Chapter

Advances in Spatially Faithful (3D) Telepresence

Seppo Valli, Mika Hakkarainen and Pekka Siltanen

Abstract

Benefits of AR technologies have been well proven in collaborative industrial applications, for example in remote maintenance and consultancy. Benefits may also be high in telepresence applications, where virtual and mixed reality (nowadays often referred as extended reality, XR) technologies are used for sharing information or objects over network. Since the 90's, technical enablers for advanced telepresence solutions have developed considerably. At the same time, the importance of remote technologies has grown immensely due to general disruption of work, demands for reducing travelling and CO2, and the need for preventing pandemics. An advanced 3D telepresence solution benefits from using XR technologies. Particularly interesting are solutions based on HMD or glasses type of near-eye-displays (NED). However, as AR/VR glasses supporting natural occlusions and accommodation are still missing from the market, a good alternative is to use screen displays in new ways, better supporting e.g. virtual meeting geometries and other important cues for 3D perception. In this article, researchers Seppo Valli, Mika Hakkarainen, and Pekka Siltanen from VTT Technical Research Centre of Finland describe the status, challenges, and opportunities in both glasses and screen based 3D telepresence. The writers also specify an affordable screen based solution with improved immersiveness, naturalness, and efficiency, enhanced by applying XR technologies.

Keywords: 3D telepresence, spatial faithfulness, remote interaction, XR technologies, AR/VR glasses

1. Introduction

This article compiles lessons learned by the writers from more than a decade of telepresence related research. The article is a review by nature, but due to the number of the described topics, the presentation is not tutorial in all parts, but relies on prior knowledge by its readers, and/or their interest to learn more e.g. from the given references. Based on the reviewed status and enablers, the writers reason and define a practical and affordable 3D telepresence solution based on screen displays. In this solution, efficient 3D capture and low bitrate streaming is an important enabler both for communication and XR functionalities. Essentially the same technical solutions can also be used for remote support applications in industry, e.g. for 3D monitoring, maintenance, control, analysis, and augmentation.

The outline of the paper is as follows. In Chapter 2, we introduce the 3D telepresence topic, describe main factors of spatial faithfulness, and give few examples of existing approaches. Several of the references are to our own patent publications,

which have not been published as papers. Chapter focus is in the requirements and challenges of supporting 3D geometries and perception, as perceived in real-world encounters (face-to-face).

In the future, glasses type of displays will likely be the best to support immersion, mobility and freedom of viewpoint. However, still today, glasses are still lacking many important properties, and have many defects limiting perceived quality, time of use, and user acceptance. At the same time, using screen displays is the most common way of supporting visual interaction in teleconferencing solutions. Correspondingly, we wanted to find out whether a simple screen-based telepresence solution could support 3D perception and XR functionalities with improved naturalness, quality, and usability. The answer seems to be positive, and in Chapter 3, we give a draft specification for such a system.

An important enabler both for communication and AR functionalities is efficient 3D capture and streaming. Further, in Chapter 3, an implementation applying existing coding methods is described together with some simulation results. Despite our demarcation to screen based solutions, we also discuss the possibilities and future of glasses based approaches. In Chapter 4, we describe ways of enhancing 3D perception and XR functionalities of the basic solution. Future improvements may include also supporting natural eye-focus by accommodative displays. Finally, Chapter 5 summarizes our findings.

2. Basics and examples of spatially faithful telepresence solutions

2.1 Spatial faithfulness supports naturalness in perception

3D telepresence solutions aim to support natural perception in 3D – sc. spatial faithfulness – better than video conferencing systems [1]. Basic problem in video-conferencing systems is the lack of support for eye contact [2]. In flat screen based solutions, it stems for example from the displacement or offset between a display (showing a counterparty's face) and camera (counterparty's eyes) of a videoconferencing terminal. Note, that although eye contact is one of the early goals for 3D telepresence solutions, it is still not supported in most of the existing telepresence systems.

Hydra system (**Figure 1**) is an early approach for supporting spatial faithfulness in telepresence [3, 4]. With a mesh of connections and a separate (proxy) terminal for each remote counterpart, it aims to support virtual lines-of-sight between participants. With small displays and small camera-display offset(s), participants are

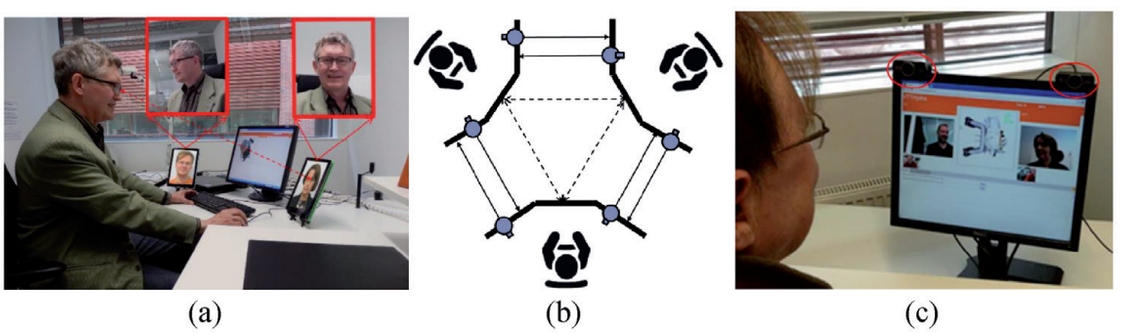

Figure 1.
a) VTT idea for a Hydra system using tablets for a communication space and a computer display for a collaboration space, b) corresponding connections between cameras and displays, c) practical implementation showing all contents on one display (cameras on top corners are indicated by red circles) [3].

able to get an approximate eye contact with each remote participant, and in certain conditions, even have a shared understanding of each other's relative positions.

Note that perceiving eye contact does not mean full gaze awareness, i.e. understanding of also other, intermediate eye directions. In a traditional Hydra system, terminals for each remote party can be placed independently by each local participant, which easily results with inconsistent meeting geometries across meeting sites, and thus also inconsistent positions and eye directions of the parties.

Note that in principle, any position of displays can support virtual lines-of-sight between participants. The situation can be compared to private house residents seeing their neighbors through windows, which, however, is the more unlikely the more of neighbors like to communicate with each other.

The easiest solution unifying a virtual meeting geometry between participants is to position proxy terminals in the same relative order into vertices of an equilateral polygon (e.g. a triangle, square, pentagon, hexagon, etc.). This naturally restricts the seating of participants more than in a face-to-face meeting.

A more recent screen based solution is Viewport by Zhang et.al [5]. In the Viewport system, high-quality 3D models are formed for each user in real time, and extracted and embedded into a common virtual geometry. Using 3D models enables correcting camera-display offset, and supporting depth perception by stereoscopy. The system supports eye contact between three sites, with one user at each site. In particular, limiting the number of sites into only few is a factor limiting the usability of corresponding solutions.

Natural perception of depth and distances belongs to the factors of spatial faithfulness. Note that this is not possible using 2D displays lacking depth, and strictly speaking not even with stereoscopic 3D (S3D) displays, whether multiplexed, polarized, or autostereoscopic, due to their incapability to support natural focus/accommodation (suffering from the sc. vergence-accommodation conflict, VAC [6]).

2.1.1 Advances by XR technologies and computer games

Note that spatial faithfulness is an inherent requirement in XR visualization, which aims at replacing, in a seamless way, parts of a physical view with virtual elements (or vice versa). XR visualization can as well be used for rendering models (cf. avatars) or visual reconstructions of human participants into a participant's view. Correspondingly, for more than two decades, developing enablers for XR has also advanced 3D telepresence solutions. These enablers include e.g. sensors for 3D capture, coding and streaming methods, low latency networks, tracking and detection for XR, camera and user positioning, motion capture and tracking methods, new display technologies, and general advances in algorithms and processing power.

In the same way, developing game technologies has advanced 3D telepresence solutions based on virtual modeling and rendering, here denoted as Virtual World (VW) approaches. Traditionally, in VW approaches, visual content has been modeled/produced in advance, and rendering of the content is based on real-time transfer of parameters for viewpoint and object positions, dynamic 3D shapes and poses (animation), etc. Although VW approaches are rather common and have their specific benefits, their description is omitted in our presentation, focusing on rendering of photorealistic real-time captures. This focus is reasoned in more detail in Chapter 2.5.

A recent example of a photorealistic telepresence solution based on advanced 3D displays is Google Project Starline (https://blog.google/technology/research/project-starline/). A good example of a 3D telepresence system based on AR/VR (XR) visualization is MS Holoportation [7] (cf. https://www.youtube.com/watch?v=7d59O6cfaM0). Both of them are quite impressive but obviously also

complicated and costly. In this article, we aim to define a more economical solution with good sides of both photorealism and XR visualization. Note that even a lone talking head on a screen - whether camera captured or 3D modeled - may well be a value-adding functionality. Communication and attractiveness may namely be supported by using e.g. a speech-controlled, look alike or anonymous virtual head (cf. https://remoteface.ai/ and a video at: https://www.youtube.com/watch?v=prpPqwV5Weo).

