LookAtChat: Visualizing Eye Contacts for Remote Small-Group Conversations

ZHENYI HE, New York University, United States
RUOFEI DU, Google, United States
KEN PERLIN, New York University, United States

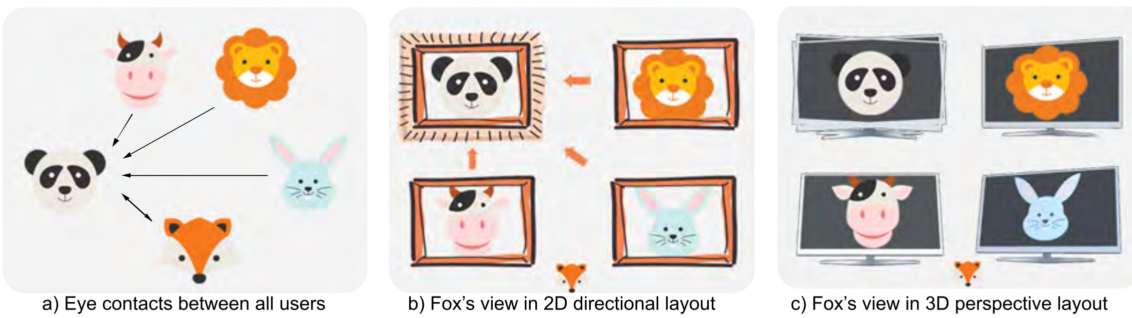


Fig. 1. We present *LookAtChat*, an online group chat system that takes advantage of eye-tracking technology available with ordinary webcams to visualize eye contacts in group chats. LookAtChat implements and evaluates both 2D directional and 3D perspective layout to inform users with spatial cues. To convey eye contacts in a), b) directional layout renders arrows between video streams, while out-glowing the video window of users who are looking at you. c) Perspective layout transforms video streams to gaze targets, while slightly shaking the video window of users who are looking at you. As a proof-of-concept, LookAtChat explores the potential of using gaze information during online video chat.

Video conferences play a vital role in our daily lives. However, many nonverbal cues are missing, including gaze and spatial information. We introduce LookAtChat, a web-based video conferencing system, which empowers remote users to identify eye contact and spatial relationships in small-group conversations. Leveraging real-time eye-tracking technology available with ordinary webcams, LookAtChat tracks each user's gaze direction, identifies who is looking at whom, and provides corresponding spatial cues. Informed by formative interviews with 5 participants who regularly use videoconferencing software, we explored the design space of eye contact visualization in both 2D and 3D layouts. We further conducted an exploratory user study (N=20) to evaluate LookAtChat in three conditions: baseline layout, 2D directional layout, and 3D perspective layout. Our findings demonstrate how LookAtChat engages participants in small-group conversations, how gaze and spatial information improve conversation quality, and the potential benefits and challenges to incorporating eye contact visualization into existing videoconferencing systems.

CCS Concepts: • **Human-centered computing → Collaborative interaction**.

Additional Key Words and Phrases: eye contact, video conferencing, video-mediated communication, gaze interaction

ACM Reference Format:
Zhenyi He, Ruofei Du, and Ken Perlin. 2021. LookAtChat: Visualizing Eye Contacts for Remote Small-Group Conversations. In . ACM, New York, NY, USA, 23 pages.

1 INTRODUCTION

Video conferencing is becoming the dominant medium for remote discussion and collaboration during the global COVID-19 pandemic. However, *eye contact information for more than two people* in mainstream video chat tools such as Skype, Zoom, and Google Meet is insufficiently conveyed through this medium. Hence, it is almost impossible to use eye gaze as a nonverbal cue to infer where attention is directed in conventional video conferences.

Prior art in remote small-group collaboration has leveraged multi-view cameras, customized displays, or mixed-reality settings to solve this problem. For example, GAZE-2 [50, 51] employs an array of cameras with an eye tracker and selectively transmits the preferred video stream to remote users. MMSpace [33] introduces novel physical kinetic displays to support eye contacts between every pair of participants; Holoportation [32] leverages full-body reconstruction and headset-removal technologies to achieve immersive small-group telepresence with Microsoft HoloLens. However, it is still unclear how to embed eye contact in a conventional videoconference with an ordinary laptop; or to determine the potential benefits and drawbacks of visualizing eye contacts in remote small-group conversation.

In this paper, we present and evaluate LookAtChat, a video conferencing system which visualizes eye contacts for remote small-group conversation. LookAtChat consists of three components: a WebRTC server to support video-conferencing and logging, an eye-tracking module powered by `WebGazer.js` [37] to recognize gaze positions, and a visualization module implemented with the `three.js`[1].

As initial work, our research questions are exploratory: How do people perceive eye contact in conventional video conferences? Can visualization of eye contact improve remote communication efficiency? In what context will users prefer to see eye contact? What forms of eye-contact visualization may be preferred by users?

To create LookAtChat, we conducted formative interviews with five people who use videoconferencing with colleagues on a daily basis. Our research is inspired and informed by prior small-group communication systems that demonstrate the potential of visualizing eye contacts and spatial information. To improve the generalizability and replicability of the system, we only require each user to use a laptop with a webcam. We further extend the design space of visualizing gaze and spatial information to a total of 11 layouts.

To evaluate LookAtChat, we conducted four three-session user studies with 20 remote participants (ages 24-39, 6 female and 14 male). In our analyses of video recordings, post-activity questionnaires, and post-hoc interviews, we found that LookAtChat can effectively engage participants in small-group conversations by visualizing eye contact and providing spatial relationships. Gaze and spatial information can improve the conversation experience, bring greater social presence and richness, and provide better user engagement. Our main contributions in this paper are:

(1) Conception, development, and deployment of LookAtChat, a video conferencing system that can visualize eye contact for remote small-group conversations.
(2) Enumerating design implications through formative interviews and extending the design space of visualizing gaze and spatial information in video conferences.
(3) Reporting evaluation results and reflections about the opportunistic use of eye contact visualization in video conferencing systems - benefits, limitations, and potential impacts to future remote collaboration systems.
(4) Open-sourcing[2]. We plan to make our software available to facilitate future development in video conferencing systems with visualization of nonverbal cues.

[1] three.js: JavaScript 3D library, http://www.threejs.org.
[2] The Github repository containing the source code will be available in the camera-ready revision. A live demo is available at https://eye.3dvar.com

2 RELATED WORK

To understand how gaze information is integrated into video conferences and to justify our design decisions, we review prior art on multi-user experience in distributed collaboration and gaze tracking technologies for videoconferencing. Many researchers have contributed to investigating future workspaces such as improving individual productivity like HoloDoc [24], reconstructing multi-user experiences like "the office of the future" [39], and cross-device interaction [52]. Furthermore, remote conferencing shows its potential for geologically dispersed users and is efficient for group discussion [27]. In scenarios requiring certain levels of trust and judgement with non-verbal communication, non-verbal cues are highly important for effective communication [40]. Gaze support and feeling of face-to-face [31] play a central role in those scenarios. With the increasing development of gaze tracking devices and technology, gaze-assisted interaction are becoming popular in the fields of text entry [53], video captions [19], and video conferences.

2.1 Multi-user Collaboration in Distributed Environments

Distributed multi-user collaboration has been widely researched from the perspective of locomotion, shared proxies, and life-size reconstruction as well as different purposes including communication, presentation, and object manipulation. Your Place and Mine [45] creates experiences that allow everyone to real walk in collaborative VR. Three's Company [49] presents a three-way distributed collaboration system that places remote users either on the same side or around a round table. In addition, Three's Company provides non-verbal cues like body gestures through a shared tabletop interface. Remote users' arm shadows are displayed locally on a tabletop device, which is beneficial for collaborative tasks with shared objects. Tan *et al.* [48] focus on presentation in large-venue scenarios, creating a live video view that seamlessly combines the presenter and the presented material, capturing all graphical, verbal, and nonverbal channels of communication. Tele-Board [12] enables regionally separated team members to simultaneously manipulate artifacts while seeing each other's gestures and facial expressions. The concept of Blended Interaction Spaces [30] is proposed to providing the illusion of a single unified space by creating appropriate shared spatial geometries. TwinSpace [41] is a generic framework discussing brainstorming and presentation in cross-reality that combines interactive workspaces and collaborative virtual worlds with large wall screens and projected tabletops. Physical Telepresence workspaces [22] is a shaped display providing shape transmission that can manipulate remote physical objects. Cameras are widely used for above alternatives to capture users, besides, 360 video has recently been researched. SharedSphere [21] is a wearable MR remote collaboration system that enriches a live captured immersive panorama based collaboration through MR visualisation of non-verbal communication cues.

Immersive collaborative virtual environment (ICVE) and Augmented Reality (AR) can be used to develop new forms of teleconferencing, which often leverages multiple cameras setup and 3D reconstruction algorithms. EyeCVE [46] uses mobile eye-trackers to drive the gaze of each participant's virtual avatar, thus supporting remote mutual eye-contact and awareness of others' gaze in a perceptually coherent shared virtual workspace. Jones *et al.* [16] design a one-to-many 3D teleconferencing system able to reproduce the effects of gaze, attention, and eye contact. A camera with projected structure-light is set up for reconstructing the remote user. Billinghurst and Kato [5] developed a system that allows virtual avatars and live video of remote collaborators to be superimposed over any real location. Remote participants were mapped to different fiducial markers. The corresponding video images were attached to the marker surface when markers are visible. Room2Room [38] is a telepresence system that leverages projected AR to enable life-size, face-to-face, co-present interaction between two remote participants by performing 3D capture of the local user with RGBD cameras. Holoportation [32] demonstrates real-time 3D reconstructions of an entire space, including people, furniture and objects, using a set of depth cameras. Gestures are preserved via full-body reconstruction and headset removal algorithms are designed to convey eye contact. However, "uncanny valley" remains a challenging problem in this domain.

