Communications of the Association for Information Systems

Volume 37 Article 45

11-2015

# Mutual Gaze Support in Videoconferencing Reviewed

Holger Regenbrecht
*University of Otago*, holger@infoscience.otago.ac.nz

Tobias Langlotz
*University of Otago*

## Recommended Citation

Regenbrecht, Holger and Langlotz, Tobias (2015) "Mutual Gaze Support in Videoconferencing Reviewed," *Communications of the Association for Information Systems*: Vol. 37 , Article 45.
DOI: 10.17705/1CAIS.03745
Available at: https://aisel.aisnet.org/cais/vol37/iss1/45

Communications of the Association for Information Systems

Research Paper ISSN: 1529-3181

# Mutual Gaze Support in Videoconferencing Reviewed

**Holger Regenbrecht**
Information Science, University of Otago, New Zealand
*holger@infoscience.otago.ac.nz*

**Tobias Langlotz**
Information Science, University of Otago, New Zealand

**Abstract:**

Videoconferencing allows geographically dispersed parties to communicate by simultaneous audio and video transmissions. It is used in a variety of application scenarios with a wide range of coordination needs and efforts, such as private chat, discussion meetings, and negotiation tasks. In particular, in scenarios requiring certain levels of trust and judgement non-verbal communication, cues are highly important for effective communication. Mutual gaze support plays a central role in those high coordination need scenarios but generally lacks adequate technical support from videoconferencing systems. In this paper, we review technical concepts and implementations for mutual gaze support in videoconferencing, classify them, evaluate them according to a defined set of criteria, and give recommendations for future developments. Our review gives decision makers, researchers, and developers a tool to systematically apply and further develop videoconferencing systems in "serious" settings requiring mutual gaze. This should lead to well-informed decisions regarding the use and development of this technology and to a more widespread exploitation of the benefits of videoconferencing in general. For example, if videoconferencing systems supported high-quality mutual gaze in an easy-to-set-up and easy-to-use way, we could hold more effective and efficient recruitment interviews, court hearings, or contract negotiations.

**Keywords:** Video Conferencing, Telepresence, Eye Gaze, Eye Contact, Eye-to-eye Contact.

This manuscript underwent peer review. It was received 01/20/2015 and was with the authors 4 months for 1 revision. The Associate Editor chose to remain anonymous.

# 1 Introduction

In an increasingly globalized world, real-time communication and collaboration become more and more important. The on-going upgrade and extension of network infrastructure allows us to virtually connect with anyone in the world. Together with the increased performance of video encoding and decoding algorithms needed to compress the high resolution signals of integrated cameras and microphones, video conferencing (VC) is now possible at high-quality standards. This includes high-resolution video and audio recording, streaming and playback, filtering of environment noise, and, often, focus-and-context cameras. Videoconferencing solutions are nowadays offered by many manufacturers, such as Cisco[1], Polycom[2], and Lifesize[3], and with different prices and feature sets ranging from off-the-shelf solutions to custom business solutions exceeding USD$100,000.

Videoconferencing has the potential to provide many significant benefits over traditional face-to-face meetings. For instance, Davis & Weinstein (2005) list:

1) Faster decision making and shorter time to market for products and services, which have enabled dispersed teams to collaborate easily, solved problems, and speed coordination (ultimately delivering faster time-to-consensus and, hence, a shorter time-to-market for new products and services).
2) Higher productivity / efficiency from a scheduled environment to an ad-hoc, unscheduled work style.
3) Higher impact and focused, shorter, more effective meetings with minimal workflow disruption (videoconferencing meetings tend to be shorter than in-person meetings).
4) Competitive advantage with, for example, recruitment: interviews with more people, from more locations, in less time, and with less cost and disruption; also, better hiring decisions.
5) Enhanced quality of life / decreased stress: business travel negatively impacts life, sleep, and general welfare.
6) Increased reach by personal touch between company and client
7) Improved management of dispersed teams by allowing impromptu, face-to-face meetings.
8) The reduction of travel costs in the form of direct expenses for airfare, hotels, meals, taxis, car services, and so on, and extended expenses for hours of downtime and days away from the office.

Given all these advantages, one might wonder why VC isn’t used more widely in all kinds of contexts. Two main factors might come into play here: 1) the lack of appropriate (technological) support for the varying purposes of meetings, and 2) the lack of support for communication aspects beyond simple head-and-shoulder views.

The nature of a VC meeting is determined by the effort to initiate and coordinate a meeting, which heavily depends on the task and group situation (Cornelius & Boss, 2003): informal communication, such as a chat, does not require a high communication effort; Idea-generation tasks require at least some protocol; Problem-discussing tasks do have a defined goal but require substantial effort and communication; judgment tasks (decision making and problem solving) require the highest coordination effort; and, finally, negotiations require the highest coordination effort in combination with a trust-enabling environment (face-to-face negotiations as the “gold standard“). Hence, one’s videoconferencing system needs to address the effort required to coordinate the particular meeting’s task. For instance, while a Skype-like system might be appropriate for a private chat or for a well-defined and brief decision making task, it is inappropriate for, for example, a legal contract negotiation meeting among new business partners. This would afford support for non-verbal communication cues and for integrating collaboration tools (e.g., interactive document sharing).

In this paper, we focus on those “serious” conferencing settings that require a certain degree of communication effort and trust support. Such settings are also characterized by a need to integrate meeting artefacts and/or collaboration tools, to support gestural communication cues (body language

[1] www.cisco.com/web/telepresence/index.html
[2] www.polycom.com
[3] www.lifesize.com

availability; Teoh, Regenbrecht, & O'Hare, 2011), and, with this, to provide a certain interaction space in front of the screen. Of particular interest in settings relying on levels of trust is providing eye-to-eye contact and gaze awareness. Gaze awareness tells the communication partners where a person is looking (e.g., at the documents discussed or at another partner). Eye-to-eye contact, also known as mutual gaze or simply eye contact, is a special form of gaze awareness to detect whether a person is directly looking at the partner. Eye-to-eye contact forms the basis for forming empathy and trust-building in a lot of situations (Teo, Regenbrecht, & O'Hare, 2010). Of particular interest are so-called telepresence systems that give participants the feeling of being in one shared room or space. This feeling is expressed with the concepts of co-presence and social presence.

Often, the principal problem for the lack of eye-to-eye contact is the positional offset between the capturing camera and the display of the partner's video image. Ideally, the camera should sit between the displayed eyes of the videoconferencing partner (see Figure 1).

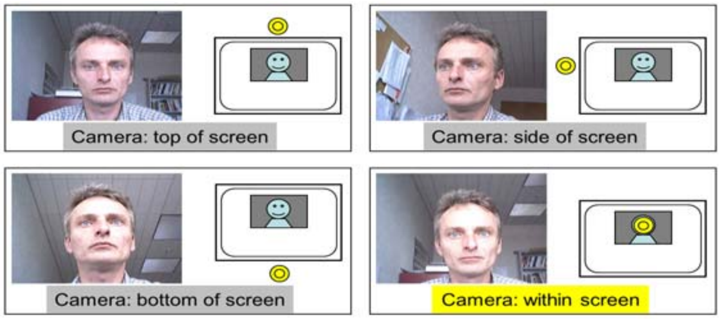

**Figure 1. Eye-to-eye Contact: Separation of Camera and Screen**

Unfortunately, placing a camera at this position would normally block one's view of the partner, which makes the solution unsuitable. As a best practice approach, most non-consumer videoconferencing systems try to place the camera as close as possible to the displayed partner video (see, e.g., top left in Figure 1). This can be implemented in desktop videoconferencing and in room-like systems. The size of the screen and video image, the distance from the user to the camera and screen, and the position of the video image on the screen are the parameters to be considered here. Because of the practical spatial limitations in most environments, true eye-to-eye contact cannot be achieved with this approach. Other technical solutions have to be applied to achieve a real sense of mutual eye contact.

As such, we need to analyze how a videoconferencing system be set up to allow for maximizing empathy- and trust-building needed in many business and other relevant communication situations. We argue that the size of the displayed face of the partner and providing mutual eye gaze are important factors to build trust and deliver a basis for high communication quality that is combined with non-verbal communication queues.

In this paper, we review the literature on the characteristics of gaze and eye contact in videoconferencing and video-based telepresence, survey different approaches for implementing eye contact win videoconference systems, and discuss their pros and cons. While videoconferencing technology is also available and extensively used for private communication between relatives and friends using applications such as Skype[4] or Google Hangout[5], we focus on professional and business-oriented videoconferencing solutions.

## 2 The Videoconferencing Experience

In most cases, one holds a videoconferencing meeting because they can't come together with one or more people in one place physically. Such a meeting can be improved by enhancing the video or audio quality and the interface to an extent that the virtual coming-together is done in a seamless way, with a disappearing interface and in a quality hardly distinguishable from a physical meeting. The videoconference may even provide advantages over a physical meeting. Furthermore, in business environments in particular, the shared access to files and virtual artefacts becomes increasingly important.

[4] www.skype.com
[5] www.google.com/tools/dlpage/res/talkvideo/hangouts/

The ultimate technology for videoconferencing would deliver an immersive experience that leads to social presence—the perceived sense of being together in one (virtual or real) place. In that context, the literature also often refers to telepresence videoconferencing. We can consider social presence as the central defining feature for a (professional) videoconferencing system. It requires a certain technical and setup fidelity comprising factors of space requirements, video quality, the system's complexity, the flexibility in posturing oneself in front of the screen/camera, hardware and software requirements, price, availability, and the ease of integration into existing environments.

Biocca, Harms, and Gregg (2001) develop a framework with three theoretical dimensions to describe and measure what determines social presence: co-presence, psychological involvement, and behavioral engagement. They associate several factors to those dimensions; namely, the feelings and perceptions of isolation/inclusion and mutual awareness (co-presence); mutual attention, empathy, and mutual understanding (psychological involvement); and behavioral interaction, mutual assistance, and dependent action (behavioral engagement). When looking at these factors, even if only intuitively, one obtains a good idea on what makes videoconferencing an artificial, sometimes frustrating (or smooth), immersive experience. While these dimensions are often naturally supported in face-to-face situations, videoconferencing solutions don't yet support the same experience. We are adopting Bondareva & Bouwhuis's (2004) and Bondarevera, Meesters, and Bouwhuis's (2006) social presence criteria. The main factors influencing the experience can be expressed as: mutual gaze (direct eye contact is preserved), a wide field of view (FOV), the display of a life-sized upper body, a high video and audio quality (including a high-quality image and correct color reproduction, audio with high signal-to-noise ratio, a directional sound field, and minimized or nonexistent video and audio signal asynchrony), the availability of a shared working space, and the way the videoconferencing partners are positioned in relation to the equipment and their partners (see Figure 2).

