Are You Gazing at Me? Eye Contact and Mini Virtual Room in Web Video Conferencing

Beibei Xiong
University of British Columbia
Vancouver, Canada
bear233@student.ubc.ca

Kaseya Xia
University of British Columbia
Vancouver, Canada
zxia0101@student.ubc.ca

ABSTRACT

Web Video Conferencing (WVC) system has become an essential tool for remote-learning and remote-working due to the pandemic. However, the communication efficiency of the current WVC systems is hindered by the lack of eye contact. This problem remains largely unsolved at the consumer level. This paper introduces a new way to achieve eye contact for multi-person teleconferencing. Our WVC prototype, *VirtualGazer*, simulates eye contact with a 2D eye model and a mini-virtual-room VR model. We conducted empirical user study and semi-structured interviews to investigate how including eye contact in current WVC systems affects users' online meeting experience. We found that eye contact can enrich interactive experiences, and enhance engagement level and focus level. Our VR model is found to generally attract more attention from users than the eye model. Finally, we highlight limitations such as rendering quality and additional features of avatars for future improvements.

KEYWORDS

Human Computer Interaction, Attentive User Interfaces, VR, Web Video Conferencing, Teleconferencing, Eye Contact, Gaze Tracking

1 INTRODUCTION

Web video conferencing (WVC) has become an essential tool in remote-learning and remote-working situations. It allows users to see and hear one another in real time. However, this technology currently does not do well in conveying eye contact. Users tend to look at the video feed of their partners while talking, but they appear to be looking at somewhere else due to the position of camera and screen.

In distant-learning classes, presenters (e.g. professors, teachers, students) often feel disengaged or less interactive when there is only standard audiovisual feedback coming from the audience. For example, in the Gallery View in Zoom (and other WVC applications), the audience is represented by black boxes and name tags, as shown in Figure 1(a). If the participant has the video turned on, the webcam video stream replaces their box. We will refer to the space each participant takes up on the screen as their footprint. Notice that everyone has the same and uniform footprint, regardless of whether they are paying attention to the meeting or not.

CPSC 544Y '21W2, April 25, 2021, Vancouver, BC

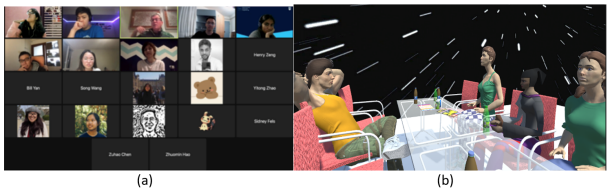


Figure 1: (a) Current WVC model: each participant stays in their grid with no eye contact interaction. (b) Proposed model: students can look at each other in a mini virtual room to create virtual eye contact.

Current systems neglect important cues presenters use to moderate their lecture. The lack of interactivity is one reason why online lectures are less effective than in-person lectures [25]. In one of the first experimental studies [8] on the effects of traditional instruction versus online learning. Researchers [8] found modest evidence that the traditional format trumps the online format in engagement. Many people also find it odd to see their faces during conversations. Conventional WVC services only offer standard visual and audio communication, and they do not support intuitive and personalized eye contact between users. Therefore, people still prefer face-to-face meetings because of the highly interactive meeting environment[10]. Lastly, some people are camera-shy and do not want to reveal themselves in WVCs. Thus, we decide to explore the effect on participation and engagement by using an avatar instead of a live video.

We propose *VirtualGazer*, a WVC system that adds eye contact and gaze amongst participants by replacing the gallery view with a mini virtual room. Our project explores whether adding eye contact to the current WVC system will enhance the interaction among users. Figure 1 shows what we intend to build in contrast to the existing WVC platforms like Zoom.

We created 3 different mock scenarios with scripted 3D avatars. To test our system, we recruited 13 volunteers to study the effects of the additional eye contact and gaze cues by attending our mocked meetings.

The key factors we want to explore in this study are changes in participants' attention, immersion and interactivity. To explore parameters that affect these factors, we consolidate these ideas into three core research questions:

By enabling eye contact and adding a mini virtual room to current WVC system:

RQ1: Can a person tell if they are being looked at and how can 3D avatars enhance this experience?

RQ2: How does a person's nervous level, focus level, and engagement level change?

RQ3: Can a person tell if other participants are looking at each other?

In summary, the major contributions of this paper include:

- Implementing A proof-of-concept: *VirtualGazer*, a WVC platform that simulates eye contact with 2D eye model and VR avatar model.
- Conducting an empirical 13-person user study on *VirtualGazer* to test its ability to provide eye contact and to examine the effects of enabling eye contact.
- Identifying opportunities for future HCI research: from supporting eye contact for WVC systems, to exploring design and prototype challenges.

2 RELATED WORK

2.1 Web Video Conferencing

Web Video Conferencing (WVC) is a real-time teleconferencing system that offers audio and visual communication among two or more participants. Examples of modern WVC services are Zoom, Collaborate Ultra, Microsoft Teams, etc. Zoom's usage surged exponentially due to the impact of COVID-19 pandemic, resulting in over 100-fold of profits compared with two years ago [19]. Researchers studied students' satisfaction level after the majority of the classes transitioned into online learning. [31, 32] show that WVC has become one of the most popular online teaching methods and has gained higher satisfaction scores than other platforms. The unique breakout room feature in Zoom creates a more collaborative and engaging experience for students [1]. However, in the study by [6], 80% of the students felt they would be more engaged in a classical classroom setting, and 57% of the students thought WVC technology is a barrier to their interaction with instructors. Conventional WVC services (e.g. Zoom), as shown in Figure 1(a), offer standard audio and visual communication but lack innovation in bringing participants' social hints such as intuitive and personalized eye contact to the audiences.