2.2 Remarks on mobility and serving with viewpoints

In above, Hydra system was described as an early attempt towards spatially faithful 3D telepresence. Using such full-mesh approach, and by making restricting assumptions on participant positions ("seating order"), all participants may perceive eye directions and participant positions consistently, however, apart from solving the disturbing camera-display offset. Furthermore, a regular setup with fixed camera and display positions naturally limits the mobility of participants within their meeting sites.

Note, that although a participant position is fixed, a whole meeting room with its occupant may move virtually. In [8], a solution is described for compiling captures of regular sensor and display setups into a landscape, enabling participants to mingle together with their meeting spaces within each other, like people in a cocktail party (**Figure 2**). For example, a capture setup in a square or hexagonal formation can be used. Writers of this paper are however not aware if someone has implemented and tested such arrangement.

Let us consider a Hydra setup a little bit further. If a Hydra system with all its terminals was in one large hall or open space, a participant is able to switch (walk) between different sites and perceive spatial faithfulness in each of them separately, i.e. participant mobility is supported in discrete locations.

User mobility may be supported in principle at any viewpoint when using near-eye glasses (NEDs) for viewing. Limitations set by fixed camera positions can be relieved using a setup of multiple 3D sensors, or multiple cameras in an array. For the latter, solutions based on wall-mounted camera arrays or moving cameras are described in [9, 10], correspondingly. Arbitrary viewpoints can be supported to remote spaces, provided that complete and real-time enough 3D reconstructions of those spaces are available, or that one of the multiple cameras provides the viewpoint along a desired line-of-sight (**Figure 3**).

For serving viewpoints from varying positions, i.e. receiving viewpoints on-demand, the system needs to deliver participant positions between sites in unified coordinates (a unified geometry). Renderings of remote participants need to be

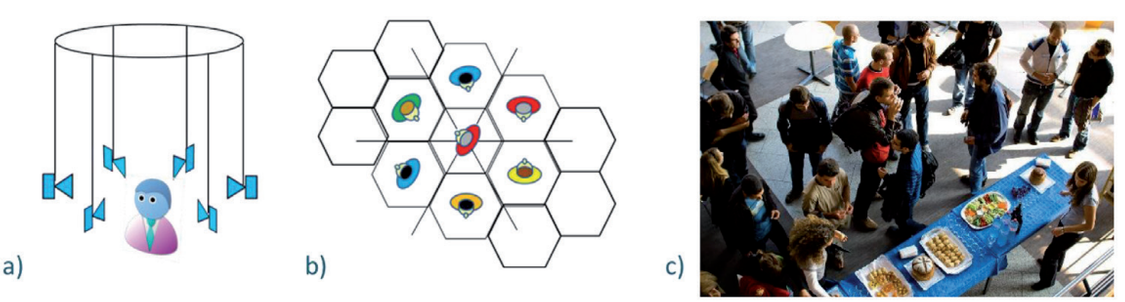

Figure 2.
Fixed capture setups can support collaboration in dynamic 2D landscapes: a) capture setup in hexagonal grid, b) captures arranged into a tessellated landscape, enabling c) moving with user spaces like people in a cocktail party [8] (image c) is creative commons image by Lucas Maystre from Renens, Switzerland - 053/365: Apéro au forum).

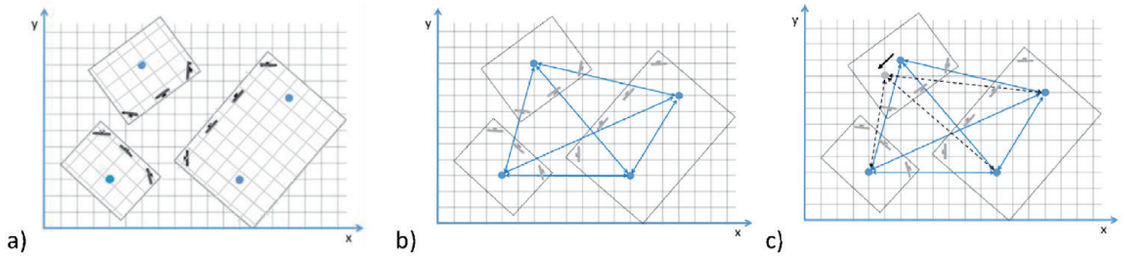

Figure 3.
Idea of bringing real-world meeting sites into a common geometry: a) three sites with users (dots) captured by RGB-D cameras in local coordinates, b) lines-of-sight between users in a unified coordinate system, and c) supporting lines-of-sights (new viewpoints) for a moving participant.

compiled in each participant view correspondingly. A viewer's head orientation needs to be detected and tracked to serve him/her with a correct part (frustum) of the compiled scene.

The writers have disclosed several inventions using the above described approach [8–10]. Note that instead of delivering visual information as large and bitrate consuming 3D volumes, a visual stream may be a video, a stereoscopic video, or a video-plus-depth stream (V + D, [11]) allowing forming a stereoscopic video. When using the viewpoint-on-demand approach in this way, the bitrate requirement may be much lower than when streaming 3D volume data. Note that Chapter 2.4 will describe the viewpoint-on-demand approach in more detail.

As a summary, spatially faithful telepresence solutions aim to support real-world-like geometries among participants. However, as will be explained in Chapter 3.2, both virtual and photorealistic approaches may be feasible even without such support, and even much easier. For example, perception of depth and motion parallax can be supported without forming and maintaining perfectly unified virtual meeting geometries.

2.3 Transmitting and displaying volume videos (3D streams)

Ideally, 3D reconstructions are coded, delivered and displayed as real-time 3D streams. However, this is very challenging e.g. due to high computation power, and very high bitrate it requires. The benefits of volume videos include more freedom in choosing one' viewpoint (cf. motion parallax and alternative viewpoints) and support for multiple local viewers. However, both capturing participant spaces and supporting viewing with glasses bring considerable complexity to this approach.

Viewing from various viewpoints may be supported also using multi-view streaming and display methods. However, without capture and delivery of user positions (cf. knowledge of a mutual meeting geometry), users need to choose their position accurately among a priori specified locations e.g. in order to perceive correct eye contact(s). Several approaches are using this approach, although simplified by reducing the volume being supported. Multi-view video coding methods and standards are available and applicable for this [12], but more advanced (real-time) 3D coding method are still under development e.g. by MPEG. Special 3D displays are already available, supporting different 3D viewpoints for multiple local viewers.

In the future, by advances in transmission and display (e.g. using light fields), real-time streaming and display of 3D volumes becomes more feasible. An apparent benefit of these solutions is that simultaneous viewers can see the 3D content from their individual viewpoints, like in the real world.

2.4 Viewpoint-on-demand – simplifying spatially faithful solutions in low latency network

5G seems to provide enough bitrate with low latency for future 3D telepresence services. According to Ronan McLaughlin (Ericsson, Ltd.), the 5G system design parameters specify a system capable of delivering an enhanced mobile broadband (eMBB) experience, in which users should experience a minimum of 50–100 Mbps everywhere, and see peak speeds greater than 10 Gbps, with a service latency of less than 1 ms, while moving at more than 300 miles/h! (https://broadbandlibrary.com/?s=5G+Low+Latency+Requirements).

Spatial faithfulness requires a shared geometry between meeting participants. In order to form and maintain such geometry, user positions need to be detected, tracked, and delivered at each moment. In addition, defined by the geometry, 3D data from several remote sites needs to be streamed to each local viewer, and compiled in a unified 3D representation. If each of the 3D captures is a full reconstruction of the corresponding 3D space, the overall bitrate requirement for rendering each view becomes huge. This may be even too much for the emerging 5G network (or at least costly). In addition to high bitrate, a very important potential of 5G network is its low latency (cf. the above figures by Ronan McLaughlin).

Most of the existing approaches for 3D telepresence are aiming to capture, encode and stream visual data of at least partial 3D volumes, which then can be seen from various viewpoints in the receiver. However, a person is able to see only from one (binocular) viewpoint at a time, which means that at each moment, there is need to see only one projection to a 3D volume. Assuming that a viewer's motions are moderate, and that a low latency network like 5G is available for data streaming, the complexity of a 3D telepresence system can be considerably reduced by streaming only video-plus-depth (V + D) projections from tracked viewpoints. Valli and Siltanen have made several telepresence inventions using this sc. viewpoint-on-demand (VoD) approach [8–10]. In particular, applying augmented reality to 3D telepresence is described in inventions [13, 14]. Note that an example of our recent 3D streaming implementation is given later in Chapter 3.4, and using the solution for supporting XR functionalities is described in more detail in Chapter 4.