2.2 Eye Contacts and Gaze Correction Technology in Video-mediated Conversation

Various hardware setups have been explored for gaze correction including hole in screen, long distance, and half-silver mirror. The hole in screen concept is about drilling a hole in the screen and placing a camera. Long distance uses a screen at a far distance while placing the camera as close as possible [47]. Half-silver mirror allows a user to see through a half-transparent mirror while being observed by a well-positioned camera at the same time. This idea was adapted in ClearBoard [13, 15] and Li *et al.* 's transparent display [23]. Despite their advantages in terms of system complexity and costs, such solutions are rarely used outside of laboratory due to the availability of hardware. In the meantime, quite a few 2D video-based (or image-based) approaches are proposed for eye contact including eye correction with a single camera [1, 2] and multiple cameras [8] while applying image-based approaches like texture remapping and image warp [10]. However, the technology is not sufficiently accurate to avoid visual artifacts and the uncanny valley. 3D video-based solutions including 3D reconstruction is another trend for maintaining eye contact while the head is reconstructed. RGB camera [54], depth camera [55], Kinect [20], or motion capture system [26] are used for 3D reconstruction.

Eng *et al.* [9] propose a gaze correction solution for a 3D teleconferencing system with a single color/depth camera. A virtual view is generated in the virtual camera location with hole filling algorithms. Compared to single camera setup, multiple cameras are popularly used for providing gaze [3] in videoconferencing. True-view [54] was implemented with two cameras (one on the left and the other on the right). The synthesised virtual camera view image at the middle viewpoint is generated to provide correct views of each other and the illusion of close proximity. GAZE-2 [50, 51] utilizes an eye tracker with three cameras. The eye tracker is used for selecting a proper camera closest to where the user is looking. GAZE-2 prototypes an attentive virtual meeting room to experiment with camera selection. In each meeting room, each user's video image is automatically rotated in 3D toward the participant he is looking at. All the video images are placed horizontally so the video image turns left or right when the corresponding camera is chosen. Likewise, MultiView [28, 29] is a video conferencing system that supports collaboration between remote groups of people with three cameras. Additionally, MultiView allows multiple users to be co-located in one site by generating a personal view for each user even though they look upon the same projection surface, which they achieve by using a retro-reflective material. Photoportals [4, 18] groups local users and remote users together through a large display. All users are tracked and roughly reconstructed through multiple cameras and then rendered within a virtual environment. MMSpace [33–36] provided realistic social telepresence in symmetric small group-to-group conversations through "kinetic display avatars". Kinetic display avatars can change pose and position by automatically mirroring the remote user's head motions. One camera is associated with one transparent display. Both camera and display can be turned to provide corresponding video input image and output angle. Sirkin *et al.* [44] developed a kinetic video conferencing proxy with a swiveling display screen to indicate which direction in which the satellite participant was looking for maintaining gaze and gestures to mediate interaction. Instead of rendering a video image on a rectangular display, a cylinder display is proposed in TeleHuman [17] with 6 Kinects and a 3D projector.

LookAtChat is designed to be used with *a minimum requirement of a laptop/PC and a single webcam.* While multi-view cameras and external hardware may yield better eye tracking and 3D rendering solutions, such systems typically require very high computational power and exclusive hardware setups. Since it is possible for users with low-cost video conferencing setup to learn to interpret gaze direction to a very high degree of accuracy [11], we decided not to apply extensive image-based manipulation on video streams but rather to focus on the design of a widely accessible online system to empower video conferencing users with real-time visualization of eye contacts.

3 FORMATIVE INTERVIEWS

To inform the design of LookAtChat and understand whether and how gaze information affects video conferencing, we conducted five formative interviews with videoconferencing users (2 female and 3 male, labeled as I1 to I5) to learn the advantages and disadvantages of current videoconferencing software compared to real-life meetings as well as people's expectation of videoconferencing. We asked participants about their recent video conferencing experience under different scenarios. Our takeaways are summarized below.

3.1 Feedback on the Existing Videoconferencing Systems

Good for multi-tasking and information sharing.

Software such as *Zoom* and *Skype* allows participants to work on multiple tasks at the same time while video conferencing, such as walking on a treadmill while listening to a talk. Users benefit from sharing screen or notes through videoconferencing software, as it allows any participant to instantly share their own document or presentation. Although participants in offline meetings can share information through whiteboarding or printed documents, video conferencing software allows a large number of people to concentrate on the same document and work on different sections of it.

Bad for white-boarding and body gestures.

For group discussions in which all participants may need to contribute their thoughts, a physical whiteboard is very popular. And yet, shared free sketch software is not well integrated into video conferencing software or available as stand-alone software for now, though quite a few researches have focused on that in immersive environments. Similarly, body gestures are partially missing due to the small view area of cameras and missing/different spatial information of participants.

Bad for finding the speaking up timing.

P1 and P4 thought it was more difficult to know when to speak in online meetings because not all the participants' gaze and body information are perceived well through the camera. It is not clear who is talking to whom, whether the speaker is waiting for an answer from a specific person, or if a speaker is pausing or is ending the conversation during group discussion.

Bad for controlling meeting length.

The length of physical meetings are usually well controlled since the meeting rooms are usually booked throughout the day and participants are aware of those who are gazing through the window, waiting to use the room next. However, participants in virtual conferences often cannot find the best time to exit for the next meeting. I2(M) commented that "in virtual video conferences, very few people strictly follow the proposed length of the meeting and oftentimes delay the next meetings. People just keep talking when the meeting goes beyond the scheduled time".

3.2 Expectations of Future Videoconferencing Systems

Improve the control of conversation. It is difficult to use words such as "you" in video conferencing contexts because participants barely know who the speaker is talking to, while "you" is natural in co-located conversations. I1(M) felt "less involved" because of the lack of this information. There also exists more simultaneous speech in video conferences. People start talking together and stop together to wait, which causes participants to lose track of the conversation.

Provide spatial information. I3(F) wished to select a seat the way they would normally enter a meeting room: "Everyone has their own perspectives and maintain spatial relationship with each other".

3.3 Thoughts of Visualizing Gaze Information in Videoconferencing System

Good for natural discussion. Interviewee (I2(M) and I5(F)) think it is helpful especially if the discussion requires feedback, attention, and interaction. Also, it is helpful for branching ideas. It is easy to suggest what topic one participant is following by looking at the proposer directly in offline meetings, but not easy to show to the group information in videoconferencing software.

Different for small group and large scale. For presentations or lectures, presenters or teachers may benefit from participants' gaze information that helps them adjust content in real-time. P1 elaborated: "teachers know the topic is difficult or get distracted when quite a few students' gaze focuses are shifted."

Concerns for privacy. Some interviewees (I3(F) and I5(F)) mentioned that they feel pressured when being looked at or looking at others. Displaying anonymous gaze information or aggregated data and reporting the result afterwards may be helpful.

4 LOOKATCHAT

Informed by formative interviews and inspired by prior systems, we formulate our design rationale, explore the design space, elaborate on two specific layouts for natural integration with conventional videoconferencing, and discuss potential use cases.

4.1 Design Rationale

We constrain our design scope to remote small-group conversations in which all participants are physically dispersed. This setting is motivated by the circumstances of COVID-19, where everyone is working remotely. Users mostly participate in these conversations on a laptop with a built-in frontal camera, or a workstation with a USB webcam. Scenarios with two or more people co-located in front of the camera for video conferencing are out of our design scope. Depth camera[38], multiple cameras[32], motion capture systems[26], professional eye-tracking systems[51], or head-mounted displays[32] are not considered as alternatives in our design due to their constraints of cost and availability. Although the above devices allow richer social engagements and more accurate gaze detection over a single webcam, we desire to make our platform accessible to most users. Taking these factors into account, we constructed a web application to prototype LookAtChat so that any device with a camera can access our website via a modern Internet browser such as Google Chrome.

4.2 Design Space

Elicited from formative interviews, we prototype LookAtChat to explore the design space visualizing gaze in video-conferencing. Considering popular video conferencing software is using 2D flat layouts for video image placement, we explore the design space on top of the traditional 2D flat layouts in order to expand the potential of 3D as well as hybrid layout alternatives. Hybrid dimension alternatives are proposed to combine 2D and 3D representations for taking advantage of both categories. The short description of our designs are listed in Table 1 with corresponding illustrations in Figure 2. Each sub-figure in Figure 2 illustrates a five-user scenario: user "panda" is speaking and looking at user "fox" while all the rest participants are listening to "panda" and looking at "panda".

Table 1. Proposed visualization of eye contact for remote video conferences

Category	Name	Description
2D flat layout	directional layout	Depict arrows between video streams to indicate sources and targets of eye contacts, while out-glowing the video window of users who are looking at the current user.
	animated flows	Render dynamic flows instead of arrows to convey eye contacts; the sizes of the flows are proportional to the duration of the gaze actions.
	text overlay	Overlay the text of "[who] is looking at [whom]" directly on captions of the video window.
	color highlights	Change the color of the video border to indicate the eye contacts while each distinct color is assigned to each participant.
	icon overlay	Append users' profile pictures to the caption area of the video window to convey eye contacts.
3D immersed layout	perspective layout	Apply perspective transformation of the video window to imply eye contact between other participants and gently shake the video of users who are looking at the current user.
	avatar / top-view	Render a top-view of 3D avatars of all users alongside with their video streams and change their rotation according to gaze actions.
	avatar / first-person	Warp live video streams to the 3D avatars of all users positioned along a curve; rotate the avatars to reflect their gaze actions; present a first-person perspective for the current user
	avatar / third-person	Based on "avatar / first-person", present a third-person perspective with isometric projection[6].
hybrid layout	split-view	Present a 2D flat layout and a 3D immersed layout side by side. Hence the 3D avatar layout doesn't need to wrap video streams to the avatars to avoid "uncanny valley" effects.
	picture-in-picture	Depict a 2D flat layout with video windows in full while a 3D avatar layout is rendered as an overview thumbnail at the screen center to convey eye contacts.

4.2.1 *2D flat layout.*

We first explore how to impose eye contact on 2D flat layout. The visualization of eye contact on 2D flat layouts could be illustrated in a direct or indirect manner. Direct design delivers straightforward signals with less cognitive load, while indirect design may result in less interference with the video streams. Figure 2(b) and (c) are illustrated how *directional layout* and *animated flows* convey eye contact. From perspective of user 'fox', 'panda' is looking at itself so the video frame is highlighted with outer glows. In the meantime, all other participants are looking at 'panda' so a static directional arrow is shown in Figure 2(b) and dynamic flow from observer to observee is rendered in Figure 2(c). *directional layout* applies fade-in and fade-out to indicate the start and end of eye contacts in a smooth transition.