2.2 Eye Contact in Current WVC Systems and Existing Solutions

WVC systems create a 2D face-to-face communication for people in remote locations. Despite more engaging in the breakout rooms, the loss of eye contact among participants still remains as a major unresolved problem [5]. Lower attention and memory retention rate were also reported when participants join a larger meeting, where the side effect of lacking direct eye contact becomes more obvious [11]. Prior work in [4, 20] shows that eye contact plays an vital role in human communication either in person or through WVC system. Thus, eye contact is a non-negligible element if we were to reconstruct WVC system to imitate real-world communication [9, 24]. However, perceiving eye contact is difficult in the existing video conferencing systems and hence limits their effectiveness [4].

The setup of webcam and user interface of current WVC systems systematically limits the perception of mutual gazes. Users tend to look at the part of the screen where the other participant is rendered. But the location of the webcam for most laptops are on the top of the screen, which makes it impossible to create simple eye contact among users. [34] shows that, the loss of eye contact is noticeable if the vertical distortion angle between the line from camera to the eyes and line from the eyes to the screen is more than 5 degrees. With average sized desktop computer displays, this angle is usually between 15 and 20 degrees [36], which results in inevitable loss of eye contact. This problem emerged as early as in 1969 when the video conferencing first started and has not yet been addressed properly for consumer-level systems.

Solutions like building a system of mirrors to change the perceived position of the webcam had been attempted by researchers [13, 28]. But altering the hardware has not been populated into the consumer level yet due to its high expenses. Researchers also tried software algorithms to tackle this problem by allowing users to change view angle in the scene by using the footage made by limited number of static cameras [17]. Synthesizing an image from a novel viewpoint different from that of the real camera has been fully explored. This method normally proceeds in two stages, first they reconstruct the geometry of the scene and in second stage, they render the geometry from the novel viewpoint. However, those methods still rely on using multiple cameras and cannot run real-time due to its complexity in graphics computation [17, 21, 22, 30, 37].

Some gaze correction systems are also proposed, targeting at a peer-to-peer video conferencing model that runs in real-time on average consumer hardware and requires only one hybrid depth/color sensor such as the Kinect [16]. However, when multiple people need to present, the gaze correction creates an illusion that everyone in the meeting is looking out of the screen. This illusion directly causes users to feel that everyone else is looking at them all the time and results in nervousness.

2.3 Eye Contact in Multi-person Conversation

Most studies of eye contact focused on situations where there are only two persons presenting in the conversation [2]. However, multi-person conversational structure becomes more complicated when a third speaker is introduced. It has long been presumed that eye contact is a powerful tool to convey additional information in multi-person conversations. Isaacs and Tang [12] performed a usability study of a group of five participants using a WVC system. They found that participants tended to use eye gaze to indicate who they were addressing in the offline setting. But this pattern was replaced by directly calling out other people's names in online meetings. Roel et al. [35] found that without eye contact in a meeting with multiple parties, 88% of the participants indicated they had trouble perceiving whom their partners were talking to. [33] was one of the first to formally investigate the effects of eye contact on the turn taking process in four-person video conferencing. Unfortunately, the author found no effects because accurate eye contact could not be conveyed using the video conferencing system they built.

2.4 Gaze Tracking

Classic gaze tracking methods estimate where a user is looking at by using high-end eye tracking cameras, which are expensive and not robust across different environments and poses [7, 23, 27].

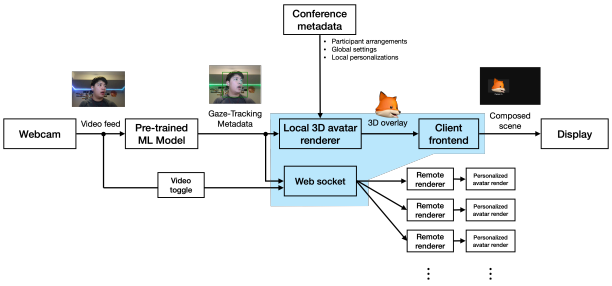


Figure 2: High level data block diagram for the idealized *VirtualGazer* application

(a) (b)

Figure 3: (a) The eye models: speaker with highlighted color; (b) The VR models.

The model-based techniques (e.g. [26] [3]) take advantages of the geometry structure of human eyes to directly determine the gaze direction. Those model-based methods provide better accuracy, but the calibration normally requires either two light sources or multiple cameras [15]. The regression-based techniques (e.g.[14]) leverage machine learning and neural networks to map the eye gaze position to the coordinates on the screen, which can be calibrated by a simple webcam. In this project, we focused on investigating the latter method with the network introduced in [29].

3 PROTOTYPE DESIGN OVERVIEW

This section briefly outlines the technical design overview of our prototype. We cover the framework and the programming of the application on a high level. The high level diagram for the full proposed application (outside the scope of this paper) is shown in Figure 2.