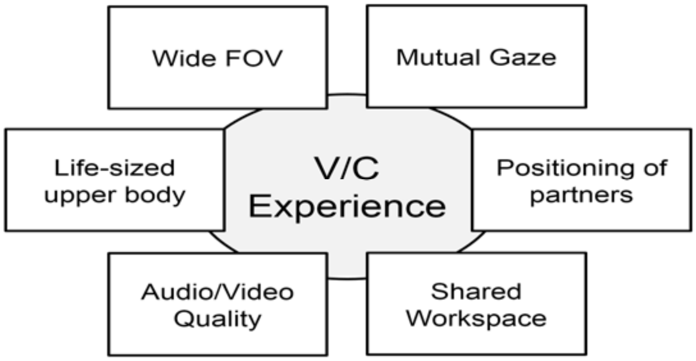

Figure 2. Factors Influencing the Videoconferencing Experience

The relative importance of these criteria depends on the communication task and the nature of the videoconferencing meeting. For instance, a directional sound field might be very important in a situation with a high number of participants but be less important for an ad-hoc, face-to-face meeting. Nevertheless, direct eye contact plays an important role in a wide range of videoconferencing situations. Indeed, many researchers who have investigated the use of different multi-party videoconferencing systems (Buxton, Sellen, & Sheasby, 1997) list "establish eye contact with other participants" as one of the most important criteria.. However, we should also consider influencing factors such as individuals' life-size appearance or the availability of a shared working space when it comes to non-verbal, gestural communication. Meeting situations with higher levels of collaboration effort require the integration of the communication with the collaboration space. For instance, the meeting table space found in physical meetings is used to define individuals' physical and social positioning, to support gestural and postural communication, and to share meeting artefacts such as documents or physical prototypes. The displayed and perceived size of an individual and the collaboration environment and the way people are positioned in that (virtual) environment determine the way that social presence develops. Also of interest here are gender and cultural factors and the familiarity with videoconferencing meetings and setups in general.

## 2.1 Life-sized Upper Body and Positioning of Partners

Related to the issue of eye-to-eye contact is an individual's perceived scale of their communication partner. This scale can range from miniature- to poster-size, which will influence whether the partner will be perceived as convincing and whether eye contact can be maintained. On one hand, Okada, Maeda, Ichikawa, and Matsushita (1994) found that the size of one's communication partner on screen is an important factor for achieving a sense of reality. If one's partner is presented smaller than life-size, that individual might be perceived as far away. Also, it is difficult to read facial expressions or gestures. On the other hand, a larger-than-life-sized communication partner in videoconferencing implies dominance. Buxton (1992) suggests that social relationships, such as power, may be more balanced and natural in life-sized video conferencing. Detenber and Reeves (1996) found that the display size has an effect on people's arousal, and Lombard (Lombard, 1992, cited in Detenber & Reeves, 1996) found that people evaluated others more positively when presented on large screens (in the right size). There are commercial systems that support life-sized videoconferencing (e.g., business solution such as LifeSize[6] or the Cisco Telepresence[7] series).

Related to the perceived size of the videoconferencing partners is their seating arrangement in group videoconferencing sessions. While in 1:1 settings, individuals would normally be facing each other directly, potentially with some interaction space (virtual and/or real) between them, in group settings, the individuals' positioning influences the support for non-verbal communication, partner and workspace awareness, possibilities for interaction and collaboration, and social-psychological balance of dominance. Yamashita, Hirata, Aoyagi, Kuzuoka, and Harada (2008) investigated two different versions of seating positions in 2 (local) x 2 (remote) table-centered, life-sized video communication. They found, among other aspects, that a side-by-side seating (remote-remote, local-local) is preferable regarding balance of turn-taking and sense of unity.

Fish, Kraut, and Chalfonte (1990) present an early room-sized VC system built to provide an ad-hoc, informal, as-if face-to-face communication channel. They do not describe implementing gaze and eye contact in technical detail, and they likely didn't achieve these features at all. However, they did achieve a life-sized video image with a wide NTSC camera and view. Also, Gibbs, Arapis, and Breiteneder (1999) present their early TELEPORT concept and partial implementation of a room-to-room videoconferencing system that allows for eye contact and gaze awareness in a combination of a virtual environment with captured video feeds, compositing, tracking, and 3D projection. More recently. other research groups have tried to achieve or support eye contact through collaborative virtual environments. Wolff et al. (2008) provide gestural communication and gaze-contact in an avatar-based, collaborative CAVE environment. Users equipped with head- and eye-trackers are communicating with each other in that CAVE system.

Gamer and Hecht (2007) emphasize the importance of observer distance, head orientation, visibility of the eyes, and the presence of a second head at the perceived direction and width of the gaze cone in videoconferencing. Also, we are less sensitive to eye contact when people look below our eyes than when they look to the left, right, or above our eyes. Additional experiments support a theory that people are prone to perceive eye contact. That is, we will think that someone is making eye contact with us unless we are certain that the person is not looking into our eyes (Chen, 2002). These aspects help to mitigate the effects of poorly configured videoconferencing.

Vertegaal (1999) and Regenbrecht et al. (2004) use video planes in three-dimensional space to indicate gaze direction but do not allow for actual eye contact. Almost eye contact could be achieved by the "reciprocal video tunnel" (Buxton & Moran, 1990) using a half-silvered mirror and a miniature user node.

Ishii, Konayashi, and Grudin (1993) support eye-contact between two parties with their Clearboard system, where users are standing in front of an acrylic glass drawing board with cameras and projections placed behind. Here, the emphasis is on the collaborative work on the clear board as such and less on the quality of the video communication; this approach can even tolerate significant video artefacts.

## 2.2 Other Confounding Aspects

Implementing and perceiving eye contact is complex. Some aspects can be controlled technologically or organizationally, while others have to be considered but cannot be so controlled. Heaton (1998), for instance, explores cultural aspects in Japan of eye gaze-allowing CSCW systems; namely, ClearBoard and MAJIC (unfortunately, MAJIC was never actually used, not even in the laboratory). Given the constantly fluctuations and redefinitions involved in any activity out of the ordinary, they view the task of

[6] www.lifesize.com

[7] www.cisco.com/en/US/products/ps7060/index.html

trying to support “delicate” communication, such as negotiation, as an impossible one. Eye-contact might also be considered to be rude.

Swaab and Swaab (2008) explore gender aspects and eye contact and found in their study that unacquainted females have a better agreement with eye-contact in contrast to males, for whom this leads to the opposite effect (no eye-contact leads to better agreement). Also, Teoh et al. (2011), Teoh, Regenbrecht, and O’Hare (2012) and Hauber, Regenbrecht, Cockburn, and Billinghurst (2012) found significant gender effects in videoconferncing sessions.

In general, if people are using videoconferencing frequently, they might learn to interpret gaze direction to a very high degree of accuracy if the equipment is configured optimally (Grayson & Monk, 2003). This is helpful when addressing objects in the environment but does not necessarily provide the perception of eye-to-eye contact.

# 3 The Impact of Mutual Gaze

In Section 2, we demonstrate the importance of mutual gaze for the quality of a videoconferencing experience. But we need to know what influence mutual gaze has on different dimensions of interest in video-mediated communication and collaboration.

Psychology researchers identified the importance of gaze in the second half of the 20th century. Indeed, as Cook (1977) notes, “How long—and when—we look “in the eye” is one of the main signals in non-verbal communications”. Thereby, gaze refers to looking at another person’s upper body or, sometimes, more specifically between the eyes. Consequently, mutual gaze refers to looking into each other’s eyes (Cook, 1977). In this paper, we use the term mutual gaze and eye-to-eye contact synonymously. We further consider gaze in a wider sense of looking at the other’s upper body but also including the importance of gaze awareness, which sometimes can be close to eye contact. In this section, we investigate the effect and importance of gaze in telepresence applications (see Figure 3).

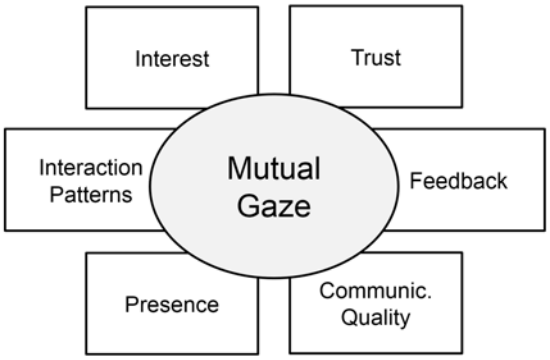

Figure 3. Factors Impacted by Mutual Gaze

## 3.1 Interest and Trust

Gaver (1992) points out that “video is anisotropic, interfering with the design of communicative gesture and with gaze awareness”. Apart from the lack of a shared media space in VC, we need to compensate for the lack of eye contact (to facilitate turn-taking, indicate interest, and reflect social relations) in them.

Fox (2005) stresses the importance of eye gaze to indicate another’s person intentions, interest in the conversation, and so on. In business meetings (and other “non-chat” situations), such factors are highly important. Relationships involving complex tasks can be maintained by increasing the frequency and flow of communication (McKinney & Whiteside, 2006), which requires gaze and mutual eye contact. Mukawa, Oka, Arai, and Yuasa (2005) show that systems using communication with eye contact induced behavior similar to face-to-face communication. In interview situations, for instance, researchers have identified perceived eye contact and mental workload as issues when using videoconferencing (Ferrán-Urdaneta & Storck, 1997).

In experiments with half-silvered mirrors, Quante & Muehlbach (1999) show that users had significantly more often the feeling of being addressed (i.e., of being looked at and recognized that they were addressed when eye-contact was provided). We look at individuals’ faces to continuously make sure that they are still with us in the meeting and that we can trust them. Predictive valid faces (targeting the partner) appear more trustworthy (Bayliss & Tipper, 2006). Bekkering & Shim (2006) found that the absence of eye-to-eye contact in videoconferencing systems is the main factor for the lack of trust (as defined by Wheeless & Grotz, 1977): “People associate poor eye contact with deception” (p. 103). Furthermore they argue that this is a main reason for why organizations have not adopted the technology at a large scale. Also, Teoh (2012) shows that eye contact and the availability of body language affect perceptions of trust, social presence, dominance, and impression management.

