Are You Looking at Me? Eye Gazing in Web Video Conferences

Muchen He
University of British Columbia
Vancouver, Canada
mhe@ece.ubc.ca

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

Recently, Web Video Conferencing (WVC) system has been exacerbated by the COVID-19 pandemic and has become an essential tool in remote-learning and remote-working situations. However, the efficiency of the communication using current WVC systems is obstructed by the lack of eye contact due to the disparity between the position of the camera and the position of the eyes on the screen. There exists some high-end expensive WVC systems that can partially solve this problem, but still it is not solved for consumer-level. This paper introduces a new way to achieve eye contact for multi-person teleconferencing. Our proof-of-concept research prototype, *FutureGazer*, leverages Processing IDE, JavaScript, and Unity Game engine to build a mocked WVC environment with eye model and head model. We conducted usability study and semi-structured interview study with 15 participants to investigates how including eye-contact in current WVC systems affects user online meeting experience particularly in terms of their nervous level, focus level, and engagement level. Our overall findings indicate that involving eye-contact can enrich interactive experiences and enhance engagement level and focus level. Our head model also generally attracts more attention from users than the eye model, but there is also trade-offs in using them for talking to audience or listening to a talk. We also highlight limitations such as rendering quality and additional features of avatars for future improvements that aim at better supporting eye-contact in teleconferencing.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; **Collaborative and social computing systems and tools**.

KEYWORDS

Eye-tracking, gaze-tracking human-computer-interaction, web video conferencing, online-learning

1 INTRODUCTION

Web video conferencing (WVC), exacerbated by the COVID-19 pandemic, are an essential tool in remote-learning and remote-working situations. Unfortunately, the same technology has lacked innovation in push human-computer-interaction that enhances WVC user-experience, such as amplifying engagement through eye-contact.

In distant-learning classes, for example, presenters (e.g. professors, teachers, students) often feel distracted or disengaged when there are no audiovisual feedback coming from the audience. These feedback include eye contact, gaze direction, and other body language cues.

For example, in the Gallery View in Zoom (and other WVC applications), the audience consists of black boxes and name tags, as shown in Figure 1(a). If the participant has 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 if they are paying attention to the meeting or the active participant.

Thus, 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 [23]. In one of the first experimental studies [9] on the effects of traditional instruction versus online learning, students attend live lectures instead of watching the same lectures online while supplemental materials and instructions were the same. Researchers [9] found modest evidence that the traditional format trumps the online format in engagement. Many people also think it is odd to see their faces during conversations, and it is hard to look away — significantly distracts participants during WVCs. 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 from using an avatar as an alias instead of a live video.

We propose FutureGazer, a WVC system that simulates eye-contact and gaze amongst the participants in a WVC meeting room to enable a highly interactive environment. Our project explores whether adding eye-contact to the current WVC system will enhance the sense of interaction and presence of the users. 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[11].

To test our system, we recruit friends and students are participants to study the effects of the additional eye contact and gaze cues in online meeting environments. Figure 1 shows what we intend to build in contrast to existing WVC platforms like Zoom. Figure 2 depicts the personalized eye-contact simulation enabled by our system. For more details, please refer to our technical report.

The key metrics we want to observe in this project are: participant's attention, engagement, and the feeling of connection. To explore parameters that effect these metrics, we consolidate these ideas into three core research questions (hereafter will be refer to as RQ1, RQ2, and RQ3):

CPEN 541 HIT'21, April 2021, Vancouver, BC, Canada

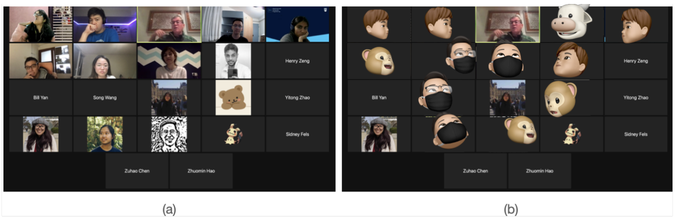

Figure 1: (a) Current WVC model: each participant stays in their grid and no eye contact interaction. (b) Proposed model: students can look at each other to create virtual eye-contact.