Our modified prototype version was developed mostly in Unity game engine for their advanced, yet easy-to-use 3D graphics, UI, and multimedia capabilities. For the 3D avatars, we created multiple different 3D humanoid characters using Unity Multipurpose Avatar (also referred as VR models). The characters were further integrated with audio to create vivid mouth movement using Salsa Lip-sync library. The gazing and glimpsing were achieved by animating the head movement with a series of animation figures. For the eye avatars , we made an eye mask consisting of shape and texture to render realistic pupils. Each eye avatar (also referred as eye models) contained a set of target coordinates, which were used to define where they look at for performing gazing and glimpsing (Figure 3).

The evaluation forms consist of binary questions, Likert scale questions, and short text input questions. All the forms were integrated into the Unity game as a pop-up window for the user to fill out after the corresponding scenarios. The recorded results

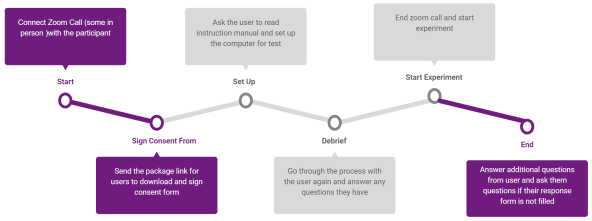


Figure 4: User Study Flowchart

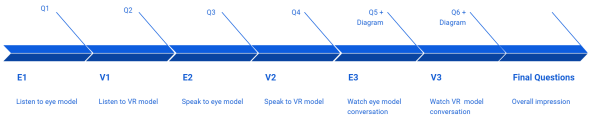


Figure 5: Application Flowchart

were exported to excel format with the help of QuestionaireToolKit Library.

4 USER EXPERIMENTS AND EVALUATIONS

We validated our system with a preliminary user study and semi-structured interview. The goal of the study is to evaluate the usability of our prototype and to identify limitations or opportunities for future improvements. We use an existing popular WVC application, Zoom, as our control variable in our experiments.

Firstly, the interviewer explains the purpose of our system and demonstrates all of the basic setup of our prototype system (Figure 4). Participants started by reading and signing the **consent form**(./ConsentForms). Then the participants were asked to perform three main experiments (Figure 5). Each experiment includes two tasks for testing two types of avatars (2D eyes and VR). The VR avatar experiments are denoted with prefix **V** and the eye avatar with prefix **E**. After each task, we asked the participants to give feedback about the system by interacting with a **questionnaire form** (./Questionnaire.pdf) embedded inside our system (See partial questions in Figure 8). After participants finished all three experiments, we conducted interviews to ask them to give additional feedback. The interview questions generally supplement the questions in the questionnaire where the user did not provide a lot of explanations. The entire experiment session lasted approximately 30–45 minutes.

4.1 Participants

We recruited a total of 13 participants (8 male and 5 female) to partake in our study from our friend-circle and fellow students in the UBC EECE department. The average age of all participants was 25 years old with a range from 23 to 31 years old. All participants study in post-secondary education institutions and all of them are competent in reading, writing, listening to English, and using computers and other online services such as Zoom.

4.2 Procedures

Experiment 1 (E1 & V1)

Experiment 1 (E1 & V1) involves the participant joining a lecture as a listener. One presenter talks throughout this experiment and the participant receives visual and audio output from our prototype app. After each task, the participant fills out a questionnaire to help us explore **RQ1**.

We initialized the WVC meeting room with six mock avatars including a presenter avatar. The test participant joins the meeting session as the seventh person, who is not visible on the screen. All avatars are programmed to look into random directions for the first ten seconds to simulate the idle phase before the lecture starts. This also allows participants to gain some familiarity with the scene to be better prepared for the task. While the presenter keeps talking, selected avatars would occasionally execute "Glimpse" and "Gaze" where they look towards the participant (look out of the screen). We define "Glimpse" as someone looking at you for less than 3 seconds and "Gaze" as someone looking at you for more than 3 seconds.

Finally we compare and correlate the participants' response, such as perceived number of gazes and glimpses. We compare their response with the ground truth, which is logged in the prototype application.

Experiment 2 (E2 & V2)

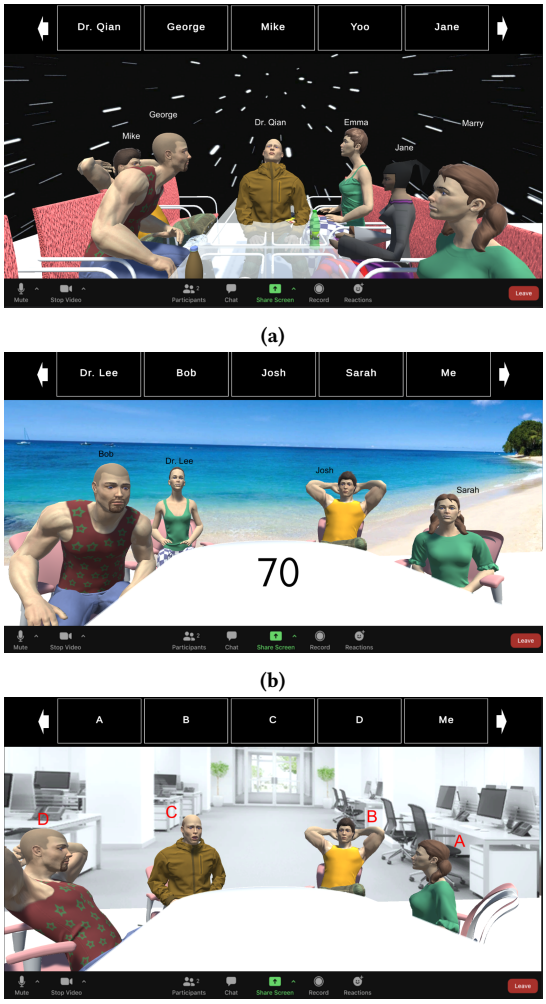
Experiment 2 (E2 & V2) involves the participant presenting in front of 4 mock avatars. The participant joins the meeting session as the fifth person — who is not visible on the screen. Participants are asked to watch a short video (2-3 minutes) and summarize the content to the audiences with the mock avatars looking at the participant. Each of the avatars can randomly toggle between two modes: Paying attention (PA) and Not paying attention (NPA). During the experiment, avatars randomly toggle between PA and NPA modes. These events are generated/logged for us as ground truth. Participants fill out a questionnaire after they present to indicate how many avatars they think were paying attention over 70% of the time. We compare the participants' observations with the ground truth.