Figure 2(d), (e), and (f) demonstrate how *text overlay*, *color highlights* and *icon overlay* visualize eye contact indirectly. Figure 2(d) shows text overlays at the bottom of video window. The names of observers are displayed following FIFO rule. The first name in the list is the one who looks at the observee earliest. Figure 2(e) renders different color borders

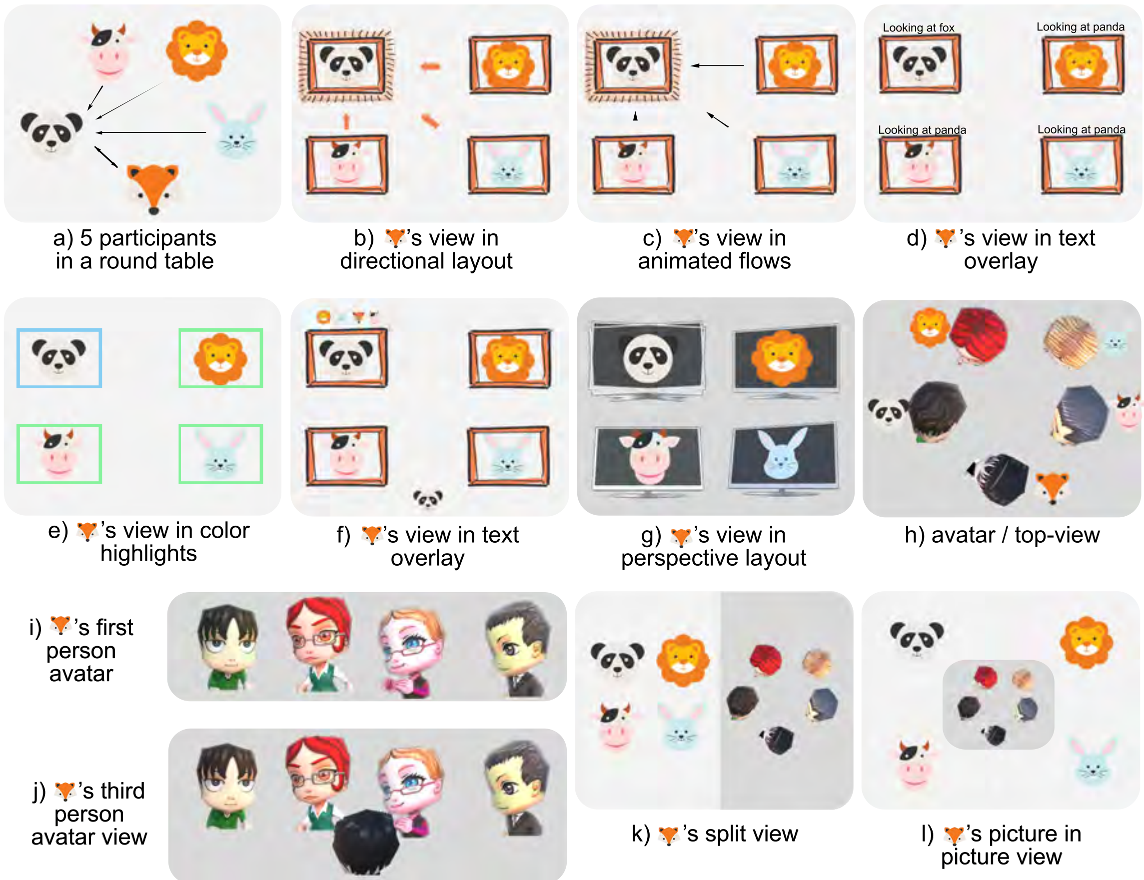


Fig. 2. Design space of LookAtChat. To convey eye contacts in a), 2D flat layouts (b – f), 3D immersed layouts (g - j), and hybrid layouts (k – l) illustrate the mutual gaze between "panda" and "fox" and how other participants gaze at "panda".

when different participants are looking at others. Likewise, Figure 2(f) illustrates the profile at the bottom of the video window. The profile is a thumbnail image of the corresponding user. We take the first frame of video as a thumbnail reference.

2D flat layouts are widely adopted in commercial video conferencing software. We provided two levels of eye contact visualization: direct and indirect. Direct eye contact options demonstrate eye contact to users intuitively, so it helps users immediately understand gaze information on a subconscious level. Indirect eye contact options imply the eye contact in a subtle way so users need to interpret the UI elements while the visual effects of elements are minimized so as not to be distracting.

4.2.2 3D immersed layout. We further investigate providing eye contact on 3D immersed layouts. The 3D immersed layouts are proposed to introduce spatial cues between participants as well as gaze information. The video image of all participants is attached to a monitor frame per user in the view. Instead of placing the video image directly in the scene and applying perspective transformation, we employ the "physical kinetic displays"[33] metaphor and attach the video image to a monitor frame. This helps users to perceive the video as a 3D display. Other alternatives such as painting frames or mirror frames can also replace the monitor frames. We also designed the 3D immersed layout with two different perspectives: first-person view and third-person view. Figure 2(g) and (i) are first-person perspective designs. In *perspective layout*, users see a billboard representation of the video stream shaking if being looked at or turning if looking at other participants. Thus, we see user "panda" is shaking slightly in "fox"s view and other users are turning to look at "panda" in Figure 2(g). In *avatar/first-person*, users see other participants as avatars representing their heads. The avatar will turn to the corresponding user when the user behind the webcam is looking at that user. So avatar 'panda' is facing to the viewer ('fox') and other avatars are turning to look at 'panda' in Figure 2(i).

Figure 2(h) and (j) show the third-person perspective designs in the 3D immersed layouts. In *avatar / top-view*, all participants' avatars are rendered from above and a real-time video texture is shown alongside. Users can infer the spatial relationships from the orientation of the avatars. In *avatar / third-person*, users' cameras are placed behind their own avatar so that each user can see other participants' head orientation as well as their own gaze cues.

The 3D immersed layout is designed to emphasize the spatial cues between participants. The effect is similar to 3D collaborative gaming experiences. We provide two levels of perspective and introduce 3D personalized avatars for user representation. Avatar representation is personalized according to real-time video textures (detailed in next section). The spatial cues and gaze information is designed to be natural and similar to real-life scenarios, though it may introduce some sense of the uncanny valley when warping the video image to fit the 3D avatars.

4.2.3 hybrid layout. We next explore the potential of hybrid layout designs that combine the 2D flat layout and 3D immersed layout. The hybrid layout is investigated to show a large video image and provide 3D gaze cues as well. We consider two rendering approaches: *split-view* (Figure 2(k) and *picture in picture* (Figure 2(l)). *split-view* organizes the 2D layout and 3D layout side-by-side. Users are able to perceive eye contact from the 3D layout while simultaneously viewing the other participants in the 2D layout. *picture in picture* allows users to acknowledge spatial cues at the center of their entire view. *split-view* allows users to choose the focus on either video texture or eye contact and spatial information. *picture in picture* prefers to present both pieces of information as a whole to users.

4.3 Directional Layout and Perspective Layout

As the first step towards visualizing eye contacts for remote small-group conversations, we chose to implement and experiment with three conditions: baseline layout, 2D *directional layout* and 3D *perspective layout*. We have several considerations for selecting *directional layout* and *perspective layout* for comparison. Our overarching goal is to explore how gaze and spatial information facilitate video conferences. *directional layout* and *perspective layout* both show eye contact in a direct manner, so that it is easy and straightforward for users to see and understand the system without further cognitive load. In addition, we chose not to include avatar designs in the first experiment because it would have required higher graphics processing capabilities of users' computers than designs not including avatars. Also, it is likely that the personalized avatars could introduce "uncanny valley" effects, which might negatively impact users' conversation quality. Lastly, hybrid layout designs are not selected for our exploratory user study, since we primarily want to understand how 2D flat layout and 3D immersed layout individually work for users.

Fig. 3. Video conferencing with varied numbers of participants in LookAtChat in 2D flat layout category. Size and placement of video image is updated in real-time according to the observer’s gaze.

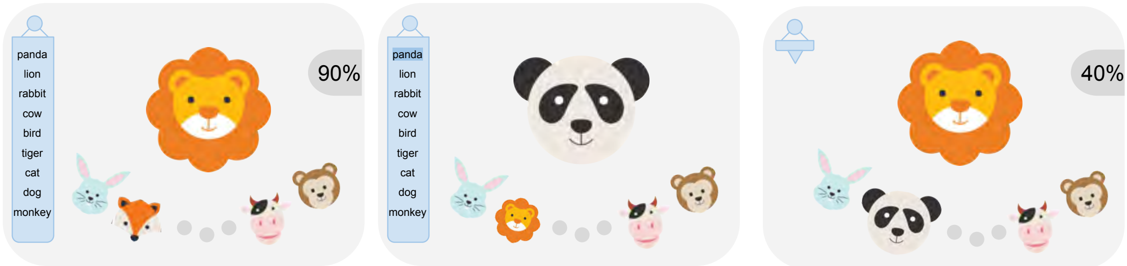
4.4 Use Cases

Figure 2 demonstrates all the designs in a small-group discussion scenario. Meanwhile, video conferences are widely utilized in a large variety of use cases. LookAtChat is designed to be easily adapted to different video conferencing requirements.

4.4.1 Small-group discussion. Small-group discussion is one of the majority use cases in video conferencing, either for working or for entertainment purposes such as brainstorming, playing games, etc. To save network bandwidth and decrease the cognitive load of users, LookAtChat provides a docked sidebar to show a full list of all participants. Users are free to choose a few participants of interest. The selected participants are rendered in a medium size video image at the beginning. The size will grow or shrink depending on the user’s focus. Users can select or de-select participants at any time during the video conference (shown in Figure 3(a) and (b)). The participants who received focus from the user are gradually moved to the center Figure 3(b) (c).

4.4.2 Presentation. Slides presentation is another strong use case in video conferences. During presentations, presenters focused mostly on the slides or the shared windows and on feedback from other participants. From formative interviews, people report being more interested in watching the slides than in watching the presenter. So LookAtChat may visualize participants’ focus with heatmaps. Presenters will see other participants’ gaze positions on the shared screen and listeners will see presenter’s focus instead.