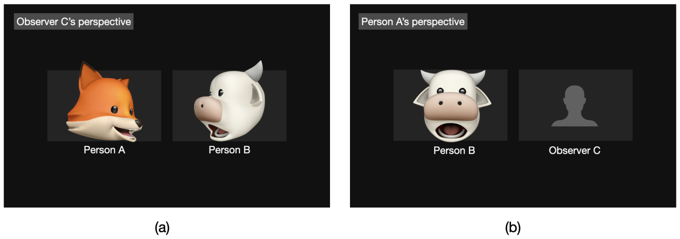

Figure 2: (a) Observer's perspective in a meeting room with person A (fox) and person B (cow) looking at each other. (b) Person A's perspective in the exact same meeting room at the exact same time.

(1) Can a person tell if they are being looked at in a WVC and how can 3D avatars be augmented to enhance this experience.
(2) Can a person tell if other participants are looking at each other in a WVC and how using 3D avatars can be augmented to increase engagement.
(3) How does a person's attention change as the avatars augmented with WVC enables eye-contact and gaze.

We intend to modify design parameters to our prototype user-interface (UI) of our mock WVC program to study the behaviour of participants.

2 BACKGROUND RELATED WORK

This section provides an overview of Web Video Conferencing (WVC), gaze tracking studies, eye-contact in current WVC system and eye-contact in multi-person communications.

2.1 Web Video Conferencing

WVC is a synchronous model that provides verbal and visual communication between two or more participants. Examples of WVC services include Zoom, Collaborate Ultra, Microsoft Teams, and others. When the COVID-19 pandemic emerged, and in-person classes transitioned to online-learning, researchers evaluated students' satisfaction with WVC-based learning and social activities.

WVC generally provides a more collaborative and engaging experience for students using interactive breakout rooms [3]. Some also suggest WVC provides higher satisfaction scores than other tools and has become one of the most popular online teaching methods [27, 28].

However, in the study by [7], 80% of the students felt they would be more engaged in a standard class setting, and 57% of the students thought WVC technology is a barrier to their interaction with instructors.

Since WVC hinders eye-contact in larger meetings, participants also observe lower attention and memory retention, a side-effect of lack of direct eye-gazes [12]. Lastly, a study observes an increase in participants' pro-social behaviour when being watched by deceptive video conferencing manipulation [5].

2.2 Gaze Tracking

Classical gaze-tracking methods estimate where a user is looking, but these implementations require expensive hardware and are not robust across different environments and poses [8, 21, 24]. Conventional WVC services (e.g. Zoom), such 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 audience. NVIDIA Maxine uses GANs to infer facial expressions and reconstruct a photorealistic feed where a presenter can look in arbitrary directions. However, their implementation only ensures direct-eye contact to the screen's centre and does not support larger meeting rooms [1].

The mixed reception of WVC and lack of non-verbal human interface forms the primary motivation for us to close the gap between teleconferencing and traditional F2F meetings. Moreover, we investigate the relationship between direct eye-gazing and pro-social behaviour in a WVC environment.

2.3 Eye Contact in Current WVC Systems

A large body of prior work has explored that eye contact is a critical aspect of human communication. [6, 18] Eye contact plays an important role in both in person and a WVC system. [10, 22] Therefore, it's critical and necessary to preserve eye contact in order to realistically imitate real-world communication in WVC systems. However, perceiving eye contact is difficult in existing video-conferencing systems and hence limits their effectiveness. [6] The lay-out of the camera and monitor severely restricted the support of mutual gaze. Using current WVC systems, users tend to look at the face of the person talking which is rendered in a window within the display(monitor). But the camera is typically located at the top of the screen. Thus, it's impossible to make eye contact. People who use consumer WVC systems, such as Zoom, Skype, experience this problem frequently. This problem has been around since the dawn of video conferencing in 1969[30] and has not yet been convincingly addressed for consumer-level systems.