## 3.2 Feedback and Communication Quality

In a study on rural psychotherapy in Norway, Sorlie, Gammon, Bergvik, and Sexton (1999, p. 458) compared different types of distant delivery: “Most participants reported reduced eye contact, less nuances in mimics and other nonverbal cues, and a corresponding increase in dependency on verbal cues under V/C conditions. Some trainees felt that this reduced their ability to monitor how the supervisor reacted towards their presentations.”. However, as O’Malley, Langton, Anderson, Doherty-Sneddon, and Bruce (1996) point out: in comparison to only audio, video can lead to interruptions in communication flow. Co-presence and high bandwidth are important. A lack of or low degree of social presence might lead to an increase in verbal and non-verbal communication as a form of overcompensation for missing confidence in mutual understanding. Gaze serves a feedback function (i.e., communication partners try to elicit feedback from their listeners). The interplay between audio and video, with or without eye-contact, is much more complex though. For instance, people sometimes over-use the visual channel, which leads to an increase in cognitive load that results in redundant verbal communication.

In a remote, collaborative design situation, Olsen, Olsen, and Meader (1995) found that video with gaze awareness versus only video and audio, gaze video was superior in terms of discussion quality and time (less clarifying discussions). Also, Vertegaal and Ding (2002) experimentally investigated the effect of different types of gaze and tasks on the (positive) quantity of turn-taking with positive results.

## 3.3 Interaction Patterns and Presence

Joiner, Scanlon, O’ (2002) conducted three experiments in remote learning and discussion situations (physics and statistics) as part of a larger project investigating technologically mediated collaboration in learning. They found that eye contact influenced problem solving and interaction patterns and that eye contact facilitated conceptual understanding. Mukawa et al. (2005) present a study comparing eye contact versus non-eye contact and found that eye contact induced a similar behavior to face-to-face communication.

Even if we cannot see a real or video-mediated representation of our partners (for instance, in Second Life-like conferencing), gaze awareness with avatars (here 3D, cartoon-like representations of humans) is important (Garau et al., 2003). Garau, Slater, Bee, and Sasse (2001) conducted an experiment testing the influence on avatar gaze on different measures. Informed (inferred, synchronized, conversational) gaze (male and female avatars) outperformed non-gaze and random gaze in many measures, such as involvement and co-presence. Bailenson, Beall, and Blascovich (2002) present another experiment investigating the importance of gaze in avatar-based collaborative virtual environments (CVE). They found that avatar head-movements play an important role in terms of higher levels of reported co-presence and positively changed patterns of interacting with each other.

Mutual gaze (as close as this can be achieved with standard desktop videoconferencing) has a significant effect on social presence and communication experience for children when playing games. The absence of mutual gaze dramatically decreases the interaction quality (Shahid, Krahmer, & Swerts, 2012).

Fullwood and Doherty-Sneddon (2006) conducted two experimental studies using a sales pitch for a cosmetic product and found that the absence of gaze had a negative impact on information recall. This finding might have consequences in many areas, such as remote teacher-student learning environments, and might lend itself to the conclusion that it is better to have an only audio channel compared to a video-audio system without eye-contact (or one needs to use workarounds such as artificially speaking into the camera).

Carville & Mitchell (2000) investigated videoconferencing used in teaching and learning in early childhood studies with early years’ tutors. Students and tutors developed skills and strategies to deal with VC’s shortcomings: “Many of the tutors identify the need to remember to keep looking into the camera to make eye contact with the…[other side]”. Birden & Page (2005) report a similar workaround in situations of remote health education where they instructed the participants to look into the camera when speaking: “The lecturer must remember that to give students at the far site the impression that s/he is making eye contact; they must look into the lens of the camera, not at the screen.”. In the same realm, as part of their “twelve tips for teaching using videoconferencing”, Gill, Parker, and Richardson (2005) advise “eye contact with the distant site is important; this can be simulated by talking to the lecturer camera… We have found a bright balloon tied to the lecturer camera encourages the lecturer to remember the camera is there.”. In a VC environment to teach and learn business French, students were told to look into the camera while speaking to compensate for the lack of eye-contact. This workaround was mainly successful, but, in a few cases, it also lead to distraction (McAndrew, Foubister, & Mayes, 1996).

In summary, videoconference systems offer high-quality audio and video performance but still cannot transport non-verbal communication queues such as eye-to-eye contact and have perception issues (e.g., an incorrect scale). All these factors add to the artificial experience that can arise from existing videoconference solutions.

## 4 Implementation Approaches

In this section, we present seven different categories that represent different approaches of how to implement videoconferencing solutions that support eye-to-eye contact. To our knowledge, these categories virtually account for all systems that support mutual gaze that research and the market report on today. For each category, we describe the core idea and technology, give representative references where appropriate, and discuss advantages and disadvantages including guidelines for their application.

We discuss the systems with respect to several aspects that are crucial when setting up such a system: for example, the system’s space requirements and the VC quality that can be achieved. We further discuss the hardware and software requirements, the support of a flexible posture, setup price, integration effort, and the availability of the components.

One of the most obvious criteria for deciding on a particular videoconferencing technology is the space available and need for it. These requirements range from a desktop space in an office or simply a mobile or portable device (e.g., laptop computer) to full room-sized spaces. We consider systems as positive if they do not require a lot of space even if, under some circumstances, this might be seen as a negative and vice versa; for instance, delivering an often desirable life-sized video representation cannot be achieved on a small screen. However, usually, the less space required the better.

The achievable quality depends on many factors and is mainly influenced by a system’s ability to deliver high resolution, high frame rate, and correct color video, preferably life-sized, with no or only few visual artefacts, and a high-quality audio channel optimized for speech frequencies and speaker awareness.

We consider more complex technology more negatively than less complex technology. The simpler the better. More complex systems tend to be more expensive, less mature, more error-prone, harder to maintain, and harder to set up, configure, and use.

The hardware and software requirements are linked to a system’s complexity and describe how demanding the particular solution is. For instance, on the one hand, a PC-based videoconferencing system delivering eye contact by a complex algorithmic software solution would have very little hardware requirements (a PC will do), but rather specialized software is required to achieve this. On the other hand, a hardware solution for eye contact such as a beam-splitter would require special silvered glass arranged in a specific way but would not require any special software—any standard videoconferencing application would do.

Some solutions require users to sit in a well-defined position in front of the camera or screen. We consider such a requirement as negative. Other systems would allow for a wide range of flexible movements. We consider such flexibility as positive.

We consider a low investment price, market availability, and ease of integration into existing videoconferencing solutions and infrastructures as positive.

We can divide the approaches to implement mutual gaze into three groups: (1) custom hardware dependent setups for videoconferencing (hole in screen, long distance/small angle, half-silvered mirror), (2) custom software dependent setups for videoconferencing (2D video-based techniques, 3D video-based techniques), and, finally, (3) videoconferencing setups dependent on both custom hardware and custom software (shuttered screen, unshuttered screen).

## 4.1 Hole in Screen

The naïve, obvious solution is to drill a hole in the screen exactly at the desired position and place a camera there for eye-to-eye contact. Obviously, this is not a suitable technique for CRT or LCD monitors but can be implemented with a screen canvas and a projector (Figure 3). Back projection is generally not possible because of the size and position of the camera: it would cast shadows on the back projection screen resulting from the back-mounted camera.

The opening and the camera should be as small in diameter as technically possible and the rim of the camera should be painted in the canvas screen color to minimize the camera's visibility, however, the hole will still be visible for the user. Even worse, the hole is visible at a position that users pay most of their attention to - between the eyes of the videoconferencing partner. Even with small (i.e., in the order of 5 mm in diameter) cameras, the user has to be positioned decently far from the screen to mitigate the disturbing effect of the hole-between-the-eyes effect, which directly affects the overall size of the setup.

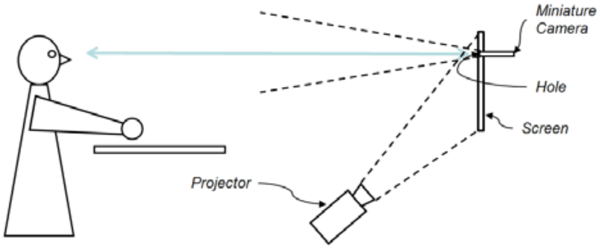

Figure 4. Schematic of Hole-in-Screen Approach

Despite the need for a frontal projection, the hole-in-the-screen approach allows for an unrestricted interaction space in front of the screen (see Figure 4). It can also be easily implemented: all it needs is a projector, an inexpensive canvas (should be inexpensive because one has to drill a hole in it), and a small camera. The main disadvantages with respect to the quality are (a) the visible artefact of seeing a (black) spot between the eyes and (b) the user has to be rather far away from the screen, which requires much more space and also determines the achievable size of the video face display, to minimize the spot effect.

Table 1. Overview of Hole in the Screen Approach

| | Size | Quality | Flexible posturing | Complexity | Hardware requirements | Software requirements | Price | Availability | Integration | Examples |
|---|---|---|---|---|---|---|---|---|---|---|
| Hole in screen | - | - | + | ++ | + | ++ | + | + | + | No example in literature |

In Table 1 and all following tables, we use "+" and "-" to represent each system's advantages and disadvantages. They are not meant to be absolute measurements; that is, readers should not interpret "++" to mean as twice as good as "+" but rather as "significantly better". Also "++" should be read as "most positive" and "--" as most negative; hence, if choosing a VC solution, more "+" and fewer "-" is desirable.

## 4.2 Long Distance/Small Angle

If the room size permits, eye-to-eye contact can also be implemented by viewing the screen from a far distance and by placing the camera as close as possible (i.e., at the edge of the screen) to the displayed video stream. If the angle between the viewing axis and the eye axis (ß in Figure 5) is small enough, then the offset between eyes and camera isn't noticeable.

