[KSBG00] B. Koleva, H. Schnädelbach, S. Benford, and C. Greenhalgh. Traversable Interfaces between Real and Virtual Worlds. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 233–240, 2000.

[KSF10] E. Kruijff, J.E. Swan, and S. Feiner. Perceptual Issues in Augmented Reality Revisited. In *Proceedings of the IEEE Symposium on Mixed and Augmented Reality (ISMAR)*, pages 3–12, 2010.

[KV19] O. Kudina and P.-P. Verbeek. Ethics from Within: Google Glass, the Collingridge Dilemma, and the Mediated Value of Privacy. *Science, Technology, & Human Values*, 44(2):291–314, 2019.

[LBHW18] M. Lee, G. Bruder, T. Höllerer, and G.F. Welch. Effects of Unaugmented Periphery and Vibrotactile Feedback on Proxemics with Virtual Humans in AR. *IEEE TVCG*, 24(4):1525–1534, 2018.

[LBW17] M. Lee, G. Bruder, and G.F. Welch. Exploring the Effect of Vibrotactile Feedback Through the Floor on Social Presence in an Immersive Virtual Environment. In *Proceedings of the IEEE Conference on Virtual Reality (VR)*, pages 105–111, 2017.

[LDC+00] M. Lombard, T.B. Ditton, D. Crane, B. Davis, G. Gil-Egui, K. Horvath, J. Rossman, and S. Park. Measuring Presence: A Literature-based Approach to the Development of a Standardized Paper-and-Pencil Instrument. In *Proceedings of the International Workshop on Presence (PRESENCE)*, volume 240, pages 2–4, 2000.

[LDW09] M. Lombard, T.B. Ditton, and L. Weinstein. Measuring Presence: The Temple Presence Inventory. In *Proceedings of the International Workshop on Presence (PRESENCE)*, pages 1–15, 2009.

[Lee01] M.R. Leek. Adaptive Procedures in Psychophysical Research. *Perception & Psychophysics*, 63(8):1279–1292, 2001.

[Lew82] C. Lewis. Using the 'Thinking-Aloud' Method in Cognitive Interface Design. Technical Report RC-9265, IBM, 1982.

[LFKD01] J. Lessiter, J. Freeman, E. Keogh, and J. Davidoff. A Cross-Media Presence Questionnaire: The ITC-Sense of Presence Inventory. *Presence: Teleoperators & Virtual Environments*, 10(3):282–297, 2001.

[LHS08] B. Laugwitz, T. Held, and M. Schrepp. Construction and Evaluation of a User Experience Questionnaire. In *Proceedings of the Symposium of the Austrian HCI and Usability Engineering Group (USAB)*, pages 63–76. Springer, 2008.

[LHW$^+$10] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross. Nonlinear Disparity Mapping for Stereoscopic 3D. *ACM Transactions on Graphics (TOG)*, 29(4):75, 2010.

[Lig19] Design Tools for Projection. `http://www.lightform.com/`, November 2019. Accessed: 2020-02-04.

[LJFKZ01] J.J. LaViola Jr, D.A. Feliz, D.F. Keefe, and R.C. Zeleznik. Hands-Free Multi-Scale Navigation in Virtual Environments. In *Proceedings of the ACM Symposium on Interactive 3D Graphics and Games (I3D)*, pages 9–15, 2001.

[LK03] J.M. Loomis and J.M. Knapp. Visual Perception of Egocentric Distance in Real and Virtual Environments. In *Virtual and Adaptive Environments*, pages 21–46. Erlbaum, 2003.

[LKD$^+$16] M. Lee, K. Kim, S. Daher, A. Raij, R. Schubert, J. Bailenson, and G.F. Welch. The Wobbly Table: Increased Social Presence via Subtle Incidental Movement of a Real-Virtual Table. In *Proceedings of the IEEE Conference on Virtual Reality (VR)*, pages 11–17, 2016.

[LKK96] L.M. Lix, J.C. Keselman, and H.J. Keselman. Consequences of Assumption Violations Revisited: A Quantitative Review of Alternatives to the One-Way Analysis of Variance F Test. *Review of Educational Research*, 66(4):579–619, 1996.