Some researchers aim to solve this by using custom-made hardware setups that change the position of the camera using a system of mirrors [14, 25]. These setups are usually too expensive for a consumer-level system. Software algorithms solutions have also been explored by synthesizing an image from a novel viewpoint different from that of the real camera. 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. [16, 19, 20, 26, 32] Those methods usually require a number

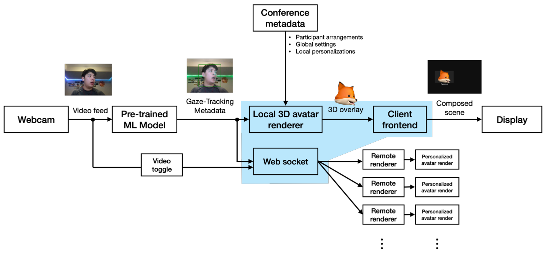

Figure 3: High level data block diagram for the idealized FutureGazer application

of cameras and not very practical and affordable for consumer-level. Besides, those methods also have a convoluted setup and are difficult to achieve in real-time.

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. [15] However, when there are more than two persons involved in a web video conference, even with gaze corrected view, users still cannot tell whether a person is looking at him or someone else in the meeting. With the gaze correction, it will create the illusion that everyone in this meeting is looking out of the screen. This could cause a serious confusion.

2.4 Eye Contact in Multi-person Conversation

Most studies of eye contact during conversations focused on two-person communication argyle [4]. However, multi-person conversational structure becomes more complicated when a third speaker is introduced. It has long been presumed that eye contact provides critical information in conversations. Isaacs and Tang [13] performed a usability study of a group of five participants using a desktop video conferencing system. They found that during video conferencing, users addressed each other by name and started explicitly requesting individuals to start talking. In face-to-face interaction, they found people used their eye gaze to indicate whom they were addressing. [29] was one of the first to formally investigate the effects of eye contact on the turn taking process in four-person video conferencing. Unfortunately, she found no effects because the video conferencing system she implemented did not accurately convey eye contact [29]. [31] found that without eye contact, 88% of the participants indicated they had trouble perceiving whom their partners were talking to.

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 3 For details regarding our prototype, please refer to our technical report.

Our prototype is developed mostly in Processing[2] and Unity game engine for their advanced, yet easy-to-use 3D graphics, UI, and multimedia capabilities. For the 3D head avatar, we load a .obj 3D model from file and the same model is re-rendered multiple times for each avatar, but with different transformations, to save memory and computation. For the eye avatar, the render consists of the eye mask that makes up the over all shape, and the texture — where pupil and the iris is drawn.

Each avatar has a set of target coordinates. These target coordinates define where the avatar should be looking at, and are often updated every frame. Using the target coordinates, we can program the avatars to look at an arbitrary scree-space point, at other avatars, or towards the participant. The target coordinates allow each avatars on a meeting participant's screen to be unique, as shown in Figure 2(a) and Figure 2(b).

4 USER EXPERIMENTS

In this section, we discuss the user experiments designed to study and measure the effects of eye contact and gaze in online meetings.

4.1 Participants

We recruited total of 15 participants to partake in our study from our friend-circle and fellow students in the department. The participants are mostly aged between 18-25 who studies in post-secondary education and all of them are competent in using computers and other online services such as Zoom. We acknowledge the limitation regarding the homogeneity of our sample participants: as it is unclear how the effects of this technology translates to a more general-represented population.

4.2 Experiments Procedures

In this subsection, we outline our preliminary strategy to perform user experiments and collect quantifiable data for our evaluations of how FutureGazer prototype affects user behaviour. We use an existing popular WVC application, Zoom, as our control variable in our experiments.