4.4.3 Large-scale meeting. Videoconferencing with a large audience is popular and useful for remote seminars and all-hands meetings. By default, the lecturer is rendered to all the participants (Figure 4(a)). Participants are free to see other participant as shown in Figure 4(b). We proposed to use aggregated data in this form. For example, Figure 4(a) shows how lecturers perceived other participants’ looking at themselves and Figure 4(c) indicates when participants other than the lecturer receive focus.



a) lecturer's view: percentages of students who are paying attention to the lecture are listed by the lecture

b) a student (*fox*)'s view when panda volunteers to answer questions and obtains the most attention from the audience

c) lecture's view: users who receive more attention from the audience (e.g., panda) are highlighted with larger video frames.

Fig. 4. Presentation with a large audience in LookAtChat. An aggregated number of gaze-received is shown to the lecturer (see the top-right percentages in a) and c)). b) Audience members can choose to watch others in a large view. Audience members who receive more eye contact have larger size video frames than others for the lecturer to pay attention to.

5 IMPLEMENTATION

LookAtChat is designed and implemented for both video conferencing users and researchers to conduct remote user studies. LookAtChat comprises three major parts: a WebRTC server to support videoconferencing and hosting remote user studies, a real-time eye-contact detection module, and a WebGL-based renderer to visualize the gaze information.

5.1 Workflow

As Figure 5 demonstrates, LookAtChat employs a WebRTC server as well as peer-to-peer networking. For each newly-joined client, it talks to the WebRTC server (including Internet Connectivity Establishment server and Signaling) first to establish peer-to-peer connection with existing clients. Hence, the clients can send and receive video and audio streams with each other. Next, LookAtChat server maintains the identifier of each client after the WebRTC connection is established. For each client, gaze and audio level information are processed locally and sent to LookAtChat server. Afterwards, the server broadcasts the information to all of the clients and the renderer on the client side will locally visualize the gaze and audio information.

5.2 Host Mode to Support Remote User Study

Due to the challenges of the global COVID-19 pandemic, it is not encouraged to recruit and gather participants in a controlled lab environment. Hence, we implement a **host mode** to monitor different clients from their own perspective and record gaze and video data. The host is a special client that does not participate in the actual study but can act as one of the participants for assisting them with technical issues. By default, the renderer will visualize all participants together with their eye contacts. The host can observe any client by sending an “observe X” command to the LookAtChat server. The server then returns the layout of all video streams as observed by the designated user. The renderer will visualize that user’s gaze information so that the host can verify that the eye tracking modules are correctly calibrated.

To minimize bandwidth, the host does not send video and audio streams to other clients but only receives the streams from others. For post-study analysis, we record and save all the video, audio, and gaze data from the host machine.

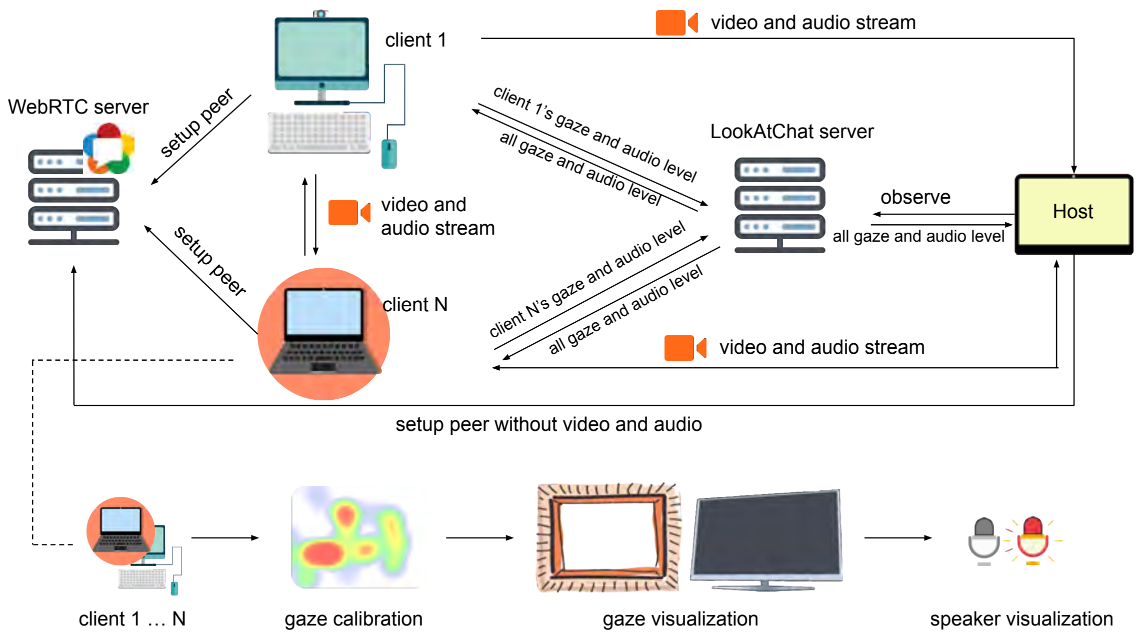


Fig. 5. LookAtChat workflow. As a regular client, video and audio streams are transmitted to each other in peer-to-peer manner. Each client is required to calibrate gaze first and then send individual gaze and audio levels to the LookAtChat server after joining the group. The LookAtChat server broadcasts received data to every client. Then each client renders the gaze according to the current layout and the microphone symbol.

Traditional video conferencing tools may allow the user to privately share their screens with the host. However, our approach reduces unnecessary communication efforts and allows the host to asynchronously examine remote user setups.

5.3 Detection of Eye Contacts

We leverage WebGazer [37] to calibrate and obtain raw gaze positions in each client. Constrained by the webcam setup with ordinary laptop or PC, WebGazer is a state-of-the-art, off-the-shelf software for eye tracking technologies. Different from WebGazer, which reports gaze coordinates, LookAtChat focuses on "who is looking at whom" for video conferencing. We smooth the coordinate outputs from WebGazer with 1€ Filter [7] and then classify the data to understand which client is being looked at.

Our system expects users to reach an accuracy of 80% during the calibration session and ensure that the size of the face is larger than 25% of the video frame. To improve accuracy, we fine-tune the parameters of 1€ Filter and video stream placement with a Tobii eye tracker.

First, we tune parameter *mincutoff* in 1€ filter to ensure the gaze coordinates are not jittering and *beta* to ensure the result is not introducing too much latency. An animated dot moves from the start to the end of a line. The animated dot is a circle with a 10-pixel radius (from experimental data). We move our cursor to follow the dot several times and record the cluster of cursor positions. The sum of the average distance between cursor and animated dot is set as a reference value. We next move our gaze to follow the animated dot and apply the same calculation. *Mincutoff* is tuned to ensure that the distance of gaze is comparable to the reference value. *Beta* is tuned to ensure that the latency is shorter than 5 ms between the raw gaze position and smoothed gaze position on average. We set *mincutoff* to be 0.3 and *beta* to be 0.3 for smoothing the raw gaze positions.

Second, we adjust the placement of the video stream. We calculate a zone ID to specify which video stream is being looked at. For ground truth (data from Tobii), we calculate a valid zone ID when the gaze dot is in the center of a video stream. The size of the central zone is defined as 1/4 of a regular video zone. With smoothed gaze coordinates, we calculate a valid zone ID when it is in the video zone. We reach 95% after changing the distance between two video stream areas horizontally and vertically. Then the proportion of the distance between two video streams (both in x and y) and the entire screen is recorded. In this way, we can ensure that LookAtChat behaves the same on different screens.

Eye contacts data is constructed as a pair of source ID and destination ID (which could be null) on each client. The data is sent to the LookAtChat server every 16 ms via web sockets to achieve real-time performance.

5.4 Rendering

LookAtChat renders two types of data for each client. As we integrate WebRTC into our system for peer-to-peer video communication (including video stream and audio stream), the video stream is rendered as a video texture through Three.js. Additionally, we retrieve the local audio level on each client. The audio level is sent to LookAtChat server and broadcast to all the clients. Thus, each client can see a microphone icon "on" or "off' based on the received audio level data (see "author1" in Figure 6(f)). We also record the audio level data for further data analysis.

Gaze data is rendered differently according to the layout. For directional layout, the out-glowing effect and the arrows are rendered with increasing opacity. Users can feel the action dynamically from the fade in and out effects. Figure 6(a) to (d) show the view of the same user "self" in directional layout. For example, the video stream of user "author1" in (a) is outglowing because user "author1" is looking at user "self". Regarding perspective transformation, the video image of the gaze source is transformed to facing to the video image of the target so that users can feel the movement of [who] is tuning and looking at [whom]. If the viewer is being gazed at, video texture of the gaze source will be slightly shaken. As Figure 6(e) illustrates, user "author 1" is looking at "self". Accordingly, user "author1" is looking at "fox" on the right (Figure 6(f)), at "egg" underneath (Figure 6(g)), and "minions" at the corner (Figure 6(h)). Interpolation is applied between different transformations for a smooth experience and also to simulate a "turning" action.

6 EVALUATION: SMALL-GROUP DISCUSSION

We conducted a user study in order to examine how the different layout variants perform in terms of conversation, subjective feedback, and user preferences, compared with a "baseline" layout where no gaze information is visualized. The user study follows a within-subject design in three conditions: baseline, directional layout, and perspective layout. The three conditions were counterbalanced to avoid bias in the following combinations: baseline–directional–perspective, directional–perspective–baseline, perspective–baseline–directional. The DVs were conversation experience defined in Sellen's work [43], user experience defined in Schrepp *et al.* 's work [42] and Hung *et al.* 's work [14], and Temple Presence Inventory (TPI) [25]. We processed the data through an analysis of variance (ANOVA). All tests for significance were made at the $\alpha = 0.05$ level. The error bars in the graphs show the 95% confidence intervals of the means.