[LLKN16] J. Lee, S. Lee, Y. Kim, and J. Noh. ScreenX: Public Immersive Theatres with Uniform Movie Viewing Experiences. *IEEE Transactions on Visualization and Computer Graphics*, 23(2):1124–1138, 2016.

[LNB$^+$18] M. Lee, N. Norouzi, G. Bruder, P.J. Wisniewski, and G.F. Welch. The Physical-Virtual Table: Exploring the Effects of a Virtual Human's Physical Influence on Social Interaction. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology (VRST)*, page 25, 2018.

[LSIG$^+$03] M.A. Livingston, J.E. Swan II, J.L. Gabbard, T.H. Höllerer, D. Hix, S.J. Julier, Y. Baillot, and D. Brown. Resolving Multiple Occluded Layers in Augmented Reality. In *Proceedings of the IEEE/ACM International Symposium on Mixed and Augmented Reality (ISMAR)*, page 56, 2003.

[LT09] D. Lanman and G. Taubin. Build Your Own 3D Scanner: 3D Photography for Beginners. In *Proceedings of the ACM Conference on Computer Graphics and Interactive Techniques Courses (SIGGRAPH Courses)*, pages 1–87, New York, NY, USA, 2009.

[Man02] S. Mann. Mediated Reality with Implementations for Everyday Life. *Presence Connect*, 1, 2002.

[Mat97] G. Mather. The Use of Image Blur as a Depth Cue. *Perception*, 26(9):1147–1158, 1997.

[MBSG09] F. Martino, R. Baù, A. Spagnolli, and L. Gamberini. Presence in the Age of Social Networks: Augmenting Mediated Environments with Feedback on Group Activity. *Virtual Reality*, 13(3):183–194, 2009.

[McD12] K. McDonald. YCAMInterlab - ProCamToolkit - mapamok. `http://www.github.com/YCAMInterlab/ProCamToolkit/wiki/mapamok-(English)/`, February 2012. Accessed: 2020-02-04.

[MG96] T. Mazuryk and M. Gervautz. Virtual Reality - History, Applications, Technology and Future. 1996.

[MHTW00] A. Majumder, Z. He, H. Towles, and G. Welch. Achieving Color Uniformity Across Multi-Projector Displays. In *Proceedings of the IEEE Visualization Conference (VIS)*, pages 117–124, 2000.

[Mik06] T.A. Mikropoulos. Presence: a Unique Characteristic in Educational Virtual Environments. *Virtual Reality*, 10(3-4):197–206, 2006.

[MNW14] R. McGloin, K.L. Nowak, and J. Watt. Avatars and Expectations: Influencing Perceptions of Trustworthiness in an Online Consumer Setting. *PsychNology Journal*, 12, 2014.

[MS15] S. Marschner and P. Shirley. *Fundamentals of Computer Graphics*. CRC Press, 2015.

[MSWT14] M.R. Marner, R.T. Smith, J.A. Walsh, and B.H. Thomas. Spatial User Interfaces for Large-scale Projector-based Augmented Reality. *IEEE Computer Graphics and Applications*, 34(6):74–82, 2014.

[MTPC08] N. Magnenat-Thalmann, G. Papagiannakis, and P. Chaudhuri. Applications of Interactive Virtual Humans in Mobile Augmented Reality. In *Encyclopedia of Multimedia*, pages 362–368. Springer, 2nd edition, 2008.

[MTUK95] P. Milgram, H. Takemura, A. Utsumi, and F. Kishino. Augmented Reality: A Class of Displays on the Reality-Virtuality Continuum. In *Telemanipulator and Telepresence Technologies*, volume 2351, pages 282–292. International Society for Optics and Photonics, 1995.

[MvBG+12] M.R. Mine, J. van Baar, A. Grundhofer, D. Rose, and B. Yang. Projection-based Augmented Reality in Disney Theme Parks. *Computer*, 45(7):32–40, 2012.