We setup FOUR main experiments (1, 2, 3, 4) with varying parameters to test our prototype. Each of the four experiments also has two variants to test the two types of avatar (eyes and heads). The head avatar variant experiments shall have the suffix **H** and the eye variant of the experiments have **E**.

Experiment 1 and 2 (E1, E2, H1, H2) involves the participant passively join a meeting. They watch and listen for the visual and audio feedback from the prototype app. We choose to use this experiment to explore **RQ1**, and study if a person can tell if they're being looked at in an online meeting, and how much.

Experiment 3 (E3, H3) involves involves the participant to speak in a room of mock-avatars. In this case we explore **RQ3** and attempt to gauge how participants feel, including nervousness, focus, and engagement with the audience using our prototype.

Experiment 4, (E4, H4) involves the participant to join passively as an observer again; however, instead of a single presenter speaking (such as in the case with lectures), the participant watches a conversation. We intend to answer **RQ2**, and see if with the help of gaze, participant is more able to identify relationships in a conversations.

For the sake of not being redundant, we do not perform both eye and head variants for experiment 1 and 2. Instead for experiment 1, we only use eye avatar. Similarly, for experiment 2, we only use head avatar. In other words, we **omit** experiments H1 and E2.

Originally, our plan was to initiate a pop-up window that prompts the test participant to answer whether they think they are being looked at. However, due to complications regards to deploying the prototype executable to people (further complicated by online-only experiments), we decided to aggregate these stare events and ask the test participant questions in the end.

The next subsection outlines the detailed procedures of each of the experiments.

4.2.1 Experiment E1. Begin by setting up nine mock-avatars in the WVC window, each with a unique name as seen in Figure 4(a). The mock-avatars does not correspond to real users in the WVC room and are programmed and controlled prior to user testing. Note that in Figure 4, head avatars are used, but as mentioned in 4.2, only eye avatars are used for experiment A.

The test participant joins the meeting session as the tenth person — who is not visible on the screen. Initially, the mock-avatars move randomly for several seconds (Figure 4(b)). Meanwhile an audio track of a lecture or a podcast plays. One of the mock-avatar, hereafter called *presenter-avatar*, is programmed to be synced with the audio track for the sake of realism. Throughout the meeting, a set of specific pre-programmed mock-avatars (that is not the presenter-avatar) will look at the test participant (look out from the screen) intermittently for several seconds at varying frequencies without disrupting the presenter-avatar or the audio (Figure 4(c)). We call this event a stare. The participant does not know which mock-avatars are selected to look at them before the experiment to preserve the validity of the results.

Finally we compare and correlate the participants' response, such as perceived number of gazes (avatars that "stared" for longer than 3 seconds), glimpses (avatars that "stared for less than 3 seconds). We compare their response with the ground truth which is logged in the prototype application. The correlation tests the hypothesis set in RQ1.

4.2.2 Experiment H2. Experiment B explores RQ3 by observing whether a person who is paying attention to a presenter can notice another person who starts to look at them (i.e. the gaze target change to the subject).

The procedure is identical to experiment E1, and as mentioned in Section 4.2, this experiment is only performed using 3D heads as avatars. At the end of the experiment, we ask the participant the same question as experiment A. Additionally, we ask how the experience differs from experiment E1 — in particular, how much more attention has the heads garnered compared to experiment E1.

4.2.3 Experiment E3, H3. In E3 and H3, we attempt to test whether the presenter can tell if the audience is paying attention to their speech and tackles both RQ1 and RQ3.

We first ask the participant to observe a short film or review a concept they would like to talk about. Once they're ready, We set up five mock-avatars in the WVC window and the participant will join the session as the sixth user. The participant will then summarize the short film, or talk about a concept for one to two minutes while the mock-avatars are looking at the participant. Each of the mock-avatars can randomly toggle between two modes: Paying attention (PA) and Not paying attention (NPA). During the experiment, we program the prototype app such that random mock-avatars is selected and it can toggle between PA and NPA modes at random times. These events are generated/logged for us to compare with.

