Project Starline: A high-fidelity telepresence system

JASON LAWRENCE, DAN B GOLDMAN, SUPREETH ACHAR, GREGORY MAJOR BLASCOVICH, JOSEPH G. DESLOGE, TOMMY FORTES, ERIC M. GOMEZ, SASCHA HÄBERLING, HUGUES HOPPE, ANDY HUIBERS, CLAUDE KNAUS, BRIAN KUSCHAK, RICARDO MARTIN-BRUALLA, HARRIS NOVER, ANDREW IAN RUSSELL*, STEVEN M. SEITZ, and KEVIN TONG, Google Research, USA

We present a real-time bidirectional communication system that lets two people, separated by distance, experience a face-to-face conversation as if they were copresent. It is the first telepresence system that is demonstrably better than 2D videoconferencing, as measured using participant ratings (e.g., presence, attentiveness, reaction-gauging, engagement), meeting recall, and observed nonverbal behaviors (e.g., head nods, eyebrow movements). This milestone is reached by maximizing audiovisual fidelity and the sense of copresence in all design elements, including physical layout, lighting, face tracking, multi-view capture, microphone array, multi-stream compression, loudspeaker output, and lenticular display. Our system achieves key 3D audiovisual cues (stereopsis, motion parallax, and spatialized audio) and enables the full range of communication cues (eye contact, hand gestures, and body language), yet does not require special glasses or body-worn microphones/headphones. The system consists of a head-tracked autostereoscopic display, high-resolution 3D capture and rendering subsystems, and network transmission using compressed color and depth video streams. Other contributions include a novel image-based geometry fusion algorithm, free-space dereverberation, and talker localization.

CCS Concepts: • **Computing methodologies** → **Computer graphics**; **Mixed / augmented reality**; **Perception**.

Additional Key Words and Phrases: videoconferencing, telecopresence, eye contact, parallax, stereopsis, spatialized audio, 3D capture

ACM Reference Format:
Jason Lawrence, Dan B Goldman, Supreeth Achar, Gregory Major Blascovich, Joseph G. Desloge, Tommy Fortes, Eric M. Gomez, Sascha Häberling, Hugues Hoppe, Andy Huibers, Claude Knaus, Brian Kuschak, Ricardo Martin-Brualla, Harris Nover, Andrew Ian Russell, Steven M. Seitz, and Kevin Tong. 2021. Project Starline: A high-fidelity telepresence system. *ACM Trans. Graph.* 40, 6, Article 242 (December 2021), 16 pages. https://doi.org/10.1145/3478513.3480490

1 INTRODUCTION

Improvements in telecommunications have steadily increased both the fidelity and availability of synchronous communication over long-distance networks [Sterling and Shiers 2000]. Video-based systems like Skype, FaceTime, Zoom, Meet, and Teams are a recent step forward in bringing people closer together who are far apart. At the far end of this spectrum is *telepresence*, i.e., enabling remote participants to feel copresent, as if they are occupying a shared physical space [e.g., Draper et al. 1998; Gibbs et al. 1999; Kuster et al. 2012; Maimone et al. 2012; Zhang et al. 2013].

Fig. 1. Our system enables two people to communicate at a distance as if they were physically together. Users report a strong sense of presence and connection with the remote participant.

Telepresence presents tremendous opportunities to bring together the world's increasingly distributed organizations and social groups. However, achieving its full potential poses three grand challenges across multiple research areas:

(1) Capture and render a **3D audiovisual likeness** of a remote person, so realistic that one forgets it is not real.
(2) Create a **comfortable display** with retinal resolution, wide field of view, stereopsis, and motion parallax.
(3) Achieve **copresence** — the feeling that two people are together — including proximity, eye contact, and interaction.

We demonstrate a telepresence system representing a significant milestone along these different dimensions. Notably, user studies demonstrate an improved experience over traditional 2D videoconferencing.

Our unencumbered, bidirectional, 3D communication system is designed for face-to-face meetings. It renders a remote participant as if they were physically copresent, with mutual eye contact (Figure 1). We carefully design and engineer the physical layout, lighting, 3D capture, compression, rendering, display, and audio subsystems to eliminate as many hints as possible that the remote participant is not in the same room as the user.