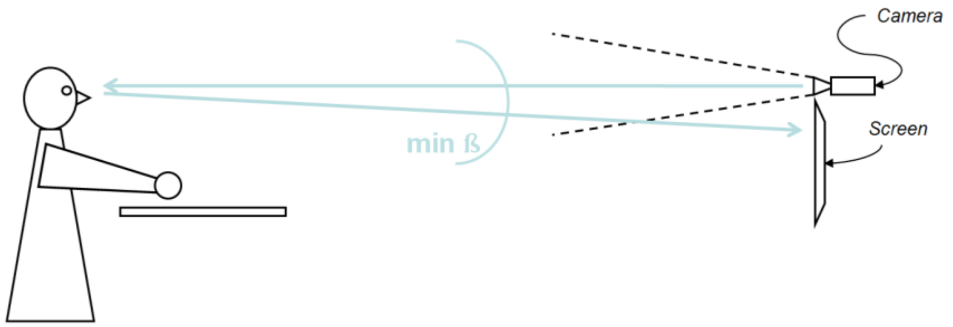

**Figure 5. Schematic of Long-distance Approach**

Tam, Cafazzo, Seto, Salanieks, and Rossos (2007) found that, in medical tele-consultation, doctors sit rather close to the monitor (around 1m). The authors investigated the influence of eye-gaze angle on perceived eye contact. Comparing scenarios with ß = 7deg vs. ß = 15deg and different VC environments (ranging from mobile devices to room-sized environments), they conclude that smaller angles/greater distances are better for perceived eye contact. Chen (2002) discusses the influence of eye-gaze angle on perceived eye contact more thoroughly. He summarizes that the eye-gaze angle is an asymmetric issue in a way that it depends on the general head position with respect to one's conference partner. Consequently, different angles have been reported in the literature. Researchers often state that the acuity of eye contact is as good as the visual acuity (1 minute of arc), which means that the eyeballs are rotated by ~2.8°. These results match experiments in our laboratory that show that ß should be not much greater than 3 degrees, and, with this, a rather long distance is needed to achieve the desired effect. Hence, if one wants to present a life-sized head (and only the head) and the camera is placed as close as possible to the rim of the display, a distance of at least 3 meters is required (atan (160mm/3000mm) = 3.05°) if we assume an average adult head size. Even if one can free that much space, the face of the videoconferencing partner appears rather small. Increasing the screen size to display a bigger face or a bigger portion of the partner's body would increase the angle ß and, consequently, requires an even longer distance. However, because of the asymmetric character, one can demonstrate that the vertical threshold is up to 5° before a deviation can be noticed (Chen, 2000). Also, Gale and Monk (2000) show that users are quite accurate about guessing gaze from videoconferencing streams with an error of estimation of only a couple of degrees.

Nguyen and Canny (2007) present a special form of the long-distance approach in their work on multiview group videoconferencing. They use a screen and camera with a rather large distance to the viewer. However, they focus mainly on generating a personal view for each user even though they see on the same projection surface, which they achieve by using a specific retro-reflective material reflecting the projection only back in the horizontal direction of the source while using a vertical diffuser to allow a varying vertical height.

In summary, the long-distance technique is the most affordable one and does not require specialized equipment and calibration. Similar to the Hole in screen approach there is enough room for interactions in front of the display and there are no visual artefacts. However, this technique requires a lot of space. The distance needed between the user and the display does not allow for close communication between the partners, this can only be achieved by much bigger than life-size displays, which is undesirable in most cases.

**Table 2. Overview of Long Distance/Small Angle Approach for Establishing Eye-to-Eye Contact**

| | Size | Quality | Flexible posturing | Complexity | Hardware requirements | Software requirements | Price | Availability | Integration | Examples |
|---|---|---|---|---|---|---|---|---|---|---|
| Small angle | -- | + | - | ++ | ++ | ++ | ++ | ++ | ++ | Chen (2002) |

## 4.3 Half-silvered Mirror

Using a half silvered-mirror or any other kind of an optical beam-splitter is probably the most commonly used solution in research and the market to achieve eye-to-eye contact. In the literature, several similar systems that use a half-silvered mirror exist for creating a video conferencing experience with eye-to-eye contact that can be implemented by either placing the camera above or below the mirror and the screen behind the mirror or the other way round by placing the screen above or below and the camera behind the mirror. Figure 6 shows one of the possible but most common configurations.

With his courtroom conferencing system, Kannes (1990, 1995) presents an early system that uses a half-silvered mirror. This system allowed courts to interview remote defendants and witnesses while maintaining eye-to-eye contact. The basic idea is that a user can see through half-transparent mirror while being observed by a well-positioned camera at the same time. Nelson and Smoot (1992) describe a similar system but add polarizers that mitigate the effect of light entering the camera from the screen used to display the conferencing partner. Large, Rosenfeld, and Emerton (2009) also use polarizers to avoid light passing from the display into the camera, but, contrary to the existing solutions, they do not use one large half-silvered mirror but many smaller ones that (similar to a lens array) are attached to the display and reflect a portion of the incoming light to the camera mounted in front of the screen. This greatly reduces the required space but at the cost of visual quality (discontinuities caused by not properly aligned mirrors) and production costs for the mirror array.

McNelley and Machtig (1999) present a series of improvements for setups not using polarizers. First improvements improved the size of the overall setup by introducing more optical elements (beam splitter, mirrors) to the system together with a projector that consequently allowed them to reduce the distance of the projector to the mirror. In later setups, they used a screen that was in front of the mirror. This setup made the system even smaller but required a beam splitter that avoided direct view into the screen (McNelley, & Machtig, 2001). Another approach integrated one’s conferencing partner into the environment by also capturing the environment and blending it with the displayed image of the conferencing partner for the cost of an increased system size (McNelley, & Machtig, 2004). These systems were static setups optimized for eye-to-eye contact. McNelley (2000) also introduced a dual mode system that permits normal use of the display because the eye-to-eye component (half-silvered mirror and camera) could be conveniently folded away when not used. Similarly, Libbey (2004) created a system integrating a half-silvered mirror and a normal mirror into a box that can be attached to a normal desktop. Assuming that the camera is placed on top of the screen, it aligns a portion of the screen showing one’s conference partner with the view of the camera. The advantages of this system were its size and the fact that the system could be easily added to normal hardware and be removed when not used, while its main disadvantage were the limited size of the video.

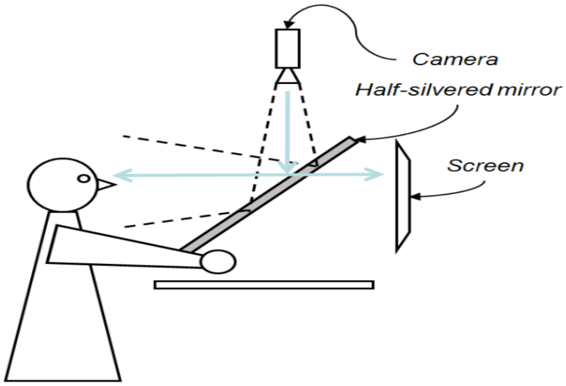

Figure 6. Schematic of Half-silvered Mirror Approach (Kannes, 1990, 1995)

The space in front of the user (where usually the desk is) provides only limited access because of the half-silvered mirror placement. However, careful positioning, akin to ReachIn[8] setups, can produce an interesting interaction space where virtual objects can be blended with manual interaction (augmented reality interaction space).

The main advantage of half-silvered mirror systems lies in the simplicity of the setup: it needs only the mirror and a standard monitor and camera. It needs careful calibration though and usually produces optical artefacts due to the fact that the camera captures only half (or a certain percentage) of the user’s

[8] en.souvr.com

true image. The use of beam splitters can reduce the effect but not avoid it while also further reducing the brightness of the camera image. In fact, besides unwanted reflections, the maximum achievable brightness and contrast levels might be problematic. The other main disadvantages are the high price and limited availability of large enough half-silvered mirrors needed for life-size displays that is shared by all approaches but those who use only small screens. Furthermore, the space in front of the user is occupied with the display setup. While this is not so important with the previously introduced approaches, here and with the following systems, the user should be placed in a way that the camera directly captures the eyes without too much deviation from the ideal spot.

Nowadays, there are also commercial solutions available, normally with smaller screen sizes (e.g., iris2iris[9]), that make use of half-silvered mirrors.

**Table 3. Overview of Key Techniques Using a Half Silvered Mirror Approach for Establishing Eye-to-Eye Contact**

| | Size | Quality | Flexible posturing | Complexity | Hardware requirements | Software requirements | Price | Availability | Integration | Examples |
|---|---|---|---|---|---|---|---|---|---|---|
| Half-silvered mirror | + | + | -- | + | - | ++ | - | - | + | Kannes (1990, 1995), Libbey (2004), Mukawa et al. (2005), Quante & Muehlbach (1999) |
| Half-silvered mirror and polarizers | + | + | -- | - | -- | ++ | - | -- | - | Nelson & Smoot (1992) |
| Half silvered mirror and beam splitter | + | + | -- | - | -- | ++ | -- | -- | - | McNelley, & Machtig (1999, 2001, 2004) |
| Note: we focus on systems supporting life-sized videoconferencing. While reducing the size decreases the costs mainly driven by the large half-silvered mirror, ALS heavily affects the video conferencing experience. | | | | | | | | | | |

## 4.4 2D Video-based Techniques

So far, the presented approaches to support eye-to-eye contact are hardware solutions that require software for calibration. Contrary to this, 2D video-based approaches apply a software-based approach to support eye-to-eye contact. In this section, we discuss 2D approaches that work without 3D scene knowledge. We present 3D video-based techniques in the next section.

The literature shows different approaches for implementing eye-to-eye support using 2D video approaches. The first category of existing works uses a single camera. They compute the position of the pupil using computer vision techniques and apply image operations to remodel the pupil in such a way that it appears that they look straight into the camera. Example images, if provided, illustrate this approach but are not convincing because of the amount of visual artefacts. Andersson, Chen, and Haskell (1996) were the first to claim a conceptual solution for this kind of approach. Their idea was to use light reflections to detect the iris. Once detected, a 3D ellipsoid facial model is textured using the camera feed and finally re-oriented to simulate eye-to-eye contact. Andersson presents a simplified idea of the previous work that reorients the eyes by only shifts iris and eyelids in the 2D camera image (Andersson, 1997). Jerald & Daily (2002) also shift the iris but apply a non-linear warp. However, their work requires prior calibration to determine the maximum iris offset. This calibration creates a sweet spot in which the user has to stay and, therefore, greatly reduces one’s ability to have a flexible posture in front of the screen.

[9] www.iris2iris.com

Cham, Krishnamoorthy, and Jones (2002) present an image-based approach using epipolar geometry. However, their approach was limited to small corrections because larger corrections introduced artefacts due to the lack of knowledge about the face's shape.

Some approaches use multiple cameras while still relying on image-based operations. Lewis (1994) conceptually introduced this basic idea in 1994. Here, two or more cameras are positioned on the sides or corners of the screen (Figure 7). The closer the cameras can be positioned to the targeted screen position, the better the results (usually).

Criminisi, Shotton, Blake, and Torr (2003) used two cameras and applied a smart blending without 3D reconstruction. All pure image-based approaches require carefully calibrated cameras and are very sensitive to lighting effects in the user's image. Furthermore, it is difficult to obtain convincing results for larger deviations in the viewer's gaze. The results are also not backed up by studies that provide evidence of the results and applicability.