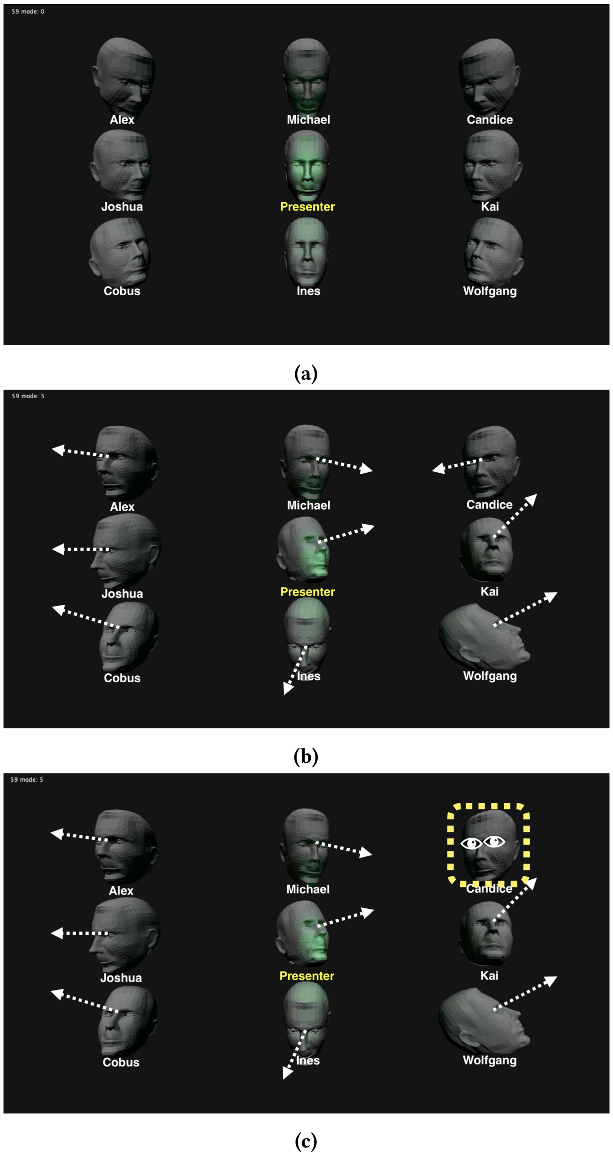

Figure 4: (a) The initialized WVC meeting room with eight mock-avatars including a presenter-avatar. (b) All avatars programmed to look into random directions. (c) Selected mock-avatars would occasionally execute "stare" where they look out of the screen and towards the participant. Note the dashed lines are for visualization only and cannot be seen by the participant.

After the participants are done talking, we ask the participant to rate whether they think they are being paying attention to, based on how many avatars they think that is paying attention. We also assess participants' nervousness, focus, and engagement level as they were speaking throughout the session. The participants report these as a rating from 0 to 100% *compared to* as if they were to perform the same task using traditional WVC apps such as Zoom.

In the end we compare the participants' observations of how many mock-avatars are paying attention versus the logged values. A strong correlation implies that RQ1 and RQ3 are likely true. We also aggregate the response data and observe effects on the participants as presenters.

We repeat the process for the other avatar type.

4.2.4 Experiment E4, H4. In experiments E4, H4, we attempt to test RQ2 in a small-group WVC environment as we assume eye contact amongst two or more people can incite a closer and intimate relationship to an observer [12]. Inspired by the body sheets as a method to collect user responses in La Delfa et al.'s work in Drone Chi [17], we intend to use a relationship matrix sheet to study the effect of 3D avatars in

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 take turns talking. Meanwhile, the other three mock-avatars who are not talking will look at the avatar who is talking. Occasionally and randomly, the non-presenting avatars can choose to look at another avatar, but not the participant. Thus, we can describe the engagement and interaction between the four avatars as a relationship matrix:

$$\mathbf{P} = \begin{bmatrix} 0 & p_{a,b} & p_{a,c} & p_{a,d} \\ p_{b,a} & 0 & p_{b,c} & p_{b,d} \\ p_{c,a} & p_{c,b} & 0 & p_{c,d} \\ p_{d,a} & p_{d,b} & p_{d,c} & 0 \end{bmatrix}$$

Where $p_{a,b}$ is the probability mock-avatar a is looking at/paying attention to b and all columns and rows adds up to 1.0.