[NB03] K. Nowak and F. Biocca. The Effect of the Agency and Anthropomorphism on Users' Sense of Telepresence, Copresence, and Social Presence in Virtual Environments. *Presence: Teleoperators & Virtual Environments*, 12(5):481–494, 2003.

[NBB+19] N. Norouzi, G. Bruder, B. Belna, S. Mutter, D. Turgut, and G.F. Welch. A Systematic Review of the Convergence of Augmented Reality, Intelligent Virtual Agents, and the Internet of Things. In *Artificial Intelligence in IoT*, pages 1–24. Springer, 2019.

[NCSS18] D. Neves Coelho, S. Schmidt, and F. Steinicke. Kategorisierung und Evaluierung von Transitionen für CAVE Umgebungen. In *Proceedings of the GI Workshop on Virtual and Augmented Reality (GI VR/AR)*, pages 169–176, 2018.

[NKH+18] N. Norouzi, K. Kim, J. Hochreiter, M. Lee, S. Daher, G. Bruder, and G.F. Welch. A Systematic Survey of 15 Years of User Studies Published in the Intelligent Virtual Agents Conference. In *Proceedings of the International Conference on Intelligent Virtual Agents (IVA)*, 2018.

[OBO94] R.P. O'Shea, S.G. Blackburn, and H. Ono. Contrast as a Depth Cue. *Vision Research*, 34(12):1595–1604, 1994.

[ODK+12] M. Obaid, I. Damian, F. Kistler, B. Endrass, J. Wagner, and E. André. Cultural Behaviors of Virtual Agents in an Augmented Reality Environment. In *Proceedings of ACM International Conference on Intelligent Virtual Agents (IVA)*, volume 7502, pages 412–418. Springer, 2012.

[ONP11] M. Obaid, R. Niewiadomski, and C. Pelachaud. Perception of Spatial Relations and of Coexistence with Virtual Agents. In *International Conference on Intelligent Virtual Agents*, volume 6895 of *Lecture Notes in Computer Science*, pages 363–369. Springer Berlin Heidelberg, 2011.

[Ope] Camera Calibration and 3D Reconstruction. `http://docs.opencv.org/2.4/modules/calib3d/doc/camera_calibration_and_3d_reconstruction.html/`. Accessed: 2020-02-04.

[Opta] OptiTrack General FAQs. `http://www.optitrack.com/support/faq/general.html/`. Accessed: 2020-02-04.

[Optb] Rigid Body Tracking. `http://v21.wiki.optitrack.com/index.php?title=Rigid_Body_Tracking/`. Accessed: 2020-02-04.

[Ott15] Longnecker M. Ott, R. *An Introduction to Statistical Methods and Data Analysis.* Cengage Learning, 7 edition, 2015.

[Pal99] S.E. Palmer. *Vision Science: Photons to Phenomenology.* A Bradford Book, 1999.

[Pan] Mapping Pantocrator Sant Climent de Taüll. `http://www.pantocrator.cat/`. Accessed: 2020-02-04.

[PML$^+$10] S. Pessoa, G. Moura, J. Lima, V. Teichrieb, and J. Kelner. Photorealistic Rendering for Augmented Reality: A Global Illumination and BRDF Solution. In *Proceedings of the IEEE Conference on Virtual Reality (VR)*, pages 3–10, 2010.

[Pre01] M. Prensky. Digital Natives, Digital Immigrants. *On the Horizon*, 9(5), 2001.

[PS02] M. Paranandi and T. Sarawgi. Virtual Reality in Architecture: Enabling Possibilities. In *Proceedings of the International Conference on Computer Aided Architectural Design Research in Asia (CAADRIA)*, pages 309–316, 2002.

[Put] PuttView - Upgrade Your Game. `http://www.puttview.com/`. Accessed: 2020-02-04.

[Raf95] A.E. Raftery. Bayesian Model Selection in Social Research. *Sociological Methodology*, pages 111–163, 1995.

[RC11] R.B. Rusu and S. Cousins. 3D is here: Point Cloud Library (PCL). In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–4, 2011.