Note that synthesizing viewpoints e.g. for supporting motion parallax and correcting camera offset for eye contact may be possible without ordering new data and experiencing a corresponding two-way delay as a result. An obvious way of reducing the need for delivering data for new viewpoint is to use multiple-viewpoint video coding instead of video-plus-depth (V + D) data [15]. This allows more freedom for viewpoint changes within the received stream. For examples of reducing viewpoint orders, see also several inventions by Valli and Siltanen on synthesizing stereoscopic or accommodative (MFP) content for small viewpoint changes [16, 17].

2.5 Photorealistic vs. virtual world (VW) approaches

Note that serving with arbitrary viewpoints may be easier in VW approaches, where virtual camera views are formed to a shared virtual meeting space (VW), using the knowledge of each viewer's pose (tracked by VR glasses, or defined by a participant e.g. by a mouse). However, virtual environments with animated avatars are less natural, and may even alienate a participant by causing the sc. uncanny valley effect. On the other hand, using modeled avatars for participants provides the possibility for their anonymity or role-play, which is an obvious benefit in some use cases and services.

Using a virtual world approach is a viable option used by several service vendors (see e.g. references in https://en.wikipedia.org/wiki/Virtual_world). In VW approaches, meeting spaces are typically modeled in advance, and as much as possible, also delivered to the participants in advance. Coding and delivering corresponding 3D information may be based e.g. on hierarchical volume coding methods like OctoMap by Hornung et al. [18]. Despite partly in-advance delivery for the meeting space, a lot of accurate motion and animation parameters remain to be delivered, and graphical processing to be made for local renderings (e.g. for forming viewing frustums). As a result, although possibly lighter than photorealistic approaches, VW approaches are by far not simple either.

In a photorealistic approach, capturing, forming, and delivering reconstructions of 3D volumes and human participants is more challenging, although the meeting spaces can likely be modeled in advance. Once formed, 3D reconstructions can be viewed like in VW approaches, using NEDs or HMDs. Naturally, hybrid solutions combining photorealistic and VW approaches are also possible. For example, 3D modeled environments may be used instead of captured meeting spaces, and XR functionalities can be used for augmenting avatars.

Figure 4 (cropped screenshots of YouTube videos by the courtesy of Oliver Kerylos) gives examples of hybrid (XR) approaches, showing real-time captured participants in 3D modeled meeting spaces.

As seen in **Figure 4**, a particular challenge in this approach is that a glasses display covers a person's face, which prevents a viewer from seeing his face and perceiving eye contact. Solutions for this have however since been described in literature, based on real-time manipulation of facial areas [19].

As a short summary, main approaches for 3D telepresence can be classified into the following four classes (**Table 1**).

Note that the quadrants of the table correspond to the classical reality-virtuality continuum by Milgram and Kishino [20]. Current videoconferencing and Virtual World approaches correspond to the real and virtual ends of this continuum, and hybrid approaches respectively to intermediate positions labeled as augmented reality (AR) and augmented virtuality (AV). Note that in parallel to the commonly used term "Mixed Reality" (MR), also the term "Hybrid Reality" (HR) was discussed in [20]. Recently the term "Extended Reality" (XR) has gained popularity much in the same meaning.

In an augmented reality (AR) approach, a virtual avatar is representing each remote participant in a local environment. This requires capturing a remote participant's facial and body gestures and animating the avatar correspondingly. Respectively, in an augmented virtuality (AV) approach, photorealistic 3D captures of participants are made and delivered in real-time to a virtual meeting space (VW).

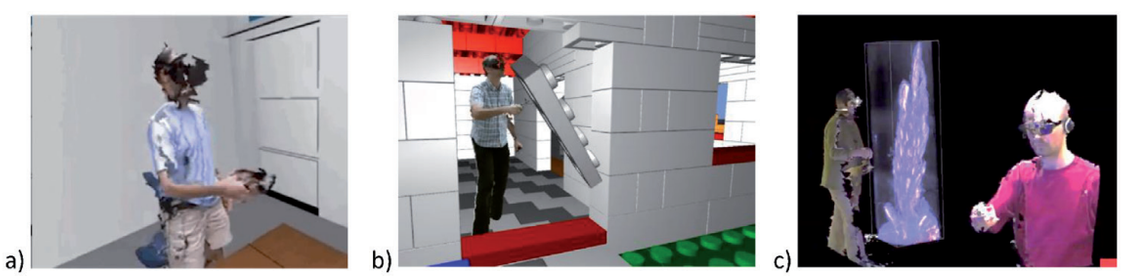

Figure 4.
Examples of XR approaches: a) person capture in a virtual space (2014), b) interaction in virtual space (2016), c) XR collaboration in virtual space (2012). (see https://www.youtube.com/channel/UCj_UmpoD8Ph_EcyN_xEXrUQ).

	Real space	Virtual space
Real human	**Photorealistic 3D approaches:** • Real-time capture of participants and spaces (cf. videoconferencing) • Challenges in supporting: ○ Shared geometry and depth for natural perception ○ Participant/user mobility ○ Avoiding use of HMD/glasses	**Hybrid (AV) approaches:** • 3D participant captures in VW • Challenges: ○ Real-time 3D capture, reconstruction, and delivery of participants ○ HMD/glasses obstructing faces
Virtual human (avatar)	**Hybrid (AR) approaches:** • Avatars in physical spaces Challenges: ○ Capturing participants for animating avatars with natural motions and facial features ○ HMD/glasses obstructing faces	**Virtual (VR) approaches:** • Scalable, spatially faithful virtual (VW) solutions exist • Challenges: ○ Avoiding unnaturalness ○ Supporting photorealism ○ HMD/glasses obstructing faces

Mutual challenge/opportunity while showing people and spaces: **Supporting remote XR interactions**

Table 1.
Main approaches for 3D telepresence.

In our case, hybrid approaches are particularly interesting. Compared to local (traditional) XR visualizations, combining real and virtual elements over distances (i.e. remote XR) causes particular challenges. These are discussed in more detail later in Chapter 2.7 and 4.1.

2.6 About using glasses and screen displays

AR/VR glasses or head-mounted displays (HMD) – together referred to as near-eye displays (NED) – can in principle support full (sc. 6DoF) freedom in choosing ones viewpoint to the displayed 3D content. Naturally, this requires also enough physical space, and precise tracking of user's motion and orientation (sc. pose).

HMDs (VR glasses) are well accepted for playing immersive and interactive computer games, and are commonly used when using VW-based telepresence and online platforms. However, they are still challenged by resolution, weight, and lack of support for natural focus (accommodation), causing discomfort and nausea when viewing stereoscopic content [6, 21, 22].

Optical see-through (OST) AR glasses (cf. MS HoloLens) have succeeded best in XR applications, but in addition to sharing the above challenges of HMDs, they are lacking natural occlusions, e.g. the ability to block a real background by augmentations, which makes AR objects to appear translucent. When augmenting natural views with 3D objects or other visual elements, closer objects should in general occlude those further away. Real and virtual objects may be in any order, meaning that foreground objects need to occlude the background whether they are

real-world captures or virtual renderings [23]. These sc. mutual occlusions are especially difficult to support by optical-see-through (OST) glasses - like MS HoloLens.

In telepresence use, a serious drawback of both HMDs and OST AR glasses is that they block their user's face, which makes it difficult to see a participant's facial features and eye-directions, either when animating an avatar in virtual approaches, or when viewing photorealistic captures. Correspondingly, although advanced considerably in recent years, NEDs are not yet good enough to be generally accepted and applied to 3D telepresence.

On the other hand, screen displays have developed by size, accuracy, and economy. While usually seen from some distance, they are e.g. less prone for perceiving VAC in stereoscopic viewing. Naturally, they restrict choosing ones viewpoints, and in XR, make it about impossible to mix remote and local content naturally in depth dimension (except when mirroring a local environment with augmentations). In short, screen displays are less immersive, but more easy to view than NEDs.

While the size and accuracy of screen displays is growing, and the distance of viewing them is reducing, supporting accommodation may become necessary also with screen displays. However, existing external accommodative displays support viewing either from very fixed viewpoints, or suffer from other severe limitations in rendering (cf. lack of colors, brightness and occlusions when using e.g. holographic volume displays).

2.7 About XR and its role in 3D telepresence

Spatial faithfulness is an inherent requirement in XR visualization, which aims at replacing, in a seamless way, parts of a physical view with virtual elements. XR visualization can as well be used for rendering 3D models (cf. avatars) or visual reconstructions of human participants into a participant's view. Correspondingly, for more than two decades, developing enablers for XR has correspondingly advanced also 3D telepresence solutions. VTT has made a lot of research in these topics, and examples of results can be found e.g. at web (http://virtual.vtt.fi/virtual/proj2/multimedia/) and in YouTube (www.youtube.com/user/VTTAugmentedReality).