When the experiment is complete and all mock-avatars finished taking turns speaking, we give the relationship matrix as shown in Figure 5 to the participant to articulate which avatar-pair is more intimate, as well as which avatar is talking with which. Evaluation:

We ask the participants to mark each directional arrow, as shown in Figure 5, of the relationship matrix, to indicate which avatar is engaging with which. We may also ask the participant to annotate each arrow with a confidence score (0.0 - 1.0). These scores can be normalized and compared with the probability matrix **P** that was pre-programmed into the mock-avatars. A strong correlation of participants' response and would imply RQ2 is likely true.

We repeat the process for the other avatar type.

5 RESULTS

In E1 experiment, all participants reported that they had been looked at. But in experiment H2, 14/15 participants reported that they had been looked at. The Figure 6 shows the gaze and glimpse number in both experiments E1 and H2 . The glimpse result is more sparse with the highest reported number being 9 (For E1), and the lowest being 1.5. This result matches our expectation.

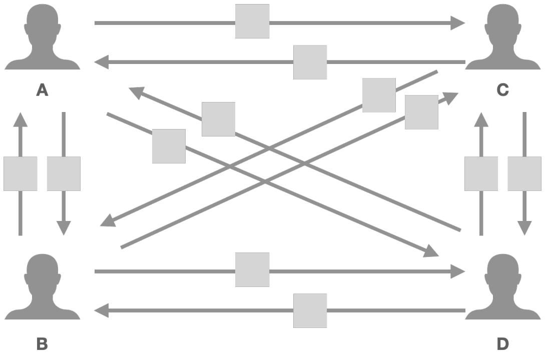

Figure 5: The diagram we ask the participant to fill out which corresponds to the relationship matrix P

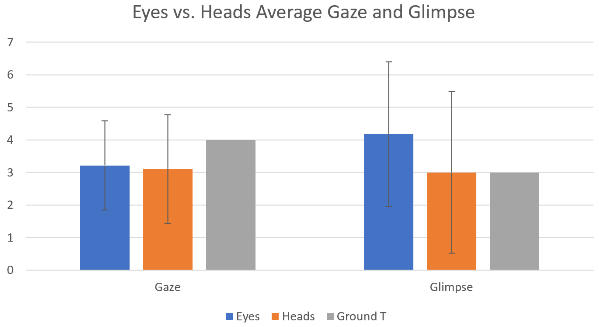

Figure 6: The eye model and head model gaze and glimpse times comparison with ground truth

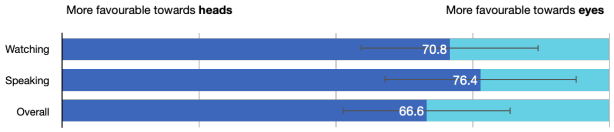

Figure 7: The attention distribution of heads and eyes in watching, speaking and overall tasks

We did not reveal the questions before the experiments because we think it will cause the participants to pay extra attention to finding the answers, which will corrupt the original experiment purposes. 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, P11 is the only one who reported he had not been looked at in the H2.

We asked the participants "Which one attracted your attention more: eyes(0) or heads(100)?" three times during the experiments. "Watching" is E1 and H2, "speaking" is after E3 and H3, and "overall" is in the final question section. Results show that with all the tasks, participants felt their attention was attracted by the head model more than the eye model. Both the eye model and the head model universally make participants notice that they were being looked at. Participants have a sense of being looked with an average of 70.86% due to the head and 29.14% due to eyes (i.e. 3D heads make it more obvious to feel the glimpse and gaze), shown in Figure 7.