The primary contribution of this paper is the first telepresence system that achieves measured improvements in meeting experiences and behaviors compared to 2D videoconferencing. User-study

*Now at NVIDIA.

0730-0301/2021/12-ART242
https://doi.org/10.1145/3478513.3480490

participants rated our system as significantly better at fostering various elements of communication including presence, attentiveness, reaction-gauging, eye contact, engagement, and personal connection. They also had greater meeting recall and demonstrated more nonverbal behaviors (hand gestures, head nods, and eyebrow movements) than in 2D videoconferencing.

Outperforming 2D videoconferencing is more challenging than it sounds, for several reasons. First, 2D video is highly realistic, whereas existing real-time 3D capture technologies are all known to suffer visual artifacts, putting them at an inherent disadvantage. Second, compared to 2D displays, most stereoscopic technologies introduce quality trade-offs such as lower resolution, tracking latency, or accommodation-vergence issues, which degrade the experience for many viewers. The fact that our system shows statistically significant user preference over standard videoconferencing despite these challenges is noteworthy.

Additional contributions in our telepresence system include:

- the first use of head-tracked audio crosstalk cancellation, creating the perception that audio originates from the remote user's mouth even as both users move,
- a rendering method that merges multiple depth and color images using an image-based formulation of geometry fusion,
- a 3D facial feature tracking subsystem that combines 2D facial landmark estimation, 3D triangulation, and double exponential filtering to yield accurate predictions at 120Hz.

Please see the accompanying video that approximates the experience of using our system.

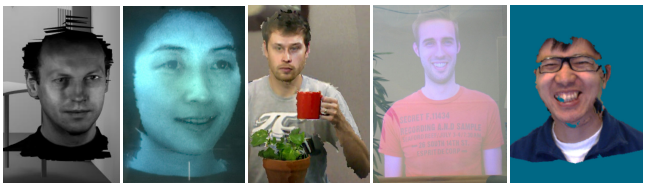
2 RELATED WORK

Videoconferencing. A number of commercial products use custom furniture and specially designed configurations of displays, cameras, microphones, and speakers to heighten the sense of sharing a common space with a remote site [e.g., Cisco Systems, Inc. 2011; DVE 2014; Hewlett-Packard 2005; Plantronics Inc. 2019; Sony 2008; Szigeti et al. 2009].

3D telepresence. Enabling a richer set of 3D depth cues (e.g., stereopsis, motion parallax, and natural scale) provides a stronger sense of immersion and copresence [Gibbs et al. 1999; Muhlbach et al. 1995]. An important goal is mutual eye gaze, a crucial nonverbal cue in human communication [Argyle and Cook 1976; Macrae et al. 2002; Pan and Steed 2014, 2016]. Researchers have explored telepresence systems for decades [e.g., Baker et al. 2002; Chen et al. 2000; De Silva et al. 1995; Dou et al. 2012; Fuchs et al. 2014; Kauff and Schreer 2002; Lanier 2001; Maimone et al. 2012; Maimone and Fuchs 2011; Majumder et al. 1999; Mulligan et al. 2004; Pejsa et al. 2016; Raskar et al. 1998; Schreer et al. 2001; Towles et al. 2002; Yang et al. 2002; Zhang et al. 2013].

Jones et al. [2009] achieve both stereo and parallax depth cues along with natural eye contact by using a polarized beamsplitter, a high-frequency projector, and a fast spinning mirror to create a volumetric display. However, 3D capture is only performed for one user, so the effect of telepresence is asymmetric.

Maimone et al. [2012] perform 3D capture using 5 Kinect units. To create new stereo images for an autostereoscopic display, they rasterize each Kinect view as a triangulated depth map, then combine



[Gibbs et al. 1999] [Jones et al. 2009] [Maimone et al. 2012] [Kuster et al. 2012] [Zhang et al. 2013]

Fig. 2. Screenshots from prior telepresence research systems.

the rendered images at each pixel using a normal-based weighting of the views seeing the nearest surface. Their experiments with a single system do not demonstrate symmetric communication.