VTT made also an early implementation of MR telepresence (MR Conferencing) in 2008–2009 [24]. The implementation supported participation to telepresence sessions using normal videoconferencing terminals and screens, and a VR space (SecondLife) with avatars. Registration of avatars to a real space used visual markers, which at that time was the main approach in AR visualization. Note that today's MR telepresence implementations do not necessarily differ too much from the VTT example, except using feature based tracking instead of markers (**Figure 5**).

Traditionally, making 3D captures at the target location ("on the spot") has been the only way to *support precisely* either positioning or viewing AR content in the location. By precisely, we mean that AR content can be bound to both shapes and textures (i.e. to precise visual context). This has meant making 3D capture and reconstruction locally, as an offline and in-advance process. Correspondingly, remote production and positioning of XR objects - without knowledge of local visual context - has been based on: 1) assumed textures (e.g. on known/assumed markers/pictures/objects/color patterns), 2) locally scanned shapes when displaying (e.g. physical delimiters like floor and wall panes), or 3) actions by the viewer (e.g. positioning of avatars and talking heads for communication).

However, tele interaction and XR applications can be supported better if 3D capture for making augmentations is enabled also from remote and more in real-time. This is possible by using efficient 3D capture, coding and streaming methods. The importance of efficient and high quality 3D streaming and interaction is growing fast

Figure 5.
MR conferencing between virtual and real spaces: a) second life view (screenshot), b) real life (augmented video).

due to the transformation towards distributed industrial processes, and having at the same time needs for reducing physical travels. Writers of this paper have got successful results in applying standard coding methods into real-time streaming of video-plus-depth data from RGB-D sensors (including means of supporting high enough pixel dynamics for the depth sensor data). These are presented later in Chapter 3.4.

As a natural trend in 3D telepresence implementations, there is a need for increasing accuracy (pixel dynamics) and resolution in 3D reconstruction. Following the progress in industrial applications, Lidar sensors are likely to become also into use in telepresence solutions. In addition to now common point cloud coding and transmission (e.g. using octrees [18]), this will likely require new coding methods which – in addition or instead of point clouds – support efficiently real-time transmission and visualization of high-quality surfaces and color textures (cf. approaches used with RGB-D sensors).

2.8 Focus of the research and the rest of our paper

Most of existing telepresence solutions are either photorealistic or virtual, i.e. fall into the first and fourth quadrants in **Table 1**. Hybrid approaches mix real and virtual components (for either participants or spaces) meaning that they are XR approaches (cf. discussion in Chapter 2.5). Note that in telepresence, augmenting remote participants, spaces or objects occurs over network, meaning that it is about remote XR (cf. Chapter 2.7), which requires delivering more position and 3D data than in traditional AR, both for augmentation and viewing.

Further, although augmented content can be viewed also on fixed or mobile screens, the best and most immersive way of viewing 3D augmentations is by using AR glasses. Correspondingly, accurate tracking is needed both for positioning augmentations (note that in telepresence this needs to happen over network/distance), and seeing them from a correct viewpoint in the target space. Supporting the same for multiple remote sites and participants causes further complexity, especially if the goal is to support a shared understanding of participant positions (cf. face-to-face meetings).

Table 2 summarizes our exemplary focus on videoconferencing type of photorealistic telepresence approaches (cf. quadrant real human - real space in **Table 1**), with hybrid enhancements based on 3D streaming and XR visualization.

Selected focus: Photorealistic solution based on screen displays		
Challenge	**Approach**	**Solution**
• Shared geometry and depth for natural perception • Participant/user mobility • Avoiding use of HMD/glasses	• Use of (3D) screen displays • Improved support for 3D geometry and depth • Improved support for user mobility • Improved (3D) interaction by (remote) XR	• Support for motion parallax • Support for remote augmentations (remote XR) • Support for viewing XR objects in participant spaces (using mobile displays and/or AR glasses) • Future option for accommodation support

Table 2.
Defining the focus to screen based 3D telepresence solutions.

A simplified hypothesis for our study is that much of the complexity of 3D telepresence solutions can be avoided by aiming at a screen-based solution without a (fully) realistic meeting geometry. An important cue is motion parallax, supported by tracking small user motions and serving with new viewpoints accordingly. Because of this choice, tracking and exchanging user positions for maintaining a unified meeting geometry is omitted, simplifying the solution considerably. Correspondingly, although beneficial in some geometry supporting solutions, the earlier described video-on-demand approach is not needed either.

Despite the demarcation to screen displays, the solution can be enhanced with remote XR functionalities, i.e. by bringing benefits of hybrid approaches to a photorealistic screen based solution. With screen displays, it is also easier to support natural occlusions when compiling remote views (e.g. no need to use XR approaches for displaying remote views around a local participant). Using external (flat) screens is naturally also a solution to avoid (the need of) covering faces by glasses display, i.e. better supporting photorealistic capture of participants.

Correspondingly, the next chapter focuses on describing the above photorealistic approach for 3D telepresence, giving more details on its main challenges and the status of related technical enablers. Most important of those enablers is support for coding and streaming RGB-D data, for which an exemplary implementation is described with some numerical results.

3. 3D telepresence solution using screen displays and supporting XR

3.1 Introduction

In the following chapters, main choices, enablers and components are described for a photorealistic telepresence with screen displays. Features from hybrid approaches are included, e.g. possibility to replace visual captures of a remote participant by an animated avatar. Further, in addition to screen based communication, XR interactions can be supported separately by streaming 3D scanning results between meeting sites, and viewing either locally or remotely produced augmentations e.g. by AR glasses.

3.2 Serving viewpoints by screen displays

Generally, serving moving participants requires views from arbitrary viewpoints. This in turn requires tracking of participant positions and virtual meeting geometry

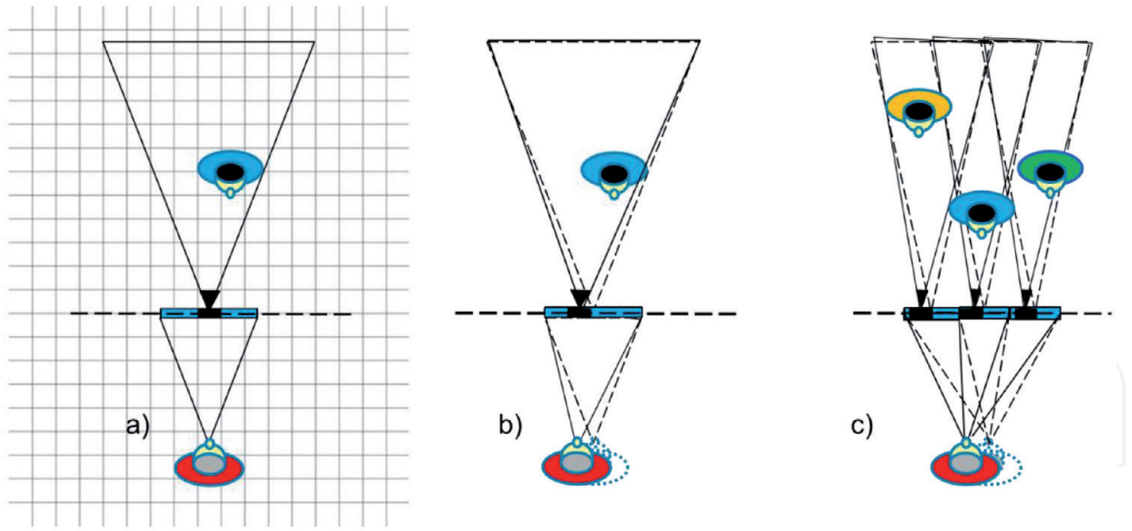

Figure 6.
Supporting motion parallax to a mosaic of 2D or 3D renderings on screen.

in real time. Further, although it may be enough to model a meeting environment in advance, photorealistic 3D capture of participants needs to be made in real time. This in turn requires a setup of multiple 3D sensors and an efficient reconstruction algorithm.

It is however possible to simplify the implementation considerably by relieving from natural geometry requirement. In the minimum, small motion parallax and even natural focus (e.g. using MFPs [16, 17, 25]) can namely be supported without forming and maintaining a virtual meeting geometry between participants. Although with more limitations than with NEDs, also flat screens can support user mobility and consistency of meeting geometries.

By relieving geometry constraints, more freedom for display arrangement and mobility can be achieved. For example, motion parallax can be supported also when compiling remote participants into a video mosaic on a display (a typical situation during video conferencing, as such) and thus to better support 3D cues (**Figure 6**). Note that it may not be that harmful even if all remote participants have their faces oriented towards a local viewer (cf. a "positive Mona Lisa effect", i.e. getting an eye contact even when not being looked at).