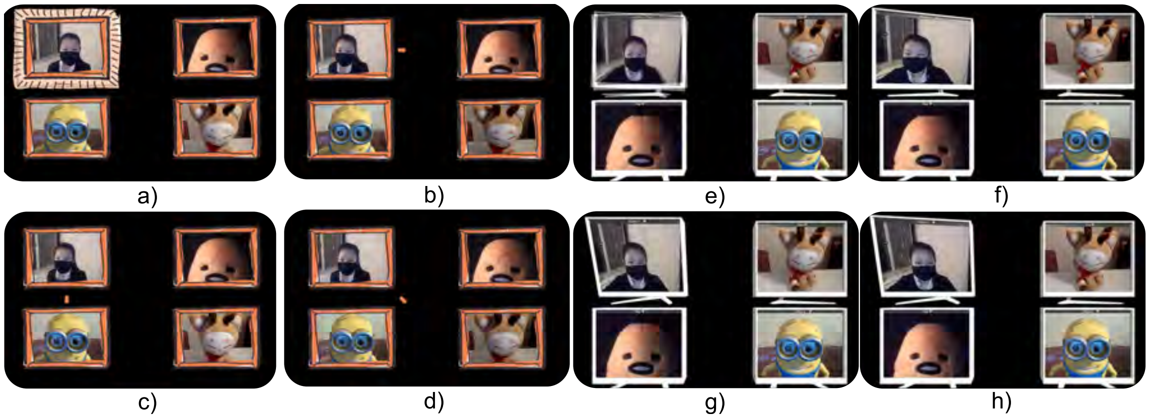


Fig. 6. a) – d) demonstrate screenshots of how the user recognize mutual gaze (a)) and eye contacts between the person on the upper left and other participants (b) – d)) in the 2D directional layout. Accordingly, e) – h) illustrate the gaze in the 3D perspective layout.

6.1 Participants and Apparatus

We recruited a total of 20 participants at least 18 years old with normal or corrected-to-normal vision (6 females and 14 males; age range: 24 − 39, $M = 28.55$, $SD = 3.62$) via social media and email lists. The participants have a diverse background from both academia and industry. None of the participants had been involved with this project before. We assign participants into four 5-person groups for the user study. The study was conducted remotely in personal homes. Participants used their personal computers with a webcam, visited the website we provided through Google Chrome browser, and experienced different conditions as instructed by the host. We instructed participants to take the user study in a quiet and brightly lit room where faces in the webcam are clearly visible from the background. For the duration of the study, participants' behavior, including their conversations, video streams, and gaze positions were observed and recorded.

6.2 Procedure

Our remote user study is scheduled using conventional calendar and videoconferencing tools (Zoom). Once all participants were online, the host briefly introduces LookAtChat system with a tutorial video and asks all participants to fill in consent forms. After the tutorial, the host instructs all participants to enter a designated layout in the user study website (https://eye.3dvar.com). Participants are instructed to mute their video&audio streams in Zoom to prevent echoing and save networking bandwidth. Meanwhile, the participants can still follow the host's instructions from Zoom and the host can monitor the experiments with the **host mode** in LookAtChat. The user study session of each condition consists of three parts: gaze calibration (∼5 min), warm-up conversation (∼3 min), and two game sessions of "who is the spy" (∼20 min). We now describe the three parts in more detail:

Gaze calibration. Participants are required to first calibrate their gaze individually. Our system adopts the calibration procedure of WebGazer[37]: A box rendered around participant's face mesh turns green when the participant is at the center of the camera view and close enough. Next, the participant calibrates 9 points on the screen and the accuracy of gaze point is reported. We suggest that the participant proceed after reaching 80%.

Warm-up conversation. In the beginning of each condition, researchers briefly describe how gaze information is visualized for the current condition. Later on, participants pick a topic (self introduction, favorite TV show, etc.) and

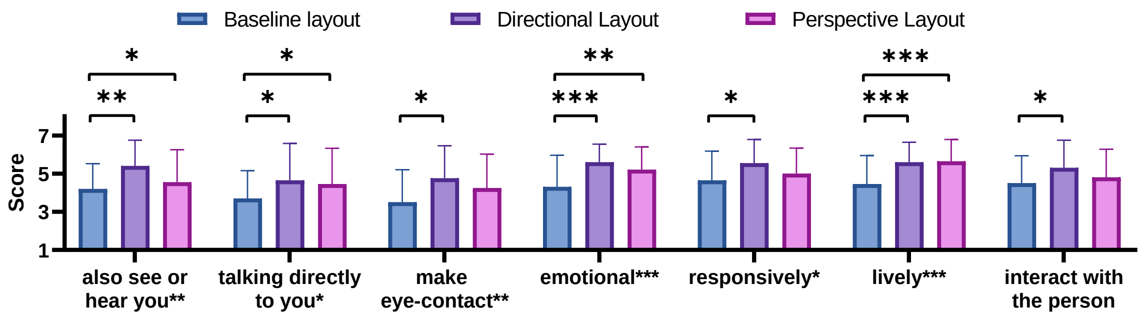


Fig. 7. Summary of significant results regarding TPI between baseline layout, directional layout, and perspective layout. *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001. 10 questions were selected in the category to be asked. We found 6 of them have significant effects through ANOVA and 7 of them have significant impacts in post hoc tests. Directional layout has notably better scores than baseline layout while perspective layout is slightly better.

one by one give a short speech for around 30 seconds. During the warm-up conversation, participants get familiar with the behavior of current condition.

Game "who is the spy". After the warm-up conversation, the group is instructed to play a party word game, *"Who is the spy"*, twice. Before the game starts, each participant receives a word: three of them act as detectives and get the same word (*e.g.*, William Shakespeare), while the other two act as spies and get a different word (*e.g.*, Leo Tolstoy). Only the spies know everyone's identities. For each round, each player needs to describe their word, talks about who may be the spies that received a different word. Spies will try to guess detectives' word and pretend they are holding the same word. Detectives will try to describe with ambiguity and infer spies with language and non-verbal cues. This game was selected because it is a conversation-based game which typically requires lots of eye contacts to tell who is lying and which two spies are teammates. At the end of each round, each player casts their vote for spy and the player with the most votes is put out of next rounds. At any time, detectives who successfully indict the spies and spies who successfully guess the detectives' word earn points. Each player has only one chance for the indictment or guess.

At the end of each study session, we ask the participants to fill an online questionnaire about conversation experience, user experience, and TPI [25] with a 7-point Likert scale for the condition they just completed (~5 min). Hence, the study session of each condition lasts for around 30 to 45 minutes. At the end of all the three conditions, we ask the participants to fill demographic information, scale of usability in general, and rank the conditions. Lastly, participants are interviewed about LookAtChat, reasons for their ranking, and gave suggestions. On average, the experiment takes about 100 to 120 minutes in total.

7 RESULTS

We validated that the data satisfies the assumptions of an analysis of variance (ANOVA). All tests for significance were made at the $\alpha = 0.05$ level. The error bars in the graphs show the standard error. Symbol * means $p <= .05$, ** means $p <= .01$, and *** means $p <= .001$.

7.1 Social experience

The results for the ratings of social experience questions that have significant results over three layouts are illustrated in Figure 7. For the question *"How often did you have the sensation that people you saw/heard could also see/hear* *you?"* ($M_{baseline} = 4.2, M_{dir} = 5.4, M_{per} = 4.55$), a one-way within-subjects ANOVA was conducted to test the influence of layout on the ratings. The main effects for layout ($F(1, 19) = 5.94, p = .006^{**}$) were significant. Post hoc t-tests with Holm correction showed a significant difference between baseline layout and perspective layout ($t(19) = -3.35, p < .006^{**}$) with a 'large' effect size (Cohen's d = .75), as well as between directional layout and perspective layout ($t(19) = 2.37, p < .046^{*}$) with a 'medium' effect size (Cohen's d = .53). The results indicate that LookAtChat with directional layout and perspective layout provided notably more bidirectional sensation than baseline layout.

For the question *"How often did it feel as if someone you saw/heard was talking directly to you?"*($M_{baseline} = 3.7, M_{dir} = 4.65, M_{per} = 4.45$), a one-way within-subjects ANOVA was conducted. The main effects for layout ($F(1, 19) = 4.93, p = .012^{*}$) were significant. Post hoc t-tests with Holm correction showed significant differences between baseline layout and other two layout (both $p < .05^{*}$) with a 'medium' effect size (Cohen's d = .52 to .66). The results suggest that LookAtChat with directional layout and perspective layout provided notably more feelings of direct conversation than baseline layout.

For the question *"How often did you want to or did you make eye-contact with someone you saw/heard?"*($M_{baseline} = 3.5, M_{dir} = 4.75, M_{per} = 4.25$), a one-way within-subjects ANOVA was conducted. The main effects for layout ($F(1, 19) = 6.71, p = .003^{**}$) were significant. Post hoc t-tests with Bonferroni correction showed significant differences between baseline layout and directional layout ($p < .002^{**}$) with a 'large' effect size (Cohen's d > .8). The results indicate that LookAtChat with directional layout provides significantly more eye contact than baseline layout.

Regarding the social richness questions including emotional ($M_{baseline} = 4.3, M_{dir} = 5.6, M_{per} = 5.2$), responsive ($M_{baseline} = 4.7, M_{dir} = 5.6, M_{per} = 5.0$), and lively ($M_{baseline} = 4.5, M_{dir} = 5.6, M_{per} = 5.7$), a one-way within-subjects ANOVA was conducted for each of them. The main effects for layout were significant on "emotional" (($F(1, 19) = 11.13, p < .001^{**}$)) with post hoc t-tests with Bonferroni correction showed significant differences between baseline layout and the other two layout (both $p < .006^{**}$) with a 'large' effect size (Cohen's d > .71); on "responsive" (($F(1, 19) = 3.69, p = .03^{*}$)) with post hoc t-tests (holm correction) showed significant difference between baseline layout and directional layout (both $p = .03^{*}$) with a 'medium' effect size (Cohen's d = .6); on "lively" (($F(1, 19) = 10.65, p < .001^{**}$)) and post hoc t-tests with Bonferroni correction showed significant differences between baseline layout and the other two layout (both $p < .001^{***}$) with a 'large' effect size (both Cohen's d > .87). The results indicate that LookAtChat with directional layout and perspective layout provides significantly more social richness feelings than baseline layout.

For other questions discussing social experience, we did not find significant effects over the three layouts. Furthermore, we found that question *"to what extent did you feel you could interact with the person or people you saw/heard?"* has "large" effect size ($\eta^2 = .144$). Post hoc t-tests with Holm correction shows baseline layout has significantly negative effects ($p < .05^{*}$) compared with the directional layout. Hence, the result suggests that LookAtChat with directional layout brings more interaction potential to users than the baseline layout.