At the end of the questionnaire, we also ask participants to rate their nervous level, focus level and engagement level from 0 to 100 to explore **RQ2**. As shown in Figure 11, 0 indicates our model is 100% less nervous, focusing and engaging than the traditional WVC. 100 indicates our model is 100% more nervous, focusing, and engaging than traditional WVC. 50 indicates our model is equivalent with traditional WVC.

Experiment 3 (E3 & V3)

Experiment 3 (E3 & V3) involves the participant joining a conversation. Instead of a single presenter constantly talking (the case of lectures), the participant watches a conversation with multiple people interacting with each other in the scene. We set up four mock avatars talking to and looking at each other with a pre-programmed sequence along with pre-recorded audio. Each mock avatar takes turns talking. Meanwhile, the other three mock avatars who are not talking will look at the avatar who is talking. Occasionally, the non-presenting avatars can randomly choose to look at another avatar or the participant.

When the conversation is over, we give the relationship matrix, as shown in Figure 7, for the participant to mark which avatar is



(a)

(b)

(c)

Figure 6: (a) Experiment 1: Watching Scenario (b) Experiment 2: Speaking Scenario (c) Experiment 3: Conversation Scenario

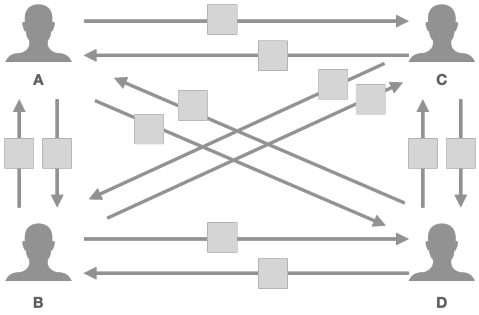


Figure 7: The relationship matrix we ask the participant to fill out

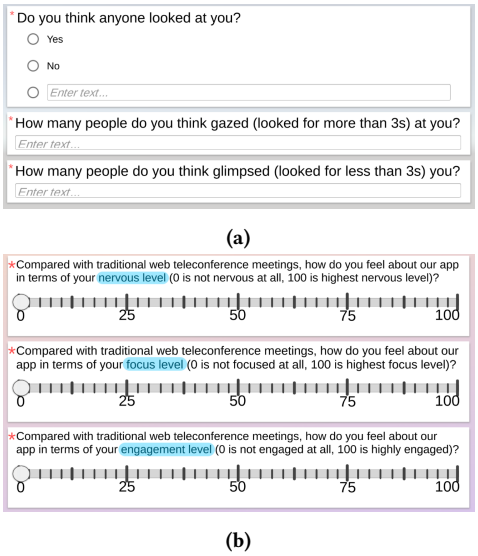


Figure 8: (a) Sample questions used for Experiment 1(E1 & V1) (b) Sample questions used for Experiment 2(E2 & V2)

paying attention to which. Inspired by the body sheets as a method to collect user responses in La Delfa et al.'s work in Drone Chi [18], we intend to use this relationship matrix to help us visualize the attention relationship among avatars observed by the participant.

We ask the participants to mark each directional arrow, as shown in Figure 7 to indicate which avatar is engaging with which. We also ask the participants to annotate each arrow with a confidence score (0.0 - 1.0) so that we could perform quantitative analysis. Besides filling out the relationship matrix, we also ask participants to leave additional comments in the questionnaire to help us explore **RQ3**.

Final Questions

We ask participants additional questions after they finish all the tasks. The detailed questions can be found in Figure 12.

5 RESULTS

Summary to Answer Research Questions

(1) **RQ1**: Users all can tell if they are being looked at after adding eye contact and a mini virtual room to traditional WVC. More vivid eye movements and better rendering quality of 3D avatars will help increase engagement level (Section 5.1 & 6).
(2) **RQ2**: Users have enhanced focus level and engagement level after adding eye contact and a mini virtual room to traditional WVC (Section 5.2).
(3) **RQ3**: No sufficient data to make valid claim on this research question (Section 5.3).