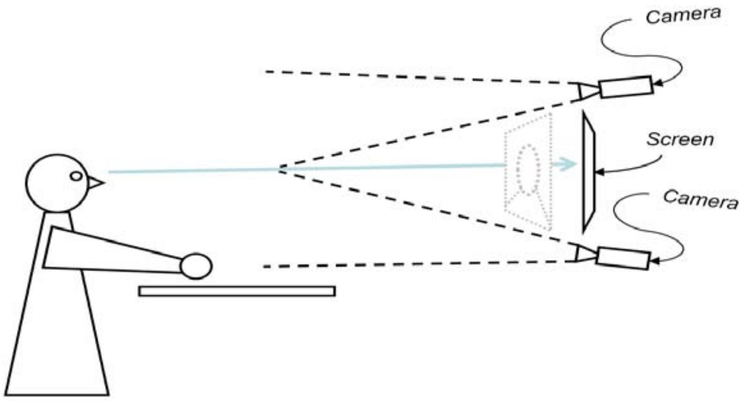

Figure 7. Schematic of 2D Video Interpolation Approach

To overcome these problems, some groups have worked with proxy geometry and model-based approaches for synthesizing images. Gemmell, Toyama, Zitnick, Kang, and Seitz (2000) and Zitnick, Gemmell, and Toyama (1999) applied image warps by using a computer-vision and image synthesis-based approach to redirect eyeballs and reorient the displayed head (Figure 8). Zitnick et al. (1999) report: "For small angles of rotation (<5 degrees) we were successful in warping a person's head to a new orientation. For larger changes in rotation, realism was lost due to distortions." (p. 6) Hence, this approach is suitable for rather small head rotations only. Similarly, Yip & Jin (2003) developed another approach that applies image-warping techniques for warping the iris into the final camera image. Later, Yip (2005) extended this system by also training an artificial neural network that gave additional input to remodel the user's head and apply image warping. While such a system makes spontaneous videoconferencing impossible, it also shows problems with rapid head movements in that they lead to an artificial experience due to poor visual results.

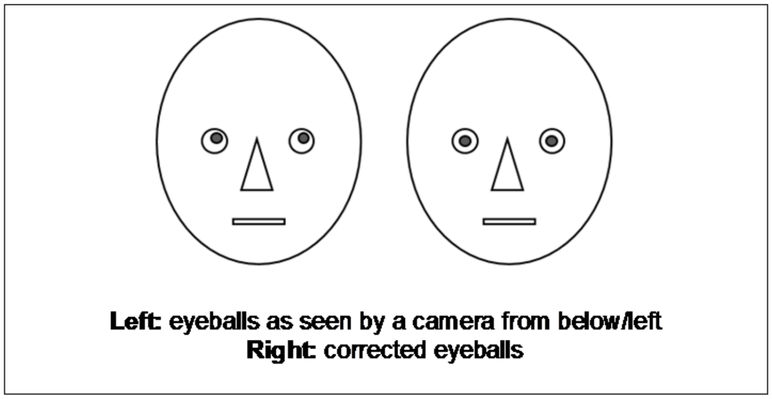
Left: eyeballs as seen by a camera from below/left
Right: corrected eyeballs

Figure 8. Eyeball Correction

Synthesizing videoconferencing images is a vivid area of research (Ott, Lewis, & Cox, 1993) that range from one or multiple camera systems that artificially replace the eye gaze in the video stream (Jerald & Daily, 2002; Gemmel et al., 2000; Tsai, Kao, Hung, & Shih, 2004; Schreer et al., 2008; Vertegaal, Weevers, Sohn, & Cheung, 2003) to systems that combine a half-silvered mirror with multiple cameras for multi-party videoconferencing (Vertegaal et al., 2003).

Test implementations and studies in our laboratory have shown that one should not expect perfect interpolation results (Wetzstein, 2005). Humans are very sensitive when it comes to subtle artificial elements in other faces. Even the slightest artefacts will be noticed and eventually destroy the eye-to-eye illusion.

Videoconference systems relying on video-interpolation to compute a synthetically frontal video allow for life-sized display and close proximity eye-to-eye contact. Due to their compact size (usually a screen with two or more cameras), they do not occupy the space in front of the screen and, therefore, are quite flexible in terms of positioning. The visual display parameters such as brightness, contrast, and color can be easily controlled, which makes it also a suitable approach for complex environments (e.g., complex lighting environments).

Contrary to many other systems that use a fixed positioned camera and, therefore, require a fixed posture, systems using video-interpolations for computing the view of a virtual camera can tolerate to some extent movements of the communication partner's head. This requires tracking the head of the communication partner and repositioning the virtual camera.

However, the movement has to be within certain limits to still be able to compute an error-minimized image of the communication partner. Furthermore, solutions using video interpolation require specialized and often expensive hardware and software. Due to the complexity of the vision-based computation of the virtual camera, these approaches also introduce perceivable artefacts while realizing eye-to-eye contact, which can be very distracting.

**Table 4. Overview of 2D Video-based Techniques For Establishing Eye-to-eye Contact**

| | Size | Quality | Flexible posturing | Complexity | Hardware requirements | Software requirements | Price | Availability | Integration | Examples |
|---|---|---|---|---|---|---|---|---|---|---|
| Image-based techniques (single camera) | + | -- | - | -- | + | -- | + | + | ++ | Andersson (1997),<br>Jerald & Daily (2002) |
| Image-based techniques (multiple cameras) | + | - | + | -- | - | -- | - | + | - | Criminisi et al. (2003) |
| Model-based techniques | + | - | + | -- | + | -- | + | + | ++ | Gemmell et al. (2000) |

## 4.5 3D Video-based Techniques

The previously presented approaches rely exclusively on 2D image operations working on one or more camera feeds to synthesize a new view. A logical step forward is to use several camera images to reconstruct a 3D model of the user's head. Some systems even go a step further by also integrating a stereoscopic display into the system, which allows for a 3D telepresence experience with eye support. In this section, we present several approaches using 3D reconstruction, some of them also with a stereoscopic 3D display.

Xu et al., (1999) present a system using two cameras for stereo tracking and stereo analysis of image features representing the user's head, through which they could build a 3D model of the head and create a synthesized view supporting eye contact. However, it required several calibrated cameras and was prone to visible artefacts resulting from falsely matched image features. Yang and Zhang (2002) present a similar system but use different image operations and also apply a Delaunay triangulation between the matched image points to compute the 3D model. However, the system shows similar remaining artefacts in particular between the background and the synthesized view of the users' head.

To improve the visual quality of reconstruction (e.g., closing holes and increasing the resolution of the underlying depth information) Zhu, Yang, and Xiang (2011) propose using stereo cameras and fuse their results with a time-of-flight depth camera.

Kuster et al. (2012) present an approach that uses a commodity 3D depth camera (MS Kinect). Using a 3D depth map and relying on a single depth camera introduces the problem that the resulting view has holes due to occlusions and errors in the measured depth. The issue is noticeable particularly in the reconstruction of the background and distance objects. In their work, Kuster et al. only create a synthesized view of the face and blend this in the original image to overcome these problems in image integrity.

Some existing works do not focus on creating eye-to-eye contact but on allowing a full telepresence experience, which usually includes reconstructing the environment in addition to reconstructing the communication partner's face. Petit et al. (2010) present one example of this approach that reconstructs the full human body of the communication partners. Because a full 3D model exists, these systems can usually also create a view allowing for eye-contact. Thereby, the quality largely depends on the used hardware (e.g., camera), system configuration (e.g., size of the overall system and distance to the camera), and used algorithms. Some approaches (e.g., visual hull computing) tend to blur the details, such as eyes, making it hard to detect gaze in the final reconstruction of the user's body.

Prince et al. (2002) created live 3D reconstructions of videoconference partners by using a camera array and visual hull algorithms. The dynamic 3D model is visualized at the partners' location and can be used in a mixed reality-based videoconferencing setup. They do not demonstrate mutual gaze but it could possibly be achieved.

Many systems for supporting telepresence rely on expensive hardware such as hardware-synced cameras running in specifically prepared environments and often even having multiple computers for the reconstruction. Maimone and Fuchs (2011) present a system for telepresence that uses off-the-shelf hardware such as multiple Microsoft Kinects. The system uses several Kinects simultaneously and merges their depth information into one model. While the results show visible artefacts resulting from error in the reconstruction, enough details are preserved to identify gaze.

A more technically complex approach combines a real-time 3D face scanner with a sophisticated projection system based on a rotating mirror, which gives a "true" 3D impression of the remote user's head in a hologram-like manner (correct for the nearest (visually tracked) viewer). The authors note convincing results, but the setup requires highly specialized equipment and knowledge. The same is true for telepresence robot systems, such as the BiReality robot surrogate with four displays arranged around a cube that Jouppi, Iyer, Thomas, and Slayden (2004) present.

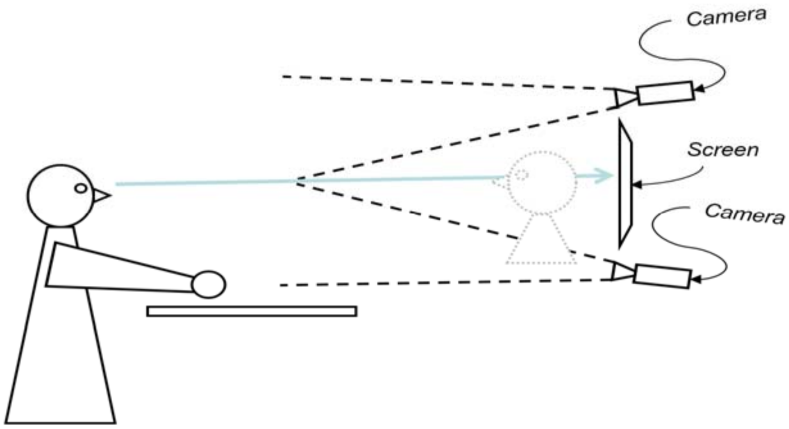

Figure 9. Schematic of 3D Interpolation Approach

As part of the VIRTUE teleconferencing research prototype, Kauff and Schreer (2002) describe an approach and implementation of a model-based approach to capture, transmit, and display participants, the environment, and artefacts in a table-based conferencing situation. Because the users' heads (and eyes) are captured in a 3D model, a view-independent presentation and with this eye-contact can be made possible. The presented pictures are impressive but still show significant signs of artefacts in the user's representation.

Most of the systems relying on 3D techniques for computing a synthesized view have similar drawbacks as 2D systems. They are sensitive to light conditions, and good results have often only been reported in controlled environments. Furthermore, many need dedicated expensive hardware such as cameras or significant computational resources. Fortunately, the rise of consumer level depth cameras such as MS Kinect have provided new research opportunities in terms of affordable systems.