Kuster et al. [2012] realize symmetric telepresence. They perform 3D capture using a single depth sensor and transmit a video stream combining both color and depth. The use of a single depth view simplifies capture, transmission, and rendering, but provides incomplete surface coverage, resulting in disocclusion artifacts.

Zhang et al. [2013] use several IR projectors and cameras to reconstruct multiple depth images. They merge the depth maps to create a sparse 3D point cloud and transmit the point cloud along with color video streams. In contrast, our system transmits depth streams and performs geometry fusion during rendering.

Compared to these prior works, our system includes many novel elements, e.g., multiple compressed depth streams, image-based geometry fusion, high-fidelity face tracking, head-tracked lenticular display, tracker-steered audio beamforming, split-frequency audio spatialization. However, the most important aspect of our work is the significant increase in overall audiovisual fidelity, e.g., comparing Figures 2 and 13. The combined improvements in spatial resolution, color fidelity, depth accuracy, audio, and refresh rate enable our system to demonstrate for the first time an immersive telepresence experience that surpasses classical videoconferencing.

Telepresence using HMDs. The benefits of virtual- and augmented-reality head-mounted displays [Maimone et al. 2013; Orts-Escolano et al. 2016; Wei et al. 2019] include a more immersive experience and a more portable, affordable device. The main difficulty is to obtain a high-quality real-time 3D capture of the user's face while it is hidden behind the headset [Chu et al. 2020; Frueh et al. 2017; Lombardi et al. 2018, 2019; Richard et al. 2021; Wei et al. 2019]. Current work aiming for photorealistic quality involves precaptured user data, unlike in our system.

Gaze redirection. Several techniques improve eye contact with faces in conventional 2D video by digitally altering their perceived gaze direction [Criminisi et al. 2003; Ganin et al. 2016; He et al. 2019; Kononenko and Lempitsky 2015; Wolf et al. 2010; Yang and Zhang 2002]. Our system achieves mutual eye gaze by accurately reproducing the 3D appearance of each user as seen from the other's vantage point, without requiring special processing of eye regions.

Immersive audio in teleconferencing. Spatialized audio in multiperson remote meetings often involves widely distributed microphones and loudspeakers [Plantronics Inc. 2019]. Zhang et al. [2013] incorporate 3D immersive audio using just two loudspeakers, as in our system. Although they mention the possibility of head-tracked

audio rendering and crosstalk cancellation, their system uses a simpler spatialization approach based on gain-and-delay panning. Our system uses talker-tracked microphone-array beamforming for enhanced audio capture, and it uses talker/listener-tracked virtual spatialization with listener-tracked binaural crosstalk cancellation to improve realism.

Autostereoscopic display. Several stereo display technologies show a different image to each eye without requiring glasses [Chen et al. 2014; Dodgson 2005; Wetzstein et al. 2012]. The lenticular display used in our system places a lens array at a precise distance in front of a 2D display [Borner et al. 2000; Matusik and Pfister 2004]. The lens array is similar to a parallax barrier, revealing a different subset of the display pixels to each eye, but the lenses are more optically efficient. A lenticular display can be combined with active head-tracking to *steer* the stereo images to a single user's eyes during head motion. This is accomplished by adjusting the interlaced mapping from the stereo images to the underlying 2D display as a function of the eye locations [Boev et al. 2008; Jurk and de la Barré 2014].

3 HIGH-LEVEL DESIGN

Design goals. Our overriding objective is unencumbered telepresence, i.e., recreating the appearance and sound of a remote user with sufficient quality to enable all conversational cues, while retaining the simplicity of just sitting down and talking with a person in real life. We identify the following requirements:

- Life-size depiction at high resolution, high framerate, and with accurate color;
- Stereopsis and parallax, with left and right views rendered from continuously moving viewpoints with low latency;
- Symmetric video experience, enabling eye contact;
- Symmetric audio experience, with speech perceived to emanate from the virtual participant's mouth;
- Absence of HMD, glasses, tracking fiducials, headphones, or lapel microphones;
- Comfortable use for typical meeting durations.