5.1 Experiment 1 (Watching)

In task E1, all participants reported that they had been looked at. But in task V1, 12 out of 13 participants reported that they had been looked at. Figure 9 shows the gaze and glimpse number in both E1 and V1 tasks. The glimpse result is more sparse (E1 highest being

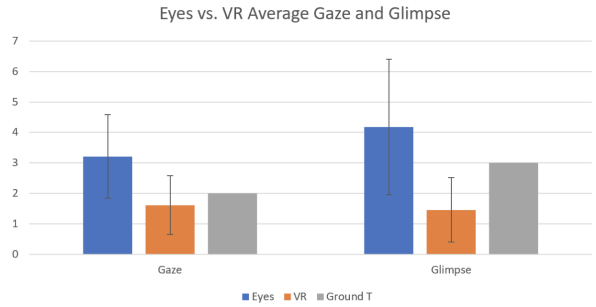


Figure 9: The eye avatar and VR avatar gaze and glimpse times comparison with ground truth (E1 & V1 Results)

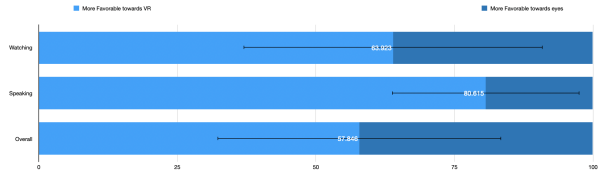


Figure 10: The attention distribution of heads and eyes in watching, speaking and overall tasks

9, lowest being 2; V1 highest being 7, lowest being 0). This result matches our expectation as described below:

We did not reveal the questions before the experiments because we think it will cause the participants to pay extra attention to find the answers, therefore corrupting the experiments. Instead we only asked the participants to observe carefully while performing the experiments. Thus, it's expected that participants could not remember exactly what they just saw when answering those questions. We believe this created some outliers. For example, P2 is the only one who reported that they had not been looked at in experiment V1.

We asked the participants "Which one attracted your attention more: eyes(0) or heads(100)?" after each experiment. Results show that with all the tasks, participants felt their attention was attracted by the VR model more than the eye model (Figure 10). Both the eye model (E1) and the VR model (V1) universally make participants notice that they were being looked at. Participants think the sense of being looked at has an average of 63.92% due to the head and 36.08% due to eyes (i.e. VR avatars make it more obvious to feel the glimpse and gaze).

5.2 Experiment 2 (Talking)

In Experment 2 (E2 & V2), participants were asked to rate their nervous level, focus level, and engagement level compared with traditional WVC when using the eye model and the VR model.

The eye model (E2) average nervous level is 41.00, and the VR model (V2) average nervous level is 53.38. This shows that the VR model makes participants more nervous (12% more) than the eye model. The focus level for the eye model and the VR model does not show significant differences (eye model is 63.08; VR model is

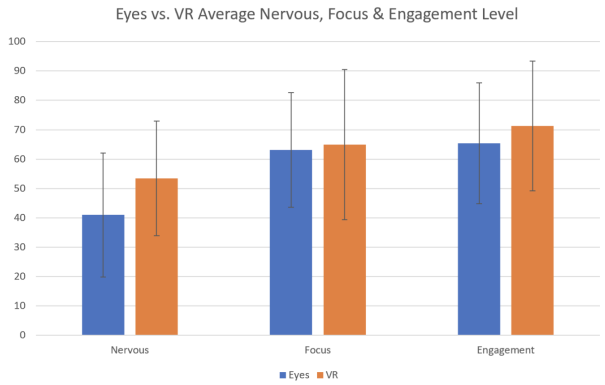


Figure 11: Nervousness, focus, and engagement levels reported by the participants as they are asked to speak/present in a room of mock avatar listeners (E3 & V3)

Additional Questions			
Questions	Suppose the avatars were more polished (with textures, colours, animations and customizations), would you use this for online meetings?	In a traditional web video conferencing app, how likely are you to voluntarily turn on your camera?	Would you replace a camera video feed with avatars from this experiment?
Average	70.07692308	36.30769231	65.23076923
STD	29.04439573	30.71206228	31.45672648

Figure 12: Additional Questions Results

64.92). The eye model (E2) average engagement level is 65.38, and the VR model (V2) engagement level is 71.31. This shows that the VR model makes participants more engaged (5% more) than the eye model. In general, participants reported their focus level and engagement level are enhanced compared with traditional WVC systems.

5.3 Experiment 3 (Conversation)

The relationship between A,B,C,D were observed and interpreted clearly (universally) when using the relationship matrix (Figure 7). However, unfortunately, while attempting to do quantitative analysis on participant-submitted relationship matrices, we observe that the values in the arrows shown in Figure 7 were more closely related to the dynamics in the dialog, rather than eye contact or gaze. Thus, the analyzed results from those two experiments do not provide valid data to investigate our **RQ3**.

Participants generally want to use VR avatars for online meetings if the avatars were more polished (Average = 70.08, σ = 29.04. Participants generally would feel comfortable replacing their camera video feed with the VR avatar in certain situations (Average = 65.23, σ = 31.46), such as when they don't want to be distracted by what people are wearing, background, or if they don't want to be seen.