In this simplified approach, accurate tracking and delivery of user positions is not needed, and neither is the definition for a unified meeting geometry. Instead, the tracking reduces to a local and rather approximate process of detecting the direction of participant motion, indicating more a viewer's qualitative desire to perceive motion parallax. Further, by satisfying to frontal 3D captures, only one capture sensor is required.

As a result, 3D cues supported by the suggested system are limited to synthesized motion parallax, true eye contact (cf. avoiding the effect of a camera-display offset), and perception of depth. All these are important improvements over existing videoconferencing solutions. As described in [16, 17, 25], supporting natural focus/accommodation is also possible, provided that practical solutions for MFP displays or alike come to market.

3.3 Simplified user tracking and geometry formation

Generally, user tracking and positioning is an important functionality of 3D telepresence solutions. User positioning is required for 1) forming a consistent virtual geometry between participants, and 2) serving a participant with viewpoints

complying his/her movements in the defined meeting geometry. A tracking device can be carried by a participant or can measure the person from outside. Visual tracking is commonly assisted by other electronic sensors (IMUs or a like) and by fusing the results for better accuracy.

For a telepresence session, most favorably, a common server makes the formation of a virtual meeting geometry. For this purpose, the server needs participant positions from each telepresence terminal (cf. varying participant positions in **Figure 3c**). Bitrates for delivering 3D positions may be reduced by a suitable coding method, e.g. differential, run-length (RL), variable length coding (VLC), or their combination.

In general, user tracking, from either outside or by wearable sensors, has evolved considerably in recent years. A good solution to provide 6DoF head motion tracking is visual-inertial odometry (VIO), which estimates the relative position and orientation of a moving device in an unknown environment using a camera and motion sensors (https://en.wikipedia.org/wiki/Visual_odometry). A big advantage of VIO is that it can be processed on glasses or HMD without external setups, i.e. sensors, markers, cameras, or lasers set-up throughout the room. A comparison of several VIO approaches is presented in [26].

Note that in our simplified telepresence approach using screen displays, the perception of a consistent geometry between participants is relieved to ease up the implementation. For the screen-based communication, there is no need to derive user positions accurately, nor to deliver them to remote sites. Correspondingly, there is no need to track a camera or cameras for 3D reconstruction either. For supporting motion parallax, rather qualitative detection of user motions is enough, i.e. to detect simply, whether a viewer is moving slightly (e.g. leaning left or right) to perceive a slightly altered view. These small viewpoint changes can be supported locally, e.g. by synthesizing the viewpoints, so that there is no need to deliver captured motions to other participants.

In case the solution is enhanced by the support for seeing augmented objects in a participant's space, the tracking needs to be more wide base and accurate. However, if the support is only for seeing XR objects locally, there is no need to deliver viewer motions to other sites.

3.4 Coding and streaming 3D data

Efficient 3D capture, coding and streaming are important for future 3D telepresence solutions [27–29]. As we introduced in Chapter 2.4, coding and delivery of 3D volume data is not reasonable nor necessary for supporting spatial faithfulness, as a viewer is able to see a 3D environment or content only from one (binocular) viewpoint at a time. This suggests a solution using viewpoint-on-demand approach, which, instead of delivering complete 3D views, serves remote viewers with video-plus-depth (V + D) perspectives from desired viewpoints. A prerequisite of this approach is that user positions are tracked and set into a unified geometry defining (virtual) lines-of-sight between participants.

Luckily, V + D format suggested above serves also well in enhancing screen based telepresence solutions, both with additional 3D cues (motion parallax, and depth, both for stereoscopy or supporting natural focus/accommodation) as well as with XR visualizations and functionalities. Although communication is based on viewing remote participants on screens, a system can also support producing and delivering XR objects, viewed by a local participant's with glasses or by looking through a mobile device.

Using video-plus-depth captures simplifies and eases-up the implementation and reduces bitrates and complexity in data coding and streaming. We applied

existing video coding methods supported by FFMPEG for encoding of RGB-D data (e.g. HEVC/X265). A basic challenge is that Kinect type of sensor produces 16 bit/sample depth values, which are not supported by video coding methods. For that reason, we rounded 16 bit depth values to closest 12 bit integers before coding.

Figure 7 illustrates the pipeline in our experiments. The quality of our video-plus-depth coding and streaming was experimented by comparing direct reconstruction result of a moving RGB-D sensor to the reconstruction made after coding and streaming the data by a HEVC/X265 (FFMPEG) codec. The reconstruction algorithm was the one provided by Open3D. The test sequence was sequence 016 (here denoted as 'Bedroom') from the SceneNN dataset at http://www.scenenn.net, obtained by using Asus Xtion PRO, a Kinect 1 type of depth sensor. The sequence consists of 1364 color and depth frames (captured in about 45 seconds), both in PNG format with 480x640pel per frame.

In **Figure 7**, video and depth sequences were transferred into two video type sequences using the sc. depth blending, modulating the original input video by a linearly weighted depth map and its inverse [25]. This results with two video-like sequences with the partition-of-unity property, meaning that the output sequence is obtained by summing up the modulated (and coded and streamed) video components in the receiver. The coded depth map sequence is obtained from the ratio of luminance(s) for the corresponding pixels. Note that the same approach is typical when forming MFPs for accommodation supportive displays. Here, we omit further details of the coding process and suggest an interested reader to study e.g. the above references.

In our experiment with the above Bedroom sequence, the average bitrate for the original video-plus-depth data from the sensor was 103Mbit/s (RGBD frames in png format, 30fps), and the average bitrate for the coded and streamed data was 567kbit/s, corresponding to about 180:1 compression ratio. Standard (RMS) deviation of the output voxels was 4.2 mm compared to the input ('original') surface, as derived from the reconstructions by the CloudCompare SW (see https://en.wikipedia.org/wiki/CloudCompare). PSNR was calculated from the differences between corresponding YCbCr pixels of the input and output sequences. Average PSNR was 50.3 dB for the luminance (Y), and 48.6 dB and 55.2 dB for Cb and Cr components. YCbCr format was chosen for being traditionally used in compression research and for better specifying obtained PSNR values. Calculations were made using Matlab (r2018b) functions for format conversions and PSNR. These above numerical results are very good, and when viewing by eye, both the video and the reconstructions appear identical.

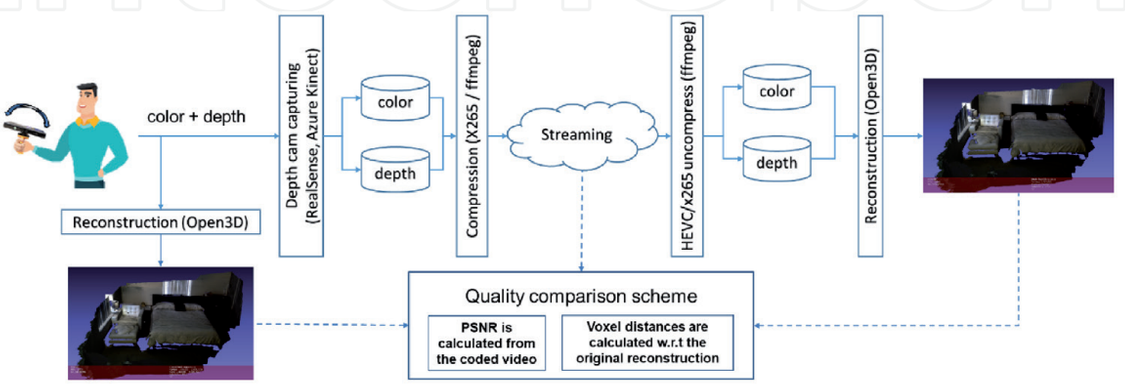

Figure 7.
3D pipeline showing data (color plus depth data from a moving RGB-D) being captured, coded, streamed, and reconstructed in our experiment. Quality comparison (PSNR) is made between original and coded videos, and reconstruction accuracy (RMS distance to the nearest voxel) is measured from 3D reconstructions using original and coded depth images.

Note that the above pipeline is still an offline implementation, using stored files for both input and output data. Correspondingly, there were no real-time limitations in the above simulations. The writers expect to complete also a real-time version of the pipeline by the autumn 2021.

3D streaming solution described above can obviously support also higher resolutions. E.g. with a fourfold resolution to the experiments, the bitrate would remain in the order of 2Mbit/s. As long as there are no better coding methods available, it will be more challenging to support higher pixel dynamics, i.e. more bits/pel (e.g. by depth blending as in our solution).

Generally, the bitrate for video-plus-depth is much less than when streaming multiple-view or volume videos. As a comparison, the approaches used in [7] resulted with the average of 1–2 Gbps transfer rate for a 30 fps stream. According to Qualcomm[1], 6DoF video demands bit-rates in the range of 200 Mbps to 1000 Mbps depending on the end-to-end latency. These figures are just indicative, as the bitrates depend heavily on the used coding scheme and many factors affecting quality (notably resolution used in 3D data capture). Interested readers may find more information from the references given in the beginning of this chapter.