[RCWS98] R. Raskar, M. Cutts, G.F. Welch, and W. Stuerzlinger. Efficient Image Generation for Multiprojector and Multisurface Displays. In *Proceedings of the Eurographics Workshop*, pages 139–144, 1998.

[RDI03] G. Riva, F. Davide, and W.A. IJsselsteijn. Measuring Presence: Subjective, Behavioral and Physiological Methods. *Being There: Concepts, Effects and Measurement of User Presence in Synthetic Environments*, pages 110–118, 2003.

[RFK$^+$98] M. Rauterberg, M. Fjeld, H. Krueger, M. Bichsel, U. Leonhardt, and M. Meier. BUILD-IT: A Planning Tool for Construction and Design. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 177–178, 1998.

[RG79] B. Rogers and M. Graham. Motion Parallax as an Independent Cue for Depth Perception. *Perception*, 8(2):125–134, 1979.

[RH17] J.S. Roo and M. Hachet. Towards a Hybrid Space Combining Spatial Augmented Reality and Virtual Reality. In *Proceedings of IEEE Symposium on 3D User Interfaces (3DUI)*, pages 195–198, 2017.

[RMSP12] J.N. Rouder, R.D. Morey, P.L. Speckman, and J.M. Province. Default Bayes Factors for ANOVA Designs. *Journal of Mathematical Psychology*, 56(5):356–374, 2012.

[ROC97] R.A. Rensink, J.K. O'Regan, and J.J. Clark. To See or Not to See: The Need for Attention to Perceive Changes in Scenes. *Psychological Science*, 8(5):368–373, 1997.

[Rou01] M. Roussou. Immersive Interactive Virtual Reality in the Museum. *Proceedings of Trends in Leisure Entertainment (TiLE)*, 2001.

[Row96] S. Roweis. Levenberg-Marquardt Optimization. *Notes, University Of Toronto*, 1996.

[RPAC17] T. Rhee, L. Petikam, B. Allen, and A. Chalmers. MR360: Mixed Reality Rendering for 360 Panoramic Videos. *IEEE Transactions on Visualization and Computer Graphics*, 23(4):1379–1388, 2017.

[RPAG95] J.J. Rieser, H.L. Pick, D.H. Ashmead, and A.E. Garing. Calibration of Human Locomotion and Models of Perceptual-Motor Organization. *Journal of Experimental Psychology Human Perception and Performance*, 21:480–497, 1995.

[RRL$^+$14] B. Ridel, P. Reuter, J. Laviole, N. Mellado, N. Couture, and X. Granier. The Revealing Flashlight: Interactive Spatial Augmented Reality for Detail Exploration of Cultural Heritage Artifacts. *Journal on Computing and Cultural Heritage (JOCCH)*, 7(2):6, 2014.

[RS99] J. Rekimoto and M. Saitoh. Augmented Surfaces: A Spatially Continuous Work Space for Hybrid Computing Environments. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 378–385, 1999.

[RSG17] B. Redwood, F. Schöffer, and B. Garret. *The 3D Printing Handbook: Technologies, Design and Applications*. 3D Hubs, 2017.

[RSS$^+$09] J.N. Rouder, P.L. Speckman, D. Sun, R.D. Morey, and G. Iverson. Bayesian t Tests for Accepting and Rejecting the Null Hypothesis. *Psychonomic Bulletin & Review*, 16(2):225–237, 2009.

[RVH13] R.S. Renner, B.M. Velichkovsky, and J.R. Helmert. The Perception of Egocentric Distances in Virtual Environments - a Review. *ACM Computing Surveys (CSUR)*, 46(2):23, 2013.

[RWC$^+$98] R. Raskar, G.F. Welch, M. Cutts, A. Lake, L. Stesin, and H. Fuchs. The Office of the Future: A Unified Approach to Image-Based Modeling and Spatially Immersive Displays. In *Proceedings of the ACM Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 179–188, 1998.

[RWF98] R. Raskar, G. Welch, and H. Fuchs. Spatially Augmented Reality. In *Proceedings of IEEE Workshop on Augmented Reality (IWAR)*, pages 11–20, 1998.