6 DISCUSSION

6.1 Nervous Level and Distraction

In general, participants rated that the eye model makes them feel less nervous than using traditional WVC systems (average nervous level is 41.00, around 9% less than neutral). However, comments from participants about nervous level are a bit polarized. P4 thinks the eye model makes him less nervous since "*Using cartoon eyes to hide the actual person also makes me less nervous.*" P5 has the opposite opinion about this, "*By only showing participants' eyes ... makes me more nervous because you can find out whether people are directly gazing at you anytime.*"

Some participants (P12, P13) also indicated that their nervous level depends on how comfortable they are with public speaking. If they're comfortable talking with a large group of people, neither eyes nor VR avatars make a difference in nervous level. P6 thinks he could not accurately rate his nervous level due to his personal preference of looking away from the screen while talking. It made us think about the possibility of implementing an optional feature to always render the presenter's view on the audience to help those people who do not have strong public speaking skills to reduce their nervous level.

P4, P7 noted that the head movement and upper body movement in the VR model are actually more distracting than the eye model. P7 also brought up that the background of the virtual conference room in the VR model could be distracting, *"In the scenario, I may shift my attention to the background, and the body movement of the participants always catches my attention."*

6.2 Immersion and Interactivity

While P5, P6 both think that compared to the eye model, the VR model is much easier to interpret the relationships among other people. P6 said "*the direction of the people's gazes made it pretty apparent that they would be paying attention to the speaker the majority of the time.*" 11 out of 13 participants agreed that our system is more interactive and immersive:

P4: *"I feel like this version is the most realistic and most immersive... I was very focused and nervous."*

P5: *"Audiences and speakers both [find it] easier to acquire other people's reaction and movement."*

P6: *"In this case, it makes the teleconference more immersive, however it also makes me more nervous."*

6.3 Body Rendering

P1 mentioned a problem that we never thought about at the design phase: *"Other people can look at your body parts rather than your head to avoid eye contact intentionally even though the rendering is still showing they are looking at you"*. It is true that this will create an delusional interactive effect for some people. We think that this issue could be addressed by rendering even fewer body parts of the participants, for example, only the neck and above. But we would like to conduct further studies for users who intend to avoid eye contact to gather more feedback. P2 also noted a similar comment, *"Sometimes, it's hard to decide who they are looking at exactly"*. Having people perceive a 3D scenario with a 2D window

does generate difficulties like this for users particularly when there are more than 4 people in the meeting.

6.4 Watching vs. Speaking

Some participants gave comments beyond our questionnaire after they finished the entire experiments. They indicated that in general, they felt more comfortable with the eye model if they are the one talking, but more comfortable with the VR model if they are listening. They also suggested that we could build a model which provides a combination of eyes and VR avatars interchangeably depending on the needs of the users. They could choose to go into VR avatar mode when they are listening to a talk and go into eye mode when they are giving a talk.

6.5 Privacy

P3 brought up a point related to privacy, "*it's a pretty interesting app, which helps keep privacy while enabling interaction.*" Privacy is one of the most controversial problems in online meetings during the pandemic period and solutions like virtual background helped to protect the meeting room privacy but not the user's appearance. Using our app, users could choose to not render their camera feed but a VR avatar version of themselves. However, in order to achieve this, real time 3D head reconstruction, WVC, and gaze tracking need to be further integrated.

7 LIMITATIONS AND FUTURE WORK

Our project leveraged Unity and Processing to build a proof-of-concept WVC system. We investigated machine-learning-based gaze tracking technologies and real-time avatar rendering. We implemented our own WVC system and we successfully integrated it with gaze tracking technology. Due to the time limit of this project, our proof-of-concept research prototype decided to mock meeting scenarios and use pre-programmed avatars to answer our research questions. The main goal of the paper is to explore the impact and usefulness of eye contact in the WVC system rather than making a working product. Achieving real-time avatar rendering is out of the scope of this paper. In the future we want to support real-time gaze tracking.

The number of participants in the meeting is a major drawback of this prototype; if there are more than 8 participants, their avatars would be arranged into more than one page. While we can overcome the arrangement issue trivially by programming a custom front-end, each participant will have a tiny grid, making the gaze-tracking component a challenge. Additionally, P7 indicated that "*The eyes of the avatars could be detailed and optimised for better attention catching for the audiences.*"

After all experiments, we asked participants for their feedback and suggestions for improving our system. Most participants complimented our system and one participant said "*It's nice enough for me to use it.*" There are two most common suggestions we gathered from the participants. Firstly, improving the rendering quality of the VR avatars (P1, 4, 5, 6, 7). P1 mentioned *"Obviously, if the avatars are more vivid, and they do represent your eye contact, your direction of looking, maybe even body gesture etc, it could be significantly improving how it is, and avoiding the nervousness and awkwardness with real images."* However, the trade-offs of implementing more realistic avatars need to be considered carefully. P4 thinks the current VR avatar model in our system is less realistic, *"I think the VR version is somehow less realistic than the eyes version. Maybe it was because the avatar is trying to be more realistic, but it is still different from how a real person looks, so it breaks immersion for me."* With more realistic avatars, uncanny valley phenomena might arise. When the VR model is closer to the realness but some tiny differences still exist, people tend to feel very uncomfortable. It also has the risk of breaking the immersion.