**Table 5. Overview of 3D Video-based Techniques for Establishing Eye-to-eye Contact**

| | Size | Quality | Flexible posturing | Complexity | Hardware requirements | Software requirements | Price | Availability | Integration | Examples |
|---|---|---|---|---|---|---|---|---|---|---|
| 3D video-based techniques relying on video cameras only | + | - | + | - | + | -- | + | + | - | Xu, Loffler, Sheppard, & Machin (1999) |
| 3D video-based techniques using professional depth cameras | + | + | + | - | - | -- | -- | - | - | Zhu et al. (2011) |
| Kinect-based 3D techniques | + | + | + | + | + | -- | + | + | + | Kuster et al. (2012) |

## 4.6 Shutter Techniques

As first shown by the blue-C system (Gross et al., 2003) and later by the HoloPort system (Kuechler & Kunz, 2006), a camera can be placed behind a back projection screen to virtually see through the screen if the screen itself and/or the projection and cameras are shuttered (Figure 10).

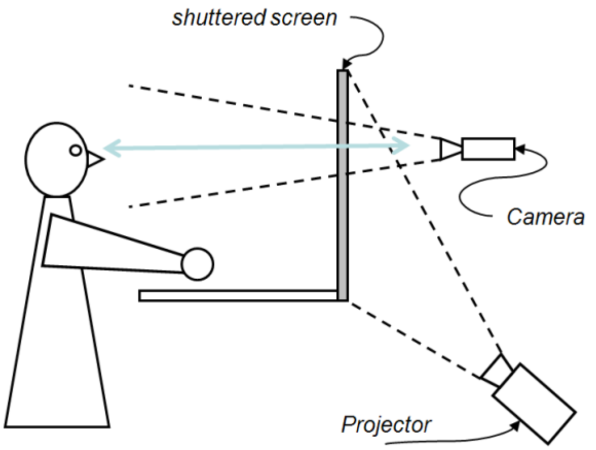

**Figure 10. Schematic of Shutter Screen Approach**

Between update cycles of the system, a black projection image is rendered and a camera image is captured. During that period of time, the screen is transparent because a shutter glass is used and triggered to transparent mode synchronously. Hartkop (2007) placed cameras behind an OLED screen. He synchronized the video camera(s) with the OLEDs illumination levels for the same purpose.

This type of installation allows for 1:1 scale videoconferencing but requires (expensive) instrumentation (i.e., shuttered glass). The shuttering (flicker) and the limited achievable transparency of the used screens introduce some artefacts in the displayed videoconferencing video though. Furthermore, the shuttered screen needs some time to fully switch its state and, depending on the screen quality, the direction and progress of switching the surface can cause visible artefacts.

In summary, all the approaches presented above are able to produce eye-to-eye contact in videoconferencing and a life-sized display of the communication partner to some degree. However, there is no optimal solution: The solution might be too expensive, require too much environmental space in front of the display, require too much distance, occupy the interaction space in front of the display, require specialized hardware and software, or produce poor visual quality (e.g., flicker, visible artefacts, and brightness).

**Table 6. Overview of Techniques for Eye-to-eye Contact Relying on Shuttered Screens**

| | Size | Quality | Flexible posturing | Complexity | Hardware requirements | Software requirements | Price | Availability | Integration | Examples |
|---|---|---|---|---|---|---|---|---|---|---|
| Shuttered screens | -- | + | + | - | -- | - | -- | -- | -- | Gross et al. (2003) |

Unshuttered Screen

Approaches in this category are usually based on the idea of the shuttered screen but try to minimize the visible artefacts caused by shuttering the screen and by the limited opacity of the screen. One approach is to replace the shuttered screen with a screen based on standard holographic optical element (HOE) screen (Tedesco, 1999). From certain predefined angles, the HOE will be opaque and can be used for projecting onto it, while, for other viewing angles, the HOE is transparent and a camera can be placed to record the user in front of the HOE screen. Similarly, Nelson & Vaning (1997) present the conceptual idea of using a transmissive layer in front of the screen with a micro-louver assembly. A similar but also not shuttered approach uses a “light transmissible screen”. A special film/material on the screen is used that allows only a certain percentage of light (e.g., MAJIC system by Okada et al. (1994)) to pass.

However, for all these systems using HOEs, the micro-louver assembly, or the ones using a film on the screen, we can still expect a considerable amount of unintended reflections and diffusions visible on the back of the screen resulting from the back projection, which the camera will capture. Regenbrecht et al. (2014) produced a conferencing system that uses a HOE-based screen but eliminates remaining reflections by using polarizing filters in front of the projector and the camera.

**Table 7. Overview of Techniques for Eye-to-eye Contact Relying on Unshuttered Screens**

| | Size | Quality | Flexible posturing | Complexity | Hardware requirements | Software requirements | Price | Availability | Integration | Examples |
|---|---|---|---|---|---|---|---|---|---|---|
| HOE or use of screen with layers/films attached | -- | + | + | - | -- | ++ | -- | -- | - | Tedesco (1999), Okada et al. (1994) |
| HOE with filters | -- | ++ | + | - | -- | + | -- | -- | - | Regenbrecht et al. (2014) |

## 5 Summary and Discussion

To date, no ideal solution that maximizes all desired qualities for a certain setting and scenario seems to exist. Half-silvered mirror techniques are well researched and studied and, therefore, lend themselves to be used in a variety of scenarios. 2D and 3D techniques are less well studied but offer good potential for future research and development. New algorithms for 2D/3D interpolation and first studies are in progress. System designers would need metrics to develop and optimize their systems though.

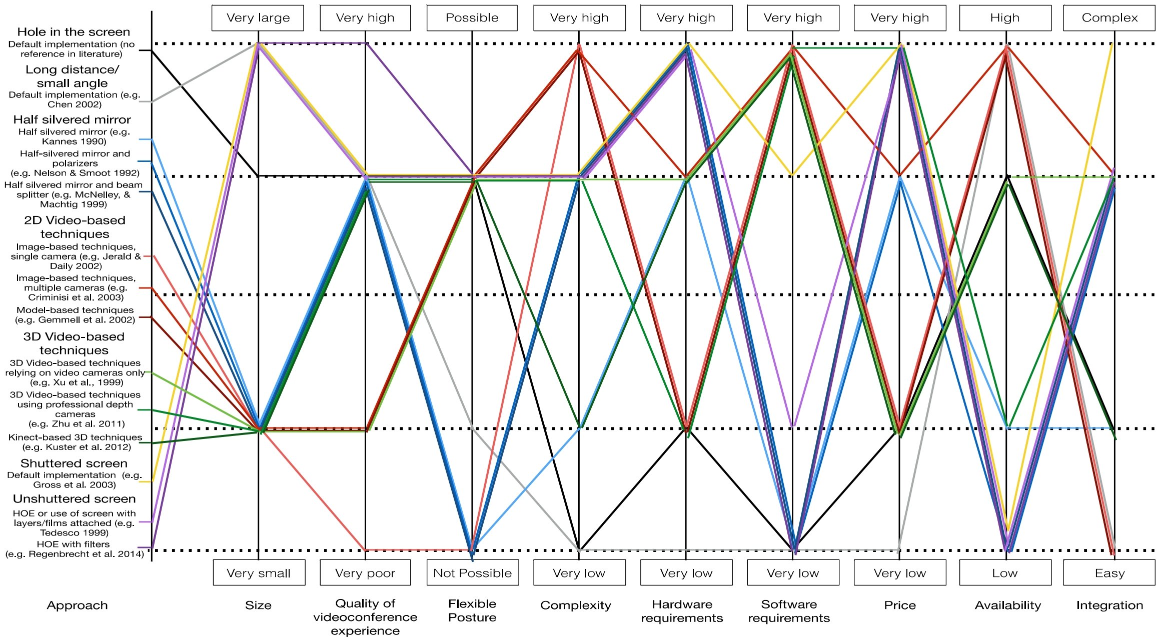

**Figure 11. Parallel Line Plot Visualizing the Main Approaches for Eye-to-eye Videoconferencing and Their Performance According to Our Criteria**

With Figure 11, we compile and overview our findings to give developers and decision makers a tool at hand for current and future technologies to implement mutual gaze in videoconferencing. This figure is not meant to replace the tables presented but rather act as a different view. The parallel line plot illustrates the diversity and relative strengths and weaknesses of the approaches. As one can see, there are a wide variety of different approaches with different strengths. However, all existing approaches have also major drawbacks. Usually, either the hardware or the software requirements are very high but still don’t guarantee perfect quality.

We show that there are different technologies for implementing eye-to-eye contact in video conferencing systems that differentiate themselves through the complexity of hardware and software. Overall, there are naïve approaches (hole in the screen and small angle) that are, despite their advantages in terms of system complexity and costs, rarely used. This might be mostly due to the general size and, in particular, to the screen distance that prohibits several use-cases. Still, in prototypical applications where the large screen distance can be compensated by a bigger screen or where the screen distance is not critical, these setups might be worth a try in particular due to their affordability.

Video conferencing setups supporting eye-to-eye contact using half-silvered mirrors are well researched and studied. They have been technically described and been used to research eye-to-eye contact and other perceptual aspects. Still, systems using half-silvered mirrors have downsides in terms of quality (usually ghosting and lower contrast) and price due to the need for the half-silvered mirror.

Video conferencing setups using video-based gaze control techniques such 2D interpolations or 3D reconstructions still have problems in terms of quality due to interpolation artefacts. We argue that these techniques probably have the biggest potential in terms of future research and possible improvements. Both of these approaches, 2D and 3D, heavily rely on real-time computer vision techniques and, consequently, benefit from new and faster hardware allowing more complex computations, and they benefit from new approaches in the area of computer vision. As such, we think that the quality of these approaches can still be significantly improved while keeping minimal hardware requirements (of usually having additional cameras), the support for flexible postures and freedom of movement in front of the screen, and a general small setup space size. However, even if the quality improved over the last couple of years, all these interpolation and reconstruction techniques are challenged mainly by two aspects: (1) even the slightest artefacts in a person's face (especially in and around the eyes) are easily noticeable; therefore, a very high and reliable synthesis quality has to be provided; and (2) related to this first aspect, developers have to overcome the “uncanny valley” known from anthropomorphic robotics research: if the eye contact looks and feels almost exactly like it does in human-human communication but not exactly, this might lead to revulsion.

To date, approaches that use shuttered or unshuttered screens offer the best quality, with several different systems described in the literature. Still, we argue that these approaches do not offer as much potential for future improvements as, for example, 2D video and 3D video techniques, and that they have the drawback of a complex hardware setup.