[RWLB01] R. Raskar, G. Welch, K.-L. Low, and D. Bandyopadhyay. Shader Lamps: Animating Real Objects with Image-Based Illumination. In *Proceedings of the Eurographics Workshop on Rendering Techniques*, pages 89–102, 2001.

[San] A. Sanjeev. Lumen: Reimagining Immersion. `http://www.arvindsanjeev.com/lumen.html`. Accessed: 2020-02-04.

[SAS19] S. Schmidt, O. Ariza, and F. Steinicke. Blended Agents: Manipulation of Physical Objects within Mixed Reality Environments and Beyond. In *Proceedings of the ACM Symposium on Spatial User Interaction (SUI)*, pages 1–10, 2019.

[Sau11] J. Sauro. *A Practical Guide to the System Usability Scale: Background, Benchmarks & Best Practices.* Measuring Usability LLC Denver, CO, 2011.

[SB19] D. Smith and B. Burke. Hype Cycle for Emerging Technologies, 2019. `http://www.gartner.com/en/documents/3956015/hype-cycle-for-emerging-technologies-2019/`, 2019. Accessed: 2020-02-04.

[SBH$^+$09] F. Steinicke, G. Bruder, K. Hinrichs, A. Steed, and A.L. Gerlach. Does a Gradual Transition to the Virtual World Increase Presence? In *Proceedings of the IEEE Conference on Virtual Reality (VR)*, pages 203–210, 2009.

[SBS15] S. Schmidt, G. Bruder, and F. Steinicke. A Layer-based 3D Virtual Environment for Architectural Collaboration. In *Proceedings of the EuroVR International Conference*, pages 79–84, 2015.

[SBS16] S. Schmidt, G. Bruder, and F. Steinicke. Illusion of Depth in Spatial Augmented Reality. In *Proceedings of the IEEE VR Workshop on Perceptual and Cognitive Issues in AR (PERCAR)*, pages 1–6, 2016.

[SBS17a] S. Schmidt, G. Bruder, and F. Steinicke. A Pilot Study of Altering Depth Perception with Projection-Based Illusions. In *Proceedings of the International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments (ICAT-EGVE) (Poster)*, pages 33–34, 2017.

[SBS17b] S. Schmidt, G. Bruder, and F. Steinicke. Moving Towards Consistent Depth Perception in Stereoscopic Projection-Based Augmented Reality. In *Proceedings* *of the International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments (ICAT-EGVE)*, pages 161–168, 2017.

[SBS18] S. Schmidt, G Bruder, and F. Steinicke. Effects of Embodiment on Generic and Content-Specific Intelligent Virtual Agents as Exhibition Guides. In *Proceedings of the International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments (ICAT-EGVE)*, pages 13–20, 2018.

[SBS19] S. Schmidt, G. Bruder, and F. Steinicke. Effects of Virtual Agent and Object Representation on Experiencing Exhibited Artifacts. *Elsevier Computers & Graphics*, 83:1–10, 2019.

[SBS20] S. Schmidt, G. Bruder, and F. Steinicke. Depth Perception and Manipulation in Projection-Based Spatial Augmented Reality. *Presence: Teleoperators & Virtual Environments*, 27(2):242–256, 2020.

[Sch01] D. Schmalstieg. *Collaborative Augmented Reality.* PhD thesis, Technische Universität Wien, 2001.

[Sch16] S. Schmidt. Interaction Techniques for Spatial Augmented Reality Setups. In *Proceedings of the IEEE Conference on Virtual Reality (VR Doctoral Consortium)*, 2016.

[SDBS15] S. Schmidt, S. Dähn, G. Bruder, and F. Steinicke. A Mobile Interactive Mapping Application for Spatial Augmented Reality On The Fly. In *Proceedings of the GI Workshop on Virtual and Augmented Reality (GI VR/AR)*, pages 1–9, 2015.

[SH16] D. Schmalstieg and T. Höllerer. *Augmented Reality: Principles and Practice.* Addison-Wesley Professional, 2016.