Our simulations on 3D streaming indicate that reconstructing 3D models from coded and streamed video-plus-depth data succeeds with an adequate quality for 3D viewing and reconstruction. Note that the same simple data pipeline enables also various remote support functionalities, including remote 3D analysis based on coded information.

The principle of using compressed data for visualization and analysis is denoted as compress-then-analyze (CTA) approach [30, 31]. According to [31], the opposite analyze-then-compress (ATC) approach may outperform at low bitrates. However, ATC limits a system flexibility, as for example normal viewing of the stream is not possible using only received visual features. Further, ATC fixes the feature selection method at a captured space, limiting applicable approaches for remote analysis. In fact, CTA provides superior flexibility in multipoint settings by enabling for example any analysis approach by multiple remote receivers. According to [30], CTA may also outperform ATC at high bitrates. It is worth noticing, that the referred studies for CTA used jpeg compression for the visual features, which wastes bitrate and lowers quality compared to our efficient spatiotemporal CTA approach.

3.4.1 Reducing the need for streaming by synthesizing viewpoints

Video-plus-depth data format supports synthesizing viewpoints without ordering and streaming new data. This in known from the sc. depth image based rendering (DIBR) approaches for stereoscopic (S3D) TV [15]. Using DIBR, stereoscopic image pairs can be formed in any desired baseline orientation. Viewpoints can also be synthesized to support 3D motion parallax, i.e. any small viewpoint changes around nominal viewpoints from which video and depth images are captured. In a telepresence solution, synthesizing viewpoints can thus be used for both reducing bitrates and avoiding possible latencies of a viewpoint-on-demand approach.

Applicable methods for synthesizing new viewpoints are virtual viewpoint generation in 3D (3D geometry calculations), which are also well supported by graphics processors for speeding up computations. Another way is used in [16, 17], where new viewpoints are formed by simple shifts of MFPs, generated also using video-plus-depth data. The latter approach is good at least if a graphics processor is not available, and if natural accommodation is supported by an MFP approach. Note

[1] https://www.qualcomm.com/media/documents/files/augmented-and-virtual-reality-the-first-wave-of-5g-killer-apps.pdf

that MFPs can also be used for virtual viewpoint generation for normal stereoscopic pairs without the aim for supporting accommodation [16, 17].

4. Enhancements by XR and naturalness

Traditionally, the gold standard for telepresence solutions has been a face-to-face meeting. While mimicking physical encounters over network is technically very challenging, the goal for telepresence solutions has even been raised to exceed the possibilities of physical meetings, referred also to as "beyond being there" [32] and "beyond being aware" [33]. Bill Buxton et al. referred to additional functionalities enhancing face-to-face collaboration as 'groupware' [4]. Correspondingly, also our 3D telepresence solution can and needs to be enhanced with additional functionalities. In our case, many of them are based on XR functionalities, which are discussed in the following.

4.1 Hybrid functionalities for human collaboration

Hybrid functionalities (cf. **Table 1**) combine real and virtual components when rendering and displaying telepresence views. There are two main options, which are described shortly in the following, namely:

1. Replacing a camera captured (and animated) participant view with a virtual avatar (augmented reality option), and

2. Replacing a camera captured participant space with a virtual space (augmented virtuality option).

There are multiple implementations and services using the first approach, e.g. https://remoteface.ai (+ YouTube https://www.youtube.com/watch?v=prpPqwV5Weo). This approach requires either capturing a participant's facial features in order to animate the avatar, or in the minimum, capturing a participant's speech to estimate underlying facial muscle movements and corresponding animation parameters.

The second approach was already illustrated in Chapter 2.5, where a real-time captured human is rendered into a typically in-advance modeled virtual space (cf. e.g. **Figure 4**). Note that a virtual space may even enable a remote participant to make virtual visits to that space (i.e. seeing to it from widely varying viewpoints) – in particular, if the viewing is supported by glasses display.

Note that using glasses for viewing makes the interaction easily nonsymmetrical or even one-way, as the glasses prevent either capturing facial movements of their wearer, or seeing his/her face and eyes. Somewhat working solutions to avoid this have however been described in literature, based e.g. on real-time manipulation of facial areas [7].

4.2 Remote XR support functionalities

When developing 3D telepresence solutions, we are particularly interested in supporting remote XR functionalities. There are two main approaches for doing it. The first approach requires only delivering of images or video to the remote site(s), and coding and streaming is supported straightforwardly by existing video coding methods. However, better support for remote interaction is provided by coding data from a depth sensor, and after streaming the data, making 3D reconstruction at a

remote site. Algorithms used for local reconstructions are applicable also for remote reconstructions.

Thus, in addition to better 3D perception, video-plus-depth data supports also forming (or copying) 3D reconstructions at remote sites. These reconstructions, which we have denoted as Visual Twins, can support various 3D remote support functionalities, e.g. 3D monitoring, control, and analysis, as well as remote augmentation with visualizations and instructions. As described in Chapter 3.4, this is feasible by applying existing coding methods.

We have tested both video-based (sc. Ad-hoc AR) and video-plus-depth based (sc. Visual Twin) approaches, and they are described in more detail in the following:

1. Local 3D reconstruction, pointed remotely for positioning augmentations (Ad-hoc AR)

In this option (**Figure 8**), a 3D reconstruction is made in a local space using e.g. an RGB-D sensor carried by a moving person or a robot. The orientation of each RGB-image is derived in a normal way in the reconstruction process (e.g. using SLAM [34] and TSDF [35]), and stored locally with the image ID (e.g. a simple timestamp). The images are coded and streamed separately (e.g. following a manual selection) or as a sequence to a remote space. In the remote space, a person selects a point (pixel) in an image to show an augmentation, and messages back the image ID, target pixel coordinates, and data (or ID, if stored on a common server) of the AR object.

At the local site, the image's orientation w.r.t. 3D reconstruction is fetched from the local memory. The point to show the augmentation is obtained by ray-tracing through the defined pixel to the known orientation. Ray-tracing defines a 3D surface point on the 3D reconstruction (and the space), and enables local participant(s) to see the augmentation from various directions. **Figure 8** illustrates the process.

2. Remote 3D reconstruction using streamed depth sensor data (Visual Twin)

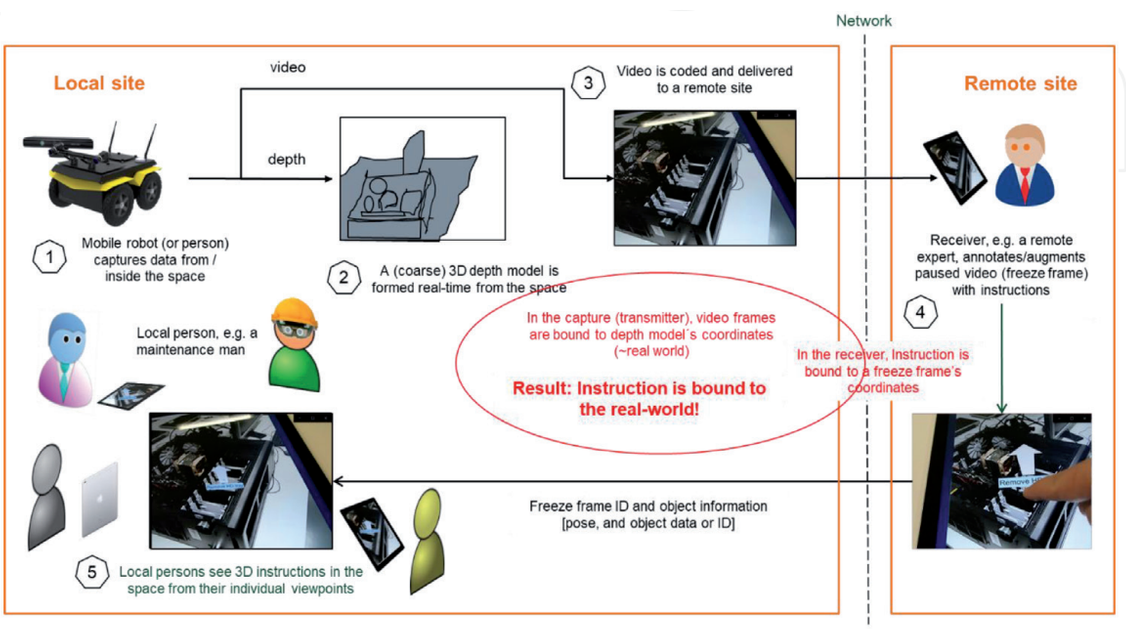

Figure 8.
Ad-hoc AR, enabling remote augmentation of a locally reconstructed space.