In the past, most research has focused on half-silvered mirror setups. Many of the later systems (2D video techniques, 3D video techniques, shuttered, and unshuttered screens) have not been studied or evaluated in terms of perception of eye-to-eye contact. In particular, no work exists that compares these existing systems in terms of user perception with respect to eye-to-eye and mutual gaze support. However, these evaluations and studies also require well-defined metrics for evaluating mutual gaze in videoconferencing, which need to be defined beforehand. To date, researchers have used no common metrics and investigated mutual gaze targeting only specific aspects (e.g., trust).

## 6 Conclusion and Future Work

In this paper, we overview videoconferencing solutions that support eye-to-eye contact between participants. We highlight the importance of eye-to-eye contact and mutual gaze for supporting communication in business-type videoconferences. We show that gaze and eye-to-eye contact are fundamental for the quality of the experience but also for trust between communication partners—an important parameter that is especially important in business meetings. Further results demonstrate that lack of eye-to-eye contact affects the interaction quality, which makes it a more artificial experience, but, if eye contact or gaze are supported, the quality of discussion can be increased in that it can lead to shorter discussion times or improve turn-taking.

We further survey and discuss existing solutions for video-conferencing supporting eye-to-eye contact that range from affordable solutions to complex systems. We develop a set of criteria as a guide for selecting research directions, developments, and investment decisions, and we evaluate all commonly known technical approaches to achieve mutual gaze according to those criteria.

Mutual gaze support in videoconferencing is still an open research topic. For instance, for future setups, one could think of a hardware arrangement where the reflections on a reflective LCD screen surface itself (normally an annoyance for users) is used as a mirror. The screen that inherently polarizes the light and a camera with a polarizing filter properly positioned near the user at a certain angle to a slightly titled screen could deliver the desired effect. This would be an affordable solution that needs little space and that leaves the space in front of the screen unoccupied. The main challenge is reducing unwanted artefacts. In a not-too-distant future, technical solutions might arise that implement eye-to-eye contact in a more elegant way (e.g.. by placing light sensors in-between light emitting elements in computer displays; Uy, 2009). In the meantime, we can and should apply one of the techniques described here and elsewhere to improve the quality of our videoconferencing experiences. In particular, software solutions using one or more 2D or depth cameras offer huge potential.

## Acknowledgments

We thank Simon Hoermann, Lavell Mueller, and Andreas Dünser for their valuable input to this paper and the reviewers of an earlier version of this paper for their feedback helping us to improve this paper. We also thank Adam LeBrocq for his constructive editorial help. This work was partly funded by a University of Otago CALT—RUT research grant (JDLJ17096).

## References

Andersson, R. L., Chen, T., & Haskell, B. G. (1996). *United States Patent 5,500,671*. Washington, DC: U.S. Patent and Trademark Office.

Andersson, R. L. (1997). *United States Patent 5,675,376*. Washington, DC: U.S. Patent and Trademark Office.

Bailenson, J. N., Beall, A.C., & Blascovich, J. (2002). Gaze and task performance in shared virtual environments. The Journal of Visualization and Computer Animation *13*, 313-320.

Bayliss, A. P., & Tipper, S. P. (2006). Predictive gaze cues and personality judgments: Should eye trust you? *Psychological Science*, *17*(6), 514-520.

Bekkering, E., & Shim, J. P. (2006). i2i trust in videoconferencing. *Communications of the ACM*, *49*(7), 103-107.

Biocca, F., Harms, C., & Gregg, J. (2001). The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity. In *Proceedings of the 4th International Workshop on Presence.*

Birden, H., & Page, S. (2005). Teaching by videoconference: A commentary on best practice for rural education in health professions. *Rural and Remote Health, 5,* 356. Retrieved from http://www.rrh.org.au

Bondareva, Y., & Bouwhuis, D. (2004). Determinants of social presence in videoconferencing. In A. Liliana & G. Semeraro (Eds.), *Proceedings of the 2004 Workshop on Environments for Personalized Information Access Working Conference on Advanced Visual Interfaces* (pp. 1-9).

Bondareva, Y., Meesters, L. M. J., & Bouwhuis, D. G. (2006). Eye contact as a determinant of social presence in video communication. In *Proceedings of the 20th International Symposium on Human Factors in Telecommunication.*

Buxton, W., & Moran, T. (1990). EuroPARC's integrated interactive intermedia facility (IIIF): Early experience. In S. Gibbs & A. A. Verrijn-Stuart (Eds.), *Proceedings of the IFIP WG 8.4 Conference on Multi-user Interfaces and Applications* (pp. 11-34). Amsterdam: Elsevier Science Publishers.

Buxton, W. (1992). Telepresence: Integrating shared task and person spaces. *Proceedings of Graphics Interface* (pp. 123-129).

Buxton, W., Sellen, A., & Sheasby, M. (1997). Interfaces for multiparty videoconferencing. In K. Finn, A. Sellen, & S. Wilber (Eds.), *Video mediated communication* (pp. 385-400). Hillsdale, NJ: Erlbaum.

Carville, S., & Mitchell, D. R. (2000). "It's A bit like Star Trek": The effectiveness of video conferencing. *Innovations in Education & Training International*, *37*(1), 42-49.

Cham, T.-J., Krishnamoorthy, S., & Jones, M. (2002). Analogous view transfer for gaze correction in video sequences. *Proceedings of the International Conference on Automation, Robotics, Control, and Vision*, *3*, 1415-1420.

Chen, M. (2002). Leveraging the asymmetric sensitivity of eye contact for videoconference. In *Proceedings of CHI 2002.*

Cook, M. (1977). Gaze and mutual gaze in social encounters. *American Scientist*, *65*, 328–333.

Cornelius, C., & Boss, M. (2003). Enhancing mutual understanding in synchronous computer-mediated communication by training. *Communication Research, 30*(2), 147-177.

Criminisi, A., Shotton, J., Blake, A., Torr, P. H. S. (2003). Gaze manipulation for one-to-one teleconferencing. In *Proceedings of the 9th IEEE International Conference on Computer Vision* (p. 191).

Davis, A. W., & Weinstein, I. M. (2005). *The business case for videoconferencing—achieving a competitive edge* (White Paper). Wainhouse Research.

Detenber, B., & Reeves, B. (1996). A bio-informational theory of emotion: Motion and image size effects on viewers. *Journal of Communication*, *46*(3), 66-84.

Ferrán-Urdaneta, C., & Storck, J. (1997). Truth or deception: The impact of videoconferencing for job interviews. In *Proceedings of the 18th International Conference on Information Systems* (183-196).

Fish, R. S., Kraut, R. E., & Chalfonte, B. L. (1990). The videowindow system in informal communications. In *Proceedings of CSCW*.

Fox, E. (2005). The role of visual processes in modulating social interactions. *Visual Cognition*, *12*(1), 1-11.

Fullwood, C., & Doherty-Sneddon, G. (2006). Effect of gazing at the camera during a video link on recall. *Applied Ergonomics*, *37*, 167–175.

Gale, C., & Monk, A.F. (2000). Where am I looking? The accuracy of video-mediated gaze awareness. *Perception & Psychophysics*, *62*(3), 586-595.

Gamer, M., & Hecht, H. (2007). Are you looking at me? Measuring the cone of gaze. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(3), 705-715.

Garau, M. (2003). The impact of avatar fidelity on social interaction in virtual environments (Doctoral dissertation). University College London.

Garau, M., Slater, M., Bee, S., & Sasse, M. A. (2001). The impact of eye gaze on communication using humanoid avatars. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, *3*(1), 309-316.

Gaver, W. W. (1992).The affordances of media spaces for collaboration. In *Proceedings of the 1992 ACM Conference on Computer-Supported Cooperative Work* (pp. 17-24).

Gemmell, J., Toyama, K., Zitnick, C. L., Kang, T., & Seitz, S. (2000). Gaze awareness for video-conferencing: A software approach. *IEEE Multimedia Magazine*, *7*(4), 26-35.

Gibbs, S. J., Arapis, C., & Breiteneder, C. J. (1999). TELEPORT—towards immersive copresence. *Multimedia Systems*, *7*(3), 214-221.

Gill, D., Parker, C., & Richardson, J. (2005). Twelve tips for teaching using videoconferencing. *Medical Teacher, 27* (7), 573-577.

Grayson, D. M., & Monk, A. F. (2003). Are you looking at me? Eye contact and desktop video conferencing. *ACM Transactions on Computer-Human Interaction*, *10*(3), 221-243.

Gross, M., Würmlin, S., Naef, M., Lamboray, E., Spagno, C., Kunz, A., Koller-Meier, E., Svoboda, T., Van Gool, L., Lang, S., Stehlke, K., Vande Moere, A., & Staadt, O. (2003). Blue-c: A spatially immersive display and 3D video portal for telepresence. *ACM Transactions on Graphics*, *22*(3), 819-827.

Hartkop, D. (2007). *United States Patent 20070002130 A1.* Washington, DC: U.S. Patent and Trademark Office.

Hauber, J., Regenbrecht, H., Cockburn, A., & Billinghurst, M. (2012). The impact of collaborative style on the perception of 2D and 3D videoconferencing interfaces. *The Open Software Engineering Journal, 6*(1), 1-20.

Heaton, L. (1998). Preserving communication context. In C. Ess & F. Sudweeks (Eds.), *Proceedings Cultural Attitudes Towards Communication and Technology '98*, (pp. 207-230).

Ishii, H., Konayashi, M., & Grudin, J. (1993). Integration of interpersonal space and shared workspace: ClearBoard design and experiments. *ACM Transactions on Information Systems, 11*(4), 349-375.

Jerald, J., & Daily, M. (2002). Eye gaze correction for videoconferencing. In *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications* (pp. 77-81).

Joiner, R., Scanlon, E., O'Shea, T., Smith, R. B., & Blake, C. (2002). Evidence from a series of experiments on videomediated collaboration: Does eye contact matter? In *Proceedings of the Conference on Computer Support for Collaborative Learning: Foundations for a CSCL Community* (pp. 371-378).

Jones, A., Lang, M., Fyffe, G., Yu, X., Busch, J., McDowall, I., Bolas, M., & Debevec, P. (2009). Achieving eye contact in a one-to-many 3D video teleconferencing system. ACM Transactions on Graphics, *28*(3), 1-8.

Jouppi, N. P., Iyer, S., Thomas, S., & Slayden, A. (2004). BiReality: Mutually-immersive telepresence. In *Proceedings of MM'04*.

Kannes, D. (1990). *United States Patent 4,965,819.* Washington, DC: U.S. Patent and Trademark Office.

Kannes, D. (1995). *United States Patent 5,382,972.* Washington, DC: U.S. Patent and Trademark Office.