Design choices. We considered both sitting and standing poses for participants, and selected a **seated** configuration to enable more comfortable conversations. Guided by proxemics work [Hall 1963], we chose a nominal eye-to-eye distance of 1.25 m, just above the boundary between personal and social space, to facilitate a range of social and business interactions[1].

Our choice to pursue a screen-based system is motivated in part by the significant weight and discomfort associated with most current AR and VR headsets. It also eliminates the difficulties of capturing a face through a headset [Wei et al. 2019]. Moreover, it dovetails with our quality objectives, as most widely available VR headsets have an angular resolution less than 20 pixels per degree, and no currently available AR headset has sufficient field of view to span the width and height of a seated human torso. An available technology that meets our combined acuity and field-of-view goals is a head-tracked **autostereoscopic** display based on a 65-inch 8K panel with 33.1M full-color pixels updating at 60 Hz. For a typical adult inter-pupil distance and an eye-to-display distance of 1.25 m, the lens array presents each eye a separate subset of the display pixels (≈5M pixels of each red, green, and blue primary), resulting in an approximate angular resolution of 45 pixels per degree.

Head-tracked autostereoscopic displays can suffer from left-right visual crosstalk, tracking latency, and vergence-accommodation conflict. The impact of these deficiencies increases with **disparity**, which in turn increases as the 3D content is rendered further from the display plane [Perlin et al. 2000]. We mitigate these issues by positioning the virtual space of the remote user such that their face — the typical focus of conversation — lies near the display plane.

Another concern is the abrupt loss of stereo at the display edges. Although 65-inch diagonal panels can comfortably display both the torso and head of most subjects, the torso and hands are clipped at the bottom of the display, giving the impression that a closer object (e.g., hand) is occluded by a more distant object (the display bezel). Such **depth conflicts** can be disorienting or even uncomfortable, pulling participants out of the illusion of presence. As a solution, we place a "middle wall" 0.59 m in front of the display to block the user's view of the display bottom. (We assume a user seated 1.25 m from the display, with seated height less than the 95th percentile or 97 cm.) The wall induces the illusion that the hands and seated legs of a remote user may exist just behind it, thereby avoiding contradictory visual cues.

In designing the remote-to-local geometry mappings, it is important to ensure **mutual eye contact**. Let S_1, S_2 denote the spaces of two users U_1, U_2. User U_1 sees a local virtual representation $T_{21}(U_2)$ where $T_{21} : S_2 \mapsto S_1$ is a rigid transformation. Similarly, user U_2 sees the representation $T_{12}(U_1)$. The virtual remote user $T_{21}(U_2)$ should appear to look directly at U_1. However, the gaze of U_2 is directed to $T_{12}(U_1)$. Eye contact is satisfied iff $T_{12} = T_{21}^{-1}$. (If T is parameterized as a rotoreflection R and translation vector t, these must satisfy $R_{12} = R_{21}^{-1}$ and $t_{21} = -R_{21}t_{12}$.) We also desire each transform to provide a level view, equalize seat heights, and position the remote face near the display plane. Many configurations satisfy all properties (Figure 3), including reflection and 180° rotation about the eye-to-display midpoint. Our system supports both these modes. Because people's features are subtly asymmetric, and moreover any text appearing on objects or clothing is obviously asymmetric, we prefer to avoid reflection and therefore choose 180° rotation by default.

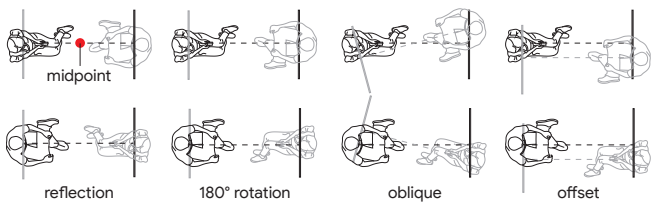


Fig. 3. Examples of geometric maps between system endpoints, showing the pair of real users (black) and their virtual counterparts (gray).

[1]Although this boundary is culturally-dependent, we chose a value appropriate for our North American user study participants.