In experiment E3 and experiment H3, participants were asked to rate their nervous level, focus level, and engagement level compared

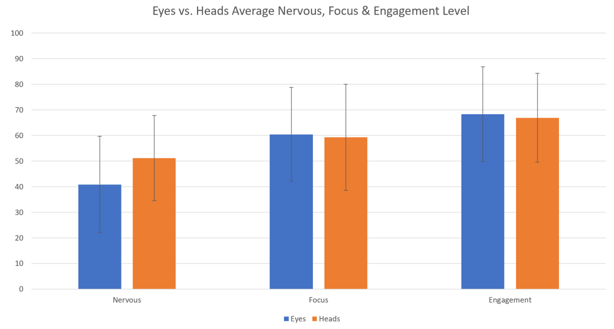

Figure 8: Nervousness, focus, and engagement levels reported by the participants as they are asked to speak/present in a room of mock-avatar listeners (E3, H3)

with traditional WVC when using the eye model and the head model. As shown in Figure 8, 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.

The eye model (E3) average nervous level is 40.86, and the head model (H3) average nervous level 51.14. This shows that the head model makes participants more nervous than the eye model. The focus level and engagement level for the eye model and the head model does not show significant differences. However, participants reported their focus level and engagement level are enhanced (average focus level is 60.43 for E3, 59.29 for H3; and average engagement level is 68.29 for E3, 66.93 for H3) compared with traditional WVC systems.

The relationship between A,B,C,D were much observed and interpreted clearly (universally) when using heads (H4). P1, P5, P9, P11, P13 all think compared to the eye model, the head model is much easier to interpret the relationships among other people. Furthermore, P1 and P5 noted that a few avatars were not participating in the mock-discussion as much. Unfortunately, while attempting to do quantitative analysis on participant-submitted relationship matrices, we observe that the values in the arrows shown in Figure 5 were more closely related to the dynamics in the dialog, rather than eye contact or gaze.

Participants generally want to use head avatars for online meeting if the avatars were more polished. (Average 81.29 STD 22.01) Participants generally would feel comfortable (Average 72.4 STD 23.05) replacing their camera video with the head avatar in certain situations, such as when they don't want to be distracted by what people are wearing, background, or if themselves don't want to be seen.

6 DISCUSSIONS

In general, participants rated the eye model makes them feel less nervous than using traditional WVC systems (average nervous level is 40.86, around 9% less than neutral). However, comments from participants about nervous level are a bit polarizing. P8 thinks the eye model makes him less nervous since "*Using cartoon eyes to hide the actual person also makes me less nervous.*" P9 has the opposite opinion about this, "*Only by showing participants' eyes ... makes me more nervous because you can find out whether people are directly gazing at you anytime.*" Some participants (P3, P5) also indicated that how nervous they felt depends on how comfortable they are with public speaking. If they're comfortable talking with a large group of people, neither eyes or heads make a difference in nervous levels. P10 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 to think 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.

Some participants (P3, 5, 9, 10, 15) noted that the head movement in the head model is actually more distracting than the eye model. P10 commented "*The movement of the head on the screen may break my train of thought.*" P9 also expressed a similar perspective, "*Looking at people's heads would make me less focused and nervous because I will pay some attention to them and find out whether they are listening or not.*" Indeed, even when users are giving a talk in real life, they would only perceive the general reactions of the majority of the audiences. They might not notice when only a few audiences start to look around as long as the majority is paying attention. But our system augmented the "looking around" movement, which made it very obvious when only a few heads start to shift attention despite that the majority are still paying attention. P5 resonated in the same way and felt disappointed when somebody starts to look around.

Focus level between eyes and heads are divided: some people (P3, P5) think that the eyes have less focus because it's hard to tell who is paying attention. On the contrary, P3, 5, 9, 10, 15 think that head is too obvious and the added animation/motion is more distracting. P1, 2, 8, 9 commented that head model is more obvious than eye model. P3, 5, 9, 10, 15 noted that the head movement is actually more distracting than the eye model.