Kauff, P., & Schreer, O. (2002). An immersive 3D videoconferencing system using shared virtual team user environments. In *Proceedings of ACM Collaborative Environments*.

Kuechler, M., & Kunz, A. (2006). HoloPort—a device for simultaneous video and data conferencing featuring gaze awareness. In *Proceedings of IEEE Virtual Reality 2006*.

Kuster, C., Popa, T., Bazin, J. C., Gotsman, C., Gross, M., & Gaze (2012). Correction for home video conferencing. *ACM Transactions on Graphics*, *31*(6), 174:1-174:6

Large, T. A., Rosenfeld, D., & Emerton, N. (2009). *US Patent App. 12/474,044*. Washington, DC: U.S. Patent and Trademark Office.

Lewis, J. P. (1994). *United States Patent 5,359,362.* Washington, DC: U.S. Patent and Trademark Office.

Libbey, K. A. (2004). *United States Patent US2004/0155956 A1*. Washington, DC: U.S. Patent and Trademark Office.

Lombard, M. (1992). *Direct responses to people on the screen: Personal space and television* (Doctoral dissertation). Stanford University.

Maimone, A., & Fuchs, H. (2011). Encumbrance-free telepresence system with real-time 3D capture and display using commodity depth cameras. In *Proceedings IEEE International Symposium on Mixed and Augmented Reality*.

McAndrew, P., Foubister, S. P., & Mayes, T. (1996). Videoconferencing in a language learning application. *Interacting with Computers, 8*(2), 207-217.

McKinney, V. R., & Whiteside, M. M. (2006). Maintaining distributed relationships: Electronic communication works best when it increases interaction and collaboration through a variety of media. *Communications of the ACM*, *9*(3), 82-86.

McNelley, S. H., & Machtig, J. S. (1999). *United States Patent 5,953,052*. Washington, DC: U.S. Patent and Trademark Office.

McNelley, S. H. (2000). *United States Patent 6,104424*. Washington, DC: U.S. Patent and Trademark Office.

McNelley, S. H., & Machtig, J. S. (2001). *United States Patent 6,243,130 B1*. Washington, DC: U.S. Patent and Trademark Office.

McNelley, S. H., & Machtig, J. S. (2004). *United States Patent 6,710,797 B1*. Washington, DC: U.S. Patent and Trademark Office.

Mukawa, N., Oka, T., Arai, K., & Yuasa, M. (2005). What is connected by mutual gaze? — User's behavior in video-mediated communication. In *Proceedings of CHI 2005* (pp. 1677-1680).

Nelson, T. J., & Smoot, L. S. (1992). *United States Patent 5,117,285*. Washington, DC: U.S. Patent and Trademark Office.

Nelson, T. J., & Vaning, B. R. (1997). *United States Patent 5,612,734.* Washington, DC: U.S. Patent and Trademark Office.

Nguyen, D., & Canny, J. (2007). MultiView: Improving trust in group video conferencing through spatial faithfulness. *Proceedings of ACM CHI'07* (pp. 1465-1474).

Okada, K., Maeda, F., Ichikawa, Y., & Matsushita, Y. (1994). Multiparty videoconferencing at virtual social distance: MAJIC design. In *Proceedings of the 1994 ACM Conference on Computer-Supported Cooperative Work* (pp. 385-393).

Olsen, J. S., Olsen, G. M., & Meader, D. K. (1995). What mix of video and audio is useful for small groups doing remote real-time design work? In *Proceedings of CHI'95* (pp. 362-368).

O’Malley, C., Langton, S., Anderson, A., Doherty-Sneddon, G., & Bruce, V. (1996). Comparison of face-to-face and video-mediated interaction. *Interacting with Computers, 8*(2), 177-192.

Ott, M., Lewis, J. P., & Cox, I. (1993). *Teleconferencing eye contact using a virtual camera*. Paper presented at INTERACT’93 and CHI’93 Conference Companion on Human Factors in Computing Systems, Amsterdam, The Netherlands.

Petit, B., Lesage, J.-D., Menier, C., Allard, J., Franco, J.-S., Raffin, B., Boyer, E., & Faure, F. (2010). Multi-camera real-time 3D modeling for telepresence and remote collaboration. *International Journal of Digital Multi. Broadcasting*.

Prince, S., Cheok, A. D., Farbiz, F., Williamson, T., Johnson, N., Billinghurst, M., & Kato, H. (2002). 3D live: Real time captured content for mixed reality. In *Proceedings of the 1st International Symposium on Mixed and Augmented Reality*.

Quante, B., & Muehlbach, L. (1999). Eye-contact in multipoint videoconferencing. In *Proceedings of the 17th International Symposium on Human Factors in Telecommunication*.

Regenbrecht, H., Lum, T., Kohler, P., Ott, C., Wagner, M., Wilke, W., & Mueller, E. (2004). Using augmented virtuality for remote collaboration. *Presence: Teleoperators and Virtual* Environments, *13*(3), 338-354.

Regenbrecht, H., Müller, L., Hoermann, S., Langlotz, T., Wagner, M., & Billinghurst, M. (2014). An eye-to-eye contact system for life-sized videoconferencing. In *Proceedings of ACM OzCHI*.

Schreer, O., Feldmann, I., Atzpadin, N., Eisert, P., Kauff, P., & Belt, H.J.W. (2008). 3Dpresence—a system concept for multi-user and multi-party immersive 3D videoconferencing. In *Proceedings of 5th European Conference on Visual Media Production* (pp. 1-8).

Shahid, S., Krahmer, E., & Swerts, M. (2012). Video-mediated and co-present gameplay: Effects of mutual gaze on game experience, expressiveness and perceived social presence. *Interacting with Computers, 24*, 292-305.

Sorlie, T., Gammon, D., Bergvik, S., & Sexton, H. (1999). Psychotherapy supervision face-to-face and by videoconferencing: A comparative study. *British Journal of Psychotherapy, 15*(4), 452-562.

Swaab, R. I., & Swaab, D. F. (2008). Sex differences in the effects of visual contact and eye contact in negotiations. *Journal of Experimental Social Psychology*, *45*(1), 129-136.

Tam, T., Cafazzo, J. A., Seto, E., Salenieks, M. E., & Rossos, P. G. (2007). Perception of eye contact in video teleconsultation. *Journal of Telemedicine and Telecare, 13,* 35-39.

Tedesco, J. M. (1999). *United States Patent 5,856,842.* Washington, DC: U.S. Patent and Trademark Office.

Teoh, C., Regenbrecht, H., & O'Hare, D. (2010). Investigating factors influencing trust in video-mediated communication. In *Proceedings of ACM OZCHI 2010* (pp. 312-319).

Teoh, C., Regenbrecht, H., & O'Hare, D. (2011). The transmission of self: Body language availability and gender in videoconferencing. In *Proceedings of ACM OZCHI 2011* (pp. 273-280).

Teoh, C., Regenbrecht, H., & O'Hare, D. (2012). How the other sees us: Perceptions and control in Videoconferencing. In *Proceedings of ACM OZCHI 2012*.

Teoh, C. (2012). Body language availability in video-conferencing (Doctoral dissertation). University of Otago, New Zealand.

Tsai, Y, Kao, C., Hung, Y., & Shih, Z. (2004). Real-time software method for preserving eye contact in video conferencing. *Journal of Information Science and Engineering, 20*, 1001-1017.

Uy. M. (2009). *United States Patent, 7,535,468.* Washington, DC: U.S. Patent and Trademark Office.

Vertegaal, R. (1999). The GAZE groupware system: Mediating joint attention in multiparty communication and collaboration. In *Proceedings of the ACM CHI'99 Conference on Human Factors in Computing Systems*. Pittsburgh, PA: ACM.

Vertegaal, R., & Ding, Y. (2002). Explaining effects of eye gaze on mediated group conversations: Amount or synchronization? In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work* (pp. 41–48). New York, NY: ACM.

Vertegaal, R., Weevers, I., Sohn, C., & Cheung, C. (2003). GAZE-2: Conveying eye contact in group video conferencing using eye-controlled camera direction. In *Proceedings of CHI 2003* (pp. 521-528).

Wetzstein, G. (2005). *Image-based view morphing for teleconferencing applications*. University of Otago.

Wheeless, L. R., & Grotz, J. (1977). The measurement of trust and its relationship to self-disclosure. *Human Communication Research*, *3*(3), 250-257.

Wolff, R., Roberts, D. Murgia, A., Murray, N., Rae, J., Steptoe, W., Steed, A., & Sharkey, P. (2008). Communicating eye gaze across a distance without rooting participants to the spot. In *Proceedings of the 12th IEEE/ACM International Symposium on Distributed Simulation and Real-Time Applications* (pp. 111-118).

Xu, L-Q, Loffler, A., Sheppard, P. J., & Machin, D. (1999). True-view videoconferencing system through 3-D impression of telepresence. *BT Technology Journal*, *17*(1), 59-68.

Yamashita, N., Hirata, K., Aoyagi, S., Kuzuoka, H., & Harada, Y. (2008). Impact of seating positions on group video communication. In *Proceedings of CSCW'08* (pp. 177-186).

Yang, R., & Zhang, Z. (2002). Eye gaze correction with stereovision for video-teleconferencing. In *ECCV* (pp. 479–494).

Yip, B., & Jin, J. S. (2003). An effective eye gaze correction operation for video conference using antirotation formulas. In *Proceedings of IEEE ICICS-PCM*.

Yip, B. (2005). Face and eye rectification in video conference using artificial neural network. In *Proceedings IEEE International Conference on Multimedia and Expo* (pp. 690-693).

Zitnick, C. L., Gemmell, J., & Toyama, K. (1999). *Manipulation of video eye gaze and head orientation for video teleconferencing* (Technical Report MSR-TR-99-46). Microsoft Research.

Zhu, J., Yang, R., & Xiang, X. (2011). Eye contact in video conference via fusion of time-of-flight depth sensor and stereo. *3D Research, 2*, 1-10.

## About the Authors

**Holger Regenbrecht** is an Associate Professor at the department of Information Science at Otago University and leads the Computer-Mediated Realities Lab. His research interests include Human-Computer Interaction (HCI), Applied Computer Science and Information Technology, (collaborative) Augmented Reality, Videoconferencing and Collaboration, Psychological aspects of Mixed Reality, Three-dimensional user interfaces (3DU) and Computer-aided therapy and rehabilitation.

**Tobias Langlotz** is a Lecturer at the department of Information Science at Otago University. His main research interest is in location-based mobile interfaces, where he works at the intersection of HCI, Computer Graphics, Computer Vision and Pervasive Computing