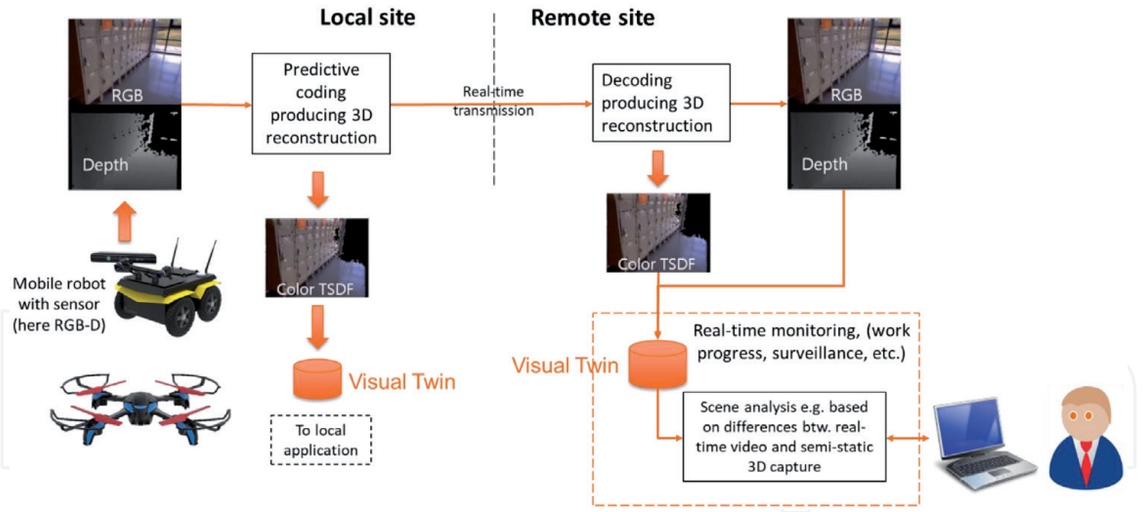

Figure 9.
Principle of visual twin, here used for remote monitoring & control.

In this approach (**Figure 9**), data for 3D reconstruction (e.g. RGB-D data) is coded and streamed over a network, and the reconstruction is made using the data decoded at the remote site. The solution described in Chapter 3.4 for the coding and streaming of video-plus-depth data supports directly this approach when made in real time (implementation for this is soon completed by VTT).

Both of the approaches have been implemented and demonstrated by VTT, and are suitable for enhancing our 3D telepresence solution with XR functionalities. The first ray-tracing based option is simpler, but being based on images/video only, does not allow viewing the data in 3D at the remote site. Note that receiving video-plus-depth data enables also the ray-tracing approach, meaning that a combination of the approaches is also possible.

4.3 Supporting lines-of-sight

A straightforward way to enhance the described solution is to support virtual lines-of-sight between participants in the same way as in the describe Hydra approach, i.e. by using an own screen based terminal for each of the (maximum number of) remote participants (cf. Chapter 2.1). This means full mesh connections between peers, but the total bitrate may remain low due to each of the streams being a low bitrate video-plus-depth stream. This enhancement would also enable also earlier referred options for increasing user (plus meeting site) mobility e.g. by grid based geometries and interactive landscapes (cf. Chapter 2.2). The downside of this approach is the increase in the complexity for a meeting setup, and support for only a limited number of participants due to the lack of space around a participant. This version may however be justified in special cases where better spatial and gaze awareness between interactive participants is particularly important.

4.4 Natural eye-focus/accommodation

Stereoscopic rendering is a further means to improve 3D perception. However, stereoscopic viewing detaches natural eye focus (accommodation) and convergence distances. The resulting vergence-accommodation conflict (VAC) causes discomfort and nausea, and restricts a person's willingness to view stereoscopic content [6, 25, 36]. A related conflict in monocular viewing is the sc. focal rivalry (FR) [37]. In addition to VAC and FR, there are various other error types caused by the

wrong blur of rendered scene components, e.g. all-in-focus virtual objects, or real objects blurred by camera optics. These types of distortions appear especially when combining content for augmented scenes. Without careful considerations, these distortions continue hindering the quality and acceptance of XR functionalities.

As described in Chapter 2.6, unnatural occlusions are another common type of distortion, which occurs especially in XR, when combining optical views (cf. OST glasses), photorealistic captures, and 3D modeled objects and spaces. In addition to occlusions, supporting natural focus is a big challenge in display design and manufacture, and has not yet been properly solved for either external screens or near-eye displays. Natural focus requires that a content is rendered into a 3D volume, instead of one or few display surfaces. There are volume displays aiming to this, but generally they suffer e.g. from the lack of occlusions, colors and brightness.

There are various means to support natural focus, including for example light field rendering [38]. Supporting occlusions in a light field display has been studied e.g. in [39], and occlusions for its simple variant multiple-focal-plane (MFP) display in [25]. MFP-rendering is essentially light field rendering to one viewpoint, and as such fits well to the above introduced viewpoint-on-demand approach. The writers have studied various ways of forming and rendering MFPs, and have even made own designs for MFP glasses [40].

Implementation of MFP or other accommodative glasses is however very challenging, as for example the efforts by Magic Leap, Inc. has us learned. However, as glasses are superior in supporting immersion and mobility, we expect that major technical problems will be solved, and high quality accommodation support will be available within about a decade. Note that their (eventual) emergence will in general mean a big change to the ways visual information is captured, processed and displayed.

5. Conclusions and acknowledgements

For using a photorealistic approach in 3D telepresence, the biggest challenges are the accuracy and cost of acquiring, delivering, and displaying 3D captures. Glasses based approaches are attractive due to their ability to support user mobility, immersion, and in future even natural focus/accommodation in 3D perception. Glasses can provide also sensors for 3D capture and user positioning. However, natural occlusions and accommodation are hard to support, and likely it will still take many years before affordable and good-enough glasses are available.

In this article, in addition to reviewing the general status and approaches for 3D telepresence, we proposed a simplified approach for spatially faithful telepresence, based on 3D screens and low bitrate streaming of video-plus-depth data. Screen displays are cheap, and can support 3D cues and mobility better than in existing teleconferencing solutions. External displays are easy to view and are a good option for improved 3D telepresence solutions. An important enabler for a simplified solution is efficient coding and streaming of RGB-D data, for which an exemplary implementation was presented with some simulation results. Described means for 3D capture and streaming support also XR functionalities, which in addition to viewing XR content on screens, may also be used with glasses or mobile displays. Features of the introduced solution can be summarized as follows:

- Photorealism with improved support for 3D cues
- Screen displays are preferred due to not obstructing faces

- Gracefully compromising mutual (real-world like) geometry for more freedom in display placement
- Supporting low bitrates by video-plus-depth (V + D) streaming
- Delivering enough 3D data for supporting remote XR functionalities
- A participant can see augmentations in his/her environment by using a mobile device (a "magic lens") or AR glasses
- Hybrid approaches can be supported, with options for:
  1. replacing a participant capture by an avatar, and
  2. replacing a captured meeting environment by a virtual space
- Easy to be modified e.g. for emerging MFP displays supporting natural eye focus

Choosing screen displays implies both limitations and benefits to the system. With one or few screen displays, the perception of a meeting geometry is not same between participants. We decided to rely on the benefits of an exaggerated perception for eye contact by collecting all remote participant renderings into a grid on (nominally) one screen display, and providing viewers with perception for depth and motion parallax.

The key enabler for the improved system is simple: support for real-time capture, coding and streaming of video-plus-depth data from the RGB-D sensor of a telepresence terminal. This choice enables low bitrates, depth perception, and support for small viewpoint changes by user motions. Currently, the most feasible option for a display is a stereoscopic (S3D) display, but it can be replaced by an accommodation supportive display as soon they come available.

Our basic assumption is that each participant has his/her own telepresence terminal in the same way as PCs and laptops are currently used in videoconferencing. Correspondingly, from appearance, the new solution does not differ too much from current videoconferencing systems. For better supporting spatial faithfulness and gaze awareness, a more complicated system with several cameras and 3D screen displays may be used (i.e. applying Hydra and Viewport type of approaches introduced in Chapter 2.1). In parallel to viewing remote participants on screens, AR glasses can be used for viewing augmented objects inside a local space.

Although we have tested only some important prerequisites of the suggested solution, we have now a good knowledge and plan for its full implementation. We hope that this article raises its readers a general interest to the status, challenges, and possibilities of 3D telepresence, as well as a specific interest to develop the described approaches and ideas even further.

We want to express our gratitude to VTT Technical Research Centre of Finland for giving us the opportunity to work on 3D telepresence and related topics. Thanks also for InterDigital Inc. for challenging us to make new inventions in this area.