7.2 User engagement and experience

The ratings of the question *"The visualization of the layout is clear and balanced."* over three layouts are illustrated in Figure 8 ($M_{baseline} = 5.8, M_{dir} = 5.7, M_{per} = 5.0$). A one-way within-subjects ANOVA was conducted to test the influence of layout on the ratings. The main effects for layout ($F(1, 19) = 3.83, p = .03^{*}$) were significant. Post hoc t-tests with Holm correction showed a significant difference between baseline layout and perspective layout ($t(19) = 2.54, p < .046^{*}$) with a 'medium' effect size (Cohen's d = .57). The results indicate that LookAtChat with baseline and directional layout is more clear and balanced than perspective layout.

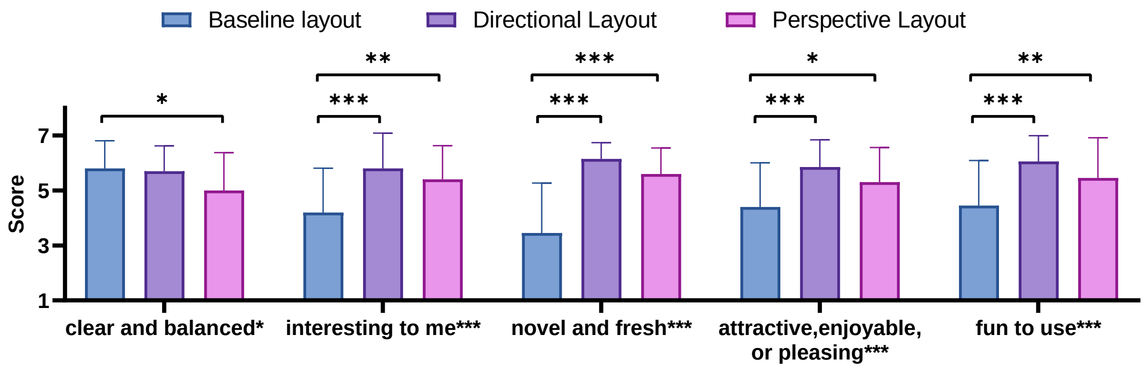


Fig. 8. Summary of significant results regarding user engagement and user experience between baseline layout, directional layout, and perspective layout. *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001. We provided 6 questions on user engagement and 4 questions on user experience for participants to fill. Results show that 3 user engagement questions and 2 user experience questions have significant effects through ANOVA. Directional layout and perspective layout both have significant effects on participants' feedback of feeling "interesting", "novel", "attractive" and "fun".

Regarding *"the content or features provided on this website were interesting to me."* ($M_{baseline} = 4.2, M_{dir} = 5.8, M_{per} = 5.4$), the main effects for layout ($F(1, 19) = 9.15, p = .001^{***}$) was significant with a 'large' effect size ($\omega^2 = 0.18$). Post hoc t-tests with Holm correction showed a significant difference between baseline layout and directional layout ($t(19) = -4.11, p < .001^{***}$) and perspective layout ($t(19) = -3.08, p < .008^{**}$). The results indicate that LookAtChat with baseline layout is significantly less interesting than directional and perspective layout.

In terms of *"the features provided by this website were novel and fresh."* ($M_{baseline} = 3.45, M_{dir} = 6.15, M_{per} = 5.6$), the main effects for layout ($F(1, 19) = 32.76, p < .001^{***}$) were significant with a 'large' effect size ($\omega^2 = 0.47$). Post hoc t-tests with Holm correction showed a significant difference between baseline layout and directional layout and perspective layout (both$p < .001^{***}$). The results indicate that LookAtChat with baseline layout is significantly less novel or fresh than directional and perspective layout.

Speaking of the overall impression of the design (attractive, enjoyable, or pleasing) with $M_{baseline} = 4.4, M_{dir} = 5.9, M_{per} = 5.3$. The main effects for layout ($F(1, 19) = 9.28, p = .001^{***}$) were significant with a 'large' effect size ($\omega^2 = 0.16$). Post hoc t-tests with Holm correction showed a significant difference between baseline layout and directional layout ($t(19) = -4.3, p < .001^{***}$) and perspective layout ($t(19) = -2.6, p = .02^{*}$). The result for ratings of *"fun to use"* over three layouts are $M_{baseline} = 4.5, M_{dir} = 6.1, M_{per} = 5.5$. The main effects for layout ($F(1, 19) = 11.49, p = .001^{***}$) were significant with a 'large' effect size ($\omega^2 = 0.17$). Post hoc t-tests with Holm correction showed a significant difference between baseline layout and directional layout ($t(19) = -4.75, p < .001^{***}$) and perspective layout ($t(19) = -2.97, p = .01^{**}$). The results suggest that LookAtChat with the baseline layout is significantly less attractive and fun than the directional and the perspective layouts overall.

7.3 Conversation experience

The results for the ratings of question *"I knew when people were listening or paying attention to me."* over three layouts are illustrated in Figure 9 ($M_{baseline} = 3.5, M_{dir} = 5.45, M_{per} = 4.55$). A one-way within-subjects ANOVA was conducted to test the influence of layout on the ratings. The main effects for layout ($F(1, 19) = 12.50, p = .001^{***}$) were significant. Post hoc t-tests with Holm correction showed a significant difference between baseline layout and directional layout

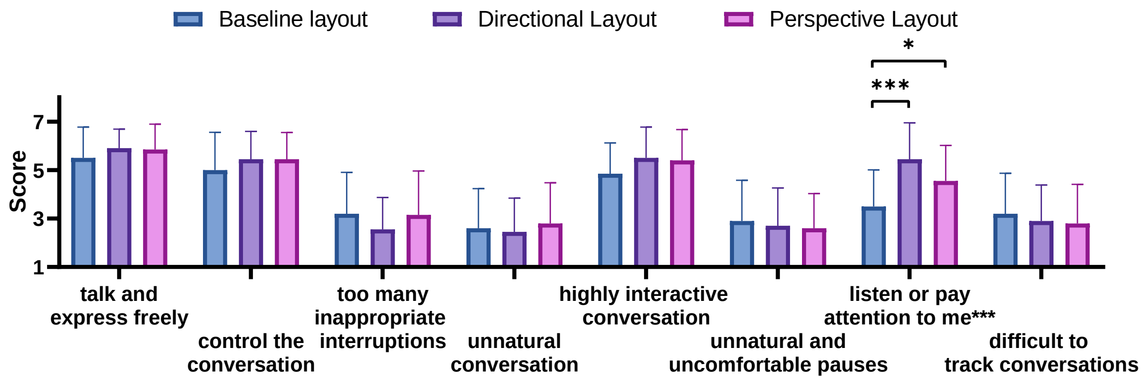


Fig. 9. Summary of the results regarding conversation experience between baseline layout, directional layout, and perspective layout. *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001. Directional layout and perspective layout both have significant effects on the feeling of "attention". For other questions, directional layout and perspective layout have better but not significant scores than baseline layout.

($t(19) = -5.00, p < .001^{***}$) and perspective layout ($t(19) = -2.68, p < .02^{*}$) with a 'large' effect size (both Cohen's d > .6). The results indicate that LookAtChat with both directional layout and perspective layout provided notably more feedback of attention than baseline layout.

We found there was no significant effect of layout on other conversation questions. Post hoc t-tests with Holm correction did not show significant differences among all questions in this category.

7.4 Usability, Preferences and Subjective feedback

Participants rated the usability of LookAtChat in general after they experienced all three layouts. Participants (like P10,M) thought the system is "*easy to understand*". The overall usability score is 72.5, which is interpreted to have higher perceived usability than 70% of all products tested.

Fourteen participants preferred the directional layout, five participants preferred the perspective layout, and one preferred the baseline layout. P6(M) thought "*arrows and highlights are easier for me to notice the other's attention*". Similarly, P11(F) agreed that "*the highlight of when people look at you is more obvious.*". For participants who preferred perspective layout, participants found it "*fun*"(P9, M) and "*having such eye contact design in video conferencing is interesting and I feel the motive of being focus on other's speeches.*" (P7, F). Additionally, P1(M) and P2(F) shared similar comments that "*more intuitive than the other designs*" and "*simple form but it also gives you info*". The only user (P4,F) who preferred the baseline layout explained that the baseline layout is "*succinct*". From the reason of preferences reported by the participants, we found that participants tended to choose the design that is "clear and easy to understand" (P18, F) and (P12, F), however, they have different perceptions on defining "clear". Providing information that is more than users expect lead to negative results. An automatic adjuster or manual one would be useful.

During the interview, we asked all the participants whether they think eye contacts are worth visualizing as well as spatial relationship. 15 out of 20 agreed it is helpful and others have concerns about being overwhelmed. We further asked how they think about eye contacts between themselves and other people, and eye contacts between other people. 19 out of 20 reported they want to know if they are being looked at as well as whether other people receive their gaze information. We believe that visualizing eye contacts data will improved users' engagement, social richness, and helpful to conversation experience while videoconferencing. Since LookAtChat only requires a consumable webcam, participants such as P4(F) and P5(M) reported that it should be feasible to be integrated into existing videoconferencing systems, although may cause extra rendering costs.

8 DISCUSSION

In general, LookAtChat is an *"easy-to-use"* system with an overall usability score of 72.5 in this user study. Participants found it *"easy to understand"* and *"fun to use"*. Compared with the baseline layout, the directional layout and the perspective layout provide better feedback on social experience, user engagement, and conversation experience.

RQ1: How do people perceive eye contact without visualization in conventional videoconferencing?

Informed by the interview feedback, participants usually don't interpret eye contacts in videoconferencing. Commented by P18(F), "*I focus more on the audio so I won't miss what other people are talking about.*". In the meantime, participants are used to interpreting gaze offset to a very high degree of accuracy [11]. Hence, providing eye contacts through user interface is worth researching in parallel with gaze correction technologies.

RQ2: Can visualization of eye contact improve communication efficiency?