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 head model if they are listening. They also suggested that for future work, we should also look into combinations of eyes and heads. They also suggested that we could build a model which provides a combination of eyes and heads interchangeably depending on the needs of the users. They could choose to go into head mode when they are listening to a talk and go into eye mode when they are giving a talk.

P7 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 an avatar head 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 also implemented our own WVC system and we successfully integrate it with gaze-tracking technology. But we realized that it is very time consuming to achieving real-time avatar rendering, especially considering this is a course project. We decided to build mock meeting scenes and implemented pre-programmed avatars in Unity to avoid spending time on achieving real-time rendering. The main goal of the paper is to explore the impact and usefulness of eye-contact in WVC system rather than making a working product.

The major limitation is that the avatar may not be as realistic as the human face. So, talking to an animated head with fake eyes may not give the participants the same eye contact experience as in real life. The number of participants in the meeting is another drawback of this prototype; if there are more than 15 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.

5 participants (P4, P6, P10, P8, P11) mentioned details on pupils that can be further improved. As P10 described, "*typically heads do not move as much during traditional online meeting apps and once someone else is speaking, eyes will shift rather than heads.*" P8 also thinks that our head model does not reflect the real life situation good enough because there are no pupils in the head. P8 said "*I think it is a good idea to represent people with fake heads, but their eyes did not have a pupil so it was hard for me to tell who is looking at whom based on the eye movement.*" P6 agreed with P10 and mentioned head models without pupils is unnatural for them to look at. Additionally, P4 noted that our eye model is not very realistic since the eyeballs in real life will not be fixed when people are paying attention; people turn their heads and keep blinking their eyes rather than having their eyes fixed. P11 also 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. First, improving the rendering quality of the avatars (P4, 7, 8, 9, 10, 11). P4 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. P8 thinks the current head model in our system is less realistic, *"I think the talking head version is somehow less realistic than the eyes version. Maybe it was because the head is trying to be more realistic, but it is still different from how a real head looks, so it breaks immersion for me."* With more realistic avatars, uncanny valley phenomena might arise. When the head model is closer to the realness but some tiny differences still exist, people tend to feel very comfortable. It also has the risk of breaking the immersion.

Second, adding more functionalities to the avatar (P2, 4, 10, 12), such as body gesture (P4), head nodding (P2), mouth animation(P10), and customization of avatars (P10) would make our system even better. Two participants (P1, 12) mentioned they'd like to see more features for our system. For example P1 commented "*give option to switch between 2d and 3d*". P6 thinks our system makes people feel more interactive. They stated that " *I'd like to see that the avatars could reflect some states of the people who are participating in a virtual conference. This will truly make people feel more interacted.*"

8 CONCLUSION

We presented FutureGazer, 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 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 includes combining head model and eye model, enhancing avatar vividness, and adding better pupil animation for the head 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 usability study and semi-structured interview. We would also like to thank Dr. Sidney Fels for his feedback adn suggestions throughout the project.

REFERENCES

[1] 2020. https://blogs.nvidia.com/blog/2020/10/05/ganvideo-conferencing-maxine/

[2] 2020. https://www.processing.org

[3] 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

[4] 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

[5] Roser Cañigueral and Antonia F de C Hamilton. 2019. Being watched: Effects of an audience on eye gaze and prosocial behaviour. *Acta psychologica* 195 (2019), 50–63.

[6] 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

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

[8] Andrew T Duchowski and Andrew T Duchowski. 2017. *Eye tracking methodology: Theory and practice.* Springer.

[9] 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.

[10] 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

[11] 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.

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

[13] 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

[14] 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

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

[16] 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

[17] 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.

[18] 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

[19] 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

[20] 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

[21] 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.

[22] 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

[23] 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.

[24] 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.

[25] 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

[26] 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

[27] 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.

[28] 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.

[29] 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

[30] 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

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

[32] 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