Secondly, adding more functionalities to the VR avatar (P1, 6, 12, 13), such as body gesture (P1), head nodding (P13), and customization of avatars (P6) would make our system even better. Two participants (P1, 13) mentioned they'd like to see more features for our system. Many participants brought up that they want to show others that they are still listening and engaging even when they have their cameras off. For instance, P1 commented *"give option to switch between 2d and 3d... This will truly make people feel more interacted."* For future work, we should also look into having the option to switch between real camera, avatar, and eyes.

8 CONCLUSION

We presented *VirtualGazer*, a WVC system that allows users to achieve multi-person eye contact in teleconferencing. The results demonstrated that involving eye contact can enrich interactive experiences and enhance engagement level and focus level. We hope this paper opens up new opportunities for interactive teleconferencing systems and inspires the HCI community to further explore eye contact element to realize the highly interactive WVC experiences. Some future implementations inspired by our participants include combining VR model and eye model, enhancing avatar vividness, and adding better pupil animation for the VR model. We hope that these insights and findings point to potential directions for designing more satisfactory WVC systems, which are actively redefining our digital social lives today.

ACKNOWLEDGMENTS

We thank the participants of our user study and semi-structured interview. We would also like to thank Dr. Dongwook Yoon and other students in CPSC 544Y for their feedback and suggestions throughout the project.

REFERENCES

[1] Hosam Al-Samarraie. 2019. A Scoping Review of Videoconferencing Systems in Higher Education: Learning Paradigms, Opportunities, and Challenges. *International Review of Research in Open and Distance Learning* 20 (07 2019). https://doi.org/10.19173/irrodl.v20i4.4037

[2] Michael Argyle, Mark Cook, and Duncan Cramer. 1994. Gaze and Mutual Gaze. *British Journal of Psychiatry* 165, 6 (1994), 848–850. https://doi.org/10.1017/S0007125000073980

[3] D. Beymer and M. Flickner. 2003. Eye gaze tracking using an active stereo head. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, Vol. 2. II–451. https://doi.org/10.1109/CVPR.2003.1211502

[4] Milton Chen. 2002. Leveraging the Asymmetric Sensitivity of Eye Contact for Videoconference. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Minneapolis, Minnesota, USA) *(CHI '02)*. Association for Computing Machinery, New York, NY, USA, 49–56. https://doi.org/10.1145/503376.503386

[5] Alex Colburn, Michael Cohen, and Steven Drucker. 2000. The Role of Eye Gaze in Avatar Mediated Conversational Interfaces. *Sketches and Applications, SIGGRAPH '00* (09 2000).

[6] Dr Doggett and Anthony Mark. 2008. The videoconferencing classroom: What do students think? *Architectural and Manufacturing Sciences Faculty Publications* (2008), 3.

[7] Andrew T Duchowski and Andrew T Duchowski. 2003. *Eye tracking methodology: Theory and practice*. Springer.

[8] David Figlio, Mark Rush, and Lu Yin. 2013. Is it live or is it internet? Experimental estimates of the effects of online instruction on student learning. *Journal of Labor Economics* 31, 4 (2013), 763–784.

[9] David M. Grayson and Andrew F. Monk. 2003. Are You Looking at Me? Eye Contact and Desktop Video Conferencing. *ACM Trans. Comput.-Hum. Interact.* 10, 3 (Sept. 2003), 221–243. https://doi.org/10.1145/937549.937552

[10] Paul Hart, Lynne Svenning, and John Ruchinskas. 1995. From face-to-face meeting to video teleconferencing: Potential shifts in the meeting genre. *Management Communication Quarterly* 8, 4 (1995), 395–423.

[11] Jari K Hietanen. 2018. Affective eye contact: an integrative review. *Frontiers in psychology* 9 (2018), 1587.

[12] Ellen A. Isaacs and John C. Tang. 1993. What Video Can and Can't Do for Collaboration: A Case Study. In *Proceedings of the First ACM International Conference on Multimedia* (Anaheim, California, USA) *(MULTIMEDIA '93)*. Association for Computing Machinery, New York, NY, USA, 199–206. https://doi.org/10.1145/166266.166289

[13] 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* (Monterey, California, USA) *(CHI '92)*. Association for Computing Machinery, New York, NY, USA, 525–532. https://doi.org/10.1145/142750.142977

[14] Qiang Ji and Xiaojie Yang. 2002. Real-Time Eye, Gaze, and Face Pose Tracking for Monitoring Driver Vigilance. *Real-Time Imaging* 8, 5 (2002), 357–377. https://doi.org/10.1006/rtim.2002.0279

[15] Khan and Lee. 2019. Gaze and Eye Tracking: Techniques and Applications in ADAS. *Sensors* 19 (12 2019), 5540. https://doi.org/10.3390/s19245540

[16] Claudia Kuster, Tiberiu Popa, Jean-Charles Bazin, Craig Gotsman, and Markus Gross. 2012. Gaze Correction for HoRN11me Video Conferencing. *ACM Trans. Graph.* 31, 6, Article 174 (Nov. 2012), 6 pages. https://doi.org/10.1145/2366145.2366193

[17] Claudia Kuster, Tiberiu Popa, Christopher Zach, Craig Gotsman, and Markus Gross. 2011. FreeCam: A Hybrid Camera System for Interactive Free-Viewpoint Video. In *Vision, Modeling, and Visualization (2011)*, Peter Eisert, Joachim Hornegger, and Konrad Polthier (Eds.). The Eurographics Association. https://doi.org/10.2312/PE/VMV/VMV11/017-024