According to the conversation experience feedback, participants have better experiences with conversation flow, including self expression, controlling the conversation, and tracking the conversation, but not significantly. Participants felt the conversation is more interactive. Meanwhile, participants reported fewer inappropriate interruptions, fewer uncomfortable pauses, in addition to feeling less unnatural when using LookAtChat with directional or perspective layouts though not significantly. Importantly, LookAtChat has a significant effect on participants' belief that other people are listening and paying attention to them. As P4(F) reported, "*visualization of eye contact is definitely helpful, knowing that there were people watching at least helped me stay focused for the entire time.*"

RQ3: In what circumstances will users prefer to see eye contact?

We found that participants have different preferences while in different roles, as a meeting host, or as a talk attender. In addition, the purpose of the video conference mattered. As P5(M) described, "business meetings may value more on the quality of video image however colorful UI designs (like arrows in directional layout) may distract people from that". For small-group discussion, P4(F) agreed that "providing such information motivates me to engage more in such videoconferencing like brainstorming." Furthermore, participants placed more value on eye contacts relevant to themselves than between other participants. P5(M) thought eye contacts between other participants "*are helpful but too much for me to process at one time*". Investigating the effects of enabling eye contacts among other participants is worth researching.

RQ4: What forms of eye contact visualization may be preferred by users?

More participants preferred the directional layout (N=14) than the perspective layout (N=5). The directional layout shows arrows and outer glows that "*are easier to notice other's attention*"(P6, M). It applies "*no change on video image*"(P18, F) and indicates the interaction "*more visually for better connection*" (P13, M). The perspective layout provides spatial relationships in 3D. Reported by P1(M), it has "*better visual presentation of gazing at somebody, more intuitive*". P7(F) Comparing directional and perspective layout, "*directional layout is somehow too obvious and may require extra effort to focus on my speech*" while "*perspective layout eliminates this issue and I got the balance between freely speaking and knowing I was listened to by others.*". Briefly, we can tell that participants prefer the the designs that are "clear" to them and with smooth transition between looking at self and others.

Design Implications:

(1) Show the eye contacts intuitively.
In video conferences, participants mainly focus on the conversation. If the information provided through the visual design is too indirect and may require additional cognitive load, participants may feel distracted and lost in the conversation. For example, out-glowing in the directional layout receives positive comments from 7 participants because it is easy to understand.
(2) Control the visualization level.
Participants' preference is affected by "how attractive the design is" and "how I want the design to be attractive". For example, P7(F) preferred the perspective layout because the directional layout is relatively more distracting to her when she wanted to focus on listening. Providing a slider for adjusting the level and automatically controlling the level with audio data is helpful.
(3) Provide the control of gaze data transmission.
As video conferencing users have the options to mute or hide the video during video conferencing, most participants want to hide or send anonymous gaze data to the host in video conferences.
(4) Scale the design for various user scenarios.
Although we only evaluate small-group setup for LookAtChat, we designed and interviewed participants about their opinions on large team meetings or presentation with a large audience. An important future direction is to adapt the system to fit different use cases and larger numbers of users.
(5) Provide host mode.
Host mode is not only helpful for researchers to understand participants from their views but essential for conducting remote user studies as well.

Limitations. While LookAtChat is designed for remote video conferencing, as a proof-of-concept, we do not support more than one participant to be co-located. Our eye-contact detection algorithm only supports one user in front of the webcam and the accuracy is limited by the algorithms and individual calibration procedure. In terms of the user study, we only evaluate small-group discussion without shared presentation. As the ages of our participants spanned 24 - 39, the results of our study may not generalize to other populations such as junior students or elder adults who may prefer more or less eye contacts in video conferences. As our user study was conducted remotely, the bandwidth of home networks may impact how the users perceive our system. A small number of participants encountered frame-dropping during the conversation due to networking issues, which may negatively impact their assessment. Furthermore, as users of our system were only able to engage with LookAtChat for casual and gaming conversations, their assessment of how such system may help in other use cases such as team meetings or lectures is not fully conclusive.

9 CONCLUSION AND FUTURE WORK

In this paper, we introduce LookAtChat, a web-based video conferencing system which supports visualizing eye contacts for small-group conversations. Motivated by missing gaze information in conventional video conferences, we investigate the demands of gaze information by conducting five formative interviews. We further explore the design space of visualizing eye contacts with video streams of small groups and propose 11 layouts by brainstorming in focused groups. As a proof-of-concept, we develop and open source LookAtChat which supports eye contact visualization for small-group conversations. We conduct a remote user study of 20 participants to examine the benefits and limitations of the interfaces, as well as the potential impacts of user engagement and experience on the conversation. The quantitative results indicate that LookAtChat with directional layout and perspective layout provided notably more bidirectional sensation, feelings of direct conversation, social experience, and engagement than the baseline layout. More participants prefer the 2D directional layout to the 3D perspective layout because it is simpler and easier to understand.

We plan to explore several future directions for improving LookAtChat. First, we plan to implement more layouts from our design space exploration stage and establish an open-source community to develop more layouts to the system. Second, we intend to integrate privacy protection filters for users to select whether or not to share their own gaze information. Third, more advanced real-time neural models may be leveraged to improve the tracking accuracy in LookAtChat and balance the trade off between accuracy and real-time performance.

As an initial step toward visualizing eye contacts in conventional video conferencing interfaces with commonly accessible hardware requirements, we believe our work may inspire more designs to convey nonverbal cues for remote conversations. Such features may eventually be integrated with video conferencing software to increase social engagement and improve conversation experience.

REFERENCES

[1] Russell Lennart Andersson and Homer H Chen. 1997. Method for Achieving Eye-to-Eye Contact in a Video-Conferencing System. US Patent 5,675,376.

[2] Russell L Andersson, Tsuhan Chen, and Barin G Haskell. 1996. Video Conference System and Method of Providing Parallax Correction and a Sense of Presence. US Patent 5,500,671.

[3] Mark Ashdown, Kenji Oka, and Yoichi Sato. 2005. Combining Head Tracking and Mouse Input for a GUI on Multiple Monitors. In *CHI'05 Extended Abstracts on Human Factors in Computing Systems*. 1188–1191. https://doi.org/10.1145/1056808.1056873

[4] Stephan Beck, Andre Kunert, Alexander Kulik, and Bernd Froehlich. 2013. Immersive Group-to-Group Telepresence. *IEEE Transactions on Visualization and Computer Graphics* 19, 4 (2013), 616–625. https://doi.org/10.1109/TVCG.2013.33

[5] Mark Billinghurst and Hirokazu Kato. 2000. Out and About—real World Teleconferencing. *BT Technology Journal* 18, 1 (2000), 80–82. https://doi.org/10.1023/A:1026582022824

[6] Deng Cai, Xiaofei He, Jiawei Han, et al. 2007. Isometric Projection. In *AAAI*. 528–533.

[7] Géry Casiez, Nicolas Roussel, and Daniel Vogel. 2012. 1€ Filter: a Simple Speed-Based Low-Pass Filter for Noisy Input in Interactive Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2527–2530. https://doi.org/10.1145/2207676.2208639

[8] Antonio Criminisi, Jamie Shotton, Andrew Blake, and Philip HS Torr. 2003. Gaze Manipulation for One-to-One Teleconferencing. In *ICCV*, Vol. 3. 13–16. https://doi.org/10.5555/946247.946637

[9] Wei Yong Eng, Dongbo Min, Viet-Anh Nguyen, Jiangbo Lu, and Minh N Do. 2013. Gaze Correction for 3D Tele-Immersive Communication System. In *IVMSP 2013*. IEEE, IEEE, 1–4. https://doi.org/10.1109/IVMSPW.2013.6611942

[10] Jim Gemmell, Kentaro Toyama, C Lawrence Zitnick, Thomas Kang, and Steven Seitz. 2000. Gaze Awareness for Video-Conferencing: a Software Approach. *IEEE MultiMedia* 7, 4 (2000), 26–35. https://doi.org/10.1109/93.895152

[11] David M Grayson and Andrew F Monk. 2003. Are You Looking at Me? Eye Contact and Desktop Video Conferencing. *ACM Transactions on Computer-Human Interaction (TOCHI)* 10, 3 (2003), 221–243. https://doi.org/10.1145/937549.937552

[12] Raja Gumienny, Lutz Gericke, Matthias Quasthoff, Christian Willems, and Christoph Meinel. 2011. Tele-Board: Enabling Efficient Collaboration in Digital Design Spaces. In *Proceedings of the 2011 15th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE, 47–54. https://doi.org/10.1109/CSCWD.2011.5960054

[13] Beverly L Harrison, Hiroshi Ishii, Kim J Vicente, and William AS Buxton. 1995. Transparent Layered User Interfaces: an Evaluation of a Display Design to Enhance Focused and Divided Attention. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 317–324. https://doi.org/10.1145/223904.223945

[14] Ya-Hsin Hung and Paul Parsons. 2017. Assessing User Engagement in Information Visualization. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 1708–1717. https://doi.org/10.1145/3027063.3053113

[15] Hiroshi Ishii and Minoru Kobayashi. 1992. ClearBoard: a Seamless Medium for Shared Drawing and Conversation With Eye Contact. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 525–532. https://doi.org/10.1145/142750.142977

[16] Andrew Jones, Magnus Lang, Graham Fyffe, Xueming Yu, Jay Busch, Ian McDowall, Mark Bolas, and Paul Debevec. 2009. Achieving Eye Contact in a One-to-Many 3D Video Teleconferencing System. *ACM Transactions on Graphics (TOG)* 28, 3 (2009), 1–8. https://doi.org/10.1145/1531326.1531370

[17] Kibum Kim, John Bolton, Audrey Girouard, Jeremy Cooperstock, and Roel Vertegaal. 2012. TeleHuman: Effects of 3d Perspective on Gaze and Pose Estimation With a Life-Size Cylindrical Telepresence Pod. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2531–2540. https://doi.org/10.1145/2207676.2208640

[18] André Kunert, Alexander Kulik, Stephan Beck, and Bernd Froehlich. 2014. Photoportals: Shared References in Space and Time. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1388–1399. https://doi.org/10.1145/2531602.2531727

[19] Kuno Kurzhals, Fabian Göbel, Katrin Angerbauer, Michael Sedlmair, and Martin Raubal. 2020. A View on the Viewer: Gaze-Adaptive Captions for Videos. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12. https://doi.org/10.1145/3313831.3376266