[SITS18] S. Schmidt, A. Irlitti, B. Thomas, and F. Steinicke. Floor-Projected Guidance Cues for Collaborative Exploration of Spatial Augmented Reality Setups. In *Proceedings of the ACM International Conference on Interactive Surfaces and Spaces (ISS)*, pages 279–289, 2018.

[SJK$^+$07] J.E. Swan II, A. Jones, E. Kolstad, M.A. Livingston, and H.S. Smallman. Egocentric Depth Judgments in Optical, See-Through Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 13(3):429–442, 2007.

[Sla09] M. Slater. Place Illusion and Plausibility can Lead to Realistic Behaviour in Immersive Virtual Environments. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 364(1535):3549–3557, 2009.

[SN18] H.J. Smith and M. Neff. Communication Behavior in Embodied Virtual Reality. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, page 289, 2018.

[SRP$^+$14] D. Schmidt, R. Ramakers, E.W. Pedersen, J. Jasper, S. Köhler, A. Pohl, H. Rantzsch, A. Rau, P. Schmidt, C. Sterz, et al. Kickables: Tangibles for Feet. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 3143–3152, 2014.

[SS17] S. Schmidt and F. Steinicke. A Projection-Based Augmented Reality Setup for Blended Museum Experiences. In *Proceedings of the International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments (ICAT-EGVE) (Poster)*, pages 5–6, 2017.

[SSUS00] M. Slater, A. Sadagic, M. Usoh, and R. Schroeder. Small-Group Behavior in a Virtual and Real Environment: A Comparative Study. *Presence: Teleoperators & Virtual Environments*, 9(1):37–51, 2000.

[Ste16] F. Steinicke. *Being Really Virtual.* Springer, 2016.

[Str06] I. Stroud. *Boundary Representation Modelling Techniques.* Springer Science & Business Media, 2006.

[Sut65] I.E. Sutherland. The Ultimate Display. In *Proceedings of the IFIP Congress*, pages 506–508, 1965.

[Sut68] I.E. Sutherland. A Head-Mounted Three Dimensional Display. In *Proceedings of the AFIPS Fall Joint Computer Conference, Part I*, pages 757–764, 1968.

[TGCH02] G. Thomas, J.H. Goldberg, D.J. Cannon, and S.L. Hillis. Surface Textures Improve the Robustness of Stereoscopic Depth Cues. *Human Factors*, 44(1):157–170, 2002.

[TI12] N.-C. Tai and M. Inanici. Luminance Contrast as Depth Cue: Investigation and Design Applications. *Computer-Aided Design and Applications*, 9(5):691–705, 2012.

[Tis11] R. Tisdale. Do History Museums Still Need Objects? *History News*, 66(3):19–24, 2011.

[TMS93] P. Thompson, K. May, and R. Stone. Chromostereopsis: A Multicomponent Depth Effect? *Displays*, 14(4):227–234, 1993.

[TYK06] T. Tsuda, H. Yamamoto, and Y. Kameda. Visualization Methods for Outdoor See-Through Vision. *IEICE Transactions on Information and Systems*, 89(6):1781–1789, 2006.

[Uni] Unity. `http://www.unity.com/`. Accessed: 2020-02-04.

[Unr] Unreal Engine: The Most Powerful Real-Time 3D Creation Platform. http://www.unrealengine.com/. Accessed: 2020-02-04.

[UUI99] J. Underkoffler, B. Ullmer, and H. Ishii. Emancipated Pixels: Real-World Graphics in the Luminous Room. In *Proceedings of the ACM Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 385–392, 1999.

[vdSWR$^+$17] R. van de Schoot, S.D. Winter, O. Ryan, M. Zondervan-Zwijnenburg, and S. Depaoli. A Systematic Review of Bayesian Articles in Psychology: The Last 25 Years. *Psychological Methods*, 22(2):217–239, 2017.

[VF17] D. Valkov and S. Flagge. Smooth Immersion: The Benefits of Making the Transition to Virtual Environments a Continuous Process. In *Proceedings of the ACM Symposium on Spatial User Interaction (SUI)*, pages 12–19, 2017.