DOI: http://dx.doi.org/10.5772/intechopen.99271

IntechOpen

Author details

Seppo Valli*, Mika Hakkarainen and Pekka Siltanen
VTT Technical Research Centre of Finland Ltd., Espoo, Finland

*Address all correspondence to: seppo.valli@vtt.fi

References

[1] Nguyen D, Canny J. MultiView: Spatially faithful group video conferencing. CHI 2005 Technol Safety, Community Conf Proc - Conf Hum Factors Comput Syst. 2005;799-808.

[2] Monk AF, Gale C. A Look Is Worth a Thousand Words: Full Gaze Awareness in Video-Mediated Conversation. Discourse Process [Internet]. 2002 May 1;33(3):257-78. Available from: https://doi.org/10.1207/S15326950DP3303_4

[3] Siltanen P, Valli S. Gaze-aware video conferencing application for multiparty collaboration. In: 2013 International Conference on Engineering, Technology and Innovation, ICE 2013 and IEEE International Technology Management Conference, ITMC 2013. 2015.

[4] Buxton W, Sellen A, Sheasby M. Interfaces for multiparty videoconferencing. In K. Finn, A. Sellen & S. Wilber (Eds.). Video Mediated Communication. Hillsdale, N.J.: Erlbaum. 1997;385-400.

[5] Zhang C, Cai Q, Chou PA, Zhang Z, Martin-Brualla R. Viewport: A distributed, immersive teleconferencing system with infrared dot pattern. IEEE Multimed. 2013;20(1):17-27.

[6] Kramida G. Resolving the Vergence-Accommodation Conflict in Head-Mounted Displays. IEEE Trans Vis Comput Graph. 2016;22(7):1912-31.

[7] Orts-Escolano S, Kim D, Cai Q, Rhemann C, Davidson P, Chou P, et al. Holoportation: Virtual 3D teleportation in real-time. UIST 2016 - Proc 29th Annu Symp User Interface Softw Technol. 2016;741-54.

[8] Valli S, Siltanen P, inventors. Spatially faithful telepresence supporting varying geometries and moving users. WO18226508 A1. PCMS Holdings, Inc.;2018 Dec 13.

[9] Valli S, inventor. System and method for augmented reality multi-view telepresence. WO17030985 A1. PCMS Holdings, Inc.;2017 Feb 23.

[10] Siltanen P, Valli S, inventors. System and method for spatial interaction using automatically positioned cameras. WO18005235 A1. PCMS Holdings, Inc.;2018 Jan 4.

[11] Smolic A, Mueller K, Merkle P, Kauff P, Wiegand T. An Overview of Available and Emerging 3D Video Formats and Depth Enhanced Stereo as Efficient Generic Solution. In: Proceedings of the 27th Conference on Picture Coding Symposium. IEEE Press; 2009. p. 389-392. (PCS'09).

[12] Chen Y, Vetro A. Next-Generation 3D Formats with Depth Map Support. IEEE Multimed. 2014;21(2):90-4.

[13] Valli S, Siltanen P, inventors. Apparatus and method for supporting interactive augmented reality functionalities. WO2017172528A1. PCMS Holdings, Inc.;2017 May 10.

[14] Valli S, Siltanen P, inventors. System and method for supporting synchronous and asynchronous augmented reality functionalities. WO17177019A1. PCMS Holdings, Inc.;2019 Sep 26.

[15] Debono, C. J., S. Faria, Luis F. R. Lucas and Nuno M. M. Rodrigues. Depth Map Coding for 3DTV Applications; 2017.

[16] Valli S, Siltanen P, inventors. Method and system for forming extended focal planes for large viewpoint changes. WO20009922A1. PCMS Holdings, Inc.;2020 January 9.

[17] Valli S, Siltanen P, inventors. Multifocal plane based method to produce stereoscopic viewpoints in a DIBR system (MFP-DIBR).

WO19183211A1. PCMS Holdings, Inc.;2019 Sep 26.

[18] Hornung A, Wurm KM, Bennewitz M, Stachniss C, Burgard W. OctoMap: an efficient probabilistic 3D mapping framework based on octrees. Auton Robots [Internet]. 2013;34(3):189-206. Available from: https://doi.org/10.1007/s10514-012-9321-0

[19] Frueh C, Sud A, Kwatra V. Headset Removal for Virtual and Mixed Reality. In: ACM SIGGRAPH 2017 Talks [Internet]. New York, NY, USA: Association for Computing Machinery; 2017. (SIGGRAPH '17). Available from: https://doi.org/10.1145/3084363.3085083

[20] Milgram P, Kishino F. A Taxonomy of Mixed Reality Visual Displays. IEICE Trans Inf Syst. 1994;77:1321-9.

[21] Hu X, Hua H. Design and Assessment of a Depth-Fused Multi-Focal-PlaneDisplay Prototype. J Disp Technol [Internet]. 2014;10(4):308-16. Available from: http://jdt.osa.org/abstract.cfm?URI=jdt-10-4-308

[22] Koulieris GA, Akşit K, Stengel M, Mantiuk RK, Mania K, Richardt C. Near-Eye Display and Tracking Technologies for Virtual and Augmented Reality. Comput Graph Forum [Internet]. 2019 May 1;38(2):493-519. Available from: https://doi.org/10.1111/cgf.13654

[23] Kiyokawa K, Billinghurst M, Campbell B, Woods E. An occlusion capable optical see-through head mount display for supporting co-located collaboration. In: The Second IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003 Proceedings. 2003. p. 133-41.

[24] Kantonen T, Woodward C, Katz N. Mixed reality in virtual world teleconferencing. In: 2010 IEEE Virtual Reality Conference (VR). 2010. p. 179-82.

[25] Akeley K, Watt SJ, Girshick AR, Banks MS. A Stereo Display Prototype with Multiple Focal Distances. ACM Trans Graph [Internet]. 2004;23(3):804-813. Available from: https://doi.org/10.1145/1015706.1015804

[26] Pfrommer B, Sanket N, Daniilidis K, Cleveland J. PennCOSYVIO: A challenging Visual Inertial Odometry benchmark. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). 2017. p. 3847-54.

[27] Bannò F, Gasparello PS, Tecchia F, Bergamasco M. Real-Time Compression of Depth Streams through Meshification and Valence-Based Encoding. In: Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry [Internet]. New York, NY, USA: Association for Computing Machinery; 2012. p. 263-270. (VRCAI '12). Available from: https://doi.org/10.1145/2407516.2407579

[28] Fu J, Miao D, Yu W, Wang S, Lu Y, Li S. Kinect-Like Depth Data Compression. IEEE Trans Multimed. 2013;15(6):1340-52.

[29] Merkle P, Müller K, Wiegand T. 3D video: acquisition, coding, and display. IEEE Trans Consum Electron. 2010;56(2):946-50.

[30] Baroffio L, Cesana M, Redondi A, Tagliasacchi M, Tubaro S. Coding visual features extracted from video sequences. IEEE Trans Image Process. 2014;23(5):2262-76.

[31] Redondi A, Baroffio L, Bianchi L, Cesana M, Tagliasacchi M. Compress-then-analyze versus analyze-then-compress: What is best in visual sensor networks? IEEE Trans Mob Comput. 2016;15(12):3000-13.

[32] Hollan J, Stornetta S. Beyond being there. Conf Hum Factors Comput Syst - Proc. 1992;119-25.

[33] Carroll JM, Rosson MB, Farooq U, Xiao L. Beyond being aware. Inf Organ [Internet]. 2009;19(3):162-85. Available from: http://dx.doi.org/10.1016/j.infoandorg.2009.04.004

[34] Izadi S, Kim D, Hilliges O, Molyneaux D, Newcombe R, Kohli P, et al. KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera. UIST'11 - Proc 24th Annu ACM Symp User Interface Softw Technol. 2011;559-68.

[35] Saputra MRU, Markham A, Trigoni N. Visual SLAM and Structure from Motion in Dynamic Environments: A Survey. ACM Comput Surv [Internet]. 2018;51(2). Available from: https://doi.org/10.1145/3177853

[36] Hoffman DM, Girshick AR, Akeley K, Banks MS. Vergence-accommodation conflicts hinder visual performance and cause visual fatigue. J Vis. 2008;8(3):1-30.

[37] Oshima K, Moser KR, Rompapas DC, Swan JE, Ikeda S, Yamamoto G, et al. SharpView: Improved clarity of defocussed content on optical see-through head-mounted displays. In: 2016 IEEE Virtual Reality (VR). 2016. p. 253-4.

[38] Iwane T. Light field display and 3D image reconstruction. In: ProcSPIE [Internet]. 2016. Available from: https://doi.org/10.1117/12.2227081

[39] Wang T-C, Efros AA, Ramamoorthi R. Occlusion-Aware Depth Estimation Using Light-Field Cameras. In: 2015 IEEE International Conference on Computer Vision (ICCV). 2015. p. 3487-95.

[40] Valli S, Siltanen P, inventors. Multi-focal planes with varying positions. WO19143688A1. PCMS Holdings, Inc.;2019 Jul 25.