[20] Claudia Kuster, Tiberiu Popa, Jean-Charles Bazin, Craig Gotsman, and Markus Gross. 2012. Gaze Correction for Home Video Conferencing. *ACM Transactions on Graphics (TOG)* 31, 6 (2012), 1–6. https://doi.org/10.1145/2366145.2366193

[21] Gun A Lee, Theophilus Teo, Seungwon Kim, and Mark Billinghurst. 2018. A User Study on Mr Remote Collaboration Using Live 360 Video. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 153–164. https://doi.org/10.1109/ISMAR.2018.00051

[22] Daniel Leithinger, Sean Follmer, Alex Olwal, and Hiroshi Ishii. 2014. Physical Telepresence: Shape Capture and Display for Embodied, Computer-Mediated Remote Collaboration. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*. ACM, 461–470. https://doi.org/10.1145/2642918.2647377

[23] Jiannan Li, Saul Greenberg, Ehud Sharlin, and Joaquim Jorge. 2014. Interactive Two-Sided Transparent Displays: Designing for Collaboration. In *Proceedings of the 2014 Conference on Designing Interactive Systems*. ACM, 395–404. https://doi.org/10.1145/2598510.2598518

[24] Zhen Li, Michelle Annett, Ken Hinckley, Karan Singh, and Daniel Wigdor. 2019. HoloDoc: Enabling Mixed Reality Workspaces That Harness Physical and Digital Content. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14. https://doi.org/10.1145/3290605.3300917

[25] Matthew Lombard, Theresa B Ditton, and Lisa Weinstein. 2009. Measuring Presence: the Temple Presence Inventory. In *Proceedings of the 12th Annual International Workshop on Presence*. 1–15.

[26] Andrew Maimone, Xubo Yang, Nate Dierk, Andrei State, Mingsong Dou, and Henry Fuchs. 2013. General-Purpose Telepresence With Head-Worn Optical See-Through Displays and Projector-Based Lighting. In *2013 IEEE Virtual Reality (VR)*. IEEE, 23–26. https://doi.org/10.1109/VR.2013.6549352

[27] Carman Neustaedter, Carolyn Pang, Azadeh Forghani, Erick Oduor, Serena Hillman, Tejinder K Judge, Michael Massimi, and Saul Greenberg. 2015. Sharing Domestic Life Through Long-Term Video Connections. *ACM Transactions on Computer-Human Interaction (TOCHI)* 22, 1 (2015), 1–29. https://doi.org/10.1145/2696869

[28] David Nguyen and John Canny. 2005. MultiView: Spatially Faithful Group Video Conferencing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 799–808. https://doi.org/10.1145/1240624.1240846

[29] David T Nguyen and John Canny. 2007. Multiview: Improving Trust in Group Video Conferencing Through Spatial Faithfulness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1465–1474.

[30] Kenton O'hara, Jesper Kjeldskov, and Jeni Paay. 2011. Blended Interaction Spaces for Distributed Team Collaboration. *ACM Transactions on Computer-Human Interaction (TOCHI)* 18, 1 (2011), 3. https://doi.org/10.1145/1959022.1959025

[31] Judith S Olson, Gary M Olson, and David K Meader. 1995. What Mix of Video and Audio Is Useful for Small Groups Doing Remote Real-Time Design Work?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 362–368. https://doi.org/10.1145/223904.223951

[32] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, Vladimir Tankovich, Charles Loop, Philip A.Chou, Sarah Mennicken, Julien Valentin, Vivek Pradeep, Shenlong Wang, Sing Bing Kang, Pushmeet Kohli, Yuliya Lutchyn, Cem Keskin, and Shahram Izadi. 2016. Holoportation: Virtual 3D Teleportation in Real-Time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST)*. ACM, 741–754. https://doi.org/10.1145/2984511.2984517

[33] Kazuhiro Otsuka. 2016. MMSpace: Kinetically-Augmented Telepresence for Small Group-to-Group Conversations. In *Virtual Reality (VR), 2016 IEEE*. IEEE, 19–28. https://doi.org/10.1109/VR.2016.7504684

[34] Kazuhiro Otsuka. 2017. Behavioral Analysis of Kinetic Telepresence for Small Symmetric Group-to-Group Meetings. *IEEE Transactions on Multimedia* 20, 6 (2017), 1432–1447. https://doi.org/10.1109/TMM.2017.2771396

[35] Kazuhiro Otsuka, Shiro Kumano, Ryo Ishii, Maja Zbogar, and Junji Yamato. 2013. Mm+ Space: Nx 4 Degree-of-Freedom Kinetic Display for Recreating Multiparty Conversation Spaces. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*. 389–396. https://doi.org/10.1145/2522848.2522854

[36] Kazuhiro Otsuka, Shiro Kumano, Dan Mikami, Masafumi Matsuda, and Junji Yamato. 2012. Reconstructing Multiparty Conversation Field by Augmenting Human Head Motions Via Dynamic Displays. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. 2243–2248. https://doi.org/10.1145/2212776.2223783

[37] Alexandra Papoutsaki, Patsorn Sangkloy, James Laskey, Nediyana Daskalova, Jeff Huang, and James Hays. 2016. WebGazer: Scalable Webcam Eye Tracking Using User Interactions. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*. AAAI, AAAI, 3839–3845.

[38] Tomislav Pejsa, Julian Kantor, Hrvoje Benko, Eyal Ofek, and Andrew Wilson. 2016. Room2room: Enabling Life-Size Telepresence in a Projected Augmented Reality Environment. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1716–1725. https://doi.org/10.1145/2818048.2819965

[39] Ramesh Raskar, Greg Welch, Matt Cutts, Adam Lake, Lev Stesin, and Henry Fuchs. 1998. The Office of the Future: a Unified Approach to Image-Based Modeling and Spatially Immersive Displays. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*. ACM, 179–188. https://doi.org/10.1145/280814.280861

[40] Holger Regenbrecht and Tobias Langlotz. 2015. Mutual Gaze Support in Videoconferencing Reviewed. *CAIS* 37 (2015), 45. https://doi.org/10.1145/2207676.2208640

[41] Derek F Reilly, Hafez Rouzati, Andy Wu, Jee Yeon Hwang, Jeremy Brudvik, and W Keith Edwards. 2010. TwinSpace: an Infrastructure for Cross-Reality Team Spaces. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*. ACM, 119–128. https://doi.org/10.1145/1866029.1866050

[42] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2017. Construction of a Benchmark for the User Experience Questionnaire (UEQ). *IJIMAI* 4, 4 (2017), 40–44. https://doi.org/10.9781/ijimai.2017.4457

[43] Abigail J Sellen. 1995. Remote Conversations: the Effects of Mediating Talk With Technology. *Human-Computer Interaction* 10, 4 (1995), 401–444. https://doi.org/10.1207/s15327051hci100_2

[44] David Sirkin, Gina Venolia, John Tang, George Robertson, Taemie Kim, Kori Inkpen, Mara Sedlins, Bongshin Lee, and Mike Sinclair. 2011. Motion and Attention in a Kinetic Videoconferencing Proxy. In *IFIP Conference on Human-Computer Interaction*. Springer, Springer, 162–180. https://doi.org/10.1007/978-3-642-23774-_16

[45] Misha Sra, Aske Mottelson, and Pattie Maes. 2018. Your Place and Mine: Designing a Shared VR Experience for Remotely Located Users. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 85–97. https://doi.org/10.1145/3196709.3196788

[46] William Steptoe, Robin Wolff, Alessio Murgia, Estefania Guimaraes, John Rae, Paul Sharkey, David Roberts, and Anthony Steed. 2008. Eye-Tracking for Avatar Eye-Gaze and Interactional Analysis in Immersive Collaborative Virtual Environments. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*. 197–200. https://doi.org/10.1145/1460563.1460593

[47] Tony Tam, Joseph A Cafazzo, Emily Seto, Mary Ellen Salenieks, and Peter G Rossos. 2007. Perception of Eye Contact in Video Teleconsultation. *Journal of Telemedicine and Telecare* 13, 1 (2007), 35–39. https://doi.org/10.1258/135763307779701239

[48] Kar-Han Tan, Dan Gelb, Ramin Samadani, Ian Robinson, Bruce Culbertson, and John Apostolopoulos. 2010. Gaze Awareness and Interaction Support in Presentations. In *Proceedings of the 18th ACM International Conference on Multimedia* (Firenze, Italy) *(MM '10)*. ACM, New York, NY, USA, 643–646. https://doi.org/10.1145/1873951.1874041

[49] Anthony Tang, Michel Pahud, Kori Inkpen, Hrvoje Benko, John C Tang, and Bill Buxton. 2010. Three's Company: Understanding Communication Channels in Three-Way Distributed Collaboration. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*. ACM, 271–280. https://doi.org/10.1145/1718918.1718969

[50] Roel Vertegaal. 1999. The GAZE Groupware System: Mediating Joint Attention in Multiparty Communication and Collaboration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 294–301. https://doi.org/10.1145/302979.303065

[51] Roel Vertegaal, Ivo Weevers, Changuk Sohn, and Chris Cheung. 2003. GAZE-2: Conveying Eye Contact in Group Video Conferencing Using Eye-Controlled Camera Direction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 521–528. https://doi.org/10.1145/642611.642702

[52] Simon Voelker, Sebastian Hueber, Christian Holz, Christian Remy, and Nicolai Marquardt. 2020. GazeConduits: Calibration-Free Cross-Device Collaboration Through Gaze and Touch. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–10. https://doi.org/10.1145/3313831.3376578

[53] David J Ward, Alan F Blackwell, and David JC MacKay. 2000. Dasher—a Data Entry Interface Using Continuous Gestures and Language Models. In *Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology*. 129–137. https://doi.org/10.1145/354401.354427

[54] L-Q Xu, A Loffler, PJ Sheppard, and D Machin. 1999. True-View Videoconferencing System Through 3-D Impression of Telepresence. *BT Technology Journal* 17, 1 (1999), 59–68.

[55] Jiejie Zhu, Ruigang Yang, and Xueqing Xiang. 2011. Eye Contact in Video Conference Via Fusion of Time-of-Flight Depth Sensor and Stereo. *3D Research* 2, 3 (2011), 5.