[18] Joseph La Delfa, Mehmet Aydin Baytas, Rakesh Patibanda, Hazel Ngari, Rohit Ashok Khot, and Florian'Floyd' Mueller. [n.d.]. Drone chi: Somaesthetic human-drone interaction. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[19] MICHAEL LIEDTKE. 2021. Zoom booms as pandemic drives millions to its video service. https://abcnews.go.com/Technology/wireStory/zoom-booms-pandemic-drives-millions-video-service-71029936

[20] C Macrae, Bruce Hood, Alan Milne, Angela Rowe, and Malia Mason. 2002. Are You Looking at Me? Eye Gaze and Person Perception. *Psychological science* 13 (10 2002), 460–4. https://doi.org/10.1111/1467-9280.00481

[21] Wojciech Matusik, Chris Buehler, Ramesh Raskar, Steven J. Gortler, and Leonard McMillan. 2000. Image-Based Visual Hulls. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '00)*. ACM Press/Addison-Wesley Publishing Co., USA, 369–374. https://doi.org/10.1145/344779.344951

[22] Wojciech Matusik and Hanspeter Pfister. 2004. 3D TV: A Scalable System for Real-Time Acquisition, Transmission, and Autostereoscopic Display of Dynamic Scenes. *ACM Trans. Graph.* 23, 3 (Aug. 2004), 814–824. https://doi.org/10.1145/1015706.1015805

[23] Carlos H Morimoto and Marcio RM Mimica. 2005. Eye gaze tracking techniques for interactive applications. *Computer vision and image understanding* 98, 1 (2005), 4–24.

[24] Naoki Mukawa, Tsugumi Oka, Kumiko Arai, and Masahide Yuasa. 2005. What is Connected by Mutual Gaze? User's Behavior in Video-Mediated Communication. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems* (Portland, OR, USA) *(CHI EA '05)*. Association for Computing Machinery, New York, NY, USA, 1677–1680. https://doi.org/10.1145/1056808.1056995

[25] Tuan Nguyen. 2015. The effectiveness of online learning: Beyond no significant difference and future horizons. *MERLOT Journal of Online Learning and Teaching* 11, 2 (2015), 309–319.

[26] Takehiko Ohno and Naoki Mukawa. 2004. A Free-Head, Simple Calibration, Gaze Tracking System That Enables Gaze-Based Interaction. In *Proceedings of the 2004 Symposium on Eye Tracking Research amp; Applications* (San Antonio, Texas) *(ETRA '04)*. Association for Computing Machinery, New York, NY, USA, 115–122. https://doi.org/10.1145/968363.968387

[27] Takehiko Ohno, Naoki Mukawa, and Atsushi Yoshikawa. [n.d.]. FreeGaze: a gaze tracking system for everyday gaze interaction. In *Proceedings of the 2002 symposium on Eye tracking research applications*. 125–132.

[28] Ken-Ichi Okada, Fumihiko Maeda, Yusuke Ichikawaa, and Yutaka Matsushita. 1994. Multiparty Videoconferencing at Virtual Social Distance: MAJIC Design. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work* (Chapel Hill, North Carolina, USA) *(CSCW '94)*. Association for Computing Machinery, New York, NY, USA, 385–393. https://doi.org/10.1145/192844.193054

[29] 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, 3839–3845.

[30] Benjamin Petit, Jean-Denis Lesage, Clément Ménier, Jeremie Allard, Jean-Sebastien Franco, Bruno Raffin, Edmond Boyer, and Francois Faure. 2010. Multi-Camera Real-Time 3D Modeling for Telepresence and Remote Collaboration. *International Journal of Digital Multimedia Broadcasting* 2010 (08 2010). https://doi.org/10.1155/2010/247108

[31] Robert J Reese and Norah Chapman. 2017. *Promoting and evaluating evidence-based telepsychology interventions: Lessons learned from the university of Kentucky telepsychology lab*. Springer, 255–261.

[32] Jeffrey J Roth and Steven Pierce, MariBrewer. 2020. Performance and satisfaction of resident and distance students in videoconference courses. *Journal of Criminal Justice Education* 31, 2 (2020), 296–310.

[33] Abigail J. Sellen. 1995. Remote Conversations: The Effects of Mediating Talk with Technology. *Hum.-Comput. Interact.* 10, 4 (Dec. 1995), 401–444. https://doi.org/10.1207/s15327051hci1004_2

[34] R. Stokes. 1969. Human Factors and Appearance Design Considerations of the Mod II PICTUREPHONE® Station Set. *IEEE Transactions on Communication Technology* 17, 2 (1969), 318–323. https://doi.org/10.1109/TCOM.1969.1090060

[35] Roel Vertegaal, Gerrit Veer, and Harro Vons. 2000. Effects of Gaze on Multiparty Mediated Communication. (12 2000).

[36] Ben Yip. 2004. Pose determination and viewpoint determination of human head in video conferencing based on head movement. 130–. https://doi.org/10.1109/MULMM.2004.1264977

[37] C. Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. 2004. High-Quality Video View Interpolation Using a Layered Representation. In *ACM SIGGRAPH 2004 Papers* (Los Angeles, California) *(SIGGRAPH '04)*. Association for Computing Machinery, New York, NY, USA, 600–608. https://doi.org/10.1145/1186562.1015766