[VKR05] A.R. Várkonyi-Kóczy and A. Rövid. Point Correspondence Matching for 3D Reconstruction Using Fuzzy Reasoning. In *Proceedings of the IEEE Conference on Computational Cybernetics (ICCC)*, pages 87–92, 2005.

[VKT$^+$01] V. Vlahakis, J. Karigiannis, M. Tsotros, M. Gounaris, L. Almeida, D. Stricker, T. Gleue, I.T. Christou, R. Carlucci, and N. Ioannidis. Archeoguide: First Results of an Augmented Reality, Mobile Computing System in Cultural Heritage Sites. In *Proceedings of the Conference on Virtual Reality, Archeology, and Cultural Heritage (VAST)*, pages 131–140, 2001.

[Vos08] J.J. Vos. Depth in Colour, a History of a Chapter in Physiologie Optique Amusante. *Clinical and Experimental Optometry*, 91(2):139–147, 2008.

[VSC13] Y. Visell, S. Smith, and J.R. Cooperstock. Interacting with Augmented Floor Surfaces. In *Human Walking in Virtual Environments: Perception, Technology, and Applications*, pages 377–399. Springer, New York, NY, 2013.

[VSL$^+$10] Y. Visell, S. Smith, A. Law, R. Rajalingham, and J.R. Cooperstock. Contact Sensing and Interaction Techniques for a Distributed, Multimodal Floor Display. In *Proceedings of IEEE Symposium on 3D User Interfaces (3DUI)*, pages 75–78, 2010.

[Wal] WALKABOUT Projection #WALKPro3D by PRICKIMAGE. http://www.walkaboutprojection.co.uk/. Accessed: 2020-02-04.

[WB15] M. Walker and B. Burton. Hype Cycle for Emerging Technologies, 2015. http://www.gartner.com/en/documents/3100227/hype-cycle-for-emerging-technologies-2015/, 2015. Accessed: 2020-02-04.

[WD13] X. Wang and P.S. Dunston. Tangible Mixed Reality for Remote Design Review: A Study Understanding User Perception and Acceptance. *Visualization in Engineering*, 1(1):1–15, 2013.

[Wel93] P. Wellner. Interacting with Paper on the DigitalDesk. *Communications of the ACM*, 36(7):87–96, 1993.

[WH05] J. Wither and T. Höllerer. Pictorial Depth Cues for Outdoor Augmented Reality. In *Proceedings of the IEEE International Symposium on Wearable Computers (ISWC)*, pages 92–99, 2005.

[Whi14] Whirlpool Interactive Cooktop at CES 2014. `http://www.youtube.com/watch?v=6frHH5OtXU4`, 2014. Accessed: 2020-02-04.

[Wil15] S. Wilkening. Beginning to Measure Meaning in Museum Experiences. *ASTC Dimensions*, 17(3), 2015.

[WLIB05] M. Whitton, B. Lok, B. Insko, and F. Brooks. Integrating Real and Virtual Objects in Virtual Environments. In *Proceedings of the International Conference on Human-Computer Interaction (HCI)*, 2005.

[Xpe] Xperia Touch Official Website. `http://www.sonymobile.com/global-en/products/smart-products/xperia-touch/`. Accessed: 2020-02-04.

[Yam98] T. Yamada. Development of Complete Immersive Display: COSMOS. *Proceedings of the Conference on Virtual Systems and MultiMedia (VSMM)*, pages 522–527, 1998.

[YBR07] N. Yee, J.N. Bailenson, and K. Rickertsen. A Meta-Analysis of the Impact of the Inclusion and Realism of Human-Like Faces on User Experiences in Interfaces. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 1–10, 2007.

[YZD+15] L. Yang, L. Zhang, H. Dong, A. Alelaiwi, and A. El Saddik. Evaluating and Improving the Depth Accuracy of Kinect for Windows v2. *IEEE Sensors Journal*, 15(8):4275–4285, 2015.

[Zha00] Z. Zhang. A Flexible New Technique for Camera Calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.

Eidesstattliche Versicherung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Ort, Datum	